

Kuinam Kim
Nikolai Joukov
Editors

Information Science and Applications 2017

ICISA 2017

Lecture Notes in Electrical Engineering

Volume 424

Board of Series editors

Leopoldo Angrisani, Napoli, Italy
Marco Arteaga, Coyoacán, México
Samarjit Chakraborty, München, Germany
Jiming Chen, Hangzhou, P.R. China
Tan Kay Chen, Singapore, Singapore
Rüdiger Dillmann, Karlsruhe, Germany
Haibin Duan, Beijing, China
Gianluigi Ferrari, Parma, Italy
Manuel Ferre, Madrid, Spain
Sandra Hirche, München, Germany
Faryar Jabbari, Irvine, USA
Janusz Kacprzyk, Warsaw, Poland
Alaa Khamis, New Cairo City, Egypt
Torsten Kroeger, Stanford, USA
Tan Cher Ming, Singapore, Singapore
Wolfgang Minker, Ulm, Germany
Pradeep Misra, Dayton, USA
Sebastian Möller, Berlin, Germany
Subhas Mukhopadhyay, Palmerston, New Zealand
Cun-Zheng Ning, Tempe, USA
Toyoaki Nishida, Sakyo-ku, Japan
Bijaya Ketan Panigrahi, New Delhi, India
Federica Pascucci, Roma, Italy
Tariq Samad, Minneapolis, USA
Gan Woon Seng, Nanyang Avenue, Singapore
Germano Veiga, Porto, Portugal
Haitao Wu, Beijing, China
Junjie James Zhang, Charlotte, USA

About this Series

“Lecture Notes in Electrical Engineering (LNEE)” is a book series which reports the latest research and developments in Electrical Engineering, namely:

- Communication, Networks, and Information Theory
- Computer Engineering
- Signal, Image, Speech and Information Processing
- Circuits and Systems
- Bioengineering

LNEE publishes authored monographs and contributed volumes which present cutting edge research information as well as new perspectives on classical fields, while maintaining Springer’s high standards of academic excellence. Also considered for publication are lecture materials, proceedings, and other related materials of exceptionally high quality and interest. The subject matter should be original and timely, reporting the latest research and developments in all areas of electrical engineering.

The audience for the books in LNEE consists of advanced level students, researchers, and industry professionals working at the forefront of their fields. Much like Springer’s other Lecture Notes series, LNEE will be distributed through Springer’s print and electronic publishing channels.

More information about this series at <http://www.springer.com/series/7818>

Kuinam Kim · Nikolai Joukov
Editors

Information Science and Applications 2017

ICISA 2017

 Springer

Editors

Kuinam Kim
Kyonggi University
Seongnam-si, Gyeonggi
Korea, Republic of

Nikolai Joukov
modelizeIT Inc., CEO and NYU
Stony Brook, NY
USA

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-10-4153-2 ISBN 978-981-10-4154-9 (eBook)
DOI 10.1007/978-981-10-4154-9

Library of Congress Control Number: 2017934217

© Springer Nature Singapore Pte Ltd. 2017

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, express 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.

Printed on acid-free paper

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

Preface

This LNEE volume contains the papers presented at the iCatse International Conference on Information Science and Applications (ICISA 2017) which was held in Macau, China, during March 20–23, 2017.

ICISA2017 will be an excellent international conference for sharing knowledge and results in Information Science and Application. The aim of the conference is to provide a platform to the researchers and practitioners from both academia and industry to meet and share the cutting-edge developments in the field.

The primary goal of the conference is to exchange, share and distribute the latest research and theories from our international community. The conference will be held every year to make it an ideal platform for people to share views and experiences in Information Science and Application related fields.

On behalf of the Organizing Committee, we would like to thank Springer for publishing the proceedings of ICISA2017. We would also like to express our gratitude to the ‘Program Committee and Reviewers’ for providing extra help in the review process. The quality of a refereed volume depends mainly on the expertise and dedication of the reviewers. We are indebted to the Program Committee members for their guidance and coordination in organizing the review process, and to the authors for contributing their research results to the conference.

Our sincere thanks to the Institute of Creative Advanced Technology, Engineering and Science for designing the conference web page and also spending countless days in preparing the final program during the time for printing. We would also like to thank our organization committee for their hard work in sorting our manuscripts from our authors.

We look forward to seeing all of you next year at ICISA.

Kuinam J. Kim
Nikolai Joukov

Organization

General Chairs

Nikolai Joukov New York University and modelizeIT Inc, USA
Nakhoon Baek Kyungpook National University, Republic of Korea

Steering Committee

Nikolai Joukov	New York University and modelizeIT Inc, USA
Borko Furht	Florida Atlantic University, USA
Bezalel Gavish	Southern Methodist University, USA
Kin Fun Li	University of Victoria, Canada
Kuinam J. Kim	Kyonggi University, Korea
Naruemon Wattanapongsakorn	King Mongkut's University of Technology Thonburi, Thailand
Xiaoxia Huang	University of Science and Technology Beijing, China
Dong-Seong (Dan) Kim	University of Canterbury, New Zealand
Nakhoon Baek	Kyungpook National University, Republic of Korea

Publicity Chairs

Hongseok Jeon	ETRI, Republic of Korea
Tomas Cerny	Czech Technical University, Czech Republic
Naruemon Wattanapongsakorn	King Mongkut's University of Technology Thonburi, Thailand
Suresh Thanakodi	National Defence University of Malaysia, Malaysia

Workshop Chair

Dong-Seong (Dan) Kim University of Canterbury, New Zealand

Publication Chair

Kyoungho Choi Institute of Creative Advanced Technologies,
Science and Engineering

Program Chair

Kuinam J. Kim Kyonggi University, Republic of Korea

Financial Chairs

WonHyung Park Institute of Creative Advanced Technologies,
Science and Engineering

Sanggyoon Oh BPU Holdings Corp, Republic of Korea

Organizers and Supporters

Institute of Creative Advanced Technologies, Science and Engineering (iCatse)

Czech Technical University, Czech Republic

River Publishers, Netherlands

Korean Industry Security Forum (KISF)

Korea Convergence Security Association (KCSA)

Kyonggi University, Korea

King Mongkut's University of Technology Thonburi, Thailand

National Defence University of Malaysia, Malaysia

University of Canterbury, New Zealand

University of Science and Technology Beijing, China

Electronics and Telecommunications Research Institute (ETRI)

Korea Institute of Science and Technology Information (KISTI)

Kyungpook National University, Republic of Korea

Program Committee

Ahmad Kamran Malik	COMSATS Institute of Information Technology, Pakistan
Ahmed Adel Abdelwahab	Qassim University, Saudi Arabia
Alessandro Bianchi	University of Bari, Italy
Alexandar Djordjevich	City University of Hong Kong, China
Ana Paula Ferreira Barroso	Universidade NOVA de Lisboa, Portugal
Andrea Ceccarelli	University of Florence, Italy
Baojun Ma	Beijing University of Posts and Telecommunications, China
Bernd E. Wolfinger	University of Hamburg, Germany
Bongkyo Moon	Dongguk University, Korea
Bruno Defude	Université Paris-Saclay, France
Chi-Man Pun	University of Macau, China
Chin-Chen Chang	Feng Chia University, Taiwan
Chittaranjan Pradhan	KIIT University, India
Christian Prehofer	Technische Universitaet Muenchen, Germany
Christof Ebert	Vector Consulting Services GmbH, Germany
Chuanyin Dang	City University of Hong Kong, China
Claudia Linnhoff-Popien	Ludwig-Maximilians Universitat Munchen, Germany
Dabin Ding	University of Central Missouri, USA
Dan Lin	Missouri University of Science and Technology, USA
Daniel B.-W. Chen	Monash University, Australia
David Liu	Indiana University, USA
David Naccache	Ecole normale supérieure, France
Dennis Pfisterer	University of Luebeck, Germany
Derek Doran	Wright State University, USA
Edward Chlebus	Illinois Institute of Technology, USA
Filippo Gaudenzi	Università degli Studi di Milano, Italy
George Okeyo	Jomo Kenyatta University of Agriculture and Technology, Kenya
Haralambos Mouratidis	University of Brighton, UK
Harikumar Rajaguru	Bannari Amman Institute of Technology, India
Helmi Zulhaidi Mohd Shafi	Universiti Putra Malaysia, Malaysia
Heming Cui	University of Hong Kong, China
Hing Keung LAU	The Open University of Hong Kong, China
Hironori Washizaki	Waseda University, Japan
Huai Liu	RMIT University, Australia
Hyoungshick Kim	Sungkyunkwan University, Korea
Hyunsung Kim	Kyungil University, South Korea
Ireneusz Czarnowski	Gdynia Maritime University, Poland

James Braman	The Community College of Baltimore County, USA
Jianbin Qiu	Harbin Institute of Technology, China
Jie Zhang	Newcastle University, UK
Jitender Grover	MMU, India
Johann M. Marquez-Barja	Trinity College Dublin, Ireland
José Manuel Matos Ribeiro da Fonseca	NOVA University of Lisbon, Portugal
Kai Liu	Chongqing University, China
Kittisak Jermstittiparsert	Rangsit University, Thailand
Longzhi Yang	Northumbria University, UK
Maicon Stihler	Federal Center for Technological Education of Minas Gerais, Brazil
Mainguenau Michel	INSA Rouen-Normandie, France
Manik Sharma	DAV University, India
Marco Listanti	University Sapienza, Italy
Marco Mesiti	Università degli Studi di Milano, Italy
Massimo Mecella	SAPIENZA Università di Roma, Italy
Mauro Gaggero	National Research Council of Italy, Italy
Michalis Pavlidis	University of Brighton, UK
Min-Shiang Hwang	Asia University, Taiwan
Mohammed M Kadhum	Queen's University, Canada
Mohd Ashraf Ahmad	Universiti Malaysia Pahang, Malaysia
Mohd Faizal Abdollah	University Technical Malaysia Melaka, Malaysia
Mohd Helmy Abd Wahab	Universiti Tun Hussein Onn Malaysia, Malaysia
Mounir Arioua	Abdelmalek Essaadi University, Morocco
Mukesh Kumar Saini	Indian Institute of Technology Ropar, India
Nasir Uddin	Oracle Inc, USA
Ng Hui Fuang	Universiti Tunku Abdul Rahman, Malaysia
Nik Bessis	Edge Hill University, UK
Óscar Mortágua Pereira	University of Aveiro, Portugal
Pablo Casaseca	University of the West of Scotland, UK
Pascal Lorenz	University of Haute Alsace, France
Pinar Kirci	Istanbul University, Turkey
Qi Shi	Liverpool John Moores University, UK
Qiao Xiang	Yale University, USA
Ramesh Rayudu	Victoria University of Wellington, New Zealand
Reza Malekian	University of Pretoria, South Africa
Robert S. Laramée	Swansea University, UK
Rossitza Ivanova Goleva	Technical University of Sofia, Bulgaria
Sakgasit Ramingwong	Chiang Mai University, Thailand
Sandro Leuchter	Hochschule Mannheim University of Applied Sciences, Germany
Sangseo Park	The University of Melbourne, Australia
Sathiamoorthy Manoharan	University of Auckland, New Zealand

Selma Regina Martins Oliveira	Federal Fluminense University, Brazil
Seung Yeob Nam	Yeungnam University, Korea
Shamim H Ripon	East West University, Bangladesh
Shao Jie	University of Electronic Science and Technology of China, China
Sharmistha P. Chatterjee	Florida Atlantic University, USA
Shimpei Matsumoto	Hiroshima Institute of Technology, Japan
Shitala Prasad	CNRS, France
Shuxiang Xu	University of Tasmania, Australia
Sira Yongchareon	Auckland University of Technology, New Zealand
Somlak Wannarumon Kielarova	Naresuan University, Thailand
Stelvio Cimato	Università degli studi di Milano, Italy
Suksan Prombanpong	King Mongkut's University of Technology Thonburi, Thailand
Susan Mengel	Texas Tech University, USA
Syh-Yuan Tan	Multimedia University, Malaysia
Terje Jensen	Telenor, Norway
Tor-Morten Grønli	Westerdals, Norway
Toyohide Watanabe	Nagoya Industrial Science Research Institute, Japan
Vasco Soares	Instituto de Telecomunicação, Portugal
Vasilis Friderikos	King's College London, UK
Vishwas Rudramurthy	Visvesvaraya Technological University, India
Wojciech Giernacki	Pozan University of Technology, Poland
Wolfgang A. Halang	Fernuniversitaet, Germany
Wun-She Yap	Universiti Tunku Abdul Rahman, Malaysia
Ximing Fu	Tsinghua University, China
Yanjun Liu	Feng Chia University, Taiwan
Yanling Wei	Technical University of Berlin, Germany
Yin-Fu Huang	National Yunlin University of Science and Technology, Taiwan
Yuen Chau	Singapore University of Technology and Design, Singapore
Zeeshan Ali Rana	University of Central Punjab, Pakistan
Zhiyong Shan	University of Central Missouri, USA
Zhiyuan Chen	University of Maryland, Baltimore County, USA
Zhiyuan Hu	Nokia Shanghai Bell, China
Zulfiqar Habib	COMSATS Institute of Information Technology, Pakistan

Contents

Ubiquitous Computing

Improving Performance and Energy Efficiency for OFDMA Systems Using Adaptive Antennas and CoMP	3
Yapeng Wang, Xu Yang, Laurie Cuthbert, Tiankui Zhang, and Lin Xiao	
Privacy in Location Based Services: Protection Strategies, Attack Models and Open Challenges	12
Priti Jagwani and Saroj Kaushik	
An Efficient and Low-Signaling Opportunistic Routing for Underwater Acoustic Sensor Networks	22
Zhengyu Ma, Quansheng Guan, Fei Ji, Hua Yu, and Fangjiong Chen	
A Novel Mobile Online Vehicle Status Awareness Method Using Smartphone Sensors	30
Dang-Nhac Lu, Thi-Thu-Trang Ngo, Duc-Nhan Nguyen, Thi-Hau Nguyen, and Ha-Nam Nguyen	
A Study on OPNET State Machine Model Based IoT Network Layer Test	38
Young-hwan Ham, Hyo-taeg Jung, Hyun-cheol Kim, and Jin-wook Chung	
A Secure Localization Algorithm Based on Confidence Constraint for Underwater Wireless Sensor Networks	46
Xiaofeng Xu, Guangyuan Wang, Yongji Ren, and Xiaolei Liu	

Networks and Information Systems

Generating Time Series Simulation Dataset Derived from Dynamic Time-Varying Bayesian Network	53
Garam Lee, Hyunjin Lee, and Kyung-Ah Sohn	

AMI-SIM: An NS-2 Based Simulator for Advanced Metering Infrastructure Network	61
Nam-Uk Kim and Tai-Myoung Chung	
Beyond Map-Reduce: LATNODE – A New Programming Paradigm for Big Data Systems	69
Chai Yit Sheng and Phang Keat Keong	
Indoor Positioning Solely Based on User’s Sight	76
Matthias Becker	
Naming Convention Scheme for Role Based Access Control in Cloud Based ERP Platforms	84
Abed Alshreef, Lin Li, and Wahid Rajeh	
Multimedia and Visualization	
Korean/Chinese Web-Based Font Editor Based on METAFONT for User Interaction	97
Minju Son, Gyeongjae Gwon, and Jaeyeong Choi	
A Highly Robust and Secure Digital Image Encryption Technique	105
Md. Anwar Hussain, Popi Bora, and Joyatri Bora	
Saliency Based Object Detection and Enhancements in Static Images	114
Rehan Mehmood Yousaf, Saad Rehman, Hassan Dawood, Guo Ping, Zahid Mehmood, Shoaib Azam, and Abdullah Aman Khan	
A Center Symmetric Padding Method for Image Filtering	124
Mengqin Li and Xiaopin Zhong	
Implementing a Stereo Image Processing for Medical 3D Microscopes with Wireless HMD	131
Cheolhwan Kim, Jiyoung Yoon, Yun-Jung Lee, Shihyun Ahn, and Yongtaek Park	
Design of OpenGL SC 2.0 Rendering Pipeline	139
Nakhoon Baek	
Saliency Detection via Foreground and Background Seeds	145
Xiao Lin, Zhixun Yan, and Linhua Jiang	
Identification and Annotation of Hidden Object in Human Terahertz Image	155
Guiyang Yue, Zhihao Yu, Cong Liu, Hui Huang, Yiming Zhu, and Linhua Jiang	

Information Visualization for Mobile-Based Disability Test Applications 164
 Jongmun Jeong, Seungho Kim, Changsoon Kang, and Mintae Hwang

Deep Convolutional Neural Networks for All-Day Pedestrian Detection 171
 Xingguo Zhang, Guoyue Chen, Kazuki Saruta, and Yuki Terata

An Augmented Reality Learning System for Programming Concepts 179
 Kelwin Seen Tiong Tan and Yunli Lee

Middleware and Operating Systems

Efficient vCore Based Container Deployment Algorithm for Improving Heterogeneous Hadoop YARN Performance 191
 SooKyung Lee, Min-Ho Bae, Jun-Ho Eum, and Sangyoon Oh

A Real-Time Operating System Supporting Distributed Shared Memory for Embedded Control Systems 202
 Yuji Tamura, Doan Truong Thi, Takahiro Chiba, Myungryun Yoo, and Takanori Yokoyama

Security and Privacy

MBR Image Automation Analysis Techniques Utilizing Emulab 213
 Gibeom Song and Manhee Lee

Detection of DNS Tunneling in Mobile Networks Using Machine Learning 221
 Van Thuan Do, Paal Engelstad, Boning Feng, and Thanh van Do

On the Security Analysis of Weak Cryptographic Primitive Based Key Derivation Function 231
 Chai Wen Chuah, Mustafa Mat Deris, and Edward Dawson

On the Security of a Privacy Authentication Scheme Based on Cloud for Medical Environment 241
 Chun-Ta Li, Dong-Her Shih, and Chun-Cheng Wang

Physical Layer Security with Energy Harvesting in Single Hop Wireless Relaying System 249
 Poonam Jindal and Rupali Sinha

A System Design for the Measurement and Evaluation of the Communications Security Domain in ISO 27001:2013 Using an Ontology 257
 Pongsak Sirisom, Janjira Payakpate, and Winai Wongthai

Timing Side Channel Attack on Key Derivation Functions 266
 Chai Wen Chuah and Wen Wen Koh

A Security Aware Fuzzy Embedded ACO Based Routing Protocol (SAFACO) in VANETs. 274
 Hang Zhang, Xi Wang, and Dieter Hogrefe

Cryptanalysis of “An Efficient Searchable Encryption Against Keyword Guessing Attacks for Shareable Electronic Medical Records in Cloud-Based System” 282
 Chun-Ta Li, Cheng-Chi Lee, Chi-Yao Weng, Tsu-Yang Wu, and Chien-Ming Chen

eDSDroid: A Hybrid Approach for Information Leak Detection in Android 290
 Hoang Tuan Ly, Tan Cam Nguyen, and Van-Hau Pham

Detect Sensitive Data Leakage via Inter-application on Android by Using Static Analysis and Dynamic Analysis 298
 Nguyen Tan Cam, Van-Hau Pham, and Tuan Nguyen

Known Bid Attack on an Electronic Sealed-Bid Auction Scheme 306
 Kin-Woon Yeow, Swee-Huay Heng, and Syh-Yuan Tan

Perceptual 3D Watermarking Using Mesh Saliency 315
 Jeongho Son, Dongkyu Kim, Hak-Yeol Choi, Han-UI Jang, and Sunghee Choi

Perceptual Watermarking for Stereoscopic 3D Image Based on Visual Discomfort 323
 Sang-Keun Ji, Ji-Hyeon Kang, and Heung-Kyu Lee

Fingerprint Spoof Detection Using Contrast Enhancement and Convolutional Neural Networks 331
 Han-UI Jang, Hak-Yeol Choi, Dongkyu Kim, Jeongho Son, and Heung-Kyu Lee

Content Recapture Detection Based on Convolutional Neural Networks 339
 Hak-Yeol Choi, Han-UI Jang, Jeongho Son, Dongkyu Kim, and Heung-Kyu Lee

Secret Sharing Deniable Encryption Technique 347
 Mohsen Mohamad Hata, Fakariah Hani Mohd Ali, and Syed Ahmad Aljunid

Improved 3D Mesh Steganalysis Using Homogeneous Kernel Map 358
 Dongkyu Kim, Han-UI Jang, Hak-Yeol Choi, Jeongho Son, In-Jae Yu, and Heung-Kyu Lee

From Sealed-Bid Electronic Auction to Electronic Cheque 366
 Kin-Woon Yeow, Swee-Huay Heng, and Syh-Yuan Tan

Enhanced Database Security Using Homomorphic Encryption 377
 Connor Røset, Van Warren, and Chia-Chu Chiang

***k*-Depth Mimicry Attack to Secretly Embed Shellcode into
 PDF Files** 388
 Jaewoo Park and Hyounghick Kim

Reconstruction of Task Lists from Android Applications 396
 Xingmin Cui, Ruiyi He, Lucas C.K. Hui, S.M. Yiu, Gang Zhou,
 and Eric Ke Wang

**Design and Evaluation of Chaotic Iterations Based Keyed
 Hash Function** 404
 Zhuosheng Lin, Christophe Guyeux, Simin Yu, and Qianxue Wang

Data Mining and Artificial Intelligence

**Intellectual Overall Evaluation of Power Quality Including
 System Cost** 417
 Buhm Lee, Dohee Sohn, and Kyoung Min Kim

**A Data-Driven Decision Making with Big Data Analysis
 on DNS Log** 426
 Euihyun Jung

A Case-Based Approach to Colorectal Cancer Detection 433
 Pedro Morgado, Henrique Vicente, António Abelha, José Machado,
 João Neves, and José Neves

**Multi-Modes Cascade SVMs: Fast Support Vector Machines
 in Distributed System** 443
 Lijuan Cui, Changjian Wang, Wanli Li, Ludan Tan, and Yuxing Peng

Deep Learning Based Recommendation: A Survey 451
 Juntao Liu and Caihua Wu

Towards Collaborative Data Analytics for Smart Buildings 459
 Sanja Lazarova-Molnar and Nader Mohamed

**English and Malay Cross-lingual Sentiment Lexicon Acquisition
 and Analysis** 467
 Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Azreen Azman,
 and Rabiah Abdul Kadir

A Novel Natural Language Processing (NLP) Approach to Automatically Generate Conceptual Class Model from Initial Software Requirements. 476
Mudassar Adeel Ahmed, Wasi Haider Butt, Imran Ahsan, Muhammad Waseem Anwar, Muhammad Latif, and Farooque Azam

The Applications of Natural Language Processing (NLP) for Software Requirement Engineering - A Systematic Literature Review. 485
Farhana Nazir, Wasi Haider Butt, Muhammad Waseem Anwar, and Muazzam A. Khan Khattak

Smart Fetal Monitoring 494
Jane You, Qin Li, Zhenhua Guo, and Ruohan Zhao

A Network-Based Approach on Big Data for the Comorbidities of Urticaria 504
Yi-Horng Lai, Chih-Chiang Ho, and Piao-Yi Chiou

Application of Automated Theorem-Proving to Philosophical Thought: Spinoza’s *Ethics* 512
Maciej Janowicz, Luiza Ochnio, Leszek J. Chmielewski, and Arkadiusz Orłowski

Improving the B+-Tree Construction for Transaction Log Data in Bank System Using Hadoop. 519
Cong Viet-Ngu Huynh, Jongmin Kim, and Jun-Ho Huh

Calculate Deep Convolution Neural Network on Cell Unit. 526
Haofang Lu, Ying Zhou, and Zi-Ke Zhang

Stepwise Structure Learning Using Probabilistic Pruning for Bayesian Networks: Improving Efficiency and Comparing Characteristics 533
Godai Azuma, Daisuke Kitakoshi, and Masato Suzuki

Differential-Weighted Global Optimum of BP Neural Network on Image Classification. 544
Lin Ma, Xiao Lin, and Linhua Jiang

Classification Model for Skin Lesion Image 553
Nontachai Danpakdee and Wararat Songpan

Software Engineering

Generation of Use Cases for Requirements Elicitation by Stakeholders. 565
Junko Shirogane

Smart Learner-Centric Learning Systems. 577
Naseem Ibrahim and Ismail I.K. Al Ani

Prioritized Process Test: More Efficiency in Testing of Business Processes and Workflows 585
Miroslav Bures, Tomas Cerny, and Matej Klima

Static Testing Using Different Types of CRUD Matrices 594
Miroslav Bures and Tomas Cerny

Extracting Test Cases with Message-Sequence Diagram for Validating the Photovoltaic Energy Integrated Monitoring System. 603
Woo Sung Jang, Bo Kyung Park, Hyun Seung Son, Byung Kook Jeon, and R. Young Chul Kim

Automatic Test Case Generation with State Diagram for Validating the Solar Integrated System 609
Bo Kyung Park, Woo Sung Jang, Hyun Seung Son, Keunsang Yi, and R. Young Chul Kim

Comparison of Software Complexity Metrics in Measuring the Complexity of Event Sequences 615
Johanna Ahmad and Salmi Baharom

Implementation of Ceph Storage with Big Data for Performance Comparison. 625
Chao-Tung Yang, Cai-Jin Chen, and Tzu-Yang Chen

Web Technology

Predicting Engaging Content for Increasing Organic Reach on Facebook 637
Natthaphong Phuntusil and Yachai Limpiyakorn

Learning Performance Evaluation in eLearning with the Web-Based Assessment 645
Cheng-Ying Yang, Tsai-Yuan Chung, Min-Shiang Hwang, Cheng-Yi Li, and Jenq-Foung JF Yao

Improving Teaching and Learning in Southeast Asian Secondary Schools with the Use of Culturally Motivated Web and Mobile Technology 652
Sithira Vadivel, Insu Song, and Abhishek Singh Bhati

Game-Based Learning to Teach Assertive Communication ClickTalk for Enhancing Team Play 660
Bah Tee Eng

Cloud Storage Federation as a Service Reference Architecture. 668
Rene Ivan Heinsen, Cindy Pamela Lopez, Tri D.T. Nguyen, and Eui-Nam Huh

Internet of Things

- A Study on the IoT Framework Design for Ginseng Cultivation** 679
Kyung-Gyun Lim and Chang-Geun Kim
- An IPS Evaluation Framework for Measuring the Effectiveness and Efficiency of Indoor Positioning Solutions** 688
Jacqueline Lee Fang Ang, Wai Kong Lee, Boon Yaik Ooi, and Thomas Wei Min Ooi
- An IoT-Based Virtual Addressing Framework for Intelligent Delivery Logistics** 698
Omar Hiari, Dhiah el Diehn I. Abou-Tair, and Ismail Abushaikha
- Context-Aware Security Using Internet of Things Devices** 706
Michal Trnka, Martin Tomasek, and Tomas Cerny
- An Energy-Efficient Transmission Framework for IoT Monitoring Systems in Precision Agriculture** 714
Peerapak Lerdsuwan and Phond Phunchongharn
- Piezoelectric Voltage Monitoring System Using Smartphone** 722
Nazatul Shiema Moh Nazar, Suresh Thanakodi, Azizi Miskon, Siti Nooraya Mohd Tawil, and Muhammad Syafiq Najmi Mazlan
- A System for Classroom Environment Monitoring Using the Internet of Things and Cloud Computing** 732
Wuttipong Runathong, Winai Wongthai, and Sutthiwat Panithansuwan
- 4th Convergence of Healthcare and Information Technology**
- Research on Design of End Site Architecture to Connect LHCONE in KREONET** 745
Chanjin Park, Wonhyuk Lee, Kuinam J. Kim, and Hyuncheol Kim
- A Study of Children Play Educational Environment Based on u-Healthcare System** 751
Minkyu Kim, Soojung Park, and Byungkwon Park
- Virtual Resources Allocation Scheme in ICT Converged Networks** 757
Hyuncheol Kim
- Enhanced Metadata Creation and Utilization for Personalized IPTV Service** 763
Hyojin Park, Kireem Han, Jinhong Yang, and Jun Kyun Choi
- A Study of Teaching Plan for the Physical Activity Using ICT** 770
Seung Ae Kang

Design and Implementation of Headend Servers for Downloadable CAS 777
Soonchoul Kim, Hyuncheol Kim, and Jinwook Chung

A Method of Modeling of Basic Big Data Analysis for Korean Medical Tourism: A Machine Learning Approach Using Apriori Algorithm 784
Jun-Ho Huh, Han-Byul Kim, and Jinmo Kim

The Study of Application Development on Elderly Customized Exercise for Active Aging 791
YoungHee Cho, SeungAe Kang, SooHyun Kim, and SunYoung Kang

Improving Jaccard Index for Measuring Similarity in Collaborative Filtering 799
Soojung Lee

Temperature Recorder System 807
Suresh Thanakodi, Nazatul Shiema Moh Nazar, Azizi Miskon, Ahmad Mujahid Ahmad Zaidi, and Muhammad Syafiq Najmi Mazlan

7th International Workshop on ICT Convergence

The Emergence of ICTs for Knowledge Sharing Based on Research in Indonesia 817
Siti Rohajawati, Boy Iskandar Pasaribu, Gun Gun Gumilar, and Hilda Rizanti Putri

Quality of Transformation of Knowledge as Part of Knowledge Management System 827
Dyah Budiastuti and Harjanto Prabowo

Author Index 835

Ubiquitous Computing

Improving Performance and Energy Efficiency for OFDMA Systems Using Adaptive Antennas and CoMP

Yapeng Wang¹✉, Xu Yang², Laurie Cuthbert¹, Tiankui Zhang³, and Lin Xiao⁴

¹ MPI-QMUL Information Systems Research Centre, Macao Polytechnic Institute, Macao SAR, China

{yapeng.wang, laurie.cuthbert}@isrcmo.org

² School of Public Administration, Macao Polytechnic Institute, Macao SAR, China
xuyang@ipm.edu.mo

³ Beijing University of Posts and Telecommunications, Beijing, China
tkzhang@gmail.com

⁴ Nanchang University, Nanchang, China
xiaolin910@gmail.com

Abstract. The research described in this paper shows how by combining CoMP with adaptable semi-smart antennas it is possible to improve the throughput of an OFDMA LTE system and at the same time reduce the power required for transmission; i.e. providing better performance at lower cost. Optimisation of antenna patterns is performed by a Genetic Algorithm to satisfy an objective function that considers throughput, number of UEs handled as well as power in the transmission. It is shown that the dominant effect is using the semi-smart antennas and that the results are not sensitive to small amounts of movement.

Keywords: OFDMA · CoMP · Adaptive antennas · Genetic Algorithm

1 Introduction

One of the well-known key principles of OFDMA is that the orthogonality within the cell eliminates intra-cell interference so that the interference is from neighbouring cells – inter-cell interference. A key technology to mitigate this effect is Multi-Input Multi-Output (MIMO), which can improve users' throughput performance significantly. It utilises time and space diversity to increase the number of communication channels between the base station (BS) and user equipment (UE). Multiple antennas are used at both ends of the link to create space diverse channels. However the usage of MIMO is limited to one base station. With synchronising between BSs, it is possible to utilise two or more BSs to transmit signals to one UE using one frequency channel without interference by utilising time and space diversity. This technology is also called Coordinated Multipoint Transmission (CoMP) [1].

In this study, a cross layer approach is proposed to mitigate the inter-cell interference (ICI) and improve energy efficiency at the same time. We combine CoMP, adaptive antenna and Artificial Intelligence (AI) technology to handle complex resource allocation problems. Radio resources are allocated to cell-edge users in an efficient and co-operative

manner. Energy costs at RF components are reduced as an optimised radio radiation pattern is produced and power is radiated towards UEs in an efficient way. In addition, it can be easily integrated into current networks to provide extra system throughput and energy efficiency. The concept was published by the authors in a Letter [2] but this paper expands the work reported there and provides a more detailed description of the technology.

In the real world, users are not uniformly distributed and this has an impact on performance. For example, in a scenario where one cell is lightly loaded (with few active UEs) and its neighbouring cells are heavy loaded (with many UEs), the free radio resources in the lightly loaded cell cannot be utilised to serve neighbouring cells. However, by using adaptive antennas to change the cell coverage, as described in this paper, some of the UEs in the highly loaded BS can be handed over to neighbouring BSs to achieve load balancing.

2 CoMP and Semi-smart Antennas

CoMP [1, 3] (also called distributed MIMO) is a technology that features joint processing to change the interference signal into a useful signal. This technology utilises a distributed multi-antenna channel to improve transmission diversity gain or spatial multiplexing gain. This can effectively mitigate inter-cell interference to improve link throughput and reliability for cell-edge users.

CoMP technology can be divided into uplink multi-point reception and downlink multi-point transmission. In this research, we only consider downlink CoMP. Downlink CoMP transmission has two categories and in this work we use CoMP JP/JT where two or more BSs can create two parallel special channels in a co-ordinated way to serve a single cell-edge. Instead of treating the other BS's signal as interference, CoMP co-ordinates the two UE transmissions from the two different BS to make them appear as if they were distributed MIMO. Tight synchronisation and co-operation is required in order to create MIMO channel coding and this remote synchronisation and co-operation makes CoMP much more complicated than MIMO.

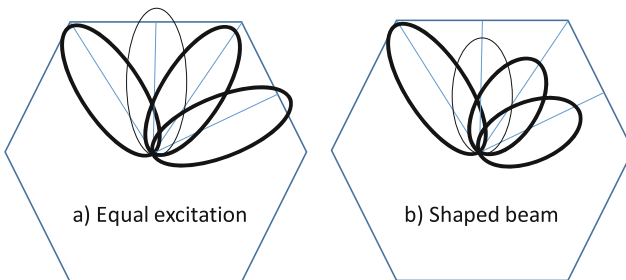


Fig. 1. Simple illustration of semi-smart antenna

Not all UEs in the system need to co-operate; those that are close to a BS or have high SINR do not need to be CoMP users: only UEs that are located at the cell edge and

suffer high interference (low SINR) require cooperation. These cell-edge users are called CoMP users. In this study, if a UE's SINR is lower than a certain threshold it is treated as CoMP user.

The adaptive antenna system (also called semi-smart antenna) used in this study is described in [4] and illustrated simply in Fig. 1.

A BS is equipped with 3 antenna sectors and each sector has 4 elements. Every individual antenna element is controlled by a power amplifier and the power level for each element can be individually controlled. By changing the gain and phase between the elements the overall pattern for that sector can be modified. In this work only the gain is changed. As there is no expensive DSP needed to track individual UEs, the cost of the adaptive system is much lower compared with a fully adaptive antenna system. The EU project SHUFFLE (IST-1999-11014) constructed such an antenna to demonstrate its feasibility.

With CoMP technology, the antenna pattern plays a key role in improving system performance as changing the antenna pattern can effectively change the channel conditions between the BSs and the UE, thereby affecting the UE throughput. Figure 2 illustrates this process: in (a) the two channels are not optimised for CoMP but changing the antenna patterns as in (b) provides overlap allowing better CoMP performance.

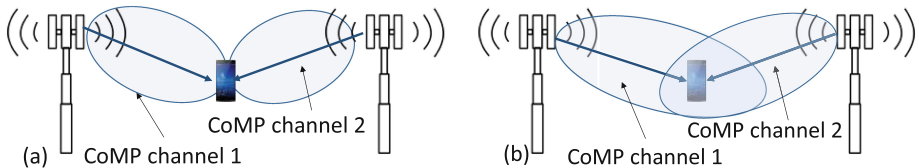


Fig. 2. Changing antenna patterns to improve cell-edge CoMP

In a multiple user scenario, the calculation of the optimal antenna pattern for all UEs is extremely difficult and almost impossible, so we use a genetic algorithm (GA) to find the optimised coverage pattern for each BS by adjusting each antenna elements' power levels. GAs [5] are known to be an effective search algorithm to find near optimal solution for many fields where little knowledge is known or there are many conflicting constraints or objectives to affect best solutions. It is acknowledged that running a GA is not fast and it is unlikely to be suitable for real-time deployment, but the results do serve as a benchmark against which other, faster, algorithms may be compared.

3 Genetic Algorithm Configuration

3.1 Encoding and Decoding

A gain vector $\mathbf{G} = [g_1, g_2, \dots, g_N]$ represents the antenna gains along N directions and in our scheme each gain value is coded as a gene. This determines one antenna beam-forming pattern. If we have M base stations, one chromosome is represented as, $[\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_M]$, which has $N \times M$ genes. Each gene is a real number with value between 0 and 1 corresponding to the minimum to maximum coverage.

We use a real-coded genetic algorithm with *BLX* – *a* crossover operator [6], distortion crossover operator, simple random mutation operator, and creep operator. The *BLX* – *a* crossover operator is widely used in real-coded GAs and has been shown to achieve very good performance in many applications [6]. The distortion crossover operator means doing a crossover between two gain values at different directions in one antenna pattern. Simple random mutation replaces randomly selected gain values with random values from the appropriate valid range. In the creep operator, gain values which are to be mutated are incremented or decremented by the creep fraction, which is a small fraction of the valid range for the attribute value [7]. Additionally Elitism [7] of a few best chromosomes, which are copied to the next population, is used to prevent losing the best found solutions to date.

3.2 Fitness Function

One general approach for solving multi-objective optimisation problems is to combine the individual objective functions into a single composite function such as the weighted sum method [8]. This is simple yet probably the most widely used classical approach.

The objective function is defined as:

$$o = \alpha \times TotalLoad + \beta \times HandledUEs + \delta \times TotalRFPower$$

TotalLoad is the sum of the traffic load in all base stations.

HandledUEs is the number of UEs receiving service.

$TotalRFPower = \sum_{i=1}^M \sum_{j=1}^N g_{ij}^2$ where *M* is the number of base stations, *N* the number of directions at each base station and *g* the antenna gain at BS *i* in direction *j*.

α, β, δ are the weights given to each objective, and their values are set (i) to make the order of each weighted objective the same (otherwise one would dominate) and (ii) to set a priority for the objectives. This may be considered to be somewhat arbitrary and a weakness of the weighted sum method, but it is easy to determine a weight to make each objective the same order of magnitude and the modification of those values to set the priority between objectives can be validated by simple experiments.

We assume that a network provider would first consider providing services to all the users in the service area but also aim to achieve maximum traffic load. The least important objective is to minimizing the transmitting power, not that saving power is unimportant, but it was given top priority there would be a tendency to reduce the power to levels where the user requirements would not be satisfied.

3.3 Constraint Handling

This work includes the constraint that there must be no gaps (no “holes”) in coverage from the BSs. As the weighted sum method combines multi-objectives into a single objective, many of constraint handling methods investigated for a single objective can be directly applied into our case. This is one reason why the weighted sum method was chosen, despite the apparently arbitrary choice of weights.

In this paper, we use the superiority of feasible points method [9, 10] to handle coverage constraint, as this approach has shown promise in many GA applications even

when the feasible region is quite small compared with whole search space. The concept is that feasible solutions have superiority over infeasible ones, and infeasible solutions are penalised to provide a search direction towards the feasible region [9, 10].

A penalty value measures the constraints violation and is added to the objective function value of the worst feasible solution in the last population. Thus, the fitness of an infeasible solution not only depends on the amount of constraint violation, but also depends on the feasible solutions in the population.

To measure the coverage constraint violation, the network makes a grid of check points (uniformly distributed in the constraint coverage area) and examines whether all the points are covered. The penalty is the number of uncovered points.

3.4 Simulation Parameters

A conventional multi-cell system level simulation is constructed for downlink multi-cell OFDMA. The baseline system is the conventional non-cooperative system with SVD (Singular Value Decomposition) precoding and MMSE (Minimum Mean Square Error) receiving and all subcarriers are transmitted with equal power. A TDD frame structure is used, the length of radio frame is 10 ms and the length of subframe is 1 ms. The detailed simulation parameters are listed in Table 1.

Table 1. Simulation parameters.

Parameters	Value
Layout	7 sites with 3 sectors each
Number of loops per GA loop	30 TTI
Inter cell distance	500 m
Carrier frequency	2.0 GHz
Bandwidth	10 MHz
Average number of users per sector	20
SINR threshold for determining CoMP	0.3 dB
Number of RBs per sector	10
No. of TX- and RX-antennas per RRU	2
Antenna gain	12 individual controlled gain for each BS. Each gain is between -6 to 14 dBi
Traffic	Full Buffer
Penetration loss	20 dB
Frequency reuse	1
Path loss	$128.1 + 37.6\lg(d)$, where d is in km, Minimum 70 dB
Scheduling method	Proportional Fairness (PF)
Channel	SCM-E

To evaluate the performance gain of the system with single user CoMP, a fixed CoMP region with 3 sectors is used, and 2 Resource Blocks (RBs) are reserved for CoMP transmission for cell-edge users. The cell-edge users are decided by the large-scale fading SINR threshold. A local precoding scheme is used in each CoMP region.

A very important decision to make in implementing a GA to solve a particular problem is to set up the values for the various parameters such as population size, crossover rate, mutation rate, creep rate and elitism rate. Discussions of parameter settings feature widely in the evolutionary computation literature, but there are no conclusive results on what is best: most people use what has worked well in previously reported cases [5]. The parameter settings for the GA here broadly follow those in [11–13] and are summarized in Table 2.

Table 2. GA parameters

Parameter	Value
Total generations	100
Number of policies in each generation	50
BLX-a: $a = 0.5$	0.2
Distortion rate	0.3
Random mutation rate	0.012
Creep rate	0.012
Elitism rate	0.1

4 Simulation Result

Tests to check the stability of the GA were performed with up to 100 generations, each with 50 policies. The best policy for each generation was recorded for further analysis. Figure 3 shows the system throughput (total throughput in Mbps for seven BSs) and RF power performance (a relative value for total base stations’ RF amplifiers) for the best policy in each generation with CoMP. We can see that stability is reached after about after 60 generations. Similar results were obtained for a system without CoMP.

To evaluate the performance gain or our approach we consider four scenarios and the results are shown in Fig. 4.

- i. Conventional network: no CoMP and fixed antenna patters (no GA);
- ii. No CoMP with GA (as in [14])
- iii. CoMP with fixed antenna patterns
- iv. CoMP with GA

Scenario 1: the network has “fixed antenna patterns”, the RF power was varied from minimum to maximum within the adjustable range (–6 to 14 dB for each sector), and the system throughput change noted. In this conventional network, the system throughput reaches a peak rate of 549 Mbps when maximum power is used for all BS antennas. This is shown in Fig. 4.

Scenario 2: with the GA and adaptive antennas, the system can find a near optimal solution for that user distribution. In Fig. 4 the optimum gives a system throughput of 566 Mbps, (3% performance gain over the system with maximum fixed antenna gain), but the RF power needed to produce the antenna pattern is reduced by about 58%. This is because with the GA-calculated optimal antenna pattern, the UEs can get the best SINR performance without wasting antenna radiation in other directions.

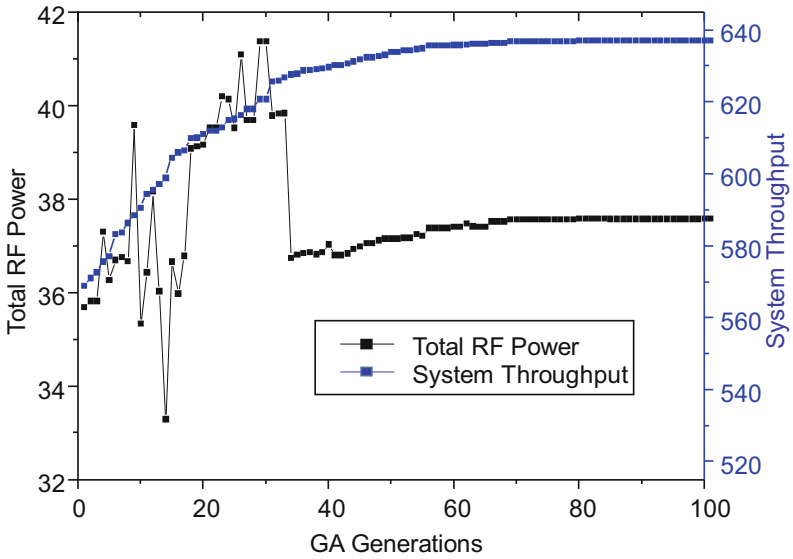


Fig. 3. System throughput and RF power for different GA generations for system with CoMP.

Scenario 3: as CoMP is mainly used for cell-edge users to improve their throughput performance, we considered mainly the cell-edge users. Without the GA, a 31% performance gain is achieved for cell-edge users in terms of throughput; at the same time, the transmission power for these cell edge users is reduced by 25% in total because, in CoMP

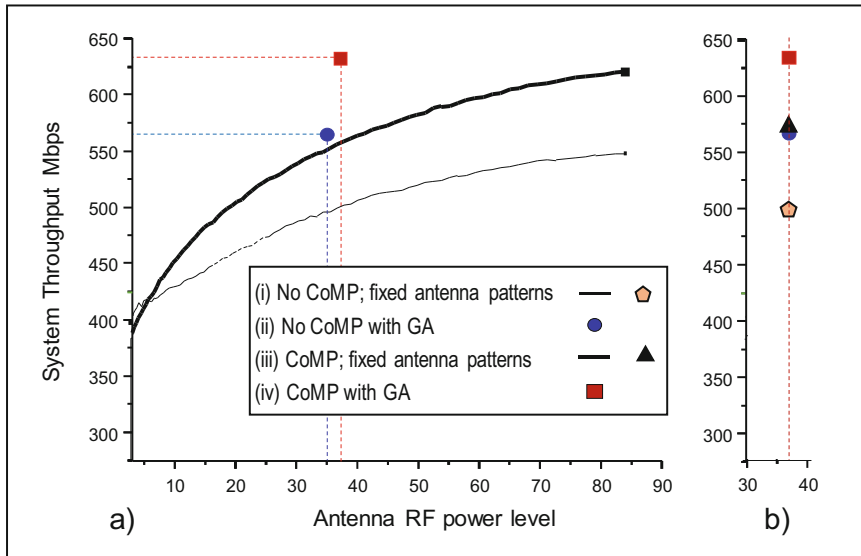


Fig. 4. System throughput against RF Power for different scenarios.

mode, the signal from a neighbouring cell is transformed from noise to useful signal, so the power required to achieve a certain throughput is reduced.

Scenario 4: the performance of a system combining GA with CoMP is also shown in Fig. 4. We can see that with GA and adaptive antenna, the system benefits with higher system throughput and much less RF antenna power required, although slightly higher power than in Scenario (ii).

In a real, deployed conventional system, the RF power is often set to a suitable level that just covers the whole service area (if it is too small it will not cover all the area and if it is too high it will cause more interference to neighbouring cells). In our simulation, with the same RF power level, the GA and adaptive antenna system offers clearly superior performance, as shown in Fig. 4(b) where the performance is shown using the power that is optimum for the GA with CoMP.

5 Conclusion

In this paper, we propose a GA based adaptive antenna system that works with CoMP for OFDMA networks. The results show that CoMP delivers a significant performance gain. However by combining CoMP with adaptive antennas whose patterns are determined by a GA optimisation, the system achieves further performance gain in terms of overall system throughput and energy efficiency. Further studies will be carried out to investigate more scenarios and real system integration issues.

References

1. Marsch, P., Fettweis, G.P.: *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge University Press, New York (2011)
2. Yang, X., Wang, Y., Zhang, T., Cuthbert, L., Xiao, L.: Combining CoMP with semi-smart antennas to improve performance. *Electron. Lett.* **47**(13), 775–776 (2011)
3. Zhou, S., Zhao, M., Xu, X., Wang, J., Yao, Y.: Distributed wireless communication system: a new architecture for future public wireless access. *IEEE Commun. Mag.* **41**(3), 108–113 (2003)
4. Nahi, P., Parini, C.G., Papadopoulos, S., Du, L., Bigham, J., Cuthbert, L.: A semi-smart antenna concept using real-time synthesis for use in a distributed load balancing scheme for cellular networks. In: *Antennas and Propagation, (ICAP 2003)*, vol. 1, pp. 168–171 (2003)
5. Mitchell, M.: *An Introduction to Genetic Algorithms*. MIT Press, Cambridge (1998)
6. Carroll, D.L.: Chemical laser modeling with genetic algorithms. *AIAA J.* **34**(2), 338–346 (1996)
7. Konak, A., Coit, D.W., Smith, A.E.: Multi-objective optimization using genetic algorithms: a tutorial. *Reliab. Eng. Syst. Saf.* **91**(9), 992–1007 (2006)
8. Deb, K.: *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, New York (2001)
9. Deb, K.: An efficient constraint handling method for genetic algorithms. *Comput. Method Appl. Mech. Eng.* **186**(2), 311–338 (2000)
10. Powell, D., Skolnick, M.M.: Using genetic algorithms in engineering design optimization with non-linear constraints. In: *5th International conference on Genetic Algorithms*, pp. 424–431 (1993)

11. Mehra, M., Jayalal, M.L., Arul, J., Rajeswari, S., Kuriakose, K.K., Murty, S.A.V.S.: Study on different crossover mechanisms of genetic algorithm for test interval optimization for nuclear power plants. *Int. J. Intell. Syst. Appl.* **6**(1), 20 (2013)
12. DeJong, K.: An analysis of the behavior of a class of genetic adaptive systems. Ph.D. thesis, University of Michigan (1975)
13. Grefenstette, J.J.: Optimization of control parameters for genetic algorithms. *IEEE Trans. Syst. Man Cybern.* **16**(1), 122–128 (1986)
14. Yang, X., Wang, Y., Zhang, D., Cuthbert, L.G.: Resource allocation in LTE OFDMA systems using genetic algorithm and semi-smart antennas. In: *Wireless Communications and Networking Conference (WCNC)*, pp. 1–6. IEEE, Sydney (2010)

Privacy in Location Based Services: Protection Strategies, Attack Models and Open Challenges

Priti Jagwani^{1(✉)} and Saroj Kaushik²

¹ School of IT, IIT Delhi, New Delhi 110016, India
jagwani.priti@gmail.com

² Department of Computer Science and Engineering, IIT Delhi, New Delhi 110016, India
saroj@cse.iitd.ac.in

Abstract. The increasing capabilities of position determination technologies (e.g., GPS) in mobile and hand held device facilitates the widespread use of Location Based Services (LBS). Although LBSs are providing enhanced functionalities and convenience of ubiquitous computing, they open up new vulnerabilities that can be exploited to target violation of security and privacy of users. For these applications to perform, location of the individual/user is required. Consequently they may pose a major privacy threat on its users. So for LBS applications to succeed, privacy and confidentiality are key issues. “Privacy protection” has become the buzz word now days for the users of location based services. This problem has gained a considerable attention among the researcher community also. A state-of-art survey of privacy in location based services containing details of all privacy protection schemes is presented. Further, attack models and their handling mechanism are discussed in comprehensive manner. Finally, some open challenges in the area of location privacy are also demonstrated.

Keywords: Location privacy · Attack models · Privacy protection strategies · K-anonymity

1 Introduction

Extensive usage of smart Phone and hand held devices brought the ubiquitous computing on the finger tips of users. With the tremendous growth of Internet and mobile phones the term “Location based services” has become a popular term now days. The GSM Association, simply defines Location Based services (LBSs) as services that use the location of the target for adding value to the service, where the target is the “entity” to be located (not necessarily the user of the service). Applications of widely used LBSs are enquiry and information services, traffic telematics, fleet management and logistics, location based advertising, and many more.

On one hand where life has entered in a zone of comfort and convenience because of LBS, on the other hand it has given rise to many issues like privacy, pricing, data availability, and accuracy in dealing with spatial information etc. Among all the issues addressed, privacy and security of clients using the LBS, is the most critical one. On the

basis of the location information, user's movement, actions, priorities, ideologies and other information can be deduced. More precisely, therefore it can be said that location information jeopardizes user's identity and integrity [26].

This work presents classification of existing location privacy approaches. An overview of different types of attacks according to the knowledge applied by attacker is also presented. Previously, researchers in [2] present the privacy attacks based on categorization of anonymity and historical anonymity only and without real life examples. Authors in [17] presented the survey of various privacy preserving approaches but not of privacy attacks. Underlined work in [33] presented upright classification of attacks but failed to provide the mechanism to handle them. Therefore, the main contribution of this paper is a comprehensive presentation of attacks along with their handling mechanisms and also the open challenges lying in that particular area.

The rest of the paper is structured as follows: Sect. 2 contains details of privacy in LBS along with its need. Various privacy protection strategies are presented in Sect. 3. Section 4 consists of various attack models while Sect. 5 contains open challenges in the area of location privacy. Finally, the work is summarized in the conclusion section.

1.1 Location Privacy and Its Need

According to the Westin [34], Location privacy can be defined as a special type of information privacy which concerns the claim of individuals to determine for themselves when, how, and to what extent location information about them is communicated to others. Precisely, key factor of location privacy is control of location information. Location privacy is the ability to prevent unauthorized parties from learning one's current or past location. All the services and the location service provider (LSP) may not be trust worthy; therefore they could misuse the user data.

A complete LBS system comprises of various players such as content providers, network operators, virtual operators, service administrators, financial parties and other service providers etc. The user has to expose its location information against the services provided by the LSPs and by this at the same time user has a risk of disclosure of its personal information also. For obtaining a complete location based service, many parties are involved and thus the personal information of user is potentially known by many different services or content providers or other parties. Thus, proliferation of personal information among the different parties is difficult to control. This requires a sophisticated access control mechanism along with an appropriate authentication system.

The consequences of a location leak vary in terms of gravity. They results uncomfortable scariness of being watched or may cause unwanted revelations of a person's activities to actual physical harm. Moreover, it is actually awkward to be seen at certain places like a female clinic, crack house, AIDS clinic or a place related to a particular political ideology [20]. A user's location privacy is affected by two factors. One, what kind of location information service providers are storing about a user? and How long do they hold onto it? Well, intrusions in location privacy can uniquely identify users, more than their names or even their genetic profile and this malicious identification may lead to unsolicited situations penetrating into one's personal space.

2 Privacy Protection Strategies and Mechanisms

Several approaches have been proposed for protecting location privacy of a user. The fundamental idea behind all techniques is to prevent revelation of unnecessary information and to explicitly or implicitly control what information is given to whom and when [24, 25]. There is an inherent tradeoff between the utility and quality of LBS that users wish to receive and the location privacy they are ready to compromise. In the following sub-sections, various strategies available for privacy protection are presented.

2.1 Regulatory Strategies

All rules regarding to fair use of personal information falls under the category of regulatory approaches to privacy. In general, regulatory frameworks aim to adequately guarantee privacy protection for individuals' users. The Location Privacy Protection Act of 2011 [1] clearly states that before collecting and sharing a customer's location one needs to take his/her explicit consent.

2.2 Policy Based

Defining privacy policies and maintaining them comes under the umbrella of another class of location privacy techniques- policies based techniques. Privacy policies are trust-based mechanisms for prescribing certain uses of location information. Privacy policies define restrictions that regulate the release of the location of a user to third parties. User's needs of privacy are satisfied by restricting the ability to manage locations and disclosing information. The biggest disadvantage of policy based measures is the lack of policy enforcement specified by service provider. So despite of regulatory and policy based frameworks, adversaries are able to intrude in one's location privacy.

2.3 Location Obfuscation

Location Obfuscation is the process of degrading the quality of information about a person's location, with the aim of protecting that person's location privacy. It is the process of slightly altering, substituting or generalizing the location in order to avoid reflecting real, precise position. The most common techniques to perform obfuscation are pseudonyms, spatial cloaking, adding random noise and dummies, Redefinition of possible areas of location.

Pseudonyms, if used and implemented properly will prove to be an effective way to protect identity of users. Authors in [15] have used pseudonym for authorization and access control. It provides same level of security as that of distributed architecture and is applicable for pull based services.

In the Spatial K-anonymity paradigm [7, 25, 27], the client sends its query to middle-ware. It then constructs an anonymizing spatial region (ASR)/cloaking region (CR) that contains the querier's location along with other K-1 client locations. This ASR along with the query request is sent to the LBS. LBS executes the query with respect to the

ASR, and returns a superset of the results to the anonymizer, which filters out the false positives. Spatial cloaking has gained a considerable attention of privacy researchers. Rectangular cloaking regions were replaced by cloaking regions based on voronoi diagrams [18]. This provides greater flexibility, security and performance gain. Also the concept of cloaking regions containing k users as well as same cloaking region for at least k users is coined by [11]. Further k -anonymity based on fuzzy context parameters was introduced in [13]. The underline concept of k -anonymity has been extended by various approaches to increase privacy protection. The most important extensions are l -diversity, t -closeness, p -sensitivity, and historical k -anonymity.

Another approach for location privacy under the category of obfuscation is generation of dummies. To add dummy locations and noise to user's position [5, 19, 37] proposed an idea of sending additional set of dummy queries along with the actual query. The obfuscation region consists of the distinct locations included in the query set sent to the LBS.

2.4 Data Transformation

In this setting the data has been transformed using some encoding methodology like Hilbert curve etc. prior to transmitting it to the LBS. An authorized client has the secret transformation keys. This client issues an encoded query to the LBS. Both the database and the queries are unreadable by the LBS. In this way location privacy is protected.

2.5 PIR Based Location Privacy

Private Information Retrieval (PIR) protocols facilitate a client to retrieve the i th block from the server, without the server discovering which block was requested (i.e., index i). These protocols safeguard against access pattern attacks [16]. They can be grouped into: (i) information theoretic, (ii) computational [21] and (iii) secure hardware [32, 35].

There is a tradeoff between privacy and efficiency in the above mentioned techniques. While anonymity/cloaking and transformation-based approaches provide competent spatial query processing, they endure various privacy implications. On the other side of the coin there are, cryptographic and PIR-based approaches that provide significantly stronger privacy guarantees but incur more costly query processing operations.

3 Common Attacks and Challenges in Location Privacy

In order to evaluate a location privacy preserving technique/mechanism accurately, the adversary against whom the protection is required must be modeled. Hence, the adversary model is actually a very vital element of a location-privacy framework. An adversary is characterized by his knowledge and type of attack(s) he can target. An adversary model comprises of two main components: (a) the information which he/she wants to target (what he wants to infer) and, (b) the background knowledge and the inference abilities available to the adversary.

Some of the location privacy attacks along with the way to handle them and, open challenges in the respective area are given below:

Spatial Knowledge attack: Assume that a user issues a LBS request from a location p and most obviously user does not want to reveal his location. Now assume that user's location p is obfuscated by region q using some geometry-based technique. Now if adversary is aware that the user is in the obfuscated location q and q is entirely contained in the spatial extent of a particular place which is publicly known, then it can be immediately inferred that user is located in that place. However, for a professional whose work is related to that place (for him/her it's a routine), such a privacy concern would not arise because the location would be related to the user's professional activity. This privacy attack has been referred as spatial knowledge attack and has been described by Lee [22]. The spatial knowledge attack arises because real semantics of the space are ignored by geometry-based obfuscation techniques.

Handling spatial knowledge attack: These types of spatial knowledge attacks can be well handle if the privacy preservation mechanism utilizes semantics of location. These semantics may be in the form of identity of location, staying duration etc.

Location dependent attacks: Location dependent attacks may be based on continuous queries while users are moving (continuous) or snapshot (one time) queries. For these queries location k -anonymity and cloaking granularity are the privacy metrics. When exact snapshot locations are unveiled, two kinds of attacks are possible: location linking attacks [11] and query sampling attacks [4]. Location linking attacks refer to the scenario where the location information included in a user query is used as a quasi-identifier to re-identify the user.

Handling location dependent attacks: The location k -anonymity model was proposed to prevent this kind of attacks by Grutser [11]. The fundamental idea is to extend an exact user location to a cloaked region that covers at least k users. Grutser used a Quad-tree based cloaking algorithm to generate cloaked regions. Ghinita [7] proposed a cloaking algorithm called hilbASR, in which Hilbert curve is used in order to approximate the spatial proximity between query requests.

Further, Cheng in [3] proposed two simple solutions, namely patching and delaying. In patching the previous cloaked area is essentially covered so that the current one is at least as large as the previous one. But obviously, drawback is increasing size of cloaking area with evolving time. The second solution, called delaying, delays the request by t time until the MBR grows large enough to fully contain the current cloaked region.

Multi query attack: As the name indicates, multi-query attack is the one where an adversary tries to identify the actual location of the query issuer with the help of a series of two or more spatial queries. All these queries involve different cloaking regions. The idea given by authors in [31] is to determine the exact location of the service requester by obtaining various cloaking regions (CR) that are shrunk or extended in succeeding queries.

Handling multi query attacks: The above mentioned problem can be addressed by ensuring reciprocity condition which ensures the users in the same anonymity set should use same CR over time. This problem can be dealt by preserving the cloaking regions for the same set of users for a specific period of time and by developing disjoint sets of users dynamically over time in order to share the common CRs.

Maximum movement boundary attack: In a maximum movement boundary attack, the adversary computes the whole area of movement of user/target, where the user could have moved between two succeeding snapshot queries or position updates. Let us assume that the initial update is sent when user was at time T1 and the other update is sent at time T2. Using this strategy the attacker can increase the precision of the update sent at T2. As only a small part of the area of T2 is reachable within the maximum movement boundary. Therefore, the remaining area of the position update can be safely excluded by the attacker.

Handling maximum movement boundary attack: Ghinita et al. [7] developed temporal and spatial transformations to sustain this type of attack. The idea of temporal transformations is to delay the requests while spatial transformations CRs are not directly generated depending on the user location, but instead are built starting from the last reported CR.

Trajectory attacks: Simply removing the identifier does not guarantee the privacy of owners while trajectory publishing. The owner might be inferred by attackers after this also. This type of attacks is called trajectory attacks [10]. The problem of trajectory anonymization is twofold. On one end the need is to preserve identity of trajectory owner and along with this the utility of published data is also to be maximized. The existing work can be classified into two categories: trajectory anonymization in free space [8, 9, 30, 35, 36] and in constrained space [9, 23]. Existing methods for location anonymization and cloaking are not applicable in this scenario.

Handling Trajectory attacks: Goal of trajectory privacy-preserving techniques is to protecting whole trajectory not to be identified by the adversary, also protecting sensitive/frequent visited locations in trajectories. This all should be done along with preserving the utility of data. However it has been shown that releasing anonymized trajectories may still have some privacy leaks. Therefore Nergiz [28] proposed a randomization based reconstruction algorithm for releasing anonymized trajectory data and also presented the adoption of this underlying techniques to other anonymity standards.

Inversion attacks: Inversion attacks are based on the situation in which a n adversary is aware of identity of k potential users of the request. Thus even after observing a specific cloaked region because of k-anonymity, adversary is not able to determine the query issuer among k users. However, if adversary knows the cloaking algorithm, he can simulate its application to the specific location of each of the candidates, and exclude any candidate for which the resulting cloaked region is different from the one in the observed request. Thus he will be able to breach the privacy of client. This type of attack

is called inversion attack. Some of the cloaking algorithms are indeed subjected to this attack.

Handling Inversion attacks: Kalnis et al. [14] show that reciprocity is the solution for this attack. Each generalization function if satisfies reciprocity will not be subjected to the inversion attack.

Query tracking attacks: In case of continuous queries, the results of query would be continuously returned for a designated time period which is called query lifetime. [4]. Query tracking attacks become possible when a user is cloaked with different users at different time instances during the query lifetime. In this way he is easily identifiable among a set of users.

Handling Query tracking attacks: Usage of memorization property is the key point to protect against query tracking attacks. According to memorization property during the whole query lifetime, the set of users being cloaked in an area should always be same [4]. Clearly, there is a line of difference between query tracking attacks and location-dependent attacks. Even if the users are prevented from query tracking attacks by applying the memorization property there is no guarantee of protection from location-dependent attacks using this.

Inference attacks: Gaining knowledge unlawfully about a subject by analyzing data is known as an “inference attack”. Inference attacks on the observed queries are basically classified into two categories: tracking and identification attacks. Such attacks can lead to two types of location-privacy breaches: presence and absence disclosure. Protection strategies always aim to reduce adversary’s information as little information about user locations makes it harder for the adversary to reconstruct their actual trajectories and to identify their real identities. But, unfortunately, doing so has its own cost in terms of reduced service quality for the user. Authors in [12, 20] examined location data gathered from volunteer subjects and apply four different algorithms to identify the subjects’ home locations and then their identities using a freely available, programmable Web search engine.

Handling Inference attacks: Inferences attacks can be handled by different obscuration methods. Further, three different obscuration countermeasures - spatial cloaking, introducing noise, and rounding, designed to halt the privacy attacks, are applied in the above mentioned case. It has been shown in [20] that how much obscuration is necessary to maintain the privacy of all the subjects.

Other attacks: Apart from the above mentioned attacks, some other remarkable studies in the area of location privacy are: [11, 12] worked with completely anonymized GPS data. They used a standard technique from multi-target tracking. On the parallel lines the approach for anonymous indoor data is given by Williams et al. [35]. They placed simple presence sensors around a house, i.e. motion detectors, pressure mats, break beam sensors, and contact switches. These sensors helped to develop a probabilistic tracking algorithm. Using observations of sensor triggers the algorithm is detect to identify

occupant of the house around which these sensors were fixed. Duckham et al. [6] presented a model of refinement operators for working around obfuscation techniques, such as assumptions about a victim's movement constraints and goal-directed behavior.

Using the strategy of query sampling attacks an adversary may still be able to link a query to its user in case user locations are publicly known. This is possible even if locations are cloaked. This kind of attacks is called query sampling attacks [4]. Idea of k sharing regions i.e. a cloaked region should not only cover at least k users, but the region is also shared by at least k of those users is the key to protect against query sampling attacks.

4 Open Challenges and Future Research Directions

In order to solve the contradiction between location privacy protection and quality of services, researchers have already come up with a number of privacy protection methods. But there are many research issues which are still open. Following is the description of open challenges in the domain of privacy in LBS.

- **Use of semantics:** In the earlier research approaches, to attain location privacy semantics of query, data, and location itself are not considered. Very few research article paid attention to the above mentioned semantics. Research is not mature enough about the use of semantics which can bring the drastic transformation in the existing data privacy techniques.
- **Privacy-preserving Location Data Collection:** Location data generated by cell phones are collected by manufactures and published/leaked to third parties for analysis. Analysis of users' location data may cause personal privacy leakage so solutions can be research on privacy-preserving location data collection.
- **Application of PIR:** The PIR-based approaches to location privacy open pathways to a novel way of protecting user's location privacy. However, to utilize complete potential of these techniques cost of computationally intensive query processing is to be beared. Therefore, the further direction of research should be reducing the costs of PIR operations. Also "use of efficient indexing technique" for spatial queries is a future research area.
- **Formalizing of LPPM (location privacy preservation mechanism):** All the works in the literature concentrates on solution of a particular problem of location privacy domain, e.g., protection mechanisms against a specific kind of attack, and therefore do not provide a generic framework that takes care of all location-privacy components in [29]. There exists a lack of a formal framework to quantify location privacy and to formalize attacker's model. Shokri et al. in [30] propose a framework in which they formalize various metrics and quantified location privacy. But from end user's point of view the above solution is not usable because of being cryptic so a candid formalization is still awaited.

5 Conclusion

The problem of privacy breach while using location based services has gain a considerable attention of the researchers community. This article demonstrate various achievements and research works accomplished in the area of location based privacy. However, despite of several measures to protect privacy, there are numerous attacks to intrude in location privacy and henceforth there are many open challenges still to be resolved. In this work, all such attacks and measures to prevent them have been integrated and also suggested future research directions.

References

1. The Location Privacy Protection Act of 2011 (S. 1223). https://www.franken.senate.gov/files/documents/121011_LocationPrivacyProtection.pdf
2. Bettini, C., Mascetti, S., Wang, X.S., Freni, D., Jajodia, S.: Anonymity and historical-anonymity in location-based services. In: Bettini, C., Jajodia, S., Samarati, P., Wang, X.S. (eds.) *Privacy in Location-Based Applications*. LNCS, vol. 5599, pp. 1–30. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03511-1_1
3. Cheng, R., Zhang, Y., Bertino, E., Prabhakar, S.: Preserving user location privacy in mobile data management infrastructures. In: *Proceedings of Privacy Enhancing Technology Workshop* (2006)
4. Chow, C.-Y., Mokbel, M.F.: Enabling private continuous queries for revealed user locations. In: Papadias, D., Zhang, D., Kollios, G. (eds.) *SSTD 2007*. LNCS, vol. 4605, pp. 258–275. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73540-3_15
5. Duckham, M., Kulik, L.: A formal model of obfuscation and negotiation for location privacy. In: *PERVASIVE 2005* (2005)
6. Duckham, M., Kulik, L.: Simulation of obfuscation and negotiation for location privacy. In: Cohn, A.G., Mark, D.M. (eds.) *COSIT 2005*. LNCS, vol. 3693, pp. 31–48. Springer, Heidelberg (2005). doi:10.1007/11556114_3
7. Ghinita, G., Kalnis, P., Skiadopoulos, S.: Prive: anonymous location-based queries in distributed mobile systems. In: *Proceedings of WWW 2007* (2007)
8. Gidofalvi, G., Huang, X., Pedersen, T.B.: Privacy-preserving data mining on moving objects trajectories. In: *Proceedings of MDM 2007* (2007)
9. Gkoulalas-Divanis, A., Verykios, V.S.: A privacy-aware trajectory tracking query engine. *SIGKDD Explor. NewsL.* **10**(1), 40–49 (2008)
10. Gkoulalas-Divanis, A., Verykios, V.S., Mokbel, M.F.: Identifying unsafe routes for network-based trajectory privacy. In: *Proceedings of SDM 2009* (2009)
11. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: *Proceedings of MobiSys 03*, pp. 31–42 (2003)
12. Hoh, B., et al.: Enhancing security and privacy in transaction monitoring systems. *IEEE Pervasive Comput.* **5**(4), 3846 (2006)
13. Jagwani, P., Kaushik, S.: Defending location privacy using zero knowledge proof concept in location based services. In: *Proceedings of MDM 2012*, Bangluru, India (2012)
14. Kalnis, P., Ghinita, G., Mouratidis, K., Papadias, D.: Preventing location-based identity inference in anonymous spatial queries. *TKDE* **19**(12), 1719–1733 (2007)

15. Kaushik, S., Tiwari, S., Goplani, P.: Reducing dependency on middleware for pull based active services in LBS systems. In: S nac, P., Ott, M., Seneviratne, A. (eds.) ICWCA 2011. LNCS, vol. 72, pp. 90–106. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-29157-9_9](https://doi.org/10.1007/978-3-642-29157-9_9)
16. Khoshgozaran, A., Shahabi, C.: Private information retrieval techniques for enabling location privacy in location-based services. In: Bettini, C., Jajodia, S., Samarati, P., Wang, X.S. (eds.) Privacy in Location-Based Applications, October 2009. ISBN: 978-3-642-03510-4
17. Khoshgozaran, A., Shahabi, C.: A taxonomy of approaches to preserve location privacy in location-based services. *Int. J. Comput. Sci. Eng.* **5**(2), 86–96 (2010)
18. Khuong, V., Zheng, R.: Efficient algorithms for K-anonymous location privacy in participatory sensing. In: IEEE Infocom Proceedings 2012 (2012)
19. Kido, H., Yanagisawa, Y., Satoh, T.: An anonymous communication technique using dummies for location-based services. In: ICPS 2005 (2005)
20. Krumm, J.: Inference attacks on location tracks. In: LaMarca, A., Langheinrich, M., Truong, Khai, N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 127–143. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-72037-9_8](https://doi.org/10.1007/978-3-540-72037-9_8)
21. Kushilevitz, E., Ostrovsky, R.: Replication is not needed: single database, computationally private information retrieval. In: FOCS (1997)
22. Lee, B., Oh, J., Yu, H., Kim, J.: Protecting location privacy using location semantics. In: KDD 2011, August 21–24, San Diego, California, USA (2011)
23. Lee, K., Lee, W.C., Leong, H.V., Zheng, B.: Navigational path privacy protection: navigational path privacy protection. In: Proceedings of CIKM 2009 (2009)
24. Liu, L.: Privacy and location anonymization in location-based services. *SIGSPATIAL Spec.* **1**(2), 15–22 (2009)
25. Liu, L.: From data privacy to location privacy. In: VLDB 2007, pp. 1429–1430 (2007)
26. Mokbel Mohammad, F.: Privacy-preserving location services. In: ICDM 2008 (2008)
27. Mokbel, M.F., Chow, C.Y., Aref, W.G.: The newcasper: query processing for location services without compromising privacy. In: Proceedings of VLDB 2006 (2006)
28. Nergiz, M.E., Atzori, M., Sayggn, Y., Gu, B.: Towards trajectory anonymization a generalization based approach. *Trans. Data Priv.* **2**(1), 47–75 (2009)
29. Shokri, R., Freudiger, J., Jadhwal, M., Hubaux, J.-P.: A distortion-based metric for location privacy. In: WPES 2009: Proceedings of the 8th ACM Workshop on Privacy in the Electronic Society, pp. 21–30, New York, NY, USA. ACM (2009)
30. Shokri, R., et al.: Quantifying location privacy. In: 2011 IEEE Symposium on Security and Privacy. IEEE (2011)
31. Talukder, N., Ahamed, S.I.: Preventing multi-query attack in location-based services. In: Proceedings of the Third ACM Conference on Wireless Network Security. ACM (2010)
32. Wang, S., Ding, X., Deng, R.H., Bao, F.: Private information retrieval using trusted hardware. In: Gollmann, D., Meier, J., Sabelfeld, A. (eds.) ESORICS 2006. LNCS, vol. 4189, pp. 49–64. Springer, Heidelberg (2006). doi:[10.1007/11863908_4](https://doi.org/10.1007/11863908_4)
33. Wernke, M., et al.: A classification of location privacy attacks and approaches. *Pers. Ubiquit. Comput.* **18**(1), 163–175 (2014)
34. Westin, A.F.: Privacy and Freedom. Atheneum, New York (1967)
35. Williams, P., Sion, R.: Usable PIR. In: NDSS (2008)
36. Xu, T., Cai, Y.: Exploring historical location data for anonymity preserving in location-based services. In: Proceedings of INFO-COM 2008 (2008)
37. You, T.H., Peng, W.-C., Lee, W.C.: Protecting moving trajectories with dummies. In: Proceedings of PALMS 2007 (2007)

An Efficient and Low-Signaling Opportunistic Routing for Underwater Acoustic Sensor Networks

Zhengyu Ma, Quansheng Guan^(✉), Fei Ji, Hua Yu, and Fangjiong Chen

Electronic and Information School,
South China University of Technology, Guangzhou, China
bcgb9382@qq.com, {eeqshguan, eefei ji, yuhua, eefjchen}@scut.edu.cn

Abstract. Underwater acoustic sensor network (UASN) has been considered as a promising technique for ocean engineering. However, existing problems like long propagation delay, multipath interference and low available bandwidth are important issues in UASN. Based on the analysis of the UASNs characteristics, we proposed a novel protocol, named DUOR (Depth-based Underwater Opportunistic Routing protocol), which directs a packet to the sonobuoy on the surface in an efficient and low-signaling method. The contribution of DUOR is twofold: (1) minimizing signaling costs; (2) solving “void area” and “the extremely long forwarding path”. Simulation results show that the proposed DUOR outperforms the existing Depth Based Routing (DBR).

1 Introduction

Underwater acoustic sensor networks (UASNs) have been proposed as an alternative solution for observing and exploring underwater environments. Data forwarding in the water is a key problem to be solved owing to the characteristics of the underwater acoustic channel such as long propagation delay, high error rate [1] and complex multipath effect [2]. These lead to high end-to-end delay, high packet loss rate, which make the design of routing protocols in UASNs very challenging [3].

Opportunistic routing (OR) [4] is preferred for data forwarding in UASNs. Unlike traditional routing protocols, which are mainly designed based on the method that the sensor nodes forward packets by looking up the predefined routing table, OR chooses a forwarding candidate set from neighboring nodes to transmit the data. All the nodes which receive the packet correctly have the chances to forward the packet. It's obvious that the data transmission through multiple nodes is more reliable than through a single node. Therefore, OR can improve the data delivery ratio and network throughput.

This work was supported in part by the National Science Foundation of China under Grant 61302058, Grant 61671211, Grant 61322108, Grant 61431005, and Grant 61671208 and in part by the Pearl River S&T Nova Program of Guangzhou.

Some existing opportunistic routing protocols forward data packets to a locally optimal next-hop node closet to the sink node. This strategy will suffer from a problem called void area. The “void area” is referred to the situation that no any other relay nodes exist between the current node and the destination in packet transmission path. There are other existing OR protocols which can solve the “void area” such as Geographic and Opportunistic Routing for Underwater Sensor Networks (GEDAR) and OVAR (An Opportunistic Void Avoidance Routing Protocol for Underwater Sensor Network). However, they can’t solve “the extremely long forwarding path” and consume more signaling exchange than our proposed DUOR. “The extremely long forwarding path” can be encountered when transmission path is much longer than the optimum route. GEDAR is based on the network topology control through depth adjustment of those void nodes. Since network nodes’ localization is needed in this protocol, abundant signaling cost and energy consumption are introduced. In OVAR, every node needs to set up a neighbor table to keep the depth, hop count and ID of neighbor nodes. So our goal in this paper is to low signaling cost, solve the “void area” and “the extremely long forwarding path”. DUOR requires less signaling cost. Different from other OR protocols which are also based on depth and hop-count for UASNs, there is no any information exchange among forwarding candidate set except overhearing packet transmissions. It just needs depth and hop count as forwarding candidate selection criterion. The main contribution of this paper is the proposal of a novel opportunistic routing protocol, named Depth-based Underwater Opportunistic Routing Protocol (DUOR), for UASNs. The novelties of DUOR are showed as follows.

- *Low signaling cost:* In existing protocols aforementioned, global topology and other signalings are required to select next hop forwarder nodes. Instead, DUOR just needs nodes’ depths and hop-counts as the essential condition for the candidate set. In addition, there is no any information exchange among forwarding candidate set except overhearing packet transmissions. The nodes with shallower depths and smaller hop-counts are self-included in the candidate set. DUOR is a receiver-based scheme where the receiver node decides forwarding candidate set. This strategy can reduce signaling exchange.
- *Avoiding the “void area” and “the extremely long forwarding path”:* These two problems would lead to high packet loss and excessive energy consumption. We adopt hop-count to the selection condition for the forwarding candidate set. The node’s hop-count is the number of hops to reach the sink node. Upon receiving and forwarding query frame from the sink node, underwater sensor nodes set up a reachable route to the sink node and acquire its own hop-count toward the sink node. Only the nodes whose hop counts and depth both satisfy certain conditions can be chosen as forwarding candidate set. This strategy can solve the problem of “the extremely long forwarding path” effectively. For example, in Fig. 1, the hop-count sensor node A is nine, while the hop-count of the sensor node B is four. Since the hop-count limitation can make the sensor node A who may take the extremely long forwarding path as a nonparticipator, the problem can be addressed. In Fig. 2, the solid line

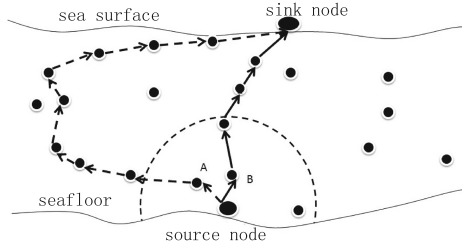


Fig. 1. The phenomenon about “the extremely long forwarding path” in underwater communication

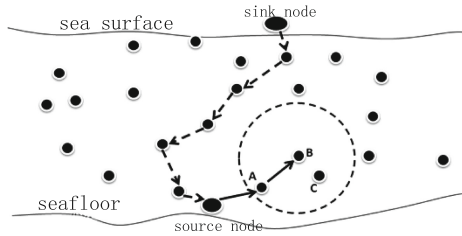


Fig. 2. The problem of “void area” in OR protocol

(source node-node A-node B) shows the transmission path of the traditional OR protocol. When the packet is transmitted to the node A, the node B whose depth is shallowest around node A will forward the packet preferentially according to traditional OR protocol. However, there is no shallower node around node B in one hop, the node B will retransmit the packet continually or discard the packet. However, in DUOR, the query frame from the sink node is transmitted along imaginary line. Because there is no shallower node around node B within one hop range, the node B cannot receive the query frame sent by the sink node from the top to the bottom. So the node B’s hop count is out of work. Since the hop-count is one of conditions to determine forwarding candidate set, so the node B won’t be involved in routing process when the node A received a packet. Definitely, the problem of “void area” is solved.

2 The Details of DUOR

The basic procedures of our proposed protocol involves four steps. Before data transmission, every sensor node keeps its local status of hop-count and depth to set up next-hop route toward the sonobuoy. Sensor nodes can acquire and update their own depth through carry-on pressure gauge. Then the source node (sender) broadcasts the packet embedding its surrounding region data. After that, neighboring nodes receiving the packet determine whether itself belongs to forwarding candidate set according to the hop-count and depth, Finally the

node with the highest priority in candidate set transmits the packet, other low priority nodes hear it and suppress forwarding to prevent redundant packet transmissions and collisions. Then repeat above forwarding processes until the sink node receives the packet.

2.1 Every Node Updates Its Own Hop-Count

- *The sink node broadcasts the query frame periodically:* The query frame only involves the query frame ID, hop-count variable (N_r) and depth variable (D_r). (The initial values of N_r and D_r are both zero.) For a query frame receiver node, N_r and D_r record the hop count and depth of the last hop. The structure of the query frame is shown in the Table 2.

Table 1. The memory of the sensor nodes

Dn	Nn	Query frames' IDs	Q1	Q2
----	----	-------------------	----	----

Table 2. The structure of the query frame head

The query frame ID	Nr	Dr
--------------------	----	----

Table 3. The structure of the packet head

Source node ID	Packet ID	Ds	Dc	Ns	Data
----------------	-----------	----	----	----	------

- *The nodes which receive the query frame check whether to store or discard it:* Each sensor node updates its local buffer with its own depth D_n , its own hop-count N_n and query frames' IDs it has received. Its local buffer structure is shown in Table 1. After receiving a query frame, a node can tell whether it has received the message through query frame IDs stored in its local buffer. When a sensor node receives a repetitive query frame, the query frame will be discarded. On the contrary, if a sensor node receives a new query frame, it will update its local information and the query frame, which we will discuss detailly in the following paragraph.
- *What to do after accepting the query frame:* The receiver updates N_r and D_r values in this query frame, stores the query frame ID and the new hop-count in its local buffer. Finally, it broadcasts the updated query frame. The depth (D_r) and hop-count (N_r) in query frame are updated as follow:

$$N_r + 1 \rightarrow N_r, D_n \rightarrow D_r, \tag{1}$$

where D_n is the current node's depth. The sensor node stores the new hop-count as.

$$N_r \rightarrow N_n, \tag{2}$$

where N_n is the current node's hop-count. So far the sensor node updates its hop-count.

Repeat the above-mentioned process until all the sensor nodes receive the query frame.

2.2 The Sensor Node Sends the Packet

The structure of the packet is shown in the Table 3, where D_s is the source node's depth, D_c is an updating depth variable which records the last hop's depth, N_s is the source node's hop-count. Those information can help the packet receivers judge whether they belong to forwarding candidate set and calculate the waiting time to coordinate the forwarding.

2.3 The Intermediate Nodes Forward the Packet

Except mentioned in the previous section, every sensor node maintains sequence Q1 and Q2. Q1 saves packets and the waiting time designed for packets, while the Q2 maintains all forwarded packets' IDs. Now we will introduce how the intermediate nodes forward the packet.

- *candidate set selection*: When the source node sends the packet, a part of neighboring nodes can receive the packet correctly. The nodes receiving the packet for the first time put the packet in the sequence Q1 and judge whether they belong to forwarding candidate set. If the sensor nodes satisfy the two following formulas, they belong to the forwarding candidate set:

$$N_r \leq m \cdot N_s, D_n < D_c, \quad (3)$$

where m is forwarding candidate set coefficient, N_r is the current node's hop count, N_s is the source node's hop count, D_n is the current node's depth. D_c is an updating depth variable which records the last hop's depth.

From forwarding candidate selection conditions, it's obvious that sensor nodes with shallower depth and less hop-count would be more likely selected to candidate set.

- *waiting time calculation*: If the nodes satisfy the forwarding conditions, they will set the waiting time in Q1 and put the packet ID into sequence Q2. That means, the candidate set will wait a certain time to forward the packet. The node with highest priority has the shortest waiting time. Utilizing hop-count and depth as waiting time calculation factors can make the sensor nodes with better link states and shallower depth have higher priority to forward packets. Firstly, to make shallower nodes with shorter waiting time, the waiting time calculation formula can be set as follow:

$$t_r = \frac{k \cdot D_r}{c} + W, \quad (4)$$

where k is waiting time coefficient, t_r is the time that forwarding candidate set should wait to transmit the packet, We set c as the propagation speed of acoustic wave. D_r is the depth of the current candidate node. W is a constant. In Fig. 3, node A has higher priority to forward the packet from the source node than node B. Node A is far from the source node than node B, so node A may receive the packet later. Node B should wait more time to eliminate the receiving time difference between node A and node B. On the

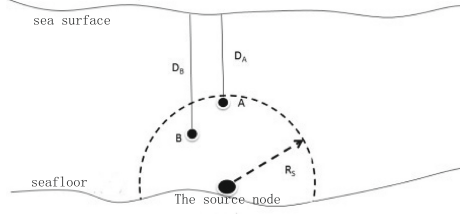


Fig. 3. Waiting time calculation

other side, when node A forwards the packet preferentially, node B will hear it after a propagation delay. So node B should wait more time to hear node A's forwarding packet and then suppress its own forwarding.

$$t_B - t_A > 2 \times \frac{D_B - D_A}{c}, \quad (5)$$

t_A is the node A's waiting time, t_B is the node B's waiting time, D_A is the node A's depth, D_B is the node B's depth. In UASNs, we consider $(D_B - D_A)$ as the distance between the node B and the node A. So in formula (4), k must be a constant greater than two to make sure formula (5). That is,

$$t_A = \frac{k \times D_A}{c} + W, t_B = \frac{k \times D_B}{c} + W, t_B - t_A = \frac{k \times (D_B - D_A)}{c}, \quad (6)$$

k is greater than two to meet the condition formula (5). Finally, to make sure the farthest node among candidate set has received the packet before another candidate node forwards the packet, we modify the waiting time calculation formula (4) as follow:

$$t_r = \frac{k \times D_r}{c} + \frac{R_s}{c}, \quad (7)$$

R_s is acoustic communication range. Finally, we hope the nodes with lower hop count have less waiting time, so we add hop count to the modified formula (7):

$$t_r = \frac{k \times D_r}{c} + \frac{N_r}{N_s} \times \frac{R_s}{c}, \quad (8)$$

The formula (7) is the reasonable waiting time calculation we designed.

- *candidate set coordination*: A candidate set node will broadcast the packet when the waiting time which is set for the packet in Q1 is over. However, low priority nodes should be able to suppress forwarding when overhearing the highest priority node's transmission.

2.4 The Sink Node Receives the Packet

Repeat the above steps until the sink node receives the packet.

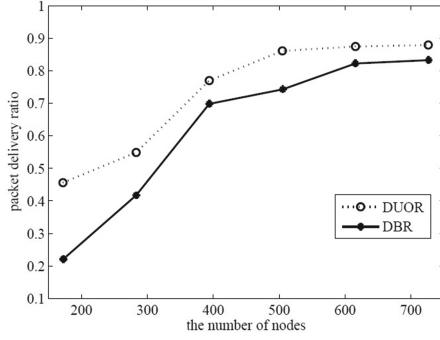


Fig. 4. The packet delivery ratio of DUOR and DBR

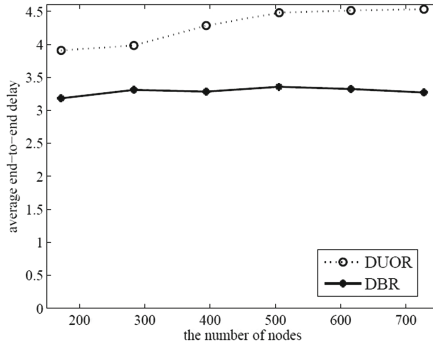


Fig. 5. The end-to-end delay of DUOR and DBR

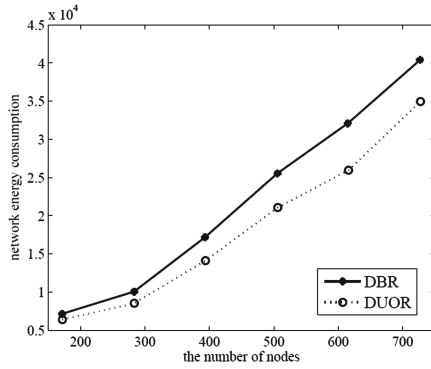


Fig. 6. The whole network consumption of DUOR and DBR

3 Simulations and Discussions

This chapter will compare DUOR with DBR using OPNET. We have used a simple media access control protocol (MAC) based on CSMA. We considered the sensor nodes randomly deployed in a 3D of $500 \times 500 \times 500 \text{ m}^3$. Every node moves toward a certain direction with velocities varying from 0 m/s to 5 m/s. The maximum transmission range is set as 100 m. We set the communication rate to 10 Kbps and the data packet size to 50 Bytes. Packets are generated randomly from nodes every one second. The sink node's initial energy is 10000 J and other sensor nodes' initial energy are 400 J. The values of energy consumption were $P_t = 2w$, $P_r = 0.75w$, $P_i = 0.08w$ for sensor operations of transmission, reception and idle. The sink node broadcasts query frames every 20 s and the sensor nodes save a query frame for 60 s. The simulation time is set as 300 s. we compare DBR [5] and DUOR since they are both the routing protocols based on depth. As shown in Figs. 4, 5 and 6, although the end-to-end delay of DUOR is longer than DBR, DUOR achieves higher packet delivery ratio and lower energy consumption than DBR. DBR maintains only the depth information at each sensor node, this strategy will cause poor packet delivery ratio in sparse networks and cannot handle "void area". However, DUOR selects the next hop forwarder opportunistically based on updated link transmission reachability (hop-count) and depth.

References

1. Coatelan, S., Glavieux, A.: Design and test of a multicarrier transmission system on the shallow water acoustic channel. In: OCEANS 1994. IEEE (1994)
2. Syed, A.A., Heidemann, J.S., et al.: Time synchronization for high latency acoustic networks. In: INFOCOM (2006)
3. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks: research challenges. *Ad Hoc Netw.* **3**, 257–279 (2005)
4. Hsu, C.-C., Liu, H.-H., Gomez, J.L.G., Chou, C.-F.: Delay-sensitive opportunistic routing for underwater sensor networks. *IEEE Sens. J.* **15**, 6584–6591 (2015)
5. Yan, H., Shi, Z.J., Cui, J.-H.: DBR: depth-based routing for underwater sensor networks. In: Das, A., Pung, H.K., Lee, F.B.S., Wong, L.W.C. (eds.) NETWORKING 2008. LNCS, vol. 4982, pp. 72–86. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-79549-0_7](https://doi.org/10.1007/978-3-540-79549-0_7)

A Novel Mobile Online Vehicle Status Awareness Method Using Smartphone Sensors

Dang-Nhac Lu^{1(✉)}, Thi-Thu-Trang Ngo², Duc-Nhan Nguyen²,
Thi-Hau Nguyen¹, and Ha-Nam Nguyen¹

¹ University of Engineering and Technology, Vietnam National University in Hanoi,
Hanoi, Vietnam

{Nhacld.dill, nguyenhau, namnh}@vnu.edu.vn

² Posts and Telecommunications Institute of Technology in Hanoi, Hanoi, Vietnam
{trangntt1, nhannd}@ptit.edu.vn

Abstract. In this paper, we proposed an efficient method with flexible framework for vehicle status awareness using smartphone sensors, so called Mobile Online Vehicle Status Awareness System (MOVSAS). The system deployed while users to put their smartphones in any position and at any direction. In our proposed framework, principal component analysis (PCA) algorithm is used to selected suitable features from set of combining features on time-base, power-based and frequency-based domain, which extracted from accelerometer sensor data. The classification model using Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) algorithms to deploy for awareness issues of vehicle status. The refining model is proposed using Artificial Neural Network (ANN) algorithm aim to improved accuracy prediction vehicle status results before. Training data sets, which are collected and the dynamic feedback also helping improved accuracy of system. A number of experiments are shown that the high accuracy of MOVSAS with vehicle kinds as bicycle, motorbike and car.

Keywords: Mobile online recognition · Vehicle status · Smartphone sensors analysis

1 Introduction

The vehicle awareness and prediction play an important role in various applications such as energy estimation, safety, healthcare, transportation, social networking, etc. [1]. This problem has the potential to impact our daily lives through extracting useful information from raw sensor data. There were many methods to understand smartphone sensors data, the most common method is using the windowing technique. For example, John J. Guiry et al. [1] used the windowed sensor data samples from phone and chest device to recognize six activities. The time domain and frequency domain are two modes used for sample analysis. The activity is inferred from the data window of one second and the frequency domain is also used to analyze sensor signals every fifteen seconds with accuracy of 98%. In the fact that, its high accuracy is resulted by the fixed position of

the subjects' phones that is in their trouser pocket. Consequently, the recorded sensor signal is more stable.

With proposed recognition system at run-time, Jiahui Wen et al. [2] suggested a method which combines a basis classifier with graphical model. They used the five-second sliding window with 50% overlap to segment the streaming data. However, they did not describe details of the features *data instances*. The assessments were employed with several existed data sets such as smartphone dataset [3], sensor activity dataset [3], UCI HAR dataset, Opportunity dataset. This paper also suggested an extra assessment model for run-time system.

In the paper of sensor-based classification, Sang et al. [4] proposed a method to recognize daily activity of a user. The data was collected from smartphones placed in users' pocket. They extracted the features from the sensor signals: auto regressive coefficient, fractal dimension, mean and standard deviation. However, they did not explain the compatible features and their results obtained from this feature set were not compared to the others. In another study, Zahid Halim et al. [5] have developed artificial intelligence techniques for driving safety and vehicle crash prediction. This data analysis included the weather conditions; the data was gathered in ten years from the vehicle sensors such as accelerometer, camera in different driving behaviors before accidents. In their study, intelligence techniques were commonly used for accident prediction problems. The classification accuracy using decision tree (DT) and ANN in this study is 95% and they are good algorithms for time series data and driving behaviors recognition.

The challenges on recognition using smartphone sensors are the noises in data, missing data, variety of signal quality from different sensors, and the change in smartphone position. Therefore, we propose a method to automatically select a set of features from each window of acceleration data when smartphone users are moving. The possible features are based on time domain and frequency domain. The principal component analysis is applied to the online choice of suitable features. Then, the system uses one of classifier algorithms such as random forest, support vector machine, k-nearest neighbor and Naïve Bayes. In our framework, the refining model with ANN algorithm is used to improve the accuracy of vehicle status prediction. The feedback module will receive label then push this status information to data training set. In this paper, the different statuses including stop, moving, acceleration, deceleration on bicycle, motor-bike and car are distinguished and the obtained results is outperformed.

2 The Mobile Online Vehicle Status Awareness System (MOVSAS)

The Mobile Online Vehicle Status Awareness System (MOVSAS) consists of three modules. The data collector module is responsible for collecting labeled smartphone sensor data of each predefined vehicle status. The signal data is then preprocessed, and a set of representative features is extracted. The Principal component analysis (PCA) is used to select the features for training model [6]. In the online training module, a classifier detects the vehicle status of smartphone users and then the model is refined to improve the accuracy from the prior detection results. It uses the recent status S_t corresponding to the window w , and combines with $k-1$ linear statuses. The set of k features

as $[S_{t-k-1}, \dots, S_{t-1}, S_t]$ aims to correctly detect the status from training data, which is a set of instances including k statuses were collected. Based on the trained knowledge, the real time vehicle status of users is detected by the Monitoring module. The MOVSAS framework is shown in Fig. 1.

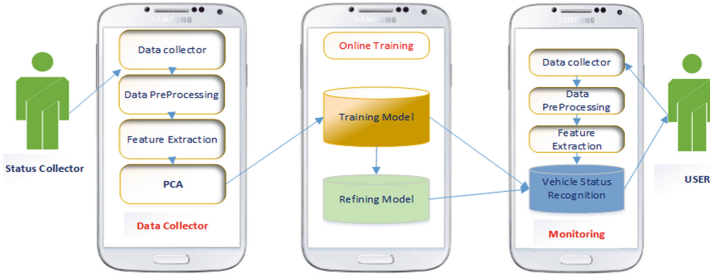


Fig. 1. The mobile online vehicle status awareness system framework

2.1 Data Preprocessing

The user information gained from smartphone sensors, especially accelerometer is very useful to recognize the vehicle status. However, while moving, the users might put their smartphones on their pocket, handbag, or in their hands, etc. As a result, the orientation of smartphones will be frequently changed. The approach to solve this issue is to transform accelerometer data from the smartphone coordinate system to the Earth coordinate system by relying on the additional data collected from magnetometer and gyroscope sensors as Fig. 2 that aims to reduce noise. For the details of this transformation, we refer readers to the work of Premerlani and Bizard [7]. Then the data is prepared for classification by a feature set based on time- and frequency domains. The PCA will be applied to select suitable features for classification.

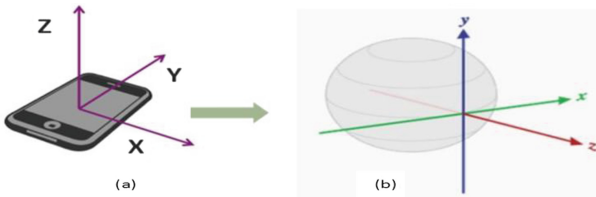


Fig. 2. The smartphone coordinate (b) the Earth coordinates

2.2 Using PCA for Feature Selection

The amount of raw data collected from smartphone sensors is various. Thus, directly analyzing such data would require a lot of either time or memory space. A popular approach to deal with this issue is to extract certain important features from such data and to select suitable features that would lead to an increase in the prediction accuracy.

In time domain, we have computed the features of accelerometer such as:

The root mean square (RMS) [8] of a signal x_i that represents a sequence of n discrete values $\{x_1, x_2, \dots, x_n\}$.

The sample correlation coefficient of axis x and y is computed by the equation below

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \tag{1}$$

The cross-correlation is a measure to compare similarity between two waveforms and computed by the following equation:

$$\text{CrossCorrelation}(x, y) = \max_{d=1}^{n-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_{i-d} \right) \tag{2}$$

The SMA feature [9] is also calculated to distinguish between a resting state and vehicle status in a classification.

The vertical and horizontal accelerometer energy features of each time window are computed by the following equation.

$$E_v = \int_{t=t_0}^{T+t_0} |a_v| dt, \quad \text{and} \quad E_h = \int_{t=t_0}^{T+t_0} |a_h| dt \tag{3}$$

where a_v, a_h are respectively vertical and horizontal acceleration values, and T is the interval of integration with $a_v(t) = a_z(t), a_h(t) = \sqrt{a_x^2(t) + a_y^2(t)}$.

Thus, in time-domain, we have extracted thirteen features as CorreCoxy, CorreCoxz, CorreCoyz, Crossxy, Crossxz, Crossyz and Xrms, SMA, E_v , E_h and Mean [10], Variance [11], Standard deviation [12] in each data samples window.

For the frequency domain features, we compute Short Time Fourier Transform on the n^{th} window including N samples $[x_n, x_{n+1}, \dots, x_{n+N-1}]$ as following:

$$X(n, k) = \sum_{m=0}^{N-1} x[n+m].w[m]. \exp(-j(2\pi/N).k.m) \tag{4}$$

with $k = 0, 1, \dots, N-1$ and $w[m]$ as window function.

The Energy of M coefficient Fourier is computed by the below formula:

$$E_M = \sum_{m=1}^M |X(m)|^2 = \sum_{m=1}^M X(m) \cdot X^*(m) \quad (5)$$

Because that, Z axis data capability different in vehicle status so that average Energy of Z axis (E_Z) also computed as:

$$\bar{E} = \frac{2 \sum_{m=2}^{N/2} |X(m)|^2}{N} \quad (6)$$

Finally, the Entropy is computed with below formula:

$$H = - \sum_{m=1}^N p_m \log_2(p_m) \text{ with } p_m = \frac{|X(m)|}{\sum_{m=1}^N |X(m)|} \quad (7)$$

Thus, we have three features in frequency domain from Eqs. (5, 6 and 7) as E_M , E_Z and H and sixteen features for our system.

An interesting approach that uses PCA for building up a set of features for activity, behavior, vehicle status recognition problems using smartphone sensor [6] has applied to choose suitable features for classification with higher accuracy.

2.3 Online Training Model

Classification is an important step in data mining problem, especially in recognition problem with smartphone sensors class. The most commonly used classifiers are decision tree, KNN, SVM and NB algorithms [13]. In practice, a classifier firstly needs to be trained by using labeled vehicle status database (called training data). There are two training approaches such as the online training method which is performed on smartphones. On the other hand, the offline training is deployed in advance, usually on a local machine. Most of studies use the offline training method because of the computational cost reduction on smartphones. Nonetheless, the modern smartphones have much better computational capacity. This advances of smartphones allows us to implement the online training in our MOVAS in Fig. 1.

In problems of the activity or behaviors or vehicle status recognition using smartphone sensor or wearable sensor data collected from any position, the prediction accuracy is usually of from 70% to 90% [7, 14]. Hence, our paper proposes a refining model using ANN algorithm on smartphone to improve the status prediction results. The training data for this collected by set of tubes includes k statuses. The tube is assigned a label by user and could be updated by the monitoring module. The value k also affects to the processing time of system. By experiment on stop, moving, acceleration, deceleration statuses, we chose the value k of 4. The method and processes of the refining model is shown in Fig. 3.

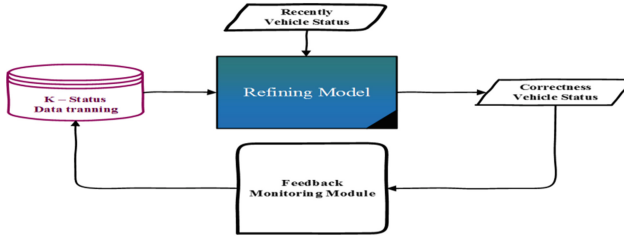


Fig. 3. The refining model for awareness vehicle status

In the traditional vehicles status recognition framework, the training data set is usually fixed and prepared in advance. Since each user might have different, characteristics and habits, for example one might drive faster than others. As a consequence that, the prediction accuracy might fall down when the system is used for another users. Gomes et al. [15] proposed an idea to incrementally update the training data set by using real-time feedbacks from users. We implement this approach in MOV SAS as following: the system provides the corresponding vehicle status prediction in each data window. Then, the user need to confirm the correctness of the result. If the prediction is correct, the data window label based on the correct prediction results is assigned to the training dataset. Continuously, it also assigns label to the tube including k status for data training of refining model. The challenge here is efficient processing with additional information.

3 Experiment and Results

3.1 Experiment Environment

We deploy MOV SAS on the Android from 4.0 to 5.0 platform. The labels of vehicles status database were collected by 30 subjects when they were driving a bicycle, motor-bike and by 20 subjects on car. They freely carried a Samsung galaxy S4, Quad-core 1.6 GHz Cortex-A15 processor, 2 GB of Ram, 2600 mAh battery, Android 4.2.2 Jelly Bean. The set of vehicle status for recognition is {stop, moving, acceleration, deceleration}.

3.2 Data Collection

In our experiment, signal data collected from three types of sensors as acceleration sensor, gyroscope sensor and magnetic sensor. Each sensor returns three values corresponding to x, y, and z coordinates. The raw data stream is first cut out one seconds at the beginning, and 3 s at the end as these periods time are usually redundant. Then, the data is segmented into a number of windows of 6 s; the overlapping time is one second. We collected 3500 samples for each status from subjects about in two months. The training data set for refining classifier is collected by each subject and it contains characteristic vehicle status.

3.3 The Accuracy of Vehicle Status Awareness

The Weka tool integrated in framework for prediction. In each case, the default setting is used. We also used 10-fold cross validation technical for classification. In the first scenario (S_1), MOV SAS predicts vehicle status with traditional method which do not use user feedbacks, PCA and refining model. In the second scenario (S_2), MOV SAS predicts vehicle status utilize PCA, refining model and user feedbacks to enhance the prediction accuracy. The obtained results are expressed in Table 1.

Table 1. The prediction accuracy(%) of MOV SAS with scenario S_1 and S_2

	Random forest		KNN		Naïve Bayes		SVM	
	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
Stop	83.00	94.10	76.00	81.00	78.00	83.00	71.25	75.65
Moving	78.65	90.85	69.19	76.00	63.15	65.15	52.00	69.52
Deceleration	74.19	86.15	63.00	74.35	60.00	71.16	57.80	61.78
Acceleration	78.66	87.75	69.16	73.86	63.00	73.45	66.25	69.75
Average	78.63	89.71	67.59	76.30	66.04	73.19	61.83	69.18

As shown in Table 1, the prediction accuracy is clearly improved when PCA, refining model and user feedbacks are applied in most of cases. Especially for the case of predicting moving status by SVM, the accuracy is increased by 17.52%. These improvements highlight the effectiveness of the PCA and refining model strategy used by MOV SAS. The RF algorithm is the most suitable for our MOV SAS framework since it always offers higher accuracy compared with the other classifiers, i.e. KNN, Naïve Bayes, and SVM. The accuracy of scenario S_2 can be up to 94.10% when Random Forest classifier is used. The accuracy for detecting deceleration status is lower than that of others. The reason is due to misinterpreting some similar patterns such as slowly moving, slowing acceleration and deceleration. We note that our MOV SAS framework allows detecting the current vehicle status in the condition that their smartphones may put at any position and in any direction.

3.4 The Processing Time

In scenarios, they usually require additional time for processing such information. We counted and compared the time for prediction on S_1 , S_2 . The experiment result shows the average time to detect each type of vehicle status by MOV SAS. The Random Forest spends the least time for detecting vehicle status as comparing to KNN, Naïve Bayes, and SVM. The average processing time is of 2.75 s for detecting the status using RF and maximum of 3.75 s using SVM.

4 The Conclusion and Future Work

In this paper, we proposed a flexible framework, called MOV SAS, for detecting current vehicle status when the smartphones are randomly placed in any position and at any

direction. Moreover, our proposed framework uses PCA to select suitable features and refining model. Following, the real-time feedbacks from users are used to increase the prediction accuracy. In the experiments, MOVSAAS can achieve on average 89.71% accuracy for detecting four predefined vehicles status, i.e. Stop, Moving, Acceleration, and Deceleration on bicycle, motorbike and car. Furthermore, RF classifier is a promising one for our framework. In the future, we are planning further improving the current framework to either increase prediction accuracy or reduce the processing time.

References

1. Guiry, J.J., et al.: Activity recognition with smartphone support. *Med. Eng. Phys.* **36**(6), 670–675 (2014)
2. Wen, J., Wang, Z.: Sensor-based adaptive activity recognition with dynamically available sensors. *Neurocomputing* **218**, 307–317 (2016)
3. Shoaib, M., Scholten, H., Havinga, P.J.: Towards physical activity recognition using smartphone sensors. In: 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC). IEEE (2013)
4. Sang, V.N.T., Thang, N.D., Toi, V., Hoang, N.D., Khoa, T.Q.D.: Human activity recognition and monitoring using smartphones. In: Toi, V.V., Lien Phuong, T.H. (eds.) 5th International Conference on Biomedical Engineering in Vietnam. IP, vol. 46, pp. 481–485. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-11776-8_119](https://doi.org/10.1007/978-3-319-11776-8_119)
5. Halim, Z., et al.: Artificial intelligence techniques for driving safety and vehicle crash prediction. *Artif. Intell. Rev.* **46**, 1–37 (2016)
6. Cui, L., Li, S., Zhu, T.: Emotion detection from natural walking. In: Zu, Q., Hu, B. (eds.) HCC 2016. LNCS, vol. 9567, pp. 23–33. Springer, Heidelberg (2016). doi:[10.1007/978-3-319-31854-7_3](https://doi.org/10.1007/978-3-319-31854-7_3)
7. Berchtold, M., et al.: Actiserv: activity recognition service for mobile phones. In: International Symposium on Wearable Computers (ISWC). IEEE (2010)
8. Randell, C., Muller, H.: Context awareness by analysing accelerometer data. In: The Fourth International Symposium on Wearable Computers. IEEE (2000)
9. Mathie, M., et al.: Classification of basic daily movements using a triaxial accelerometer. *Med. Biol. Eng. Comput.* **42**(5), 679–687 (2004)
10. Schmidt, A.: Ubiquitous computing-computing in context. Lancaster University (2003)
11. Healey, J., Logan, B.: Wearable wellness monitoring using ECG and accelerometer data. In: Ninth IEEE International Symposium on Wearable Computers (ISWC 2005). IEEE (2005)
12. Figo, D., et al.: Preprocessing techniques for context recognition from accelerometer data. *Pers. Ubiquit. Comput.* **14**(7), 645–662 (2010)
13. Chetty, G., White, M., Akther, F.: Smart phone based data mining for human activity recognition. *Procedia Comput. Sci.* **46**, 1181–1187 (2015)
14. Okeyo, G., et al.: Dynamic sensor data segmentation for real-time knowledge-driven activity recognition. *Pervasive Mob. Comput.* **10**, 155–172 (2014)
15. Gomes, J.B., et al.: Mars: a personalised mobile activity recognition system. In: 2012 IEEE 13th International Conference on Mobile Data Management. IEEE (2012)

A Study on OPNET State Machine Model Based IoT Network Layer Test

Young-hwan Ham¹(✉), Hyo-taeg Jung¹, Hyun-cheol Kim², and Jin-wook Chung³

¹ Quality Innovation Team, ETRI, 218 Gajeong-ro, Yuseong-gu, Daejeon, Korea
{yhham, htjung}@etri.re.kr

² Computer Science, Namseoul University, 21 Maeju-ri, Seonghwan-eup, Cheonan,
Chungcheongnam-do, Korea
hckim@nsu.ac.kr

³ Computer Engineering, SungKyunKwan University, 2066 Seobu-ro,
Jangan-gu, Suwon-si, Gyeonggi-do, Korea
jwchung@skku.edu

Abstract. Model based testing can enable automated test case generation for many kind of application. Even test code can be generated from the model by specialized tools. IoT protocols for network layer have many constraints for exhaustive or manual testing because of battery problem and large number of sensor nodes. To solve these testing constraints, this paper proposes an efficient State Machine based test case generation for IoT network layer by using OPNET simulation model and test case generation tool. The size of test suite is compared according to the size of State Machine model from OPNET.

Keywords: State machine · IoT · Network layer test · Test generation · OPNET

1 Introduction

IoT (Internet of Things) normally has hundreds or thousands of sensor nodes and battery constraints in case of outdoor field test. Therefore, it is necessary to efficient testing method for the IoT.

In addition, it is necessary to consider an application layer interaction which is useful for dynamics caused by mobility, failures, and dynamic power modes of IoTs. The traditional layered structure passes a limited set of information over defined interfaces between separate layers of the protocol. It is good for abstraction and development, but bad for efficiency in case that high level information is useful in over layers or vice versa. The examples are power control, overlay service, error control, aggregation, fusion, localization, service discovery, semantic addressing, etc.

In this study, we are going to use State Machine-based testing for the cost saving in test case design, systematic testing and controlling of the model coverage and the number of tests [1, 2]. It can help the early detection of flaws and ambiguities in the specification, and the conformance of implementation to the corresponding State Machine model.

For the State Machine based testing of IoT protocol, application layer and network layer should be reflected on the protocol State Machine to cover the standard

specification. It is very critical to limit the number of test case in IoT because of battery power constraints, so it is necessary to draw efficient test cases [3].

2 OPNET Modelling for Test Case Generation

2.1 IoT Network Layer and OPNET Simulation

ZigBee sensor network standard, which is a representative low-power standard for IoT applications, was modelled by OPNET [4].

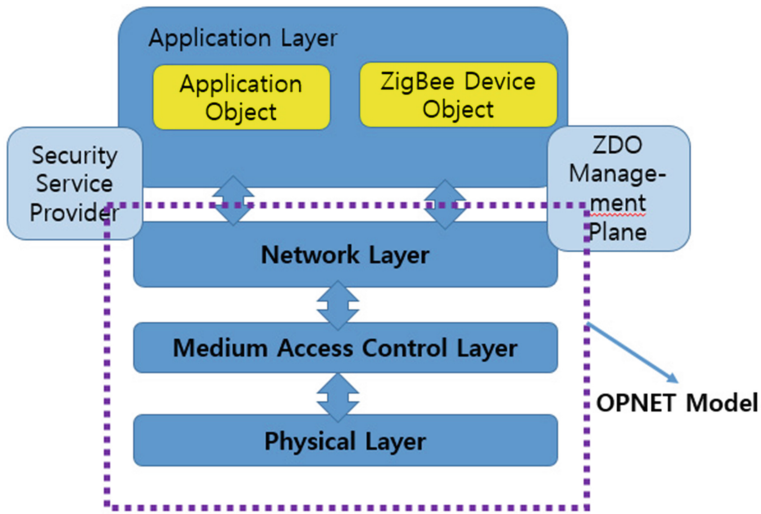


Fig. 1. The OPNET model of zigbee protocol stack

A simulation model based on OPNET was developed for the simulation of sensor networks. Through the OPNET based simulation, various parameters related to the sensor network can be set in advance to find suitable ones for the application system. The candidate technology or structure to be applied to the developed system can be evaluated in advance to receive feedback. Using the existing OPNET library, the interoperability between different protocols and systems can be verified in advance [5] (Fig. 1).

1. Physical layer: This layer is the lowest layer. It consists of two layers, operating in two separate frequency ranges [4].
2. Medium Access Control layer: The responsibility of the MAC layer is to control access to the radio channel using CSMA/CA. The MAC layer provides support for transmitting beacon frames, network synchronization and reliable transmission [4].
3. Network layer: This layer sends and receives data to and from the application layer. It performs the task of associating to and disassociating from a network. This layer network protocol allows us to extend the battery life of the nodes, allowing it to do

only the minimum work when it needs to transmit data [4]. The emphasis is on very low cost communication of neighboring devices with no other wired/wireless network infrastructure. The low cost communication results in lower power consumption, which is even more important.

The following shows the result of simulating a Beacon-enabled ZigBee Network using the ZigBee library.

- OPNET Simulation: End-to-end delay and Receiver-on time of ZigBee network
 - Mode 1: Random Beacon Slot (Beacon Enabled Mode)
 - Mode 2: Proposed Beacon Scheduling (Beacon Disabled Mode)

Figure 2 shows that the delay increases exponentially with Beacon-Disabled mode as sensor node increases. Figure 3 shows that the beacon-enabled mode has much less awake time than the beacon-enabled mode, which is much better in terms of battery consumption.

In a ZigBee application that generates traffic with a frequency lower than a certain level, such as remote meter reading and environmental monitoring, the battery usage time can be greatly improved when the beacon-enabled mode is applied. However, it can be adversely affected in a heavy traffic environment. Through the event/traffic simulation results, the correct operation of OPNET model including network layer has been verified.

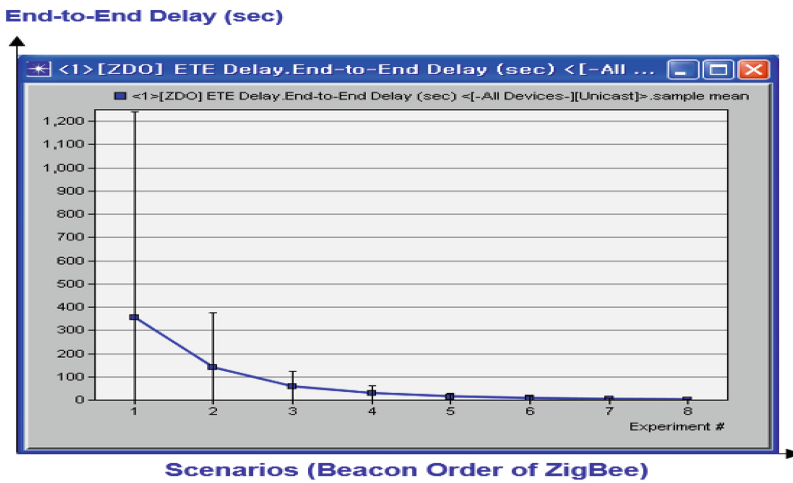


Fig. 2. End-to-End Delay simulation result by OPNET model

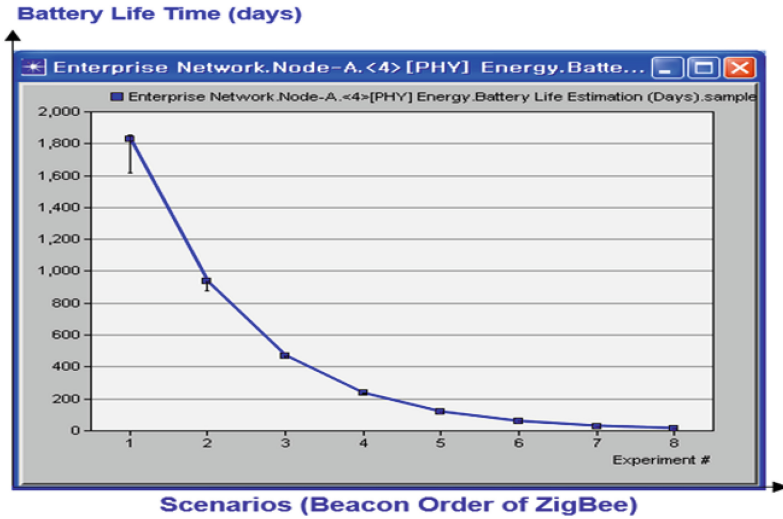


Fig. 3. Receiver-on time (Battery Life Time) simulation result by OPNET model

2.2 Network Layer State Machine for Test Case Generation

OPNET uses a state machine based modeling technique to simulate each layer. Figure 4 shows the network protocol layer state machine of the implemented model. We propose a method to efficiently generate a test case by using the state machine. The circles below represent each state, and the terms on the arrows represent interrupts.

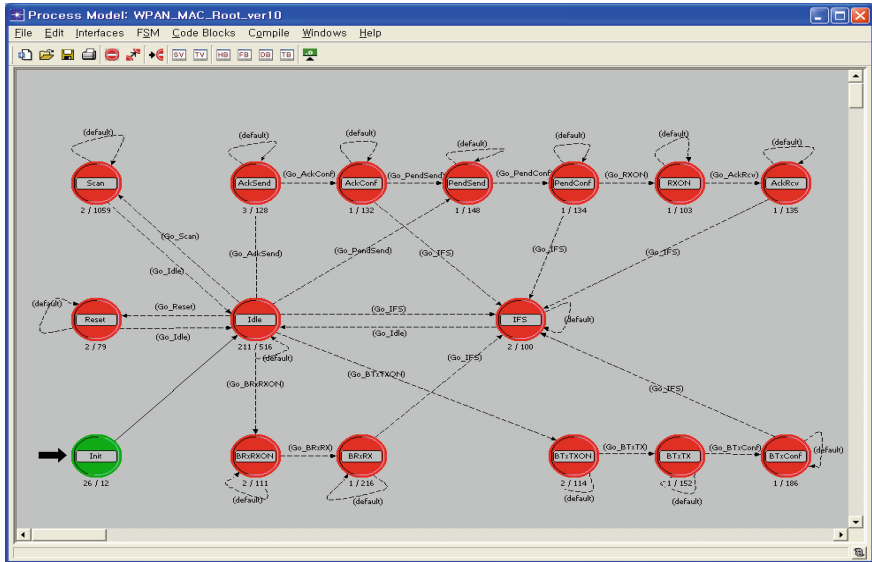


Fig. 4. The State Machine of OPNET model network layer

In this way, the finite state machine that defines the operation of the network layer can check whether the network layer of the actually implemented sensor node is operating properly. An automated tool such as ModelJunit [6] can be used to obtain a test case that can test the operation of the network using the network state transition diagram. ModelJunit is an available and prevalent State Machine based test case generation tool. It is very easy to learn and convenient because it is based on Java language.

3 Experiment Result and Analysis

The State Machine of OPNET network layer is slightly modified and simplified for excluding the meaningless interrupt in point of test case generation, such as “default”. The final state machine is as follow. The possible test cases for network layer can be generated from this state machine diagram by using appropriate test case generation tool.

The experiments for test case generation were executed by ModelJUnit, and each experiment result has averaged among 10 times executions of test.

The comparison experiment has been performed by using Random Walk & Greedy Random Walk test case generation algorithm [6]. Random Walk algorithm simply tests a system by making random walks through a State Machine model, and Greedy Random walk gives priority to transition never taken before. In addition to generation algorithm, test case coverage is also important factor and the ModelJUnit supports three kinds of coverage metrics such as state metric, event metric, and transition metric [7]. The state metric shows how many states are traveled at least once. The event metric shows how many events are triggered at least once. Transition metric represents how many transitions are exercised at least one was chosen in this study because it is important to cover every state transition to ensure correct operation of the network layer [7]. The details of experiments for comparison are as follows.

- **Experiment 1: The number of state is same as original OPNET network model**

When the state machine model had 16 states and the number of interrupt (event) was 25 (Fig. 5), test length (the number of test suite) for 100% transition coverage metric was as follows.

- State Machine Model: 16 states, 25 event
- Random Walk: 210
- Greedy Random Walk: 160

When the relatively small number of State Machine states were randomly added for test, the number of test length (Random Walk) was exponentially increased.

- **Experiment 2: The number of state is reduced by simplifying original OPNET network model**

The similar group of states are merged as follows for the simplification of state machine mode as follows.

- (ACK SEND, ACK CONF) => ACK SEND

- (PEND SEND, PEND CONF) => PEND SEND
- (RXDN, ACK RCV) => RXDN
- (BRxRXON, BRxRX) => BRxRXON
- (BTxTXON, BTxTX, BTxCONF) => BTxTXON

When the Network had 10 states and the number of interrupt (event) was 19 (Fig. 6), test length was as follows.

- State Machine Model: 10 states, 19 event
- Random Walk: 170
- Greedy Random Walk: 50

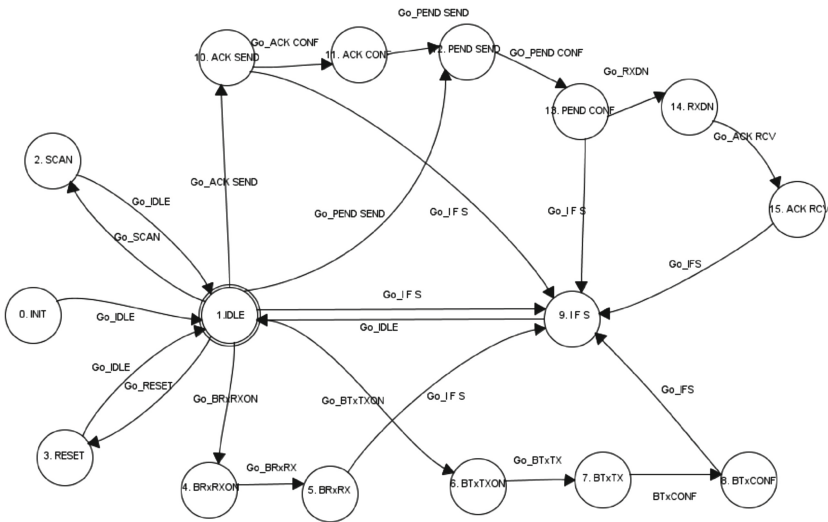


Fig. 5. The State Machine of network layer for ModelJUnit

The test length for the network testing can be very critical in such resource constrained IoT environment. The experiment 2 shows that simplified by state merging can dramatically reduce the number of test case especially in case of greedy random walk. It should be considered that how we can reduce the number of states by simplifying the state machine from OPNET model. The simplified version of state machine model can be also verified by putting & executing it in OPNET modeler. The effect of merged states can be monitored by the simulation result of OPNET. This can be a reciprocal way for an efficient test case generation and network simulation.

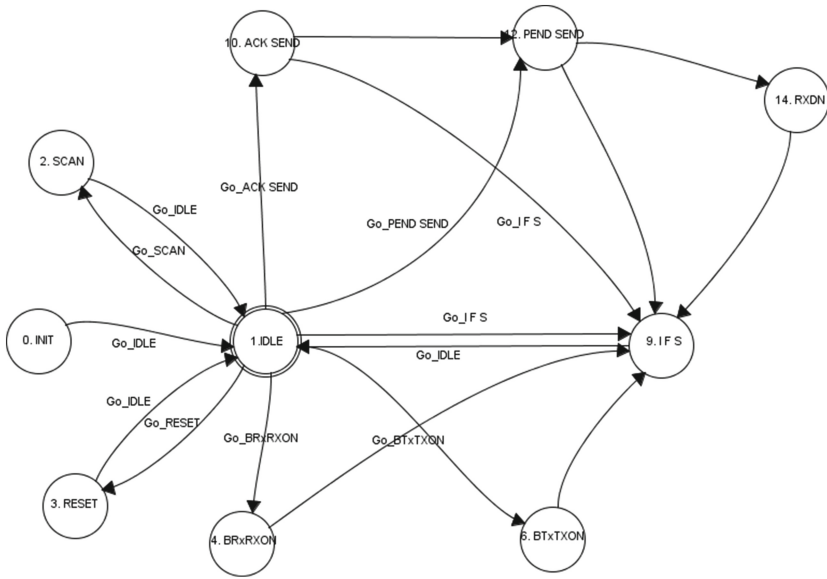


Fig. 6. The Simplified State Machine of network layer for ModelJUnit

4 Conclusions

The IoT normally has hundreds or thousands of sensor nodes and battery constraints in case of outdoor field test. Therefore, it is necessary to efficient testing method for the IoT. In addition, it is necessary to consider a network layer operation which is useful for dynamics caused by mobility, failures, and overlay modes of IoTs.

For the State Machine based testing of IoT protocol, ModelJUnit tool and OPNET ZigBee model are used. The tool generated the test cases by using OPNET network layer state machine model. By the result of these experiments, we realized that the test cases can be generated by using the state machine model of OPNET. Because the number of test length could be rapidly increased in proportion to the number of state machine, a simplification of FSM is necessary. The effect of merged states can be checked & monitored by simulating the simplified model in OPNET.

The more various test case generation experiment is also necessary for verification of the network layer design. The ModelJUnit has many benefits as a tool of State Machine based test case generation. We have used transition-tour test generation algorithm by the tool, but it doesn't support other test sequence generation methods [8, 9]. The study about overcoming above weaknesses also should be done in the future.

References

1. Gansner, E., North, S.: An open graph visualization system and its Appl. Soft. Pract. Experience **30**, 1203–1233 (1999)
2. Javed, A., Strooper, P., Watson, G.: Automated generation of test cases using model-driven architecture. In Proceedings of the 2nd International Workshop on Automation of Software Test (AST 2007), p. 3 (2007)
3. Link, J., Fröhlich, P.: Unit Testing in Java: How Tests Drive the Code. Morgan Kaufmann Publishers Inc., Burlington (2003)
4. IEEE Standards Association: IEEE standard for local and metropolitan area networks–Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) (2011). <http://standards.ieee.org/about/get/802/802.15.html>
5. Xiaolong, L., Peng, M.: OPNET-based modeling and simulation of mobile Zigbee sensor networks. Peer-to-Peer Netw. Appl. (2015). doi:10.1007/s12083-015-0349-8
6. <http://www.cs.waikato.ac.nz/~marku/mbt/modeljunit/> (2016)
7. Utting, M., Legeard, B.: Practical Model-Based Testing: A Tools Approach, pp. 157–162. Morgan Kaufmann Publishers Inc., San Francisco (2007)
8. Lelis, L., Pedrosa, C.: A new method for incremental testing of finite state machines. In: NFM2010
9. Ural, H.: Formal methods for test sequence generation. Comput. Commun. **15**(5), 311–325 (1992)

A Secure Localization Algorithm Based on Confidence Constraint for Underwater Wireless Sensor Networks

Xiaofeng Xu¹, Guangyuan Wang², Yongji Ren^{3(✉)}, and Xiaolei Liu⁴

¹ Science and Technology on Communication Information Security Control Laboratory, Jiaxing 314033, China

² Department of Military Training, Naval Aeronautical and Astronautical University, Yantai 264001, China

³ Department of Command, Naval Aeronautical and Astronautical University, Yantai 264001, China
lenglengqiuyu@sina.com

⁴ Department of Electrical Engineering, Yantai Vocational College, Yantai, 264001, China

Abstract. This paper proposed a novel secure localization algorithm based on confidence constraint for Underwater Wireless Sensor Networks (UWSNs). In recent years, UWSNs have attracted a rapidly growing interest from ocean battlefield surveillance. As essential technology, secure localization is crucial to the location-based applications. However, the localization process has been restricted by the adverse battlefield environments, e.g. the confidence problem of reference nodes and information due to disturbances or attacks, which lead to obvious degradation of localization security and accuracy. To solve this issue, we transformed the secure localization into a confidence constraint satisfaction problem. Zero-sum game method has been utilized to deal with the confidence problem of reference information. Simulation results show that our algorithm is an effective and efficient approach to localization for UWSNs.

Keywords: UWSNs · Localization · Security · Confidence constraint

1 Introduction

During the last few years, there has been a rapidly growing interest in Underwater Wireless Sensor Networks (UWSNs), which brought us a new way to sense and monitor the adverse battlefield environments [1]. As an essential technology, the localization performance significantly affects the location-based applications. In complex ocean battlefield, several kinds of adverse factors would lead to obvious degradation of localization security and accuracy [2], e.g. the potential malicious attacks, the unreliable reference nodes and reference information, etc. Extensive research has been conducted in this interesting area [3–5]. Therein, Alfao et al. considered the security of localization under limited trust anchor nodes [4]. It introduced three algorithms to enable the sensor nodes to determine their positions. But it would fail when the malicious anchor nodes are in colluding condition. Chen et al. proposed to make each locator build a conflicting-set and then the sensor can use all conflicting sets of its neighboring locators to exclude

incorrect distance measurements of its neighboring locators [5]. However, the limitation of the scheme is that it only works properly when the system has no packet loss.

Actually, the substantial reason of the above problem is that the localization has been restricted by confidence constraint of reference information. Therefore, a novel secure localization algorithm based on confidence constraint has been proposed. We transformed the secure localization problem into a confidence constraint satisfaction problem (CSP) [6]. A confidence CSP, i.e. the determination problem of secure localization, has been defined by a constraint contractor C , with an interval domain $[x]$. Then, the localization issues will be tackled in a constraint CSP framework.

2 Confidence Constraint Based Secure Localization Algorithm

2.1 Confidence Constraint of Reference Nodes

In this phase, our primary objective is to find out which anchor nodes should be employed as reference nodes so that the utilization in localization is secure. To deal with the problem, zero-sum game method will be employed [7].

Formulate game domain. Firstly, the ordinary node N_i initiates an inviting request to its neighbor or multi-hop anchor nodes, namely set X . If the anchors in X are overcommitted, they respond the abandoning ACK to N_i . Otherwise, the nodes respond the joining ACK. Then the local game domain of node N_i is created, and the anchors with joining ACK will become the game players. As a game player, there are two actions <keep, reject> to enforce for N_i . Assume that the UWSN is composed by n nodes and m game domains acting on it. Note that the m game domains could co-exist over the network so that the game-plays could be calculated in the concurrent way.

Calculate cost functions. The node N_i announces the localization information to all the players. Then, each player in the local game domain receiving the announcement calculates its cost function [8]. The cost function of game domain k is given by

$$J^k(t, x, u^k) = \int_t^{t_f} L^k(t, x, u^k) dt + \Psi^k(x_{t_f}^k), \quad 1 \leq k \leq m \quad (1)$$

with the running cost function

$$L^k(t, x, u^k) = \sum_{i \in V_k} c_i(u_i) - \sum_{i \in V_k} \sum_{j \in V_{k'}, k' \neq k} \left[a_{i,j} e^{-\theta_{ij}^k(x_{i,t_k} - x_{j,t_k})} - a_{i,j} e^{-\omega_{ij}^k(x_{i,t_n} - x_{j,t_n})} \right] \quad (2)$$

where x_i describes the running states of N_i , u^k describe the control vectors of group k , V_k is the node set and Ψ^k is the terminal cost function. c_i is control cost function of node N_i . $a_{i,j} e^{-\theta_{ij}^k}$ are attack payoffs running functions and $a_{i,j} e^{-\omega_{ij}^k}$ are information loss running functions of node N_i to N_j .

Play game and make decision. At the first time of the play, all players make their action based on their payoffs, i.e. if the payoff is positive, broadcasting a ‘keep’ message to all players, or else broadcasting ‘reject’. All localization groups wish to maximize their payoffs, i.e. minimize the respective cost functions. Let group k ’s admissible control set be u^k . As the game repeats, the admissible control combination, i.e. the actions of game-play, can be denoted as a Nash equilibrium solution if it satisfies:

$$J^k(0, x, u^k) \leq J^k(0, x, \langle \hat{u}|k \rangle), \quad 1 \leq k \leq m \quad (3)$$

On this basis, the ordinary node can adopt the players with ‘keep’ as the reference anchors, and then broadcasts a message to all players to dismiss the game domain.

2.2 Confidence Constraint Satisfaction Problem

Solving the confidence CSP in an interval analysis approach consists of finding the intersection that contains all possible solutions. The set of the intervals regarding to ordinary node N_i actually is the set of the constraints f_1, f_2, \dots, f_k . The location of node N_i can be described by $X_i = [x_i, y_i, z_i]^T$. The distance from X_u to X_i can be denoted by $\zeta_{ui}^I = [\zeta_{ui}^{I-}, \zeta_{ui}^{I+}]$. Consider the intersection of two intervals ζ_{u1}^I and ζ_{u2}^I , it can be computed by $\zeta_{u1}^I \cap \zeta_{u2}^I = [\max\{\zeta_{u1}^{I-}, \zeta_{u2}^{I-}\}, \min\{\zeta_{u1}^{I+}, \zeta_{u2}^{I+}\}]$. The admissible solutions of node X_u can be rewritten as $F_u = \bigcap_{i=1}^k \{\zeta \in \zeta^I; \zeta_{ui}^I = [\zeta_{ui}^{I-}, \zeta_{ui}^{I+}]\}$. Regarding the coordinates of all sub-boxes’ centers as samples of X_u , we can get a sample set $F_u = \{\Theta_1, \Theta_2, \dots, \Theta_n\}$, and the centre of Θ_n can be found by $\zeta_n^* = (\zeta_n^- + \zeta_n^+)/2$. Then the optimum point estimate, i.e. the desired coordinates of ordinary node X_u can be obtained by

$$\hat{W}_u = \arg \min_{W_u} \sum_{i=1}^k (\|\zeta_n^* - W_i\|_2 - d_{ui})^2.$$

3 Performance Evaluation

In our simulation experiments, 400 nodes with adjustable transmission range R are randomly distributed in a $3000 \times 3000 \times 200$ region. For comparison with classical approaches relying on secure hypotheses, different effective reference anchor percentages are considered in our simulation by varying the malicious nodes percent. Moreover, the DV-distance localization scheme has been simulated for comparison.

Figure 1 shows the accuracy comparisons with different anchors. When we varied the effective anchor percentage from 4% to 12%, i.e. the number of effective anchor nodes varying from 16 to 48, the localization error decreased by 50%. However, the DV-distance scheme only decreased by 25%, when the network connectivity is 9 and the malicious nodes percent is 5%. This suggests that our scheme can achieve higher localization accuracy and security in same malicious nodes percentage.

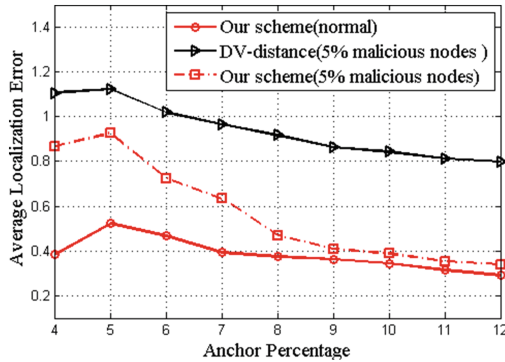


Fig. 1. Average localization error vs. anchor number

Conclusion. Proposed is a novel confidence constraint based secure localization algorithm. The advantage of our framework is that both the malicious nodes and the reference information can be treated as information uncertainty and casted into game process. Simulation results show that it is an effective and efficient approach.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (grant no. 61501488).

References

1. Erol-Kantarci, M., Mouftah, H.T., Oktug, S.: A survey of architectures and localization techniques for underwater acoustic sensor networks. *Commun. Surv. Tutorials IEEE* **13**(3), 487–502 (2011)
2. Tan, H., Diamant, R., Seah, W.K.G., Waldmeyer, M.: A survey of techniques and challenges in underwater localization. *Ocean Eng.* **38**(14–15), 1663–1676 (2011)
3. Xing, T., Jian, L.: Cooperative positioning in underwater sensor networks. *IEEE Trans. Signal Process.* **58**(11), 5860–5871 (2010)
4. Alfao, J.G., Barbeau, M., Kranakis, E.: Secure localization of nodes in wireless sensor networks with limited number of truth tellers. In: *Proceedings of 7th Annual Communication Networks and Services Research Conference*, pp. 86–93 (2009)
5. Chen, H., Lou, W., Wang, Z.: Conflicting-set-based wormhole attack resistant localization in wireless sensor networks. In: *Proceedings of the 6th International Conference on Ubiquitous Intelligence and Computing*, pp. 296–309 (2009)
6. Jaulin, L., Kieffer, M., Didrit, O., Walter, E.: *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics*. Springer, London (2001)
7. Cheng, G., Chen, H.: Game model for switch migrations in software-defined network. *Electron. Lett.* **50**(23), 1699–1700 (2014)
8. Ning, G., Yang, D., Tie, L., Cai, K.-Y.: Nash equilibrium of time-delay interaction complex networks subject to persistent disturbances. *IET Control Theory and Applications* (2012)

Networks and Information Systems

Generating Time Series Simulation Dataset Derived from Dynamic Time-Varying Bayesian Network

Garam Lee, Hyunjin Lee, and Kyung-Ah Sohn^(✉)

Department of Software and Computer Engineering, Ajou University, Suwon-si, South Korea
kasohn@ajou.ac.kr

Abstract. Numerous network inference models have been developed for understanding genetic regulatory mechanisms such as gene transcription and protein synthesis. Dynamic Bayesian network effectively represent the causal relationship between genes and gene and protein. Modern approaches employ single multivariate gene expression data set to estimate time varying dynamic Bayesian network. However, evaluating inferred time varying network is infeasible due to the absence of known gold standards. In this paper, the simulation model for time series gene expression level under certain network structure is proposed. The network can be used for assessing inferred data which is estimated based on simulated gene expression data.

Keywords: Time series data · Dynamic Bayesian network · Simulation study

1 Introduction

For the past decades, numerous network inference methods have been developed to model underlying genetic regulatory mechanisms such as gene transcription and protein synthesis. The main focus of network inference is on investigating the interactions between genes, and attempt to build descriptive models for understanding complex system. For representing causal relationship dynamic Bayesian network (DBN) is one of well-known probabilistic graphical models. While in static Bayesian network the topology of network is fixed [1–3], dynamic Bayesian network is particularly well suited to tackle the stochastic nature of gene regulation and gene expression measurement [4], thus has been widely used for its ability to recover the underlying genetic regulatory network [5]. With development of time series gene experimental expression data estimating time-varying DBN has become feasible. In [4], DBN is inferred based on a penalized likelihood maximization implemented through an extended version of EM algorithm. Also, [6] proposed temporally rewiring networks for capturing the dynamic causal influences between covariates. For estimation, kernel reweighted L_1 -regularized auto-regressive procedure is applied.

However, there has been a challenging problem due to the infeasibility to evaluate inferred time-varying Bayesian network. Traditionally, network inference model has been assessed by comparing predicted genetic regulatory interactions with those known from the biological literature [7]. This approach is controversial due to the absence of

known gold standards, which renders the estimation of the sensitivity and specificity, that is, the true and false detection rate, unreliable and difficult.

Rare attempts to generate simulated gene expression data have been developed. In [8], author proposes simulation model for biological system to try on inferred DBN resulted from the simulated gene expression data. [7] develops simulated gene expression data from a realistic biological network involving DNAs, mRNAs, inactive protein monomers and active protein dimers.

Modern approaches such as [6, 9, 10] make an assumption to fully utilize time series dataset: underlying network structure are sparse, vary smoothly across time, and models first-order Markovian. From the assumption, it is derived that temporally adjacent networks are likely to share common edges than temporally distal networks. This assumption makes it possible to reconstruct time varying network with single multivariate time series data. Intuitively, inferred network resulted from time series gene expression data which is generated from underlying network based on the assumption should be maximally equivalent to the underlying network. In other words, time-varying network made up based on the assumption gives upperbound of performance of network inference model in which gene expression data is generated from the underlying network. Therefore, in this paper totally different approach is used for assessing time varying dynamic Bayesian network. First, time varying network is built, and time series dataset is generated from the network. Then the simulated dataset can be used for measuring the performance of methodologies of which their assumption is based on first-order Markovian model.

2 Method

2.1 Preliminaries

Models describing a stochastic temporal processes can be naturally represented as dynamic Bayesian networks [11]. As defined in [6], taking the transcriptional regulation of gene expression as an example, let $\mathbf{X}^t := (X_1^t, \dots, X_p^t)^T \in \mathbb{R}^p$ be a vector representing the expression levels of p genes at time t , a stochastic dynamic process can be modeled by a “first-order Markovian transition model” $p(\mathbf{X}^t | \mathbf{X}^{t-1})$, which defines the probabilistic distribution of gene expression at time t given those at time $t - 1$. Under this assumption, likelihood of the observed expression levels of these genes over a time series of T steps can be expressed as:

$$p(\mathbf{X}^1, \dots, \mathbf{X}^T) = p(\mathbf{X}^1) \prod_{t=2}^T p(\mathbf{X}^t | \mathbf{X}^{t-1}) = p(\mathbf{X}^1) \prod_{t=2}^T \prod_{i=1}^p p(X_i^t | X_{\pi_i}^{t-1}), \quad (1)$$

where π_i is the set of genes specifying the gene i , and the transition model $p(\mathbf{X}^t | \mathbf{X}^{t-1})$ factors over individual genes. Each $p(X_i^t | X_{\pi_i}^{t-1})$ can be viewed as a regulatory gate function that takes multiple covariates and produce a single response. A simple form of the transition model $p(\mathbf{X}^t | \mathbf{X}^{t-1})$ in a DBN is a linear dynamic model:

$$\mathbf{X}^t = \mathbf{A} \cdot \mathbf{X}^{t-1} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (2)$$

where \mathbf{A} is a matrix of coefficients relating the expressions at time $t - 1$ to those of the next time point, and ϵ is a vector of isotropic zero mean Gaussian noise with variance σ^2 .

Our simulator generates time-series gene expression dataset under assumption (2):

$$x_i^t = \alpha_0 x_i^{t-1} + \alpha_1 \sum_{j \in \pi_i} \beta_j x_j^{t-1} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad (3)$$

where x_i^t is i -th gene expression level at time point t , and α_0 is the parameter to regulate the influence of the target gene expression level at previous time point on one at time point t . β_j is the degree of association that affects gene expression level at target time point. Finally, expression level of each gene at a time point is generated with a noise with 0 mean, and σ^2 variance.

At network building stage, a set of genes is grouped to generate gene expression data based on the group in which a gene is belongs to only one group. Group is made to make it possible to activate associations in the group at the same time. To represent temporal interaction between genes, degree of activation of group is varying over time, and multiple groups are activated at different time point for different time periods. The example of interaction variation is illustrated in Fig. 1.

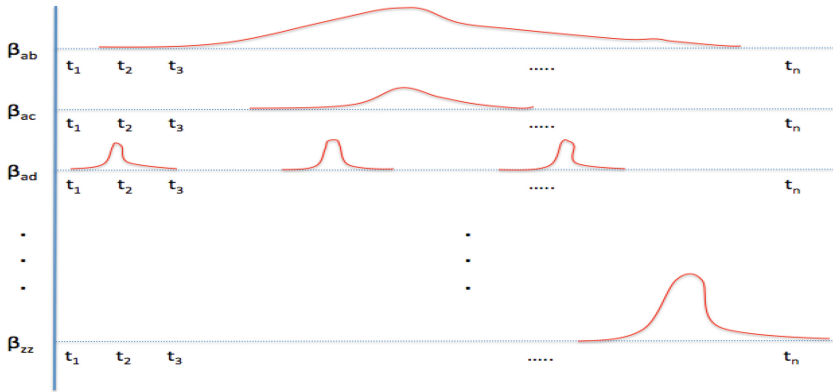


Fig. 1. The examples for variation of interactions possibly appeared in underlying network. β_{ab} is the interaction between gene a and b. It is smoothly increased and decreased in activation over time periods. β_{ad} repeats to be activated spontaneously.

2.2 Algorithm

The algorithm takes parameters the number of genes n , the number of time points m , target influence parameter α_0 . And it produces time varying network and time series gene expression data over m time points, and group information of each gene.

At the first stage, time varying Bayesian network is built from line 2 to 5. Then gene expression level is generated based on underlying network structure. At line 2, each node belongs to a group, and their interactions within the group are randomly set at line 3. Finally, activation period of each group is set randomly.

At second stage, time series gene expression data is generated. The expression levels of genes at initial time point are randomly set ranging from 0.3 to 1. $X^i[j]$ means gene expression level of j -th gene at time point t , and $G[i, j]$ is group number of interaction between i -th gene and j -th gene. Activation period and degree of activation are contained in the matrix $gInfo$ whose row represents group, and first column for the starting point of activation, and second column for ending point of activation, and third column for degree of activation.

Input #gene n , #time points m , target influence parameter α_0

Output time varying networks $\{A^1, A^2, \dots, A^m\}$, time series gene expression data $\{X^1, X^2, \dots, X^m\}$, group sets $\{G_1, G_2, \dots\}$

```

1  Begin
2    Randomly initialize  $X^1$ 
3    Randomly initialize group matrix  $G$ 
4    Randomly initialize beta coefficient matrix  $B$ 
5    Randomly initialize group activation periods  $gInfo$ 
6    for  $i = 1 \dots m$  do
7      for  $j = 1 \dots n$  do
8        for  $k = 1 \dots n$  do
9          if  $G[k, j]$  is not 0 and  $gInfo[G[k, j], 1] \leq i \leq gInfo[G[k, j], 2]$ 
10           if  $k$  equals to  $i$ , then  $X^i[j] = X^i[j] + (1 - gInfo[G[k, j], 3]) \cdot X^{i-1}[k]$ 
11           else  $X^i[j] = X^i[j] + gInfo[G[k, j], 3] \cdot X^{i-1}[k] \cdot B[k, j]$ 
12         Else
13           if  $k$  equals to  $i$ , then  $X^i[j] = X^i[j] + \alpha_0 \cdot X^{i-1}[k]$ 
14           else  $X^i[j] = X^i[j] + \alpha_1 \cdot X^{i-1}[j] \cdot B[k, j]$ 
15       End for
16      $X^i = X^i + \epsilon$ 
17   End for
18 End for
19 End

```

Algorithm 1 The procedure generates time series gene expression data, underlying time varying network, and group information of nodes on input the number of nodes n , the number of time points m , and target influence parameter α_0

3 Result

This section shows the procedure of parameter optimization to generate gene expression level smoothly varying over time. The parameter α_0 is optimized to generate smooth gene expression levels.

First, we attempted to generate small number of genes' simulated data. As shown in Fig. 2, gene expression level grows up to infinity as time increased because the number of genes having influence on target gene is large. As parameter n is increased, the expression level of target gene is not smoothly varying over time because the target gene

affected by its associated gene is changed drastically. It leads us to attempting second experiment with regulation of parameter α_0 . The configuration of setting target influence parameter to .9 generates gene expression level as shown in Fig. 3.

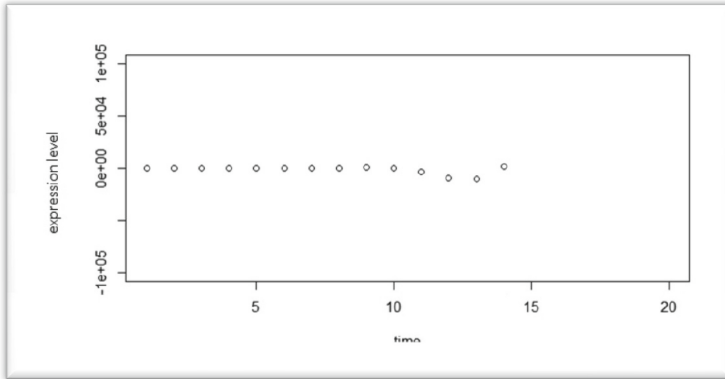


Fig. 2. This is expression level of a gene from 20 genes nodes. The initial expression level is set ranging from 0 to 1

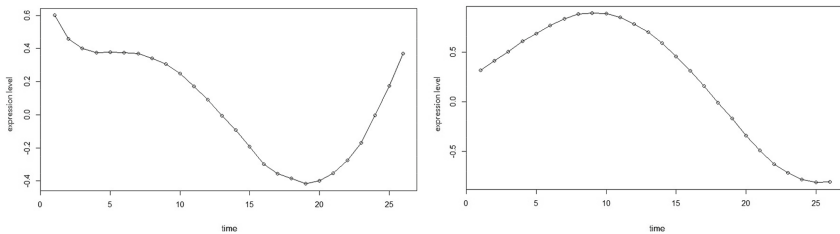


Fig. 3. Two examples among 20 genes. The expression level at initial time point is set ranging from 0 to 1. And target influence parameter α_0 is set to 0.9

In third experiment, network is built based on group. The associations between genes only appeared in group. Figure 4 illustrates simulation data generated from the group setting. Without setting target influence parameter α_0 , gene expression level does not look smooth across time.

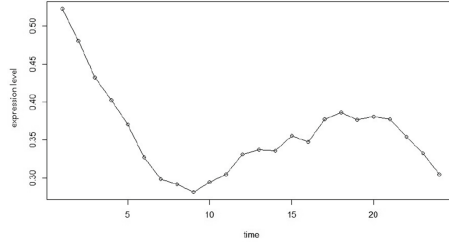


Fig. 4. Gene expression data generated from group setting. The gene expression level at initial time point is set randomly ranging from 0 to 1.

Finally, we investigate how to set α_0 to generate smooth time series gene expression data set as the number of nodes increases. The Figs. 5, 6, and 7 illustrates smooth gene expression levels.

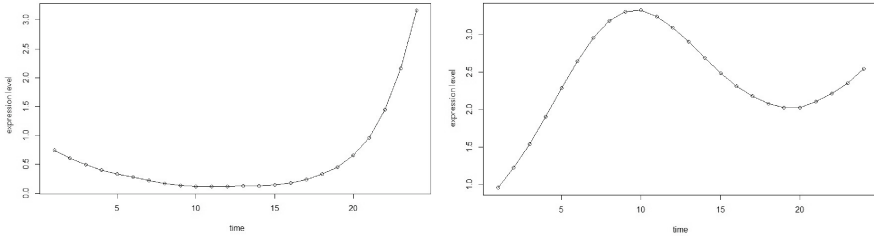


Fig. 5. Gene expression level resulted from setting α_0 to 0.8 and 0.9 for left and right figure respectively. The number of genes is 32.

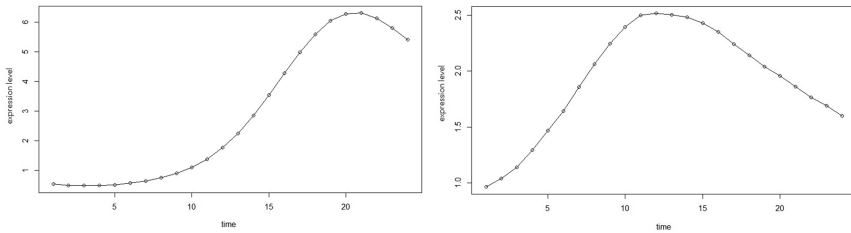


Fig. 6. Gene expression level resulted from setting α_0 to 0.9 and 0.95 for left and right figure respectively. The number of genes is 64.

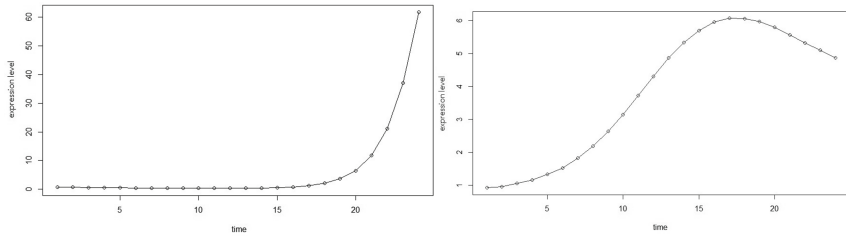


Fig. 7. Gene expression level resulted from setting α_0 to 0.9 and 0.95 for left and right figure respectively. The number of genes is 128.

4 Conclusion

Traditionally, network inference model has been assessed by comparing inferred network with associations between genes known from the biological literature. This approach is infeasible to measure false detection rate. In this paper, we propose a simulation model for the use of assessing network inference algorithm. The proposed simulator generates time varying Bayesian network, time series gene expression data resulted from the network, and group information of genes. For generating gene expression level smoothly varying across time, target influence parameter has been optimized. The simulated dataset can be used to evaluate network inference algorithms in which smoothness of temporal process is assumed. As future work, simulation model for imitating genetic regulatory system can be developed. Currently, gene expression level is affected only by expression level at previous time point. However, in genetic regulatory system, gene expression level can also be affected by protein. Simulation model that attempts to reflect real regulatory system can be widely used to evaluate network inference model under various network structure.

Acknowledgement. This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & communications Technology Promotion) (R22151610020001002).

References

1. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**(3–4), 601–620 (2000)
2. Werhli, A.V., Husmeier, D.: Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* **6**(1) (2007)
3. Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R.: A primer on learning in Bayesian networks for computational biology. *PLoS Comput. Biol.* **3**(8), e129 (2007)
4. Perrin, B.-E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., d'Alche-Buc, F.: Gene networks inference using dynamic Bayesian networks. *Bioinformatics* **19**(suppl 2), ii138–ii148 (2003)

5. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Using Bayesian network inference algorithms to recover molecular genetic regulatory networks. In: International Conference on Systems Biology (2002)
6. Song, L., Kolar, M., Xing, E.P.: Time-varying dynamic bayesian networks. In: Advances in Neural Information Processing Systems, pp. 1732–1740 (2009)
7. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* **19**(17), 2271–2282 (2003)
8. Smith, V.A., Jarvis, E.D., Hartemink, A.J.: Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* **18**(suppl 1), S216–S224 (2002)
9. Song, L., Kolar, M., Xing, E.P.: KELLER: estimating time-varying interactions between genes. *Bioinformatics* **25**(12), i128–i136 (2009)
10. Ahmed, A., Xing, E.P.: Recovering time-varying networks of dependencies in social and biological studies. *Proc. Nat. Acad. Sci.* **106**(29), 11878–11883 (2009)
11. Kanazawa, K., Koller, D., Russell, S.: Stochastic simulation algorithms for dynamic probabilistic networks. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 346–351. Morgan Kaufmann Publishers Inc. (1995)

AMI-SIM: An NS-2 Based Simulator for Advanced Metering Infrastructure Network

Nam-Uk Kim^(✉) and Tai-Myoung Chung

Department of Computer Engineering, School of Information and Communication Engineering,
Sungkyunkwan University, Suwon-si, Republic of Korea
nukim@imtl.skku.ac.kr, tmchung@ece.skku.ac.kr

Abstract. One of the main components of the smart grid is advanced metering infrastructure, which is responsible for delivering, concentrating and analyzing energy usage data. For an advanced metering infrastructure should cover extensive area like a city or a province, if organizing systems and lines were inadvertently chosen or their structure were not properly configured, they would generate huge waste of various resource or cause great damage to system itself. That's why the simulation tool for large scale advanced metering infrastructure network is needed. In this paper, we defined requirements of advanced metering infrastructure network simulator useful for choosing proper system spec in given advanced metering infrastructure network topology, and proposed simulator designed to follow these requirements. The core engine of simulator is NS-2 but original NS-2 was modified to match our goal of simulation. Several tests were performed to evaluate if the performance of a node properly have an effect on simulation result, and to evaluate accuracy of capability usage degree calculation performed by simulator. These tests show that proposed simulator is available for practical use.

Keywords: Smart grid · AMI · Simulator · NS-2

1 Introduction

One of the main components of the smart grid is Advanced Metering Infrastructure (AMI), which is responsible for delivering, concentrating and analyzing energy usage data. Because an AMI should cover extensive area like a city or a province, if organizing systems and lines were inadvertently chosen or their structure were not properly configured, they would generate huge waste of various resource or cause great damage to system itself. That's why the simulation tool for large scale AMI network is needed.

In this paper, we describe the design of AMI Simulator (AMI-SIM), which is useful for finding out proper system specifications, numbers, and configuration in given AMI network topology. The core of AMI-SIM is NS-2, but we modified it to follow the requirements we have preliminarily drawn up.

This paper is organized as follows. The Sect. 2 introduces general AMI network configuration and operations. The Sect. 3 presents the architecture of proposed AMI network simulator. In this section, we introduce the design of core engine and analyzer module of the simulator. In the Sect. 4, we execute AMI-SIM under several

conditions, and evaluate if AMI-SIM is valid for practical use. Finally, we conclude this paper in the Sect. 5.

2 AMI Network Architecture

General AMI network configuration is as shown in Fig. 1. It has a hierarchical structure like tree, in which the MDMS is root node, and Smart Meters are leaf nodes.

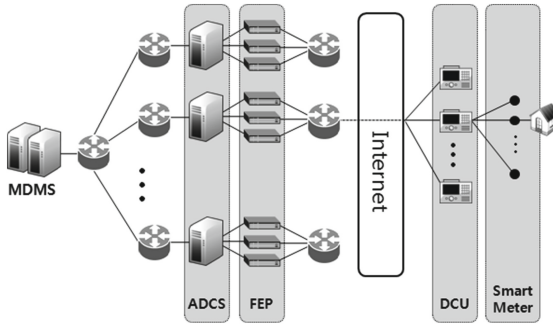


Fig. 1. AMI Network Configuration

AMI network operations are classified into 5 task groups: periodic task group, non-periodic task group, SM configuration task group, DCU/TR configuration and get-status task group, background task group. Periodic task group is for collecting metering data or Load Profile (LP) data from smart meters. The other task groups are activated only at the needed moment, for specified nodes. Therefore, periodic task group generates so much far heavier, more consistent data traffic than the other task groups. Periodic tasks take following steps.

1. An ADCS sets up task start time of periodic tasks, and notify the task start time to each DCUs in its control domain.
2. A DCU starts collecting metering data at the time notified by ADCS. These tasks are performed by DLMS/COSEM protocol.
3. The DCU makes connection to one of smart meters and requests it to send metering data.
4. Then the smart meter would send requested data.
5. The DCU continues 2 and 3 for the other smart meters until it receives all metering data from all smart meters in its control domain.
6. The DCU sends collected metering data to ADCS by SCMP protocol at the pre-defined time.

3 The Design of AMI Network Simulator

This section presents the architecture of proposed AMI network simulator, and then, the design of core engine and analyzer module of the simulator.

3.1 The Architecture of AMI Network Simulator

Overall architecture of AMI-SIM is as shown in Fig. 2. Users can access the simulator via web browser anywhere.

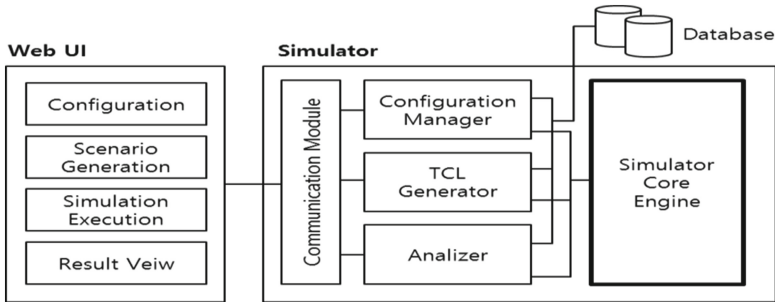


Fig. 2. AMI-SIM Architecture

Web UI provides interface for real simulator. A user can configure nodes and topologies of AMI network in detail, execute simulation, view simulation results with Web UI. Configuration manger provides store, modification, deletion of configuration information. Tool Command Language (TCL) generation module makes a TCL file for NS-2 simulation with given configuration information. Analyzer generates result using trace data after simulation, and reports it to user. Simulator core engine takes a TCL file for input, executes simulation, and generates several trace files for output. The simulator core engine is modified NS-2 simulator.

This paper focuses on the design of simulator core, its trace data and how to analyze it.

3.2 The Design of Simulator Core Engine

The NS-2 simulator includes many features for network protocol simulation, but it focuses on protocol and links, so it does not provide mechanism to make delay generated by producing data reflect system performances. Therefore it is inevitable to modify design of NS-2 node.

Typical design of NS-2 node is as shown in (a), Fig. 3. NS-2 does not provide mechanism to make delay generated by producing data reflect system performances. Also, it does not generate trace data necessary for our requirements. Finally, there's no object, in which AMI specified protocols are implemented in NS-2.

Modified NS-2 node design is as shown in (b), Fig. 3. New objects added to NS-2 node are as follows.

- **AMI_Traffic Object:** AMI_Traffic object generates trace data for total traffic, peak traffic, peak buffer, average buffer.
- **Object for NIC:** To apply performance and capacity of network interface on each AMI system to simulation, NIC_Queue and AMI_NIC objects are added to NS-2

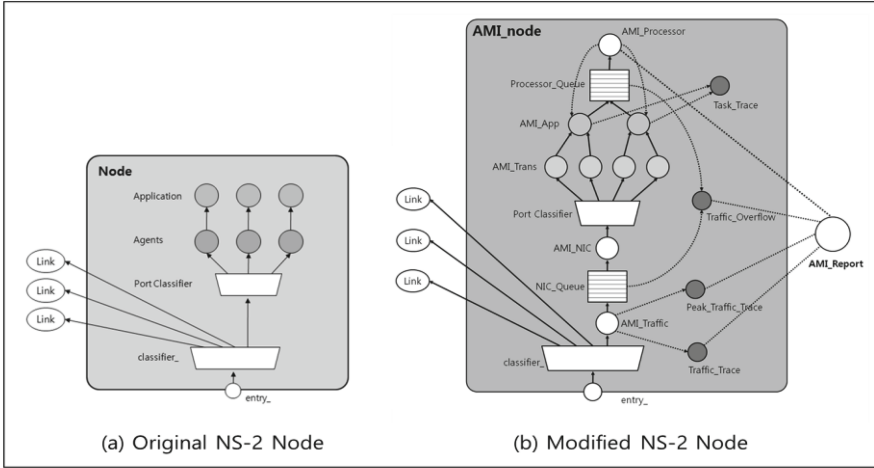


Fig. 3. Original NS-2 Node Configuration and Modified NS-2 Node Configuration

node. NIC_Queue is designed to act as real NIC buffer. AMI_NIC adds simulation time according to NIC’s performance.

- **Objects for network protocol:** AMI system uses typical protocol for transport layer or lower layer. TCP, UDP can be used for MDMS-ADCS and ADCS-DCU communication. Zigbee, PLC can be used for DCU-Smart meter communication. AMI-SIM offers all of them as an AMI_Trans object.
- **Objects for application layer protocol:** In AMI system, several task applications and protocols for several sections are used. AMI-SIM offers these as an AMI_app object.
- **Objects for system processor and memory:** In a real world, each specific job for a task in a node waits ‘during some time interval’ in a buffer for its turn, and when its turn comes, the job is processed by a processor ‘during some time interval’. In AMI-SIM nodes, AMI_Processor object calculates delay of a job in a processor, and Processor_Queue calculates wait time of a job in a memory. They periodically report the calculated data to AMI_Report object.

3.3 The Design of Analyzer

The analyzer calculates CUD(Capability Usage Degree) of each node given simulation scenario, using trace data. In this paper, the term CUD defined as following meaning. If a system’s CUD value is too high, the system performance or capacity is too low to process all AMI traffic. On the contrary, if a system’s CUD value is too low, the system performance or capacity is too high.

There are two kind of CUD for MDMS, ADCS, FEP. One is CUD_P (Processing CUD), and the other is CUD_M (Memory CUD). CUD_P gives us the information about permanent availability of the node. If we represent amount of traffic processed per second as R_{perf} and amount of incoming traffic as R_{traf} , CUD_P is calculated as following equation.

$$CUD_p = 100 \cdot (R_{traf}/R_{perf}) \quad (1)$$

If CUD_p is lower than 80, the node is decided to be stable traffic processing performance. If the value is higher than 100, it is assumed that data loss will occur when the memory is full.

CUD_M gives us the information about temporary availability in the case of traffic congestion. If we represent maximum usage of buffer as B_{max} and memory size as M , CUD_M is calculated as following equation.

$$CUD_p = 100 \cdot (B_{max} + R_{traf})/M \quad (2)$$

CUD for DCU, CUD_{DCU} tells us if a DCU collected metering data from all smart meter of its control domain during predefined time interval. If we represent finish time of task as T_{tf} , start time of task as T_{ts} , number of smart meter as N_{meter} and average collecting time for a meter as T_{ta} , CUD_{DCU} is calculated as following equation.

$$Temp_t = T_{tf} - T_{ts} \quad (3)$$

$$Temp_{dt} = T_{ta} \cdot N_{meter} \quad (4)$$

$$CDU_{DCU} = 100 \cdot Temp_{dt}/Temp_t \quad (5)$$

4 Performance Evaluation

In this section, we execute AMI-SIM under several conditions, and evaluate if AMI-SIM is valid for practical use. Tests are classified in 2 categories - selective handicapped node test and CUD calculation test.

4.1 Selective Handicapped Node Test

In this test, we choose a node as ‘handicap’ in AMI network topology, and configure that node to have very low performance. The primary goal of this test is to evaluate if the performance of a node properly have an effect on simulation result. AMI network

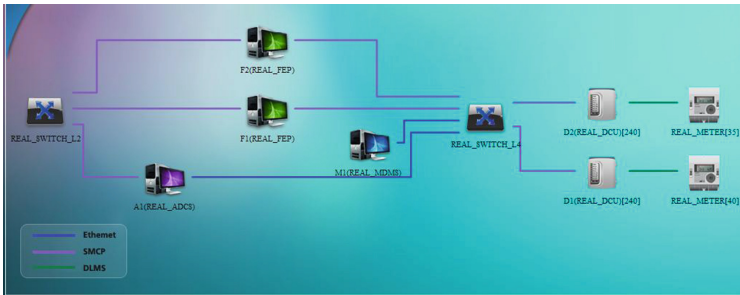


Fig. 4. AMI network topology for T-H test

topology for the simulation is as shown in Fig. 4. For each test, we choose a handicapped node and set up system performance as shown in Table 1.

Table 1. Common trace factors for each node

Test ID	Handicapped system	CPU performance (tpmC)	Memory size	NIC bandwidth
T-H1	MDMS	10000	128 MB	1024 Mbps
T-H2	ADCS	10000	128 MB	1024 Mbps
T-H3	FEP	65543	1 MB	1024 Mbps

Figure 5 is simulation result views of test T-H1, T-H2 and T-H3, respectively. Each test shows expected result and demonstrates that system performance properly affect to simulation result.

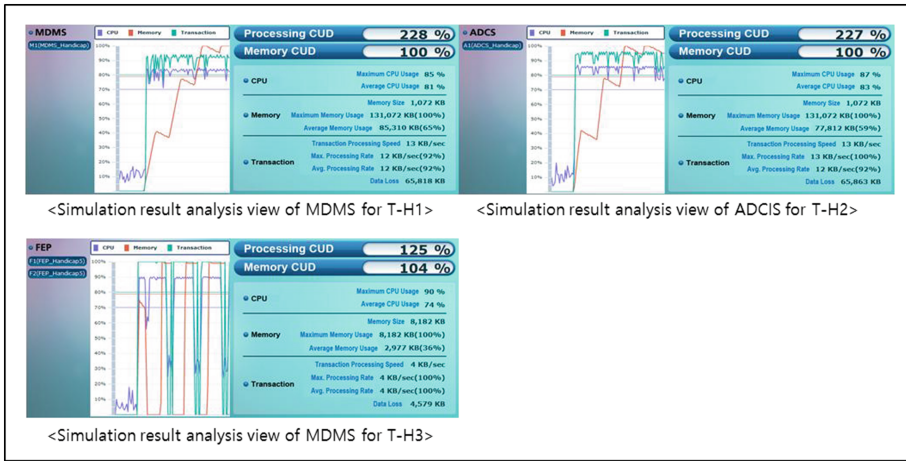


Fig. 5. Simulation result analysis view for TH-1, TH-2 and TH-3

4.2 CUD Calculation Test

The goal of this test is to evaluate accuracy of CUD calculation performed by AMI-SIM. In this test, firstly set up general AMI network topology, and then perform several

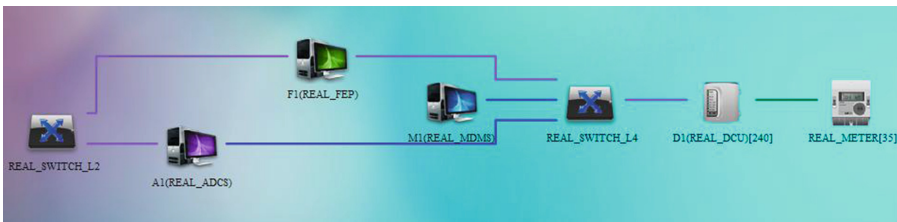


Fig. 6. AMI network topology for T-PS test

consecutive simulations steadily increasing the number of smart meters. AMI network topology for simulation is as shown in Fig. 6.

In each test, we set up the same topology and the same system performance as Table 2, and change number of smart meters and DCUs only as shown in Table 3. Figure 7 is simulation result views of test T-PS1, T-PS2, T-PS3, and T-PS4 respectively. These results show CUD calculation is quite reasonable.

Table 2. System performance setting for T-PS Test

Test ID	System	CPU performance (tpmC)	Memory size	NIC bandwidth
All	MDMS	360014	512 MB	1024 Mbps
	ADCS	108435	512 MB	1024 Mbps
	FEP	108435	64 MB	1024 Mbps
	DCU	108435	64 MB	24 Mbps
	Meter	—	—	24 Mbps

Table 3. Number of DCUs and smart meter T-PS Test

Test ID	Number of DCUs	Number of smart meters
T-PS1	240	8400
T-PS2	480	16800
T-PS3	720	25200
T-PS4	960	33600

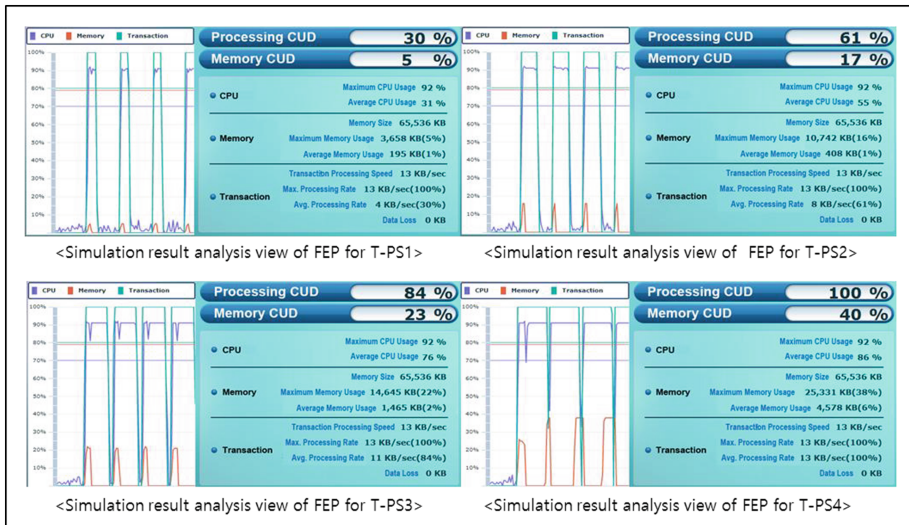


Fig. 7. Simulation result analysis view for T-PS1, T-PS2, T-PS3 and T-PS4

5 Conclusion

In this paper, we defined requirements of AMI network simulator useful for choosing proper system spec in given AMI network topology, and proposed AMI-SIM designed to follow these requirements. The core engine of AMI-SIM is NS-2 but original NS-2 was modified to match our goal of simulation. AMI-SIM needs several trace data, for example, peak traffic rate, but original NS-2 does not provide such trace. Additionally, original NS-2 focuses on protocol and links, so it does not provide mechanism to make delay generated by producing data reflect system performances.

Several tests were performed to evaluate if the performance of a node properly have an effect on simulation result, and to evaluate accuracy of CUD calculation performed by AMI-SIM. These tests show that AMI-SIM is available for practical use. However more tests for evaluating various aspects such as availability, accuracy and practicality are necessary in the future.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0184-15-1003, oneM2M The Development of oneM2M Conformance Testing Tool and QoS Technology)

References

1. Jun, W., Leung, V.: A survey of technical requirements and consumer application standards for IP-based smart grid AMI network. In: 2011 International Conference on Information Networking (ICOIN), pp. 114–119. IEEE (2011)
2. Wenye, W., Xu, Y., Khanna, M.: A survey on the communication architectures in smart grid. *Comput. Netw.* **55**(15), 3604–3629 (2011)
3. Subir, D., et al.: A key management framework for AMI networks in smart grid. *IEEE Commun. Mag.* **50**(8), 30–37 (2012)

Beyond Map-Reduce: LATNODE – A New Programming Paradigm for Big Data Systems

Chai Yit Sheng^(✉) and Phang Keat Keong

Faculty of Computer Science and Information Technology,
Universiti Malaya, Kuala Lumpur, Malaysia
yschai@siswa.edu.um.my, kkphang@um.edu.my

Abstract. The Compute Aggregate model used to model Map Reduce does not allow for dynamic node reordering once a job has started, assumes homogenous nodes and a balanced tree layout. We introduce heterogeneous nodes into the tree structure, thereby causing unbalanced trees. Finally, we present a new programming abstraction to allow for dynamic tree balancing.

Keywords: Big data · Map-Reduce · Trees · Lambda calculus · Functional programming · Programming paradigm

1 Introduction

Map Reduce [1] has been the dominant programming paradigm for building big data systems with Open Source projects like Hadoop [2] and Spark [3] built on its foundations. Map and Reduce can be abstracted as a Compute and Aggregate Model and subsequently, this allows optimizing performance by optimizing the underlying node structure to perform the aggregation [4]. These assume homogenous nodes and other assumptions (Sect. 2).

A big data system that requires scaling cannot have access to an unlimited number homogenous nodes. At some point, machines of different configurations and processing power will have to come into play.

The contributions of this paper are

- We extend the compute-aggregate model to take in account heterogeneous nodes and demonstrate its effects to the prevailing time complexity model (Sect. 3.2)
- We will demonstrate that heterogeneous nodes in the compute-aggregate model results in an unbalanced tree and articulate the conditions when this occurs (Sect. 3.3)
- We propose a new abstraction called Lambda-AVL-Tree Node (LATNODE) to represent a big data computation. This abstraction is used to overlay the compute-aggregate model (Sect. 4)
- We also demonstrate how a compute aggregate job can be expressed in LATNODE (Sect. 4).

2 Background and Related Works

Several works relating to trees with processor nodes discuss load balancing [5] and parallelization [6] but these are for more general workloads. In addition, a later work on merge trees present a new data structure to take advantage of multi-core machines [7] but also not specific to Map Reduce.

With the advent of cloud computing services like Amazon Web Services and Google Cloud Platform, data processing does not have to be tied to a particular machine. In addition to a physical machine (single-core and multi-core), a unit of compute can also be located in (1) a virtual machine, (2) a container, or (3) a serverless cloud function.

Representing a Big Data Job as a tree is a useful abstraction method to model a cluster's time complexity regardless of the implementation of (1) the big data cluster, (2) each individual node's geolocation or (3) unit of compute (Fig. 1).

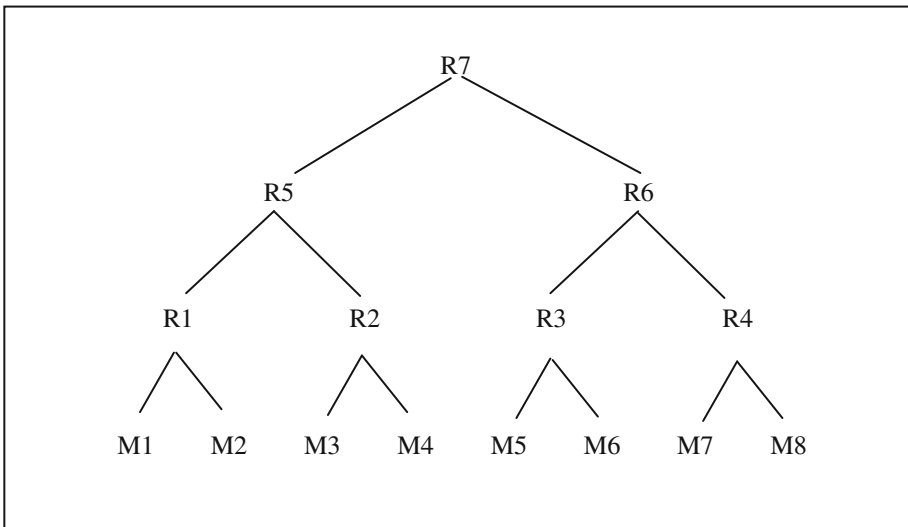


Fig. 1. The current Compute Aggregate Model showing a Map Reduce job with a fan-in of 2 represented as a tree. M nodes denote mappers and R nodes denote reducers.

A study showed that the optimum-fan in for a compute aggregate job is 2 [4]. This study also showed some techniques for performance optimization by optimizing the underlying node structure to perform the aggregation. It assumes a number of conditions, the most important of which are:

1. Homogenous nodes
2. Homogenous fan in at all nodes
3. Ignores complexity associated with communications

We will extend these works by introducing the idea of heterogeneous nodes running some of the aggregate functions.

3 Extending the Model

3.1 Existing Model (Base Case)

Consider a large dataset of tab separated values in this example showing the name and car make for every person in London and we assume each line is a unit of data (Fig. 2).

001 John Smith London Toyota
002 Jenny Ross London Jaguar
..
..

Fig. 2. Sample Data in a text file formatted in tab separated values.

We denote the path taken by data flowing through the system during a compute aggregate job, using the sequence notation {M1, R1, R5, R7} Referring to Fig. 1, data entering Node M1, data flows from M1>R1>R5>R7, data entering Node M2 flows from M2>R1>R5>R7 and so on.

We formalize these with some definition that will be used throughout this work (Table 1).

Table 1. Notation used

Token	Meaning	Unit
MX	An Element from a set consisting of [M1, M2, M3, M4, M5, M6, M7, M8, R1, R2, R3, R4, R5, R6, R7] representing each node	
t_x	Processing time taken at each node	Seconds
$t_{\text{aggregatePath}}$	Processing time taken for each path	Seconds
T	Total time for the entire job to complete	Seconds
Balance Factor	Height of right branch – Height of left branch	Integer

Table 2. Timing Values with homogenous and heterogeneous nodes

Path	$t_{\text{aggregatePath}}$ for 1 s per node (Homogenous Node)
{M1, R1, R5, R7}	4
{M2, R1, R5, R7}	4
{M3, R2, R5, R7}	4
{M4, R2, R5, R7}	4
{M5, R3, R6, R7}	4
{M6, R3, R6, R7}	4
{M7, R4, R6, R7}	4
{M8, R4, R6, R7}	4

All possible paths are shown in the table below:

Assuming 1 unit of data inserted into each M Node, T is the sum of all $t_{\text{aggregate-Path}} = 32$ s.

3.2 Adding Heterogeneous Nodes

In the existing base case, all nodes are homogenous, we introduce heterogeneity by swapping out any random node (we pick Node R6 for illustration). Adding several assumptions for R6: (1) R6 takes $3t_x$ to completely process 1 unit of data. Thus for each path that passes through Node R6, takes a longer time than for those paths that do not pass R6. Referring to Fig. 1, this changes the height of the right branch making the balance factor of +2. Therefore causing the tree to be unbalanced.

All possible paths are shown in the table below (Table 3):

Table 3. Timing Values with homogenous and heterogeneous nodes

Path	$t_{\text{aggregatePath}}$ for 1 s per node (Homogenous Node)	$t_{\text{aggregatePath}}$ if only node R6 takes 3 s (One heterogeneous Node)	Balance factor
{M1, R1, R5, R7}	4	4	0
{M2, R1, R5, R7}	4	4	0
{M3, R2, R5, R7}	4	4	0
{M4, R2, R5, R7}	4	4	0
{M5, R3, R6, R7}	4	6	+2
{M6, R3, R6, R7}	4	6	+2
{M7, R4, R6, R7}	4	6	+2
{M8, R4, R6, R7}	4	6	+2

3.3 Results in an Unbalanced Tree

In this work, we define balance in a tree using time units as the tree height instead of physical node count. We believe this to be valid due to the rise of cloud computing where the physical topology is separate from its logical layout.

We see in Table 2 that just a single heterogeneous node in R6 results in the right branch having a balance factor of +2. This tree then becomes unbalanced under the constraints of an AVL tree and results in suboptimal job performance.

Naturally multiple permutations are possible if we use one or more heterogeneous nodes.

To illustrate the effects of an unbalanced tree, we designed a simulator program written in Python. Our choice of programming language is motivated by the need for code simplicity over raw performance speed.

The simulator has several components

- A controller running a single event loop that controls data flow from one level of the tree to the next

- A starting ingestion queue for data that is read in from any data source such as Hadoop Distributed File System
- A final collector node to collect all processed results.
- A drop queue, where all data units that exceeds the ingestion rate of the system is place. Further compute & aggregation of this drop queue data is scheduled when ingestion capacity is freed up. However, in this initial simulation, we will not discuss in detail the impact of the drop queue.

3.4 Simulator Results

Referring to Fig. 1, we assume that the following data paths denoted by (Table 4)

Table 4. Tree Branches and corresponding data paths

Left branch	Right branch
{M1, R1, R5, R7}	{M5, R3, R6, R7}
{M2, R1, R5, R7}	{M6, R3, R6, R7}
{M3, R2, R5, R7}	{M7, R4, R6, R7}
{M4, R2, R5, R7}	{M8, R4, R6, R7}

And that the ingestion queue sends in a burst of 1000 data units each into the left branch and right branches.

Assuming the ingestion rate matches the capacity of the tree, both branches will result in a total of 2000 data units in the starting queue and 2000 data units in the final collector respectively.

However, when the tree is unbalanced by adding a slow node in the right branch, and the slowness factor is defined as x times of the left branch, we have the following chart (Fig. 3).

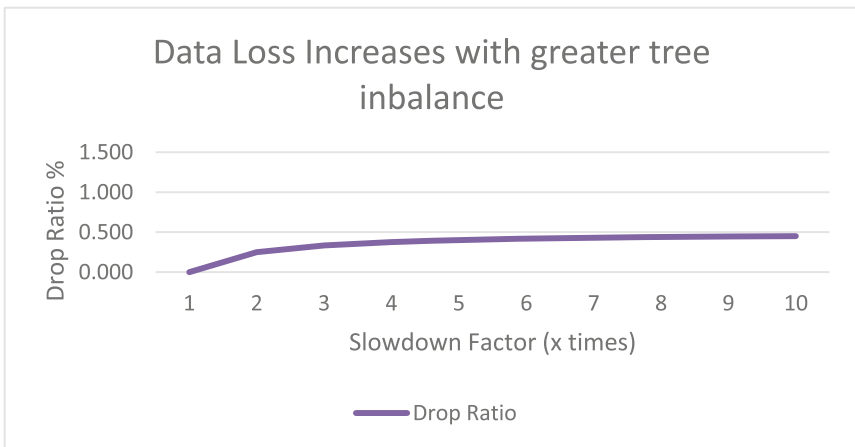


Fig. 3. Drop Ratio shows data that cannot be processed and sent to the drop queue.

The chart showed that if slowness factor is 10 times, the amount of data units that is sent to the drop queue is almost 45% of the data. This data has to be re-aggregated causing a significant performance bottleneck.

4 Balancing the Tree

To balance a tree as we have defined, it is necessary to relook at the basic unit of computation in a big data job. A Map Reduce job consists of mappers and reducers, both are however merely functions with the following steps (1) taking an input, (2) performing a computation, (3) presenting the output and (4) sending the output to the next destination.

But this programming paradigm is one drawback, the output destination cannot be changed once the job has started. In this case then, balancing the resulting tree structure dynamically during runtime is not possible.

To achieve our aim of dynamic reallocation, we put forth a new abstraction called Lambda-AVL-Tree Node (LATNODE) where each node is viewed as a single independent function. We do not care about what the function is but merely model them as having three other pieces of metadata, (1) running time for the function, (2) the substituted function and (3) the next destination.

We use Lambda Calculus to define each node as $\lambda x.x + y$ and each path as $(\lambda x.x + y)(\lambda x.x + y)(\lambda x.x + y)$, In Lambda calculus, the x is substituted with another expression or function, name or variable. We also represent the next destination with y (Fig. 4).

```

var heightOfLeftBranch = x
var heightOfRightBranch = y

run event loop
  if heightOfRightBranch - heightOfLeftBranch <=1
    transfer data to RightBranch AND Left Branch
  if heightOfRightBranch - heightOfLeftBranch <=2
    transfer data from RIGHTBranchNode to LEFTBranchNode
iterate run event loop

```

Fig. 4. Simple pseudocode to balance the tree (for $y < x$)

5 Conclusion

This idea expressed in Lambda Calculus allows us to chain a series of functions into a topology for a tree of nodes and allows dynamic reordering of the topology when each individual data unit is passed through. The nodes also do not need to have the same compute unit, allowing an extremely flexible big data topology.

Acknowledgements. This work is supported by the Ministry of Science, Technology and Innovation Malaysia [Grant No.: FP067-2015A].

References

1. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. *Communications of the ACM* **51**, 107–113 (2008)
2. Konstantin, S., Hairong, K., Sanjay, R., Robert, C.: The Hadoop distributed file system. In: *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)* (2010)
3. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., et al.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, San Jose, CA (2012)
4. Culhane, W., Kogan, K., Jayalath, C., Eugster, P.: Optimal communication structures for big data aggregation. In: *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 1643–1651 (2015)
5. Cheng, Y.C., Robertazzi, T.G.: Distributed computation for a tree network with communication delays. *IEEE Transactions on Aerospace and Electronic Systems* **26**, 511–516 (1990)
6. Hyung Joong, K., Gyu-in, J., Jang Gyu, L.: Optimal load distribution for tree network processors. *IEEE Transactions on Aerospace and Electronic Systems* **32**, 607–612 (1996)
7. Morozov, D., Weber, G.: Distributed merge trees. *SIGPLAN Not.* **48**, 93–102 (2013)

Indoor Positioning Solely Based on User's Sight

Matthias Becker^(✉)

Modeling and Simulation Group, University Hannover,
Welfengarten 1, 30167 Hannover, Germany
xmb@sim.uni-hannover.de
<http://www.sim.uni-hannover.de>

Abstract. Determination of the absolute geographical position has become every day routine, using the Global Positioning System (GPS), despite the prior existence of maps. However no equally universal solution has been developed for determining one's location inside a building, which is an equally relevant problem statement, for which GPS cannot be used. Existing solutions usually involve additional infrastructure on the end of the location provider, such as beacon installations or particular configurations of wireless access points. These solutions are generally facilitated by additional native mobile applications on the client device, which connect to this infrastructure. We are aware of such solutions, but believe these to be lacking in simplicity. Our approach for indoor positioning alleviates the necessity for additional hardware by the provider, and software installation by the user. We propose to determine the user's position inside a building using only a photo of the corridor visible to the user, uploading it to a local positioning server, accessible using a browser, which performs a classification of the photo based on a Neural Network approach. Our results prove the feasibility of our approach. One floor of the university's building with partially very similar corridors has been learned by a deep convolutional neural network. A person lost in the building simply accesses the positioning server's website and uploads a photo of his current line of sight. The server responds by generating and displaying a map of the building with the user's current position and current direction.

Keywords: Machine-learning · Scene analysis · Neural network · Indoor positioning

1 Introduction

Global Positioning System (GPS) technology has reformed human life irreversibly. In the past, a hypothetical argument could be made, that the potential utility provided by GPS would be redundant and mundane, due to the existence of countless landmarks and directional indicators, both natural and man-made, especially in an outdoor environment. Moreover, the spatial and map-reading abilities of human cognition would be a factor in such a criticism of GPS. However, such a theory would seem quaint and receive no support today, simply due to the

popularity of GPS. This is highlighted by the blatant loss of directional sense, a utility possessed by our ancestors, and also the comprehensive navigation features inherent to every smartphone entering the marketplace today. Outdoors, GPS can conveniently be used to locate the user's position and show this position on a map, along with the direction the user is heading, when using a state of the art smartphone. One GPS application is sufficient for use in the whole world, given that the right maps are available either online or offline by downloading the maps for the location beforehand. Consider then, the problem of indoor navigation, where landmarks are few, corridors appear similar, wireless signal strengths are variable, no natural heuristics of direction are prevalent and locational proximity is no guarantee of reachability. As a result, one often finds oneself in the position that one gets lost in a large unknown building, e.g. on a conference that takes place in an unknown foreign university. Even having been handed a map of the location, the map is not useful once the position on the map is lost and unknown. Since the nomenclature of rooms is already logical and intuitive, and the fact that easily accessible maps remain confusing to visitors, there exists a demand for indoor localization support for very large buildings. Moreover, it may be helpful to know the approximate duration of navigating inside the building, as is attempted with information labels and signs by certain airports. In contrast to the situation outdoors, there is no universal and easily usable indoor solution available. Of course there exist numerous special solutions, relying on additional hardware on the provider's side and additional software and/or hardware on the user's side, which will be presented in the next section.

Our approach presents a simpler solution to the indoor positioning problem, which needs no additional measures from the user side and minimal effort by the location provider. It is noteworthy that no extra building-wide hardware is required, in contrast to many other solutions (such as the use of ultra wide band transmitters, for example). Our idea uses a human centered approach, based on asking how a person to whom the building was familiar, an expert, would navigate through the building. Such a person, positioned in an arbitrary place in the building, would attempt to look to the left and right down a corridor and then recognize some landmarks, to infer the position inside the building. In our approach, the visual knowledge of a person's sight is learned by a convolutional neural network and stored on a web server. A user with an internet connection can then use his smartphone to take a photo of his current sight, upload it to the webserver with no additional software, just knowing the server's address which could be textual or accessible via QR Code at the entrance of the buildings. As a result of the upload, the response of the server, after analyzing the image by the neural network, consists of a map with the user's location and the direction the user is heading in. In case the image cannot be identified exactly, the server requests a new image, which corresponds to a real expert looking in another direction, when first facing an unknown corridor or just a corridor that looks very similar to another place and cannot be identified exactly. Analogously, the user of our system can 'look' into another direction and take a photo there, or walk down to end of the corridor and look around for some more unique looking places, corridors or landmarks.

In this paper we will present the state of the art of indoor positioning systems and then present the techniques we use for building our working prototype, which show the feasibility and usability of our approach.

2 State of the Art

In the following section we review existing approaches for indoor positioning:

- Active Badge Active Badge is one of the first localization systems, developed by AT&T Cambridge¹. This system is more a person tracking system rather than an information system for the user. The system is based on locating persons wearing an infrared badge which send signals in intervals, that are detected by infrared sensors on the walls of the building [1,2].
- Active Bats Similarly Active Bats is a system locating ultra sonic impulses of persons wearing a ultra sonic badge. Ultrasonic localization can be very exact, when a signal can be detected from several sensors, in the range of centimeters [1,3].
- Smart Floor Smart Floor has been developed from Georgia Tech² and recognizes the stepping pattern of different persons to identify the persons and their position. However this approach needs a lot of effort and costs on the provider's side [1,4].
- Easy Living Easy Living is a positioning system developed by Microsoft Research.³ The system is based on a scene analysis of the rooms by stereo cameras. Based on empty rooms, the system can later detect the persons in the room and their exact location. The system is more suitable for surveillance purposes, and is expensive due to additional hardware installation everywhere [1,5].
- RADAR RADAR is another system developed by Microsoft Research that uses WLAN signals. Two versions exist that use lateration of WLAN signals and scene analysis
- Cricket Cricket relies on ultrasonic similar to Active Bats. In contrast, the senders are mounted on the wall and the receivers are mobile. Cricket has been developed by MIT⁴ and is a device that can be used as a sensor or transmitter [1,3].
- Image Localization in Buildings The University of Berkeley has developed⁵ an image based localization system in buildings first completely scanning the building using a *data acquisition backpack* consisting of a notebook, several cameras and laser sensors. The data acquisition backpack scans the building and constructs a complete 3D model by ray-tracing. Using a kd-tree for each

¹ AT&T Cambridge (<http://www.cl.cam.ac.uk/research/dtg/attarchive/>).

² (<http://www.gatech.edu/>).

³ (<http://research.microsoft.com/en-us/>).

⁴ Massachusetts Institute of Technology (<http://web.mit.edu/>).

⁵ (<http://www.eecs.berkeley.edu/>).

picture features are stored. The features are detected by the SIFT⁶ method. SIFT is also used when an image is to be evaluated for localization. In the kd-tree the node is searched which shows the most matching SIFT features. In the second step the sensors of the smartphone are used to find the right orientation of the input image, so that the output corresponds to the orientation of the user.

3 Our Approach to Indoor Positioning

One essential part of our approach is the learning of the user’s view and relating the view to a position inside the building. Several open source frameworks are available for this purpose and we decided to use the Computational Network Toolkit (CNTK) from Microsoft Research [6]. We use the framework together with a high performance CUDA GPU, in our case an Nvidia 960 GTX [7]. In the following we describe the setup of the convolutional neural network [8] used in our approach. We describe the parameters concerning the structure of the convolutional neural network, the properties of the trainings and test set, the structure of the input and output of the neural network. The convolutional neural network has three convolutional layers and three pooling layers. After each convolutional layer a pooling layer follows. In our feasibility study we use four classes to be classified, thus the output is calculated by a fully meshed layer with dimensions $1 \times 1 \times 4$. The learning algorithm is stochastic gradient descent, the pooling layers use the Max-pooling algorithms. The activation function is based on the Rectifier Linear Unit. The neural network has been trained with picture data in form of JPEG’s⁷ The section of the building, for which the network was trained, consists of four similar corridors which form a square. The training set photos were views into the corridor, with variation in focus and direction of each view. The views of the users are taken from the corners of the square. As can be noticed, the corridors are very similar, which can be seen in the figures of two different classes, for example Figs. 2 and 3. The training sets include nine pictures for each of the four positions and views.

The network expects a JPG file from the users smartphone. The size of the picture varies due to different smartphones of the users, that might have primitive



Fig. 1. Simple App for easy use of the framework. In the upper part the users view is seen as the photo just taken, below the map is shown with users position.

⁶ Scale-invariant-feature-Transform. US 6711293B1, Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image.

⁷ Norm ISO/IEC 10918-1, describing methods for picture compression. The Joint Photographic Experts Group developed the algorithm, thus the name JPEG.



Fig. 2. Training set class one, between corridor B and A



Fig. 3. Training set class two, between corridor A and C

cameras or very good cameras in high end smartphone. However, the absolute size of the picture does not matter, since each picture serving as input will be scaled down to a small size that is needed for efficient deployment of the neural network algorithm. During the training phase, each picture is labeled with the class of the position. The output of the neural network is the number of the class, the input picture has been assigned to. In our feasibility study, each position is assigned to one of four classes. After the upload of the requested image on the server, the input is classified and the result, the class, is written into a file. The file is used afterwards in order to generate a map with the position of the users, plus the direction, in which the user is heading.

4 Application

One main feature of our approach is the simplicity, by which we mean the minimal requirements on the user’s side, which is simply the URL of the positioning server in our case. For convenience, we also developed an application for smartphones (App), which also is kept simple. The first view consists of three buttons, “choose picture file”, “upload”, and “delete”, see Fig. 1. “choose picture file” opens a menu where the user can choose to upload an existing image or upload a freshly taken photo. (Fig. 4). As soon as an image is chosen or the photo is taken, the button “upload” initiates the process of the image recognition. The image is uploaded and classified, and as a result the webserver delivers a map of the building in which the current position and orientation of the user is included, corresponding to the results of the classification, see Fig. 1. “delete” deletes the currently uploaded image, in case the classification delivers no significant results. This case corresponds to the user looking in one direction, not being able to identify the location, and thus looking into another direction. The App is kept simple, the computation takes place on the server. This client just uploads an image and receives a webpage with an image comprising of the map and the position of the user. For the realisation of the server the Django framework has been used. The web page is displayed as HTML, the functions have been realized in Python.

Django includes a Webserver, that has been implemented in Python, however that webserver should only be used during the development phase. The server is started simply on the commandline.

Note that Django is controlled by the file given in `manage.py`. The output of the console is achieved by `python manage.py runserver 0.0.0.0:80` The page can be addressed after the start of the server as IP address of the webserver

The input of a request of a user is either the upload of a picture out of the DCIM directory or a picture that is just taken by the integrated camera. Using the button “upload” the picture is saved as `Django-Modell` Additionally the Session-ID is saved for multiple users support.

5 Experiments

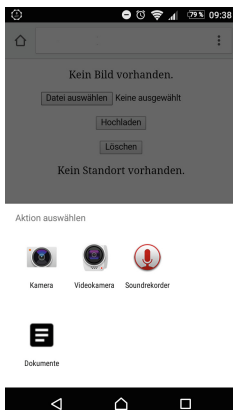


Fig. 4. Upload file or take a photo

In order to efficiently gather the training material we developed a workflow that makes it possible to walk around all corridors of a building while recording a video with a camera. Afterwards a script is used to select single frames and group the extracted pictures. Here, some manual work is need in order to have a grouping that reflects the different positions and the respective image training set. The images should be collected at different daytimes, so that day and night lighting is trained, as well as empty corridors and corridors with persons. However, it turns out that persons on pictures of the training set do not improve the recognition results. It is better to learn the features of a position without the distraction of persons. When people block some features in the recognition phase, the net can still identify the remaining visible features for the classification. People in the training set would bear the danger that the net may try to learn features of the persons, which would slow down the learning phase.

6 Results

We test our approach in a four level university building. we obtained 90 different classes, corresponding to different positions in the building. We recorded 1500 images per class and used 1440 for the training and 60 for evaluation. Epoch size was 512 minibatches and the size of each minibatch was 128. The training of one epoch took nearly 7 min. We conducted several experiments in order to gain insights on the influence of the crucial parameters of the algorithm. First, we studied the necessary size of the training set on the recognition success. Figure 5 shows the influence of the cardinality of the training set on the classification success. The blue graph shows that the training set can be learned very well, starting from 45 images per class (recognition rate 82.5%) to nearly 100% success from 180 images per class on. The recognition rate of untrained images shows

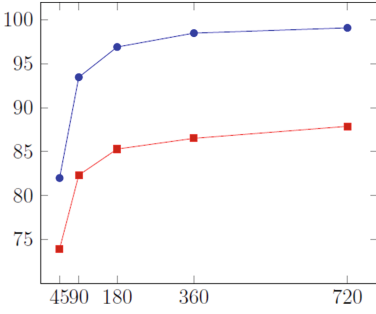


Fig. 5. Classification success versus number of images per class, blue: training set, red: evaluation set

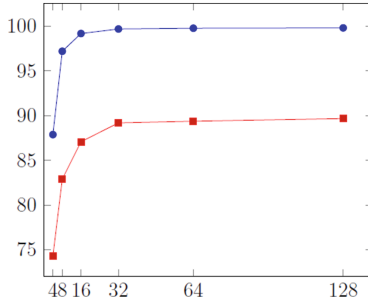


Fig. 6. Classification success versus number of training epochs, blue: training set, red: evaluation set

the same characteristics however on a lower level: starting from a recognition rate around 75% up to nearly 85%.

In a second experiment we evaluated how many training epochs are needed for an acceptable recognition success rate. As shown in Fig. 6 32 to 64 training epochs are enough to have a nearly perfect recognition of the trained images (blue graph) and to reach the maximal possible recognition success on untrained images, reaching nearly 90% (red graph).

When the system is used by real users, probably many different cameras will be used, with different resolutions and color characteristics of the images. In order to evaluate this influence on the efficiency of the approach, we conducted several experiments, with different cameras, and also with different users of different height and with different habits (e.g. holding the phone, pointing the camera to the view, etc.) As can be observed in Fig. 7, the camera used during the training performs best (green column). The yellow columns show the results for different cameras in the test phase, which were done in the setting as the training phase, concerning lighting/time of day etc. A recognition success rate between 70 and 90% is absolutely sufficient for our purposes, since the users can easily take a picture from a more characteristic part of the building, in case the current image has not been recognized correctly.

Wiko	Wiko	100
	Nexus	82,1
	LG	81,8
	OnePlus	50,4
Nexus	Nexus	98,9
	Wiko	83,7
	LG	76,6
	OnePlus	44,8
LG	LG	98,9
	Wiko	89,4
	Nexus	76,3
	OnePlus	58,3

Fig. 7. Classification success using different cameras. First row shows the cameras used for training, second and third the recognition success using a certain camera in the test

The red column shows the results for a camera that has not been used for training and additionally, the tests were conducted in a different setting than

the training. As one would expect, the success rate drops between 40 and 60%. This rate is still somewhat decent, however the user experience is not up to par any longer.

7 Conclusion

We presented an approach for an indoor positioning system relying on convolutional neural networks. In contrast to other approaches, we rely on a user-centric, simple and cost effective approach, with no need for additional hard- or software on provider's and user's side and only little effort on provider's side. The minimal requirement is a smartphone that can take a photo of the user's view and send it to a webserver, that will return a web page showing an image with the plan of the building including the user's position and direction of view. Note that our approach does *not* necessitate the installation of an 'App' (software) on the user's phone. Every other approach requires a special software on the device. Our feasibility study was very promising, an extensive study of the successful application has been done covering an entire building with four comprehensive levels. In our future work, we will evaluate how to improve the recognition success rates of different cameras. Furthermore, a navigation app currently is developed that will guide the users to a destination.

References

1. Lemelson, H.: Eine Übersicht über In- und Outdoor Positionierungssysteme, universität Mannheim (2005)
2. Want, R., Hopper, A., Gibbons, J.: The active badge location system. *ACM Trans. Inf. Syst.* **10**, 91–102 (1992). Olivetti Research Ltd., Cambridge
3. Koyuncu, H., Yang, S.H.: A survey of indoor positioning and object locating systems. *IJCSNS Int. J. Comput. Sci. Netw. Secur.* **10**, 121–128 (2010). Computer Science Department, Loughborough University
4. Orr, R.J., Abowd, G.D.: The smart floor: a mechanism for natural user identification and tracking, pp. 1–6. *ACM Press* (2000)
5. Shafer, S., Krumm, J., Brumitt, B., Meyers, B., Czerwinski, M., Robbins, D.: The new easyliving project at microsoft research. In: *Proceedings of Joint DARPA/NIST Smart Spaces Workshop*, pp. 30–31 (1998)
6. Agarwal, A., Akchurin, E., Basoglu, C., Chen, G., Cyphers, S., Droppo, J., Eversole, A., Guenter, B., Hillebrand, M., Hoens, T.R., Huang, X., Huang, Z., Ivanov, V., Kamenev, A., Kranen, P., Kuchaiev, O., Manousek, W., May, A., Nano, O., Navarro, G., Orlov, A., Parthasarathi, H., Peng, B., Radmilac, M., Reznichenko, A., Seltzer, M.L., Slaney, M., Stolcke, A., Wang, H., Wang, Y., Yao, K., Yu, D., Zweig, G.: *An introduction to computational networks and the computational network toolkit* (2014)
7. Garland, M., Skadron, K., Nichols, J., Buck, I.: Scalable parallel programming with CUDA. *Queue* **6**(2), 40–53 (2008). *ACM*
8. Nielsen, M.A.: *Neural Networks and Deep Learning*, kapitel 6: Convolutional Neural Network. *Determination Press* (2015)

Naming Convention Scheme for Role Based Access Control in Cloud Based ERP Platforms

Abed Alshreef^{1(✉)}, Lin Li¹, and Wahid Rajeh²

¹ Department of Computer Science and Technology, Wuhan University of Technology,
Wuhan, China

{alshreefaced, cathylilin}@whut.edu.cn

² Huazhong University of Science and Technology, Wuhan, China
wahid.ra@ut.edu.sa

Abstract. Cloud computing users can use at the same time the same cloud service. So, there is a need for having an access control mechanism to ensure that each user cannot access any sensitive data of other users. Several access control models have been proposed for cloud computing. However, these models need to be efficient and scalable due to increased workload (e.g., users, policies, etc.) in the cloud. This paper presents a role based access control model (RBAC) for cloud computing based on naming convention (NC) concept. The WSLA specification language is used for SLAs specification. A naming convention role based access control (NC-RBAC) is presented by modifying the standard RBAC to support the NC. Then, the proposed framework is designed based on the NC-RBAC to offer a simplified designed for the system administration of security in a large institution where there are many users is challenging to control access to resources. The proposed framework is implemented and its efficiency and scalability are measured using an experiment study. The result shows that the proposed framework provides an efficient and scalable access control for cloud computing while provides an administrator with an efficient and simple search method for classifying the cloud users.

1 Introduction

Most of the technology industries are rapidly deploying their applications to the cloud. This deployment provides a low-cost and efficient IT operations for the organization and even for the individual users. Cloud computing offer more usable access to the data that should securely stored on the cloud provider servers. There are some concern about being in the cloud which mean - in some sense - being more visible. Therefore, users must guarantee that cloud vendors are taking the appropriate security requirements and goals to protect their information. To achieve such a trust relationship, cloud providers should insure that costumers' critical information are accessible only to the authorized users.

It is important to understand the concept of cloud computing before we look at role base access control using naming convention. We can state that cloud computing where clients get networked storage area and other computing assets. These services are

rendered by the organization with the capability to store the large facts and figures in their computing apparatus [1, 2].

The cloud deployments are not only including private or public model, there are also can include community and hybrid models. The cloud services offer a usable and remote access over the internet when users or an organization outsourced their storage space by cloud providers [3].

The increased risks of cloud facilities compared to data centers call for more security since private data is at risk. Since the clients of the cloud services may be competitors, the risk of having one client accesses information of another client may result into huge losses. It is the role of the cloud providers to ensure that this possibility of system breakdown and access to other user's data platform is minimized.

Therefore, in all three access control models which include Role Based Access Control (RBAC), Mandatory Access Control and Discretionary Access Control, users and/or resources should be recognized by the end users as unique identifiers. The dynamic deployment of those identifiers give the security administrators the ability to manage privileges and authorizations easily [3].

The use of user control where users are assigned to resources and objects in the cloud is not suitable. This is because cloud platform has millions of users who would like to get access to resources. It will be almost impossible to manage each of the users. Hence, the role base access control is the suitable method for cloud computing. This is because a group of members who would like to access resources are specified based on their roles in the system. Clustering the roles based on duties within an organization is the best approach to manage the millions of users and resources [4].

The identifiers can be indicated using naming convention rule which is a pre-organized sequence of characters. This kind of sequences are frequently used in programming to categorize variables, arguments fields in order to understand source code easily. So the programmer assign an plausible naming convention rule for each data.

This paper will look at the security of cloud data by proposing a role based access model based on the naming convention (NC) concept. The NC concept is used for simplifying the access control processing and administration. Instead of processing the entire user SLA, only NC is implemented to classify the users or even to evaluate their request. Commons Lang tools are required for the implementation to process the string, which represents the NC values. Moreover, the Document Object Model (DOM) is considered for processing the XML-based documents such as WSLA, roles and polices. The proposed work is implemented and an experiment study is used to evaluate its efficiency and scalability.

This paper is structured as follow: The second section discusses the related works. Section 3 by introduces the proposed framework. Section 4 discusses and evaluates the proposed framework. The last section concludes the presented paper and recommends some future works.

2 Related Work

The cloud computing access control technical and organizational challenges were investigated by Darren Platt [5]. In [6], a policy language for semantic access control is proposed. Urquhart [7] proposed the first attempt to control access on cloud computing. This study presented what was called “The Cloud Computing Bill of Rights” which includes some cloud computing authorization requirements. In [8], a trust management model is presented where Kerberos authentication mechanism is used to authenticate users in cloud environments. The trust management is applied for establishing a trusted administration. As a weakness, the authentication process becomes a bottleneck when dealing with a large number of users since the authentication scheme is based on the Kerberos protocol where passwords migration is not automated process.

In [9], an access control framework is proposed in order to secure the distribution of the multimedia contents in the cloud platforms. In this framework, only multimedia content is considered. The new provided cloud services (SaaS, PaaS and IaaS) are not investigated. This framework is centralized which means it cannot scale well in cloud computing.

In [10], a fine-grained cloud computing access control that seems to be secure and scalable is introduced. The content is encrypted to ensure the data security but the encryption used here is not applicable for cloud computing services because it can only guarantee the data confidentiality. The scalability here is also based on the key distribution scheme used which is not sufficient for cloud computing scalability need.

In [11], an attribute-based and fine-grained access control model for cloud computing is proposed. The same approach is used in [12]. A cloud based RBAC model has also been used in [13].

The latest method in the literature used is a reference ontology which is an improvement on the current models where users are replaced by specific policies with the role. Role is defined as a named job function in the organization which describes the authorizations and responsibilities given on the member of that role. The term permission simply means granting access, authorization or privilege to the user [14]. Constrains is conditions which return a value either positive or negative. This means the value is acceptable or not. Session is also a term on the model and it is used to describe the establishment of the sessions that by the users. Tenants can be assigned authority without changing the access policies. There are two modules for the ontology role based access control model. The first one is the real service module which offers tenants with different kinds of services like the e-commerce. The second module is the security check which ensures that security is provided before the services are given to the tenants [15].

However, while our work is introducing NC concept for RBAC, it facilitates the authentication management process among many tenants because it offers more meaningful ontology reference for each user.

3 The Proposed Framework

The proposed framework has two main modules as illustrated in (Fig. 1). These modules are User Subscription Module (USM) and Cloud Controller Module (CCM). At the beginning, the user submits his subscription request to the USM which then responds with an appropriate Web Service Level Agreement (WSLA). The WSLA is used by this research as it is the more widely used SLA specification language. After that and whenever required, the user can access his services through the CCM which controls the user usage based on RBAC model. However, this research proposes a modified RBAC model by integrating the naming convention (NC) with it and as will be presented later.

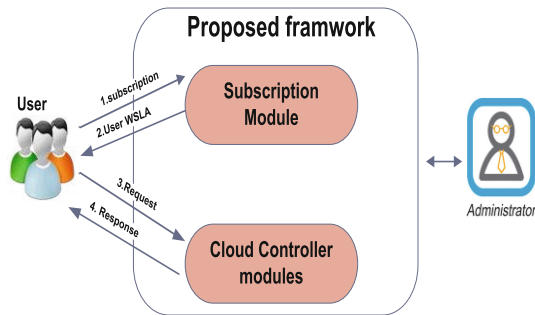


Fig. 1. The general architecture of the proposed framework.

The following sub-sections present the above modules in more detail along with their implementation and also starting with our presented and modified RBAC.

3.1 Naming Convention Using RBAC

This paper integrates the naming convention concept (NC) with the RBAC to propose a naming convention based access control framework for cloud computing. The standard RBAC [16], is chosen which has four main elements and four relation-sets (which are Permission-assignment, User-assignment, Role-sessions and User-session). For naming convention, each user NC combines some information such as user ID, country, city, and roles. Although NC-RBAC has the same elements as the standard RBAC, the new relation-set NC has been adjusted for the representing the user's ontology reference. For specifying the role and permission, the XML-based cloud policy language proposed in [17] is used since this language also applied the RBAC to cloud computing. Thus, the NC concept should be applied according to the cloud provider policies when assigning roles to the users.

3.2 The User Subscription Module (USM)

The proposed framework is applied for cloud user subscription where the user related roles are mapped to a WSLA template list to be retrieved with user WSLA. The user

WSLA is then saved into a WSLA DB. The next step is to generate an NC for user from his WSLA. The user NC is saved into the NC list.

The WSLA mapping just maps the user selected roles to a specific WSLA template. Selecting the user roles as well as choosing the suitable WSLA template is outside this research.

3.3 The Cloud Controller Module (CCM)

The second module in our proposed framework has four main components namely NC evaluator, role evaluator, policy control and NC reporting. The NC evaluator receives the user request (for example ID and password) and retrieves the required NC from the NC list to find his ID, country, city and roles. The role evaluator searches for the access control permissions related to the user's roles listed in his NC. Finally, the policy control finds the details of each user permission such as objects and operations.

Now, the administrator can use the NC reporting component to search for users. One can search for the users inside the NC list, instead of searching inside the WSLA DB which seems to need more time due to its size. The NC list component provides a list of candidate naming convention from the existing ones to the user. The NC list includes user name or ID, country, city, and roles. The proposed method utilizes "NC list" module to compare the similarity of user's NC with existing SLA database. The whole SLA data that match user's NC are collected in the NC list. Thus, user can refer to the list of their related SLA database. On the other hand, if the user is totally new, then the user must signup first to the cloud service.

Once NC sheet is generated then the roles data in NC Sheet is compared with Role list of the cloud. It is performed to discover whether those roles are available on the cloud provider or not. The proposed method needs only processing the NC list which is actually is a sheet that supported by the cloud service and discard NC with unsupported roles.

3.4 The Proposed Scenario

To implement a new security model for RBAC in the cloud, the cloud providers can offer the ontology reference module that introduced on [13] instead of building the model from the scratch. The cloud provider can employ NC-RBAC to deliver more efficient and scalable roles assignments when the tenants searching for the suitable ontology database. For example, when international information Technology Company decide to employ the RBAC among all users in one consolidate database, the role naming should follow the appropriate NC policy in order to manage the authorization and permission attributes in orderly manner. However, we assumed that IT Company is required to develop NC-RBAC for six child single roles that have been created from four parent roles.

Each child role derived from a parent role to facilitate the system scalability. The first parent role created for security administration. Then, the two child security administration roles can be derived from the parent role using NC-RBAC, one for Riyadh and the other for Beijing. In these security admin child roles, the authorization attribute and

the user assigning can be maintain to the right location easily based on the appropriate NC pattern. So each user can be build for its location by assigning them to its specific tenant group.

While the NC helps to provide a consistent naming prototype, those prototypes could be used to control the user's access and authorization to maintain roles. In our framework, we assume two naming convictions; one can be apply for the parents' roles while the other one for the child roles using the following patterns:

1-Parent roles naming convention: ZP-XX* where XX = Tenant Group.

2-Child roles naming convention:

Z-YYY-XX*

YYY = Department location -using IATA standard cities code.

(e.g.: Riyadh = RUH, Beijing = PEK, Wuhan = WUH)

However we can assume many factors that provide the required ontology reference for the role attribute such as position code, department code and even role activation date. So after setting the NC policy, the role naming can be adjusted based on its attribute as can be seen in Table 1. For example, while the NC for HR and Purchasing Officers in Riyadh is Z-RUH-41*, the role name should be Z-RUH-41PUROFF in order to apply the NC-RBAC structure. However, each role are inherited from its parent role such as ZP-41PUROFF which should be associated to role example above.

Table 1. NC-RBAC patterns description.

Role description	NC	Role name
Security admin in Riyadh	Z-RUH-41*	Z-RUH-41SECADMN
HR and purchasing officers in Riyadh	Z-RUH-41*	Z-RUH-41PUROFF
Junior purchasing officer in Riyadh	Z-RUH-41*	Z-RUH-41JPUROFF
Security admin in Beijing	Z-PEK-41*	Z-PEK-41SECADMN
HR and purchasing officers in Beijing	Z-PEK-41*	Z-PEK-41PUROFF
Senior purchasing officer in Beijing	Z-PEK-41*	Z-PEK-41SPUROFF

3.5 Implementation

The model is implemented to insure that our designing model is applicable, so the designed model is refined based on its implementation, and measure its performance, which can be used as one of the indicators in measuring the system efficiency and scalability. The Java programming language (JDK 1.7.0) is used for implementing the proposed work. In addition, the following tools (package) are also used: Document Object Model (DOM) for processing the XML-based documents such as WSLA, roles and polices; and Commons Lang tools for processing the string which represents the naming convention values.

4 Framework Evaluation and Results

This section presents and discusses the proposed framework by sitting up a case that measure the respond time. So the mapping, naming, NC evaluating, role evaluating, policy controlling, NC searching and SLA searching efficiency is measured as the response time with different number of users. Each user request is processed and evaluated before jumping to the processing of the next user request. To evaluate the scalability of the mapping, naming, NC evaluating, role evaluating and policy controlling, the response time is measured with different number of user. The users here are concurrent users. For example, when the number of users is 100 then all the requests of these users are processed at the same time. We used Java threading for establishing the multi-tasking manner in which each user's request is processed with a single Java thread. The scalability of the NC and SLA searching is not measured at all as the cloud administrator is not going to run a number of search hits at the same time. To deal with the Java garbage problem as well as making the finding softer, each result is ruined three times and, at the end, the average is taken as seen below.

4.1 Measuring Efficiency

Figure 2a, shows the efficiency result of the mapping, naming, Role evaluator and policy control components. The cost of mapping the request of 50 users is about 449.333 ms. This means that – at the beginning - the cost of processing the request of each user is about 9 ms. At the end and when the number of users is 1000. The total needed time is 4862 ms. Means at the end it cost only 4.869 ms for each user. The processing time for each user is reduced from 9 to 4.869 ms because of that Java requires more time at the beginning for parsing the Java packages and so on. For the naming component, the cost of processing a single request is 0.88 ms while at the end the cost is 1.896 ms. It is clear that naming process or generating a NC for each user and from his SLA is very low in term of time. The role evaluator takes 1.24 and 1.1 ms when the number of users is 50 and 1000, respectively. For the policy control the average time for evaluating the policies of each single user request is about 1.04 ms while at the end the cost is little bit reduced to 0.9583 ms. The cost of the NC evaluator is shown in Fig. 2b, evaluating each NC needs about 85 ms after evaluating 50 NCs. However, after processing 999 NCs or at the 1000 users, the cost for evaluating a single NC is 86.81 ms. However, the efficiency of this component is less than the efficiency of other components.

4.2 Measuring Search Improvement

The NC searching (or required time for searching about a specific consumer in the NC DB) is shown in (Fig. 2c). Compared to the SLA searching shown in (Fig. 2d), the NC searching is largely more efficient. For example, searching for a consumer when the number of consumers is 1000 takes about 1.344 ms using SLA searching and only 0.0023 ms. This means that the NC searching is 584 faster than the SLA searching. This is because that with the NC searching, we search only inside a specific string (NC) while

with the SLA searching we need first parse the SLA (which is an XM—based file) and then searching inside it.

4.3 Measuring Scalability

The scalability result of the mapping, naming, NC evaluator, Role evaluator and policy control components is shown in (Fig. 2e). It is clear that mapping process still scales well even when the number of the concurrent consumers is increased from 50 to 1000. For example, the required time for each consumer is 6.8 ms when the number of concurrent consumer is 50 and also 5.985 ms. The different in time values is due to the time spent by Java Virtual Machine (JVM) at the beginning of the running process. However, it is evidence that the mapping process is scaling very well.

For the naming scalability, generating a NC for each consumer takes 92.64 ms when the number of the concurrent consumers is 50. The same process takes 119.19 ms when the number of concurrent consumers is 1000. So, the naming process scalability gets worse and worse when the number of concurrent consumers is increased. However, the system still can work with more requests. The time cost for evaluating the role of a single consumer is 0.866 and 1.558 ms when the number of concurrent cloud consumers is 50 and 1000, respectively. As a result, the required time for evaluating the roles of each consumer is increased during increasing the number of concurrent consumers. But, the role evaluator process still scalable since it is able to serve the increased number of concurrent consumers. For evaluating the polices of a single consumer is 0.92 and

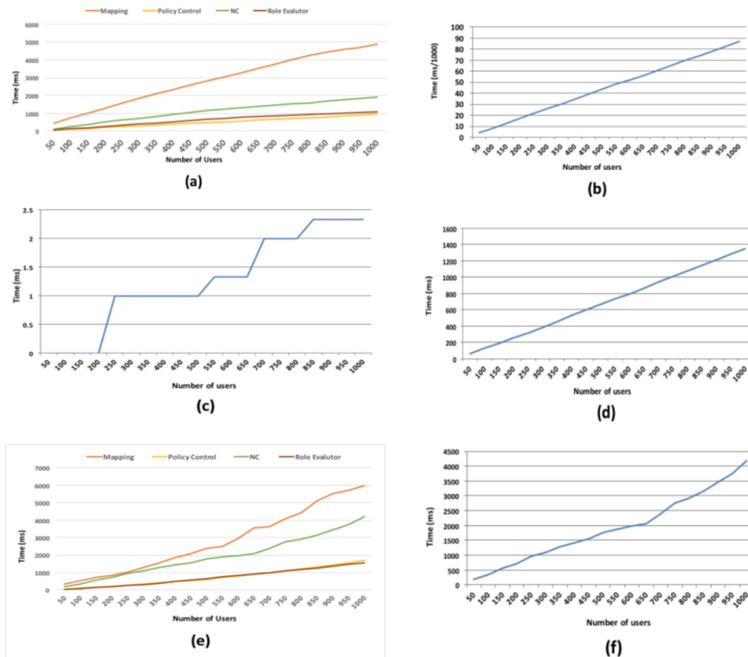


Fig. 2. Efficiency and scalability evaluation.

1.693 ms when the number of concurrent consumers is 50 and 1000. The required time is increased, or in other word the scalability is getting worse, while increasing the number of the concurrent consumers is increased.

Figure 2f, shows the scalability of the naming process. The required time for generating a NC for each consumer is 3.426 ms and 4.183 ms when the number of concurrent consumers is 50 and 1000, respectively. However, the system scalability is getting worse while increasing the number of the concurrent consumers but the process still scales with the increased number of the concurrent consumers.

Our proposed framework decreases administration density and resource consumption. A flexible and scalable solution for access control management in cloud environment had been deployed using appropriate methods of naming convention harmonizes that provide a guidance about SLA management between cloud providers and users it also Ensure that the designed model is workable and sophisticated based on its Implementation and measure its performance by measuring the system efficiency and scalability.

5 Conclusion and Future Works

This paper proposed a role based access control model for cloud computing based on the naming convention concept. The model is implemented and its efficiency and scalability are evaluated using an experiment study. The result shows that the proposed model provides an efficiency and scalable solution for cloud computing access control. Also, using the naming convention for searching and classifying user provides an efficient solution compared to WSAL.

We proposed as a future work to generalize our model to support access to multiple facilities concurrently using a central replicated database. Another direction is to utilize new cloud that supports new advanced Naming Convention techniques.

Acknowledgments. This work was funded by the ministry of higher education in Saudi Arabia. Many thanks for the people from the Center of Excellence in Information Assurance (COEIA) at King Saud University and the staff from the Saudi Culture Mission in China for their immense support towards this research work.

References

1. Mena, E., Kashyap, V., Sheth, A., Illarramendi, A.: OBSERVER: an approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distrib. Parallel Databases* **8**(2), 223–271 (2000)
2. Gao, K., Wang, Q., Xi, L.: Reduct algorithm based execution times prediction in knowledge discovery cloud computing environment. *Int. Arab J. Inf. Technol.* **11**(3), 268–275 (2014)
3. Singh, S. (ed.): A survey on cloud computing security: issues, threats, and solutions. *J. Netw. Comput. Appl.* **75**, 200–222 (2016)
4. Zhou, L., Varadharajan, V., Hitchens, M.: Achieving secure role-based access control on encrypted data in cloud storage. *IEEE Trans. Inf. Forensics Secur.* **8**(12), 1948–1960 (2013)

5. Platt, D.: Untangling access control and audit for cloud computing. In: Cloud Computing Virtual Conference (Cloud Slam 2010) (2009)
6. Hu, L., Ying, S., Jia, X., Zhao, K.: Towards an approach of semantic access control for cloud computing. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) CloudCom 2009. LNCS, vol. 5931, pp. 145–156. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-10665-1_13](https://doi.org/10.1007/978-3-642-10665-1_13)
7. Urquhart, J.: Cloud Computing Bill of Rights: 2010 edition (2010)
8. Manue, P.D., Selvi, S.T., Barr, M.I.: Trust management system for grid and cloud resources. In: First International Conference on Advanced Computing (ICAC9), pp. 176–181. Chennai (2009)
9. Ali, T., Nauman, M., Fazl-e, H., Muhaya, F.B.: On usage control of multimedia content in and through cloud computing paradigm. In: 5th International Conference on Future Information Technology (FutureTech), pp. 1–5. Busan (2010)
10. Yu, S., Wang, C., Ren, K., Lou, W.: Achieving secure, scalable, and fine-grained data access control in cloud computing. In: 29th Conference on Computer Communications, pp. 1–9. San Diego, CA, USA (2010)
11. Li, N., Mitchell, J.C., Winsborough, W.H.: Beyond proof-of-compliance: security analysis in trust management. *J. ACM* **52**, 474–514 (2005)
12. Ngo, C. (ed.): Multi-tenant attribute-based access control for cloud infrastructure services. *J. Inf. Secur. Appl.* **27–28**, 65–84 (2016)
13. Tsai, W-T., Shao, Q.: Role-based access-control using reference ontology in clouds. In: 2011 Tenth International Symposium on Autonomous Decentralized Systems, Tokyo & Hiroshima (2011)
14. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-based integration of information - a survey of existing approaches. In: IJCAI-01 Workshop: Ontologies and Information Sharing, vol. 2001 (2001)
15. Tsung-Yi, C.: Knowledge sharing in virtual enterprises via an ontology-based access control approach. *Comput. Ind.* **59**(5), 502–519 (2008)
16. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-based access control models, pp. 38–47. *IEEE Computer* (1996)
17. Halboob, W., Mahmod, R., S. Algathbar, K., Mamat, A.: TC-enabled and distributed cloud computing access control model. *J. Appl. Sci.* **14**(7), 620–630 (2014)

Multimedia and Visualization

Korean/Chinese Web-Based Font Editor Based on METAFONT for User Interaction

Minju Son^(✉), Gyeongjae Gwon, and Jaeyeong Choi

School of Computer Science and Engineering, Soongsil University, Seoul, Korea
{sonmibz, gwon, choi}@ssu.ac.kr

Abstract. Font designers need to spend the same amount of time and efforts when designing a different style of an already designed outline font like italic, bold, and so forth. On the other hand, METAFONT expresses fonts by tracing the skeleton of characters with a pen. It can easily change the style of characters only by altering the shape of a pen. However, since the METAFONT is a programming language, it is difficult to be used by font designers who are not familiar with programming languages. In this paper, we propose a web-based font editor based on METAFONT, which can be used to easily edit Korean/Chinese fonts. It extracts parameters of characters based on their anatomy and applies them to modify fonts using GUI.

1 Introduction

As digital media are now commonplace and pervasive, the attention of users for typography and calligraphy design is also increased. Some companies give impressions of products to the users by developing and distributing their own specific fonts. In this way, the influence of fonts is increased to the fields of design and industry beyond the simple usage in printing.

When designing Chinese fonts, around 8,000 characters which are widely used among the total 50,000 characters should be designed. And in the case of Korean fonts, all of 11,172 characters should be designed. When generating the font, characters are generally described as ‘outline’, and then outlines are filled in. It takes averagely more than 1 year to design one set of Korean or Chinese font with an outline font editor program. In addition, there is a drawback that it takes lot of time to change the style of an already generated font using general outline font editor. In order to complement these problems, many studies of programmable fonts have been carried out since 1980s.

METAFONT, which is a font design system for improving the quality of TeX typesetting in TeX document, is the representative programmable font [1]. METAFONT can reduce the cost of generating fonts by deriving various fonts with changing size or shape of pen. However, it is very difficult for font designers to design fonts by directly using the METAFONT which is provided as a programming language.

In this paper, we propose web-based font editor, especially for Korean and Chinese font, which can easily edit fonts in web browser. It can not only be used by font designers but also by general users who do not have the programming skills. Our system is built

on preceding work, i.e. structural font generating program based on METAFONT [2]. In this paper, we will discuss previous work related to this subject and structural font generating program, which was outcome of the preceding study. We will explain implementation of a web-based font editor and will show a few samples of generated characters. At the end we will make conclusion and the directions for future work.

2 Related Works

2.1 METAFONT

METAFONT is a programming language to define fonts. It uses “handwriting” method to draw skeleton of character, and fills the track of the skeleton with a pen to express fonts. In fact, METAFONT provides all the processes which we use to write a character on a simple white paper with our hand, like selecting a pen, grabbing the pen, and drawing character with desired directions in a human friendly way through the programming language. In addition, METAFONT can define the pen, curve, etc. to be used. It can increase the productivity of font design by code reusability.

However, there is no proper font editor for Korean and Chinese with GUI, using the METAFONT. In this paper, we address this limitation of METAFONT by providing the GUI for user interaction.

2.2 MMF

MMF developed by Adobe is an extended font format of the Type1 [3]. MMF has two or more fonts information called ‘master’ and derives various fonts by changing values such as a thickness and width within the range of ‘master’. The principle of MMF which raises the productivity of fonts by using two fonts was widely used for the study of generating the font. For example, Adobe announced the font generating application project faces [4] which is based on MMF, in the Adobe MAX 2015.

2.3 Metaflop

Metaflop is the web editor which can generate fonts with the basis of METAFONT [5]. Metaflop provides 3 kinds of basic fonts (bespoke, adjuster, fetamont) with METAFONT, and various fonts can be generated according to users’ requirement through the GUI. However, Metaflop can only design the 256 extension ASCII characters such as alphabet, number, pronunciation distinction mark, symbol, etc., and do not support Unicode such as Korean and Chinese.

2.4 Structural Font Generating Program

The programmable characteristic of METAFONT becomes more obvious when dealing with the Korean/Chinese font in which characters are constituted through the combination of radicals than single characters such as alphabet. Korean is constituted with the

structure of 3 radicals, i.e. initial, medial, and final, and the font can be generated with the method of calling the corresponding radicals after defining them to draw the skeleton of each radical then change values of the parameters for size and location. Similarly, Chinese character can be generated by using the METAFONT with the method of defining the stroke, radicals, main character, etc. in advance. Based on these characteristics of METAFONT, the structural font generating program has been implemented with METAFONT in our preceding study [2].

The font generating program was designed by considering the structure of ‘initial-medial-final’ and the ‘number of set’ of Korean. The font is automatically generated with the hierarchical method of generating radicals by combining the strokes defined in advance, and generating a character by combining the radicals. In addition, for a completed font, the thickness, tilt, size, serif, etc. of the font can be changed simply by modifying the values of parameters. Using this program, once we design a font set with certain parameters, they can be reused. We don’t need to design again and again like outline method. However, since structural font generating program is implemented with METAFONT language, therefore, knowledge for METAFONT is required. Therefore, if it is to be used by font designers who do not have any programming knowledge, learning METAFONT is a prerequisite.

In this paper, we propose GUI-based font editor which provides a convenient environment for designers to develop Korean and Chinese characters by using METAFONT easily. This will increase usability of METAFONT program and will improve existing font editing environment.

3 Extraction of Parameter for Font Style Change

3.1 Classification of Character Element

In the typography, character is divided into various design anatomies. These anatomies can be used as guideline in an editing tool when editing the font.

In the case of English characters, they are constituted with the anatomies as shown in Fig. 1 by classifying into capital letter, small letter, shape, etc. However, Korean is constituted with the structure of ‘initial-medial-final’ therefore; the character is completed by combining each of the radicals. Furthermore, in the case of Korean, there is the structure of ‘initial-medial-final’ in accordance with the style and location of medial, and existence of final. Even for the same radicals, anatomies can be different depending upon the style of character, and different values of parameters can be applied for the same radicals. Similarly, Chinese also has the basic units of strokes and radicals. It has combination rules, extraction of parameters based on Chinese anatomies, their classification and application. In this paper, we implemented 22 parameters for changing the style of Korean fonts as shown in Table 1 by considering Korean styles and anatomies as shown in Fig. 2.

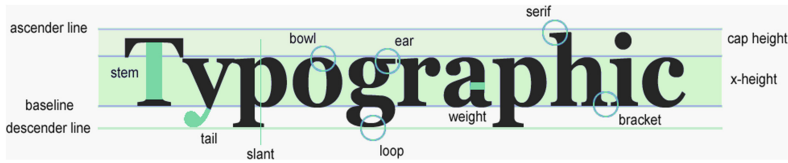


Fig. 1. Anatomies of alphabet

Table 1. Parameter materialized from METAFONT editor

Classification	Parameter name	Description	Range of value
pen	pen shape	The shape of pen	circle, rectangle, triangle
	pen width	The width of pen	0.1pt ~ 0.9pt
	pen height	The height of pen	0.1pt ~ 0.9pt
	pen rotation	The angle of pen	0 ~ 360
character	character width	width of character	5pt ~ 15pt
	character height	height of character	5pt ~ 15pt
	initial width	The width of initial	initial width * 0.0 ~ 0.2
	initial height	The height of initial	initial height * 0.0 ~ 0.2
	medial width	The width of medial	medial width * 0.0 ~ 0.2
	medial height	The height of medial	medial height * 0.0 ~ 0.2
	final width	The width of final	medial width * 0.0 ~ 0.2
	final height	The height of final	final height * 0.0 ~ 0.2
space	medial space(h)	The space between the initial and medial	0pt ~ 2pt
	medial space(v)	The space between the initial and medial	0pt ~ 2pt
	final space	The space between the initial/medial and final	0pt ~ 2pt
	fortis space	The space between fortis consonants	0pt ~ 2pt
style	serif width	The width of serif	serif width * 0 ~ 1.25
	serif height	The height of serif	serif height * 0 ~ 1.25
	hat1 style	The style of crest	style 1, style 2
	hat2 style	The style of tieut	style 1, style 2
	siot style	The style of siot	style 1, style 2
	italic	Application of tilt	yes, no



Fig. 2. Anatomies of Korean

3.2 Implementation of Korean Font Parameter

At first, we implemented parameters such as the shape, width, height, angle, pen, etc. METAFONT provides various shapes of pen such as circular, tetragonal, and triangular. The skeleton of character is to be filled with a selected pen. In Fig. 3, the shape of character is modified by changing width and height of pen. There are parameters to affect overall width, height, etc. of each character. And also there are parameters to modify width, height, etc. of each radical. The letters are unique in Korean and Chinese. Serif and slant of font, styles of crest applied to ‘ㅎ’ and the styles of ‘ㅍ’ and ‘ㅈ’ for variations of Korean font as shown in Fig. 4.

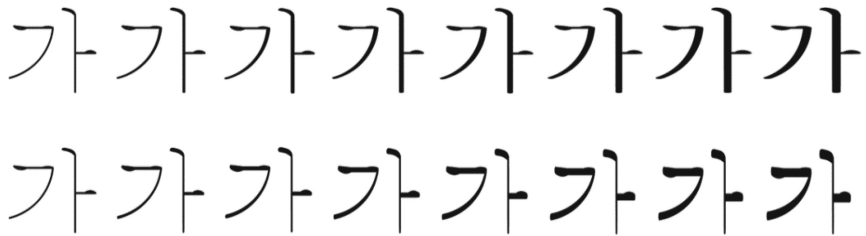


Fig. 3. The shape of Korean for which the parameters of pen are applied



Fig. 4. The style of Korean font

3.3 Applying the Parameters of Chinese Font

Among the parameters stated in Sect. 3.2, the parameters such as the width, height, shape of pen, overall size of character, serif, etc. can also be applied to Chinese font. In terms of the shape of stroke, Chinese character is similar to Korean; therefore, Chinese character is easy for applying the parameters extracted from the Korean anatomies. By

changing only the horizontal thickness of pen as shown in Fig. 5, a variety of Chinese fonts can be derived. However, the combination rule of Chinese character is more complicated than Korean, therefore, it is difficult to apply the parameters of Korean to Chinese directly. Therefore, the parameters considering combination rule of Chinese character should separately be extracted.



Fig. 5. The shapes of Chinese character for which the parameters of pen are applied

4 Implementation Korean/Chinese Web-Based Font Editor Based on METAFONT

The structure of Korean/Chinese web-based font editor based on METAFONT is shown in Fig. 6 and font designers can easily generate fonts through the change of the extracted parameters. The ‘Ming style’ is provided as the basic font in Fig. 7. Users can modify the parameters of ‘Ming style’ and can confirm change in fonts through the web interface. The ‘global.mf’ is updated by new values of parameters modified by font designers from GUI and used for the runtime application in METAFONT program. In this file, the font information is specified including a name of font, basic size, units, and data for changing the outline font file.

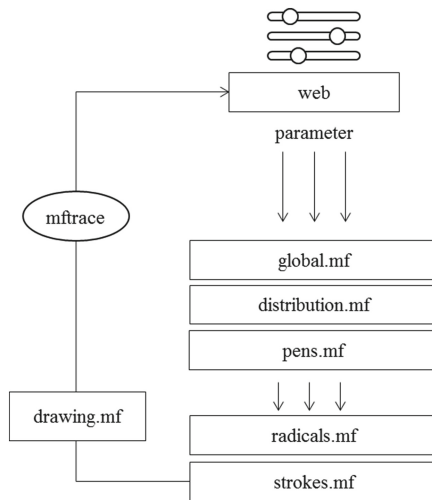


Fig. 6. Structure of font editor based on METAFONT

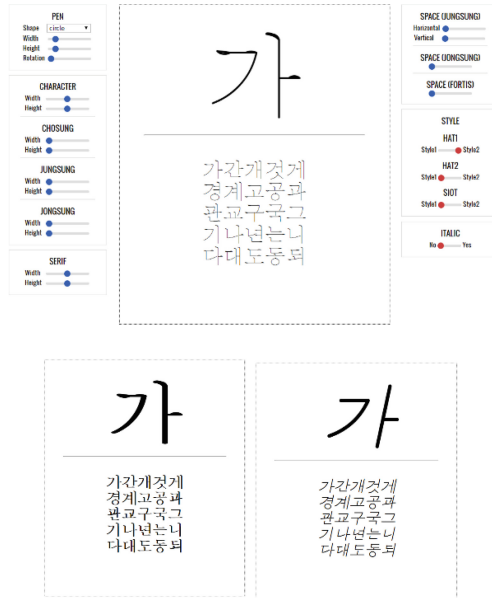


Fig. 7. User Interfaces of Korean/Chinese font editor based on METAFONT

As mentioned in Sect. 4, in order to complete the character by applying parameters or combining the stroke and radical, the style of character should also be considered. In the case of Korean, for the vertical shapes of medial such as ‘ㄱ, ㅋ, ㆁ, ㆁ, ㆁ,’ the parameters for the width between radicals should be applied separately. Same criteria should be considered for horizontal shapes of medial such as ‘ㄴ, ㄴ, ㄷ, ㄷ.’ This is because a change should be provided to the horizontal axis for the vertical shape of medial, and likewise to the vertical axis for the horizontal shape of medial.

Therefore, the parameters with appropriate values in accordance with styles are defined in ‘distribution.mf’. This is done after extracting styles of Korean for which the horizontal and vertical shape of medial, the shapes of combining the horizontal and vertical style such as ‘ㄱ’, ‘ㄱ’, and the existence of final are considered. Parameters have been sufficiently considered even for the styles of characters. And they are stored in ‘strokes.mf’ and ‘radicals.mf’. Then they are called and collectively applied to draw characters. In order to apply the fonts to the web, we have to convert the METAFONT file to outline font file such as true type, open type, etc. which are supported by browser. The mftrace is the Python program which converts the bitmap font, i.e. the output of METAFONT, into the outline font files such as the Type 1, true type, etc. We use mftrace for converting METAFONT file to outline font.

The user interface is developed using HTML5 as shown in Fig. 7; therefore, the result of style changing can be confirmed directly on the screen. On the basis of the default ‘Ming style’, completely different fonts can be generated by changing the thickness of pen, and changing the size of serif or applying the height, slant, etc. to the character.

5 Conclusions

In this paper, we propose an easy method to be used by font designers for METAFONT program, who do not have any subject knowledge and programming skills. Korean and Chinese font style can be changed using simple GUI controls, provided by Korean/Chinese web-based font editor. Lot of repeated work was required to change the font style with the existing outline font editor. In the case of Korean/Chinese web-based font editor; the styles of whole characters can be changed simultaneously by extracting parameters from anatomies of characters, and applying them. In this way, our Korean/Chinese web-based font editor can reduce the complexity of designing font and increase usability of METAFONT. This Korean/Chinese web-based font editor facilitates font design using METAFONT, based on the anatomies of characters. As it is implemented with HTML5, it can be used on PC as well as on hand held devices. Korean/Chinese web-based font editor is currently using 22 parameters to change font style. However, the study for the extraction and application of more parameters is under progress. With this study, it is expected that more convenient and detailed font editing system will be provided to font designers.

Acknowledgements. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIP) (No. R0117-16-0001, Technology Development Project for Information, Communication, and Broadcast).

References

1. Donald, K.: METAFONT: The Program. Addison Wesley Publishing Company, Boston (1986)
2. Gwon, G.J., Son, M.J., Choi, J.Y., Jeong, G.H.: Structured hangul font generator using METAFONT. In: Korea Information Science Society Conference, pp. 1134–1136 (2015)
3. Adobe Systems Inc.: Adobe Type 1 Font Format: Multiple Master Extensions (1992)
4. Adobe Systems Inc.: Adobe News. Sneak Peeks at Adobe Max 2015 (2015)
5. Metaflop (2012). <http://www.metaflop.com>. (Accessed 1 Nov 2016)

A Highly Robust and Secure Digital Image Encryption Technique

Md. Anwar Hussain^(✉), Popi Bora, and Joyatri Bora

North Eastern Regional Institute of Science and Technology, Itanagar, India
bubuli_99@yahoo.com, popibora2015@gmail.com,
bjoyatri@yahoo.com

Abstract. Robust and secure image encryption requires primarily a very large key-space and immunity to differential attack. We report in this paper a new novel encryption technique using chaotic Logistic map, a support image, and a hyper-chaotic 4-D system. The support image, which is available both with the transmitter and the receiver, is utilized to enhance the robustness and security possible from the encryption algorithms based on chaotic Logistic map and Hyper-chaotic system. The support image is used in two different ways in two schemes that provide different levels of key-space enhancement and immunity to differential attacks. The final stage of the technique uses either one round or two rounds of pixel value diffusion with hyper-chaotic system. The results reported here are encouraging to defeat attacks attempted with high computing resources.

Keywords: Chaos · Encryption · Logistic map · Hyper-chaos · Differential attack · Key-space · Support image

1 Introduction

To encrypt a plain image for secret delivery to a secret recipient the most commonly accepted scheme is to first apply confusion and then diffusion on the pixels of the plain image. The main work load of real encryption is done by the diffusion stage. Some authors apply an intermediate stage [1] for diffusion of pixels thus reducing the work load on the final stage. The confusion-diffusion scheme [2, 3] which is also termed as permutation-substitution is mostly reported with application of chaotic logistic map [4] and hyper-chaotic 3-D and 4-D system [4]. Out of all adversarial attacks on such encrypted images and communications, three very important are the differential attack, correlation analysis attack and the brute-force attack [5]. The confusion-diffusion mechanism is so designed that such attacks can be defeated. Chaotic systems are deterministic dynamical systems, first observed by Poicare with complex non-periodic behavior which is very sensitive to initial conditions such that its future time evolution is impossible for prediction. The chaotic Logistic map and hyper chaotic system, generally used for such algorithms are very sensitive to initial conditions and system parameters, show complex non-periodic behavior in future time evolution, and topological transitivity [6]. In literature, cipher system based on Chen chaotic system, Lorentz chaotic system, Jia chaotic system, and Rabinovich chaotic system are mostly used [5–8]. Bit level permutation in a particular bit plane (planes) is reported in [5, 9–11]. The diffusion stage of the ciphering

algorithm may be divided into two steps as in [1], consisting of simple substitution algorithm for fixed size blocks of bits with same size codes as reported in [12], The authors in reported digital image encryption technique using Logistic map, MSB substitution, and hyper-chaos. Authors in [13] reported a novel and robust digital encryption algorithm with Logistic map and 4-D Rabinovich Hyper-chaotic system which claim to defeat brute-force, differential and correlation analysis attacks.

In this paper we report a novel and very robust digital image encryption technique with very large key-space and highly resistant to differential attack, the technique uses a support image along with a chaotic logistic map and hyper-chaotic 4-D system. The same support image is used both by the sender and the receiver for encryption and decryption. The logistic map and the support image are used for confusion as well as to carry some load of diffusion as is explained below.

2 Proposed Encryption Technique

We formalize our technique mainly to defeat differential, correlation analysis, and brute-force attacks. Our proposed scheme is shown in Figs. 1 and 2. As shown in figures, the scheme can be implemented in two different algorithms: *Scheme A* and *Scheme B* which are based on how the plain image is pre-ciphered. We consider a plain text image of pixels $P_{i,j}$ for $i = 1, 2, 3, \dots, M$ and $j = 1, 2, 3, \dots, N$, and the Logistic map equation $x_{n+1} = 4r \cdot x_n(1 - x_n)$ as explained below. The final ciphering stage uses 4-D Rabinovich hyper-chaotic system [5] in bidirectional operation meaning two rounds of Hyper-chaotic system, although our results show that it is sufficient to have only one round.

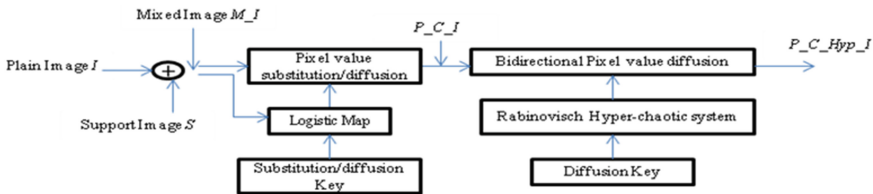


Fig. 1. Block diagram of scheme A

As shown in Fig. 1, we bitxor the plain image with the support image and then doubly diffuse/substitute the pixel values. The Fig. 2 shows the second algorithm of implementation in which the plain image is totally shuffled in pixel position, doubly

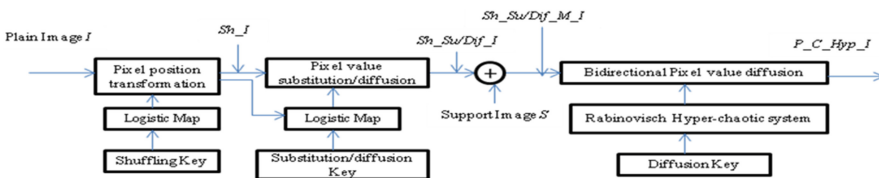


Fig. 2. Block diagram of scheme B

diffused/substituted in pixel values and then bitxored with the support image. By bitxoring we imply two image xored in the binary domain of the pixel values.

We use 4-D hyper-chaotic system in the final stage in one round or two rounds for final diffusion of pixel values with four initial values and the key-space of this stage is very strong. As we can see in Fig. 2 all the stages of Scheme B, first confusion-doubly diffusion in the first stage with plain image, intermediate stage with the support image bitxoring with the confused-doubly diffused plain image, and the final stage ciphering the processed image of the output of the intermediate, are all independent the key-space is the product of all the three stages. Similar is the observation from Fig. 1 of Scheme A. The various steps in the whole scheme is briefly explained below.

Row Shuffling: For given specific values of rand x_0 ($n = 0$) such that $0.8925 \leq x_0 \leq 1$, and $0 \leq x_0 \leq 1$, a new x_{n+1} is obtained after $n + 1$ rounds of iteration of the logistic map. Let

$$R_i = \text{mod}(x_{n+1} \cdot 10^{14}, M) \quad (1)$$

where $R_i \in [0, M - 1]$ which is used for row shuffling in the plain text image I , such that each R_i is different. The procedure is continued till all M rows are shuffled.

Column Shuffling: Pixels of columns of each row are shuffled in the same way as in row shuffling. The initial x_0 value is obtained by solving a nonlinear equation $y_1 = a_0 + a_1 y_0 + a_2 y_0^2$ after a number of iterations with system parameters a_0, a_1, a_2 , and initial value y_0 in the range $[0, 1]$ such that $0 \leq y_1 \leq 1$ and which is then put as the initial value x_n in the Logistic map to obtain x_{n+1} . Let

$$C_i = \text{mod}(x_{n+1} \cdot 10^{14}, N) \quad (2)$$

where $C_i \in [0, N - 1]$ and is used for column shuffling, each time C_i is to be different. The procedure is carried for each row of the row shuffled image, the last x_{n+1} value of a given row is used in the Logistic map equation for the next row.

Substitution/Diffusion in the pre-ciphering stage:

Scheme A: We use pixel value substitution/diffusion in pre-ciphering as well as in final ciphering stage, using logistic map and hyper-chaotic system, respectively. As shown in Fig. 1 this stage takes $M \cdot I$ image, resulting from bitxoring the plain and the support images, as the input and carries two rounds of chaotic Logistic map based substitution/diffusion, each round with a different initial state value x_0 .

The operation starts from the first pixel of the first column and proceeds from top to bottom in the first round, and the last pixel in the last column proceeding from bottom to top in the second round. For every pixel value substitution/diffusion, the Logistic map is iterated in times equal to the pixel value, multiply the resultant x_{n+1} by 10^{14} and divide by 256, and the remainder is the substituted/diffused pixel value. For every next pixel, the process takes the last x_{n+1} of the previous pixel as the starting x_0 . Mathematically this is shown in Eq. 3.

$$P_{i,j}^{subs} = \text{mod}(x_{n+1} \cdot 10^{14}, 256) \quad (3)$$

Let the output of this stage be termed as P_C_I image. The two rounds would resist any differential attack. This is because the substitution/diffusion key in every pixel is the previous iterated value x_{n+1} of Logistic map (excepting the starting first pixel) and hence any change in any pixel is going to affect the later pixels. Since this stage is of two rounds all pixels before and after the modified pixel are going to be affected. This operation may reduce the number of rounds in the final stage as we observe in our results.

Scheme B: In this scheme as shown in Fig. 2, the plain image I is first totally shuffled which outputs image Sh_I , and then pixel value substitution/diffusion is applied on Sh_I in two rounds by the logistic map. The output image Sh_Sub/Dif_I is bitxored with the support image S , and the resultant mixed image Sh_Sub/Dif_M_I is here the pre-ciphered image P_C_I . It may be noted that the logistic map is used thrice for pixel position shuffling and pixel value substitution using three different initial values. The pixel value substitution/diffusion follows the same Eq. 3 as in scheme A.

Image Pixel Value Cipherng: This is the final stage of the encryption technique. Rabinovich 4-D Hyper-chaotic system [5] used in this stage is shown in Eq. 4 below.

$$\begin{aligned} \dot{x} &= ry - ax + yz \\ \dot{y} &= rx - by - xz \\ \dot{z} &= -dz + xy + u^2 \\ \dot{u} &= xy + cu \end{aligned} \quad (4)$$

The system exhibits chaotic behavior for $a = 4$, $b = -0.5$, $c = -2.2$, $d = 1$, and $r = 8.1$. The initial state values x_0 , y_0 , z_0 , u_0 are used as key. The four Lyapunov exponents are: $L_1 = 1.090046$, $L_2 = 0.012243$, $L_3 = -3.105106$, $L_4 = -4.697183$, and the Kaplan-Yorke dimension is $D_{KY} = 2.5736132$. The final stage consists of 4 steps, which outputs the final cipher text $P_C_Hyp_I$. We follow the same 4 steps as explained in [5]. Steps are briefly explained as under.

Step 1: The rectangular pre-ciphered image P_C_I is mapped to a vector $V = (v_1, v_2, v_3, \dots, v_{M \times N})$ taking columns one at a time and from top to bottom.

Step 2: The system of equations in Eq. 4 are pre-iterated for a constant T_0 times, and then solved using fourth order Runge-Kutta method for a step value $h = 0.0005$.

Step 3: A sequence of four key k_{Qn} streams is obtained iterating the hyper-chaotic system, as shown in Eq. 5, for cipherng four pixels at a time.

$$\begin{aligned} k_{Qn} &= \text{mod}[\text{round}((\text{abs}(Qn) - \text{floor}(\text{abs}(Qn))) \times 10^{14}), 2^D] \\ \text{where } Q &\in \{x, y, z, u\} \end{aligned} \quad (5)$$

Step 4: For current four pixel values $v_{4(n-1)+m}$, where $m = 1, 2, 3, 4$ k_{Qn} is circularly left shifted by l_Q bits, where l_Q is obtained from the previously operated four pixels, as given by the following Eq. 6.

$$\begin{aligned} l_x &= \text{mod}(v_{4(n-1)}, 2^D), l_y = \text{mod}(v_{4(n-1)+1}, 2^D), \\ l_z &= \text{mod}(v_{4(n-1)+2}, 2^D), l_u = \text{mod}(v_{4(n-1)+3}, 2^D) \end{aligned} \tag{6}$$

The shifted key streams are used to cipher the current four pixel values using the Eq. 7 below. The initial cipher pixel c_0 is assumed in the range $[0, 255]$.

$$\begin{aligned} c_{4(n-1)+1} &= k_{xn} \oplus \{[v_{4(n-1)+1} + k_{xn}] \text{mod } 2^D\} \oplus c_{4(n-1)} \\ c_{4(n-1)+2} &= k_{yn} \oplus \{[v_{4(n-1)+2} + k_{yn}] \text{mod } 2^D\} \oplus c_{4(n-1)+1} \\ c_{4(n-1)+3} &= k_{zn} \oplus \{[v_{4(n-1)+3} + k_{zn}] \text{mod } 2^D\} \oplus c_{4(n-1)+2} \\ c_{4(n-1)+4} &= k_{un} \oplus \{[v_{4(n-1)+4} + k_{un}] \text{mod } 2^D\} \oplus c_{4(n-1)+3} \end{aligned} \tag{7}$$

For detail of key sequence generation and ciphering reference [5] may be seen. The resultant image is termed $P_C_Hyp1_I$. The above 4 steps are again carried out column wise from bottom to top on the cipher text $P_C_Hyp1_I$ and the final encrypted image is termed as $P_C_Hyp_I$.

3 Results and Discussion

We consider the Lena image as the plain image I to be encrypted of size 128×128 , shown in Fig. 3 and the Baboon image as the support image S of the same size as shown in Fig. 4. The initial state values in the total shuffling of the Lena image I of *scheme B* are assumed as $x_0 = 0.93412045$, $a_0 = 0.23$, $a_1 = 0.54$, $a_2 = 8.5$, and $y_0 = 0.78956$, y_0 is changed for column pixel shuffling in each row as $y_0 = 0.001M.y_l$, $M = 1, 2, 3, \dots, 128$, and $N = 1, 2, 3, \dots, 128$. Initial state values of two rounds of Logistic map based substitution/diffusion stage of *scheme B* are $x_0 = 0.93103012$ and $x_0 = 0.93152013$, respectively. For *scheme A*, the initial state values of x_0 for pixel value substitution/diffusion may be taken different from *scheme B*. The initial state values of Hyper-chaotic system are assumed as $x_0 = 5.23$, $y_0 = -5.786$, $z_0 = 6.722$, and $u_0 = 5.2654$, and the system parameters are $a = 4$, $b = -0.5$, $c = -2.2$, $d = 1$, and $r = 8.1$. We consider here 8-bit gray pixel image and hence $D = 8$. Figures 5, 6, 7, 8 and 9 show the result of *Scheme A*, and Figs. 10, 11, 12 and 13 show the result of *Scheme B*.

For plaintext sensitivity analysis, where relationship between the plaintext image and the cipher text image may be analyzed for an attack by an adversary, we carried out all the stages of the proposed encryption technique, both for scheme A and scheme B, on the Lena image with its 60th row and 75th column value ($I(60,75)$) changed by 1 and we term it I_C . Correlation between randomly chosen 4000 adjacent pairs in the vertical, and the horizontal directions are as shown in Table 1 for scheme A and in Table 3 for scheme B. For differential attack analysis, we consider two well-known

metrics: *NPCR* (Number of Pixel Change Rate) and *UACI* (Unified Average Changing Intensity). The first metric is defined as the measure of different pixel numbers in two random images and the second metric is defined as the measure of the average intensity differences in two random images. The *NPCR* and *UACI* are shown in Table 2 for scheme A and in Table 4 for scheme B, respectively.



Fig. 3. Original plain image



Fig. 4. Support image



Fig. 5. Mixed image

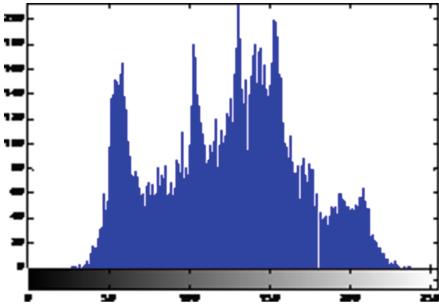


Fig. 6. Histogram of plain image

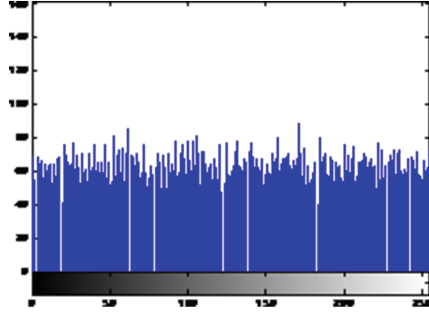


Fig. 7. Histogram of mixed-substituted image

Key-Space for Scheme A: As pixel values of I and S are bitxored in binary domain for all 128×128 pixels, the key-space of this step is $(2^8)^{128 \times 128}$, the mixed pixels are then substituted/diffused using chaotic Logistic map in two rounds and hence key-space in this step is $(10^{15})^2$. Finally, the cipher image is obtained as the output of the Hyper-chaotic system which is used in two rounds each with four initial values, thus providing a key-space of $(10^{15})^4$. Hence the overall key-space is $(2^8)^{128 \times 128} \times (10^{15})^2 \times (10^{15})^4 \sim 2^{131370}$. The key-space is really huge for brute-force attack.

Key-Space for Scheme B: As in this scheme, the plain image is first totally shuffled, and substituted/diffused in two rounds the key-space in this step is $(10^{15})^3$, after which the resultant image is bitxored with the support image in binary domain, thus providing a key-space of $(2^8)^{128 \times 128}$. Finally, the cipher is obtained using the Hyper-chaotic system in two rounds and key-space is $(10^{15})^4$. The overall key-space is $(10^{15})^3 \times (2^8)^{128 \times 128} \times (10^{15})^4 \sim 2^{131420}$, which is again really a huge key-space to defeat any brute-force attack.

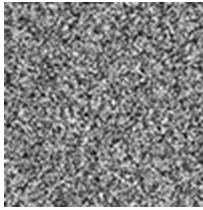


Fig. 8. Final cipher image

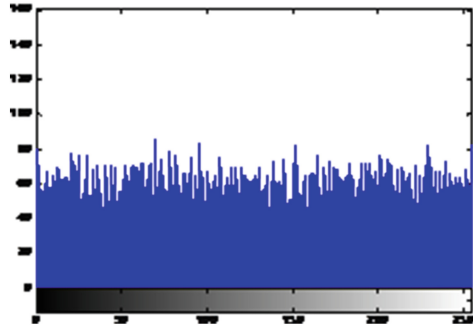


Fig. 9. Histogram of final cipher (of I)

Table 1. Results of correlation analysis of different images (Scheme A)

Direction	Plain image I	M_I	$P_C_HypI_I$	$P_C_Hyp_I$
Vertical	0.8892	-0.0012	0.0034	-0.0144
Horizontal	0.9393	0.0046	-0.0128	0.0153

Table 2. Results of $NPCR$, and $UACI$ between different images (Scheme A)

Metric	Between P_C_I of original and modified plain images	Between $P_C_HypI_I$ of original and modified plain images	Between $P_C_Hyp_I$ of original and modified plain images
$NPCR$	99.55%	99.63%	99.59%
$UACI$	33.60%	33.34%	33.22%

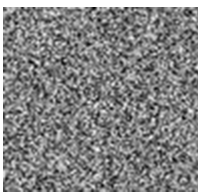


Fig. 10. Shuffled-substituted-mixed image

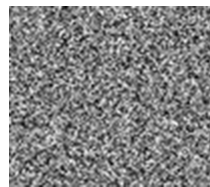


Fig. 11. Final cipher image

Table 3. Results of correlational analysis of different images (Scheme B)

Direction	Plain image	Sh_Sub/Dif_I	Sh_Sub/Dif_M_I	$P_C_HypI_I$	$P_C_Hyp_I$
Vertical	0.8892	0.0377	0.0180	-0.0381	-0.0178
Horizontal	0.9393	0.0001	0.0208	-0.0265	0.0142

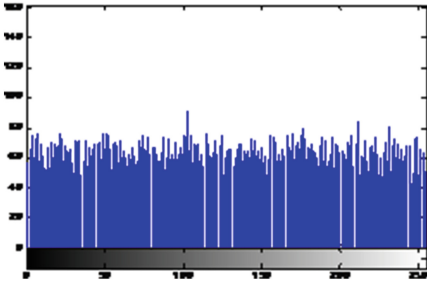


Fig. 12. Histogram of Shufld-Substd-Mixed

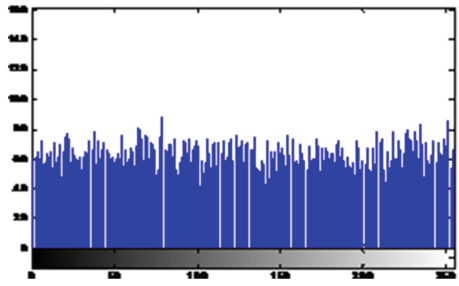


Fig. 13. Histogram of final cipher image

Table 4. Results of *NPCR* and *UACI* between different images (Scheme B)

Metric	Between <i>Sh_Sub/Dif_M_I</i> of original and modified plain images	Between <i>P_C_HypI_I</i> of original and modified plain images	Between <i>P_C_Hyp_I</i> of original and modified plain images
<i>NPCR</i>	99.55%	99.71%	99.54%
<i>UACI</i>	33.26%	33.56%	33.51%

4 Conclusion

We report a highly robust and secure digital image encryption technique with two different schemes. The novelty of the schemes is that a support image is used in the encryption technique, and the same is used for decryption also. The idea of using an additional (support) image is to make encryption techniques, which are based on Logistic map and Hyper-chaotic system, highly robust and secure by increasing the key-space to very huge values. Since the support image is known to the recipient, the decryption time is minimal. From the algorithmic analysis it is observed that both the schemes have huge key-space to defeat any brute-force attacks. The schemes are equally resistant to differential and correlation analysis attacks. It is further observed from the *NPCR* and *UACI* values that it is sufficient to apply the Hyper-chaotic system for one round only as the differential attack is defeated before application of the hyper-chaotic system. From the key-space analysis and *NPCR/UACI* values, we may consider the application of hyper-chaotic system as totally optional, as even with the first two stages in both the schemes the brute-force and differential attacks are resisted by the remaining part of the encryption technique. Hence the proposed digital image encryption technique may be considered as adaptive to requirements. The observed results are encouraging for accepting and applying the proposed technique.

References

1. Wong, K., Kwok, B.S., Law, W.: A fast image encryption scheme based on chaotic standard map. *Elsevier Phys. Lett. A* **372**, 2645–2652 (2008)
2. Fridrich, J.: Symmetric ciphers based on two-dimensional chaotic maps. *Int. J. Bifurc. Chaos* **8**, 1259–1284 (1998)
3. Scharinger, J.: Fast encryption of image data using Kolmogorov flows. *J. Electron. Imaging* **7**, 318–325 (1998)
4. Xiao, H., Zhang, G.: An image encryption scheme based on chaotic systems In: Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 2707–2711 (2006)
5. Fu, C., Huang, J., Wang, N., Hon, Q., Lei, W.: A symmetric chaos-based image cipher with an improved bit-level permutation strategy. *Entropy* **16**, 770–788 (2014)
6. Gao, T., Chen, Z.: A new encryption algorithm based on hyper-chaos. *Elsevier Physics Letters A*, pp. 396–400 (2008)
7. Vaidyanathan, S., Volos, C., Pham, V.: Hyperchaos, adaptive control and synchronization of a novel 5-D hyperchaotic system with three positive Lyapunov exponents and its SPICE implementation. *Arch. Control Sci.* **24**(4), 409–446 (2014)
8. Tong, X., Liu, Y., Zhang, M., Xu, H., Wang, Z.: An image encryption scheme based on hyperchaotic Rabinovich and exponential chaos maps. *Entropy* **17**, 181–196 (2015)
9. Fu, C., Lin, B.B., Miao, Y.S., Chen, J.J.: A novel chaos-based bit-level permutation scheme for digital image encryption. *Opt. Commun.* **284**, 5415–5423 (2011)
10. Zhu, Z.L., Zhang, W., Wong, K.W., Yu, H.: A chaos-based symmetric image encryption scheme using a bit-level permutation. *Inf. Sci.* **181**, 1171–1186 (2011)
11. Zhang, G.J., Shen, Y.: A novel bit-level image encryption method based on chaotic map and dynamic grouping. *Commun. Theor. Phys.* **58**, 520–524 (2012)
12. Hussain, M.A.: Simple encryption encoding for data hiding and security applications. In: The 3rd International Conference on Multimedia Technology, Published by Atlantis Press, Guangzhou, China, pp. 1621–1628 (2013)
13. Bora, P., Bora, J., Hussain, M.A.: A novel and robust image encryption algorithm using logistic map and hyper chaotic system. In: IEEE Conference I2CT 2017

Saliency Based Object Detection and Enhancements in Static Images

Rehan Mehmood Yousaf¹✉, Saad Rehman², Hassan Dawood¹, Guo Ping³,
Zahid Mehmood¹, Shoaib Azam⁴, and Abdullah Aman Khan⁵

¹ Software Engineering Department, UET, Taxila, Pakistan

Rehanmehmoodyousof@gmail.com,

{hassan.dawood, zahid.mehmood}@uettaxila.edu.pk

² College of Electrical and Mechanical Engineering, NUST, Islamabad, Pakistan

saadrehman@ceme.nust.edu.pk

³ Image Processing and Pattern Recognition Laboratory,

Beijing Normal University, Beijing, China

pguo@bnu.edu.cn

⁴ Gwangju Institute of Science and Technology, Gwangju, Korea

shoaibazam@gist.ac.kr

⁵ HITEC University, Taxila, Pakistan

abdokhan@hotmail.com

Abstract. Human visual system always focuses on the salient region of an image. From that region the salient features are obtained and can be collected by generating the saliency map. Natural statistics measures are used to measure the saliency from data collection of natural images. ICA filters are used to generate the saliency map that can blur the image. We have improved it by using different techniques like edge detection and morphological operations. By applying these algorithms we have successfully reduced the blur in images. That makes the salient objects more prominent by sharpening the edges. Proposed method is also compared with the state-of-the-art method like Achanta model.

Keywords: Saliency · Edge detection · Morphological image processing · AUC score

1 Introduction

The human visual system is enriched by modern technology and lot of work has been done to improve the saliency map. Improvement in saliency map can increase the object detection and tracking. Different techniques are being used by the latest researchers such as particle filters, log maps, background subtraction, feature extraction and feature description. Feature extraction and description are used for image matching and recognition [30–33]. Saliency is basically making the most prominent features salient so that the machine visual system can recognize the important information in an image. Saliency can be done through different techniques that includes difference of Gaussian, independent component analysis filters [2], spectral residual and spatial-temporal [26].

Recently many models have tried to explain the above mentioned problem for instance Hou and Zhang proposed a method which deals with spectral residual method by using the Fourier transform [26]. Then Achanta *et al.* proposed a method based on colour and luminance of the image and obtained the well-defined regions of the image [25]. Natural statistics also plays an important role to obtain the saliency map. The independent component analysis filter is used to obtain linear features, used in the saliency algorithm, obtained by applying on the natural images. By using the edge detection and morphological operations we have improved the performance of the system. The rest of the paper is organised as: Sect. 2 briefly explains the ICA. In Sect. 3 we explained our proposed methodology. Sections 4 and 5 are about edge detection and the morphological operations. In Sect. 6, we have discussed the results and compared with baseline method.

2 Independent Component Analysis

Independent component analysis [2] is used to recover independent signals from the measured signals. The measured signals are a linear combination of independent signals. Therefore an equal number of independent and measured signals are obtained.

Independent component analysis is defined as

$$X_i = a_1s_1 + a_2s_2 + \dots + a_ns_n \quad (1)$$

Or in matrix form

$$X = AS \quad (2)$$

Where X_i belongs to every measured signal, S is independent signals and A is $n \times n$ matrix called the mixing matrix. An alternative form of (2) can be obtained if and only if matrix A is invertible that is

$$W = A^{-1} \quad (3)$$

So,

$$S = WX \quad (4)$$

This means that each independent signal S_i can be expressed as a linear combination of measured signals. So, by estimating the W , independent signal S can be obtained. We assume that each signal is a random variable. The Central Limit Theorem states that if the sum of several independent random variables, such as those in S , tends towards a Gaussian distribution [3]. So $x_i = a_1s_1 + a_2s_2$ is more Gaussian than either s_1 or s_2 . The Central Limit Theorem also implies that if the combinations of the measured signals in X with minimal Gaussian properties are obtained, then that signal will be one of the independent signals. To achieve this we have to measure the nongaussianity of WX [3, 22–24]. To measure the nongaussianity the Negentropy approximation is used.

We have applied the FASTICA [2] algorithm as used by Lingyun Zhang and Tim K. Marks [1] and enhanced the images by using different operations; hence improved the salient features in saliency map to a noticeable result.

The following are results and 3D plots of ICA filters applied on three different images.

3 Proposed Methodology

Saliency map is generated by ICA has some drawbacks like edges are blur so salient features are not easy to obtain. We have improved the saliency map generated by ICA. Sobel operator is applied on the saliency map that was generated by ICA. Morphological operator (dilation) is then used to improve the edges by which salient features can be obtained efficiently. The visual results clearly show that proposed methodology improves the performance of the system.

Flow chart of proposed method is as follows (Fig. 2).

4 Edge Detection

Edge detection technique enhances the edges of an image by sharpening the image edges, Sobel operator is found to be good edge detector [10]. The Sobel operator is a discrete function computing the gradient of the intensity in an image. The operator uses two 3×3 kernels which are convolved with the original image to calculate approximations of the derivatives-one for horizontal changes, and one for vertical [10]. Image is represented as B , and H_x and H_y are two filters to compute the horizontal and vertical derivative approximations, the computations are as follows:

$$\mathbf{H}_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} * \mathbf{B} \quad \mathbf{H}_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * \mathbf{B}$$

Where $*$ is the 2-dimensional convolution operation.

Since the Sobel kernels can be decomposed as the products of an averaging and a differentiation kernel, they compute the gradient with smoothing. For example, H_x can be written as

$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} [1 \ 0 \ -1]$$

The x-coordinate is defined here as increasing in the “right”-direction, and the y-coordinate is defined as increasing in the “down”-direction. At each point in the image, the resulting gradient approximations can be combined to give the gradient magnitude, using

$$\mathbf{H} = \sqrt{\mathbf{H}_x^2 + \mathbf{H}_y^2}$$

Gradient direction can be calculated as:

$$\alpha = \text{atan2}(\mathbf{H}_y, \mathbf{H}_x)$$

Where, for example, α is 0 for a vertical edge that is darker on the right side.

5 Morphological Operation

Morphological operations are used for extracting meaningful components from the images. There are different operations like dilation, erosion, opening, closing etc.

Dilation operator is used in our experiments [11]. Dilation [11] is used to thicken or grow objects in an image. The dilation process takes two pieces of data as inputs. The first is the original image and the second one is structuring element (also known as kernel). Structuring element is the one through which thickening process is controlled in dilation operation. The line structuring element is applied on the image pixels from the start till end. The SE is applied in the form of the line in the image every time and a change in the pixel values according to the SE i.e. the change appears in places where the line affects the pixel and finally thicken the points we wanted and making the edges thicker and clear.

Dilation function is defined in term of set operation. The dilation of C and D is defined as

$$C \oplus D = \{z|(D')_z \cap C \neq \phi\}$$

Where ϕ is the empty set, D is structuring element and C is the binary image. In other words, dilation of C and D is the set consisting of all elements of D' such that its origin remain in C.

Morphological operation (Dilation) is applied to edge detection images to enlarge the boundaries of the regions of the salient features.

6 Results

To verify the proposed algorithm and with base line algorithm the image dataset is used as in [21]. For experiments 1000 images randomly selected from 10000 images and computed the results. Some images are shown in Fig. 3 after applying our proposed method.

6.1 Results by Applying Algorithm on the Dataset

7 Comparison Between ICA, Edge Detection Technique and Morphological Operation

Saliency map of original image is generated by applying ICA filter image. These saliency maps are shown in Fig. 1, with their 3D plots that clearly show the most prominent and salient features in the images. From figure it can be shown that the edges are not clear so sobel and dilation operators are used to improve the image quality. The results clearly show that the edges are sharper in whole image. This gets us to the point nearer to the object detection rather just projector out the prominent features. The histogram of the sobel operator shows us the peaks referring to the edges of the object and filter out purely the object. Results of the sobel operator is improved by applying another operation i.e. dilation; that provide the perfect results for object detection as shown in Fig. 3. The uncompleted edges have been thickened and completed by dilation. The histogram of dilation images gives the clear peaks and also those peaks that are blurred before. So, it is concluded after comparison between three of the techniques that when applied in hybrid these three techniques gives us the clear detected object (Fig. 4).

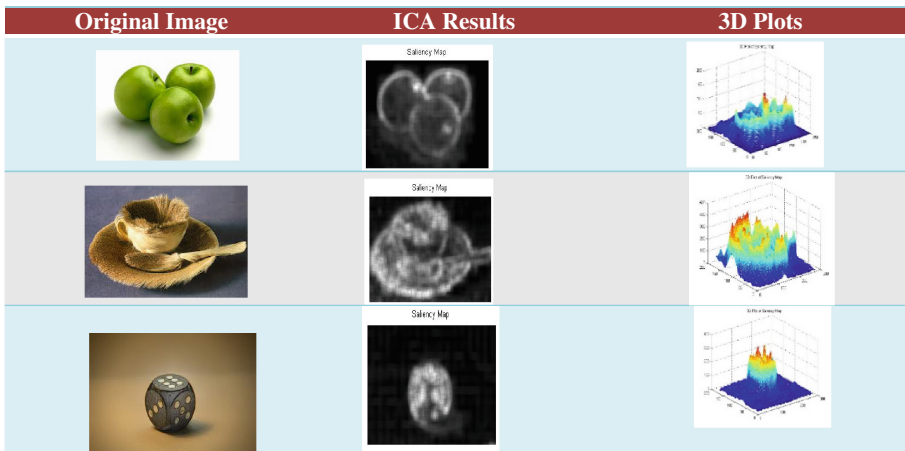


Fig. 1. Saliency map using ICA filters.

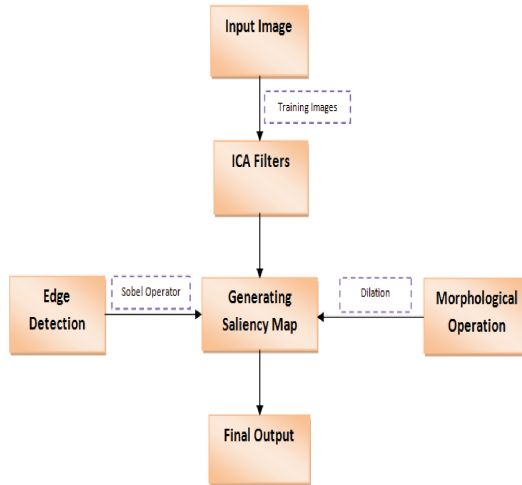
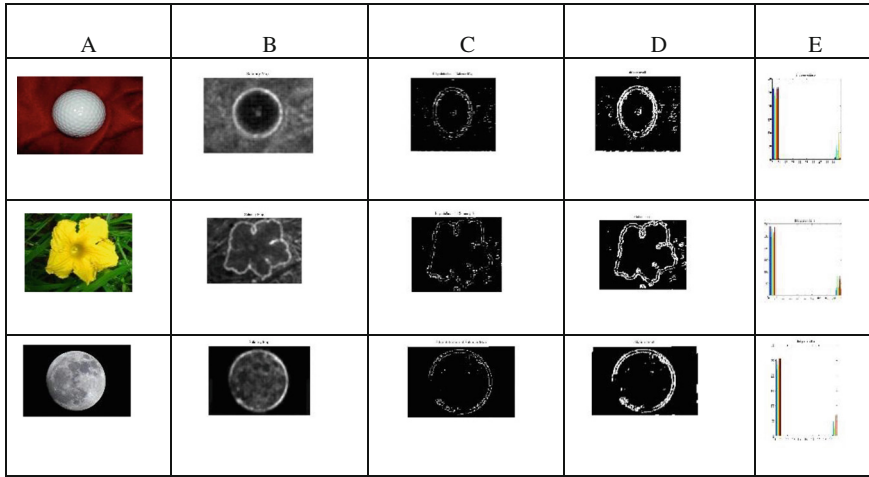


Fig. 2. Flow Diagram



Fig. 3.

The results can be visualized as under:



(a) Original Images (b) Saliency maps (c) Edge detection (d) Morphological operation (e) Histogram

Fig. 4.

8 Comparison of Two Computational Models and the AUC Score

The comparison between two models, which are the Achanta et al model and the ICA based saliency model, are based on the testing of area under the ROC curve score.

The area under the ROC curve needs a dataset that has been tested on human visual system. The dataset are used for calculating an AUC consists of random number of observers in a free viewing scenario on 135 different images [28]. The scenario is kept as to obtain the best possible results. The data for these random numbers of observations are viewed on is viewing 135 different images and their point of focus is computed. These values are used to compute the eye fixation map. Then the Achanta [28] model is used to compute the saliency maps of the same 135 images. These saliency maps are then compared with the eye fixation results and a score for each image is computed. Similarly the ICA based saliency maps are also compared to the eye fixation results and score is computed again. The score limits from 0 to 1. If the comparison results in score near to 1 the similarity is maximum thus the result is good and if it's near 0 then the similarity is minimum. We have only computed the comparison for first 8 images

The results of comparison between the AUC score is as under (Fig. 5):

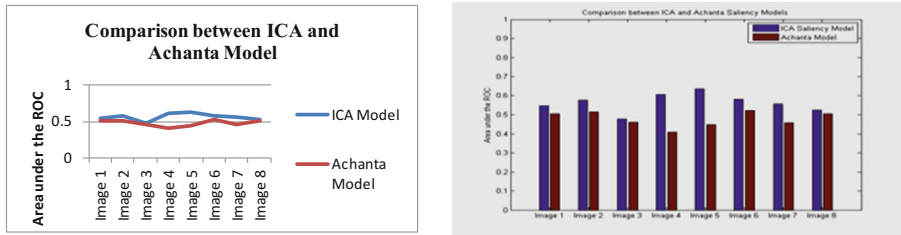


Fig. 5.

9 Conclusions

To generate the saliency map, ICA filter is used that gives the linear features used in saliency algorithm. To enhance the salient region edge detection technique with sobel operator is used which shows the accurate region of the saliency map and distinguishes from background. Using the same edge detection results, morphological operation (dilation) is applied that has brighten the edges, improves the lines and curves of edge detection result. Dilation is basically used to complete the incomplete boundaries of the region and thickens it contains. After applying all these operations the results clearly shows the effects, improving the saliency map and the edge detection has enhanced the results of saliency map. The morphological technique highlights the final result of the feature detection. The AUC ROC score also proves the models results better than many models as baseline model.

References

1. Zhang, L., Marks, T.K., Tong, M.H., Shan, H., Cottrell, G.W.: SUN: a bayesian framework for saliency using natural statistics. *J. Vis.* **8**(7), 1–20 (2008)
2. Hyvarinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9**(7), 1483–1492 (1997)
3. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. *Neural Comp.* **7**, 1129–1159 (1995)
4. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 30–43 (2010)
5. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
6. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: *IEEE 12th International Conference on Computer Vision* (2009)
7. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) *Advances in Neural Information Processing Systems*, vol. 18, pp. 155–162. MIT Press, Cambridge (2006)
8. Umbaugh, S.E.: *Digital Image Processing and Analysis: Human and Computer Vision Applications with CVIptools*, 2nd edn. CRC Press, Boca Raton (2010). ISBN 9-7814-3980-2052

9. Jähne, B., Scharr, H., Körkel, S.: Principles of filter design. In: Handbook of Computer Vision and Applications. Academic Press, Cambridge (1999)
10. Sobel, I.: History and Definition of the Sobel Operator (2014)
11. Dougherty, E.R.: An Introduction to Morphological Image Processing (1992). ISBN 0-8194-0845-X
12. Efford, N.: Digital Image Processing: A Practical Introduction Using Java™. Pearson Education, Upper Saddle River (2000)
13. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, Upper Saddle River (2008). ISBN 0-13-168728-8
14. Haralick, R., Shapiro, L.: Computer and Robot Vision, vol. 1, Chap. 5. Addison-Wesley Publishing Company, Boston (1992)
15. Jain, A.: Fundamentals of Digital Image Processing. Prentice-Hall, Upper Saddle River (1986)
16. Itti, L., Koch, C.: Computational modeling of visual attention. *Nat. Rev. Neurosci.* **2**(3), 194–203 (2001)
17. Avraham, T., Lindenbaum, M.: Esaliency: Meaningful attention using stochastic image modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **99**(1) (2009)
18. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* **9**(3), 1–24 (2009)
19. Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C.: Attentional selection for object recognition — a gentle way. In: Bühlhoff, Heinrich, H., Wallraven, C., Lee, S.-W., Poggio, Tomaso, A. (eds.) *BMCV 2002. LNCS*, vol. 2525, pp. 472–479. Springer, Heidelberg (2002). doi:[10.1007/3-540-36181-2_47](https://doi.org/10.1007/3-540-36181-2_47)
20. <http://www.ece.rice.edu/~dhj/courses/elec531/notes.pdf>
21. Cheng, M.-M., Mitra, N.J., Huang, X., Hu, S.-M.: SalientShape: group saliency in image collections. *Visual Comput.* (2013)
22. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430 (2000)
23. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis, 481 pages. John Wiley & Sons, Toronto (2001)
24. McKeown, M.J., Makeig, S., Brown, G.G., Jung, T., Kindermann, S.S., Bell, A.J., Sejnowski, T.J.: Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Mapp.* **6**, 160–188 (1998)
25. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 1597–1604, June 2009
26. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach, Department of Computer Science, Shanghai Jiao Tong University
27. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, John, K. (eds.) *ICVS 2008. LNCS*, vol. 5008, pp. 66–75. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-79547-6_7](https://doi.org/10.1007/978-3-540-79547-6_7)
28. Mancas, M., Le Meur, O.: Memorability of natural scenes: the role of attention. In: *Proceedings of the International Conference on Image Processing (IEEE ICIP 2013)*, Melbourne, Australia, September 15–18 (2013)
29. Dawood, H., Dawood, H., Guo, P.: A hybrid image feature descriptor for classification. In: *2015 11th International Conference on Computational Intelligence and Security, (CIS 2015) (EI)*, pp. 58–61 (2015)
30. Hassan, D., Dawood, H., Guo, P.: Removal of random-valued impulse noise by Khalimsky grid. In: *Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast 2015) (EI)*, vol. 1 (2015)

31. Wang, Y., Dawood, H., Guo, P., Yin, Q.: A comparative study of different feature mapping methods for image annotation. In: *The Seventh International Conference on Advanced Computational Intelligence (ICACI 2015)*. (EI) (2015)
32. Dawood, H., Dawood, H., Guo, P.: Texture image classification with improved weber local descriptor. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, Lotfi, A., Zurada, Jacek, M. (eds.) *ICAISC 2014. LNCS (LNAI)*, vol. 8467, pp. 684–692. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-07173-2_58](https://doi.org/10.1007/978-3-319-07173-2_58)
33. Dawood, H., Dawood, H., Guo, P.: Global matching to enhance the strength of local intensity order pattern feature descriptor. In: Guo, C., Hou, Z.-G., Zeng, Z. (eds.) *ISNN 2013. LNCS*, vol. 7951, pp. 497–504. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39065-4_60](https://doi.org/10.1007/978-3-642-39065-4_60)

A Center Symmetric Padding Method for Image Filtering

Mengqin Li^(✉) and Xiaopin Zhong

College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China
mengqin_li@126.com, xzhong@szu.edu.cn

Abstract. Padding the boundary is the first step in image filtering. If not appropriately handled, it often cause serious artifacts and losing information. In this paper an optimized boundary padding method is proposed to solve this problem. We analyze the pixel values near the boundary, calculate the gradient of the pixels, and pad the input image with a center symmetric padding method. Consequently, the details of the boundary will be preserved, and the number of unexpected noise will be suppressed as much as possible. We demonstrate that the center symmetric padding method is more effective than the traditional border handling methods in some computer vision applications.

1 Introduction

In computer vision, most of applications have to do some preprocessing to reduce noise or get useful information. Image filtering is a general and effective method in image processing such as non-linear filter in [1] and most of linear filter. Image filtering is generally modeled as a convolution of a pixel's neighborhood. The convolution operator makes use of not only the image in the Field of View (FOV) of the given observation, but also part of the scenery in the area bordering it. The missing pixel information outside the observed image can be synthesized by extrapolating the available image data.

In this paper, we propose a boundary padding application when dealing with the flaw detection of the backlight plate in cell phone. The picture of the backlight plate is taken from the line-scan CCD camera, whose resolution is 8192*14000, see in Fig. 3. In this case, padding the boundary which is the first step in image filtering plays an important role in such an extremely high resolution. Border handling provides approximate output pixels close to the boundary of the image where insufficient data is available to fully compute the local operator. There are many available methods to padding the boundary. The "tile" method in [2] proposed an algorithm that pads the image by replicating small tiles of pixels adjacent to the border to reduce boundary artifacts in image deconvolution. Halide*'s [3, 4] library also includes "tile". Tile padding is particularly useful for Bayer coded color images because it preserves the Bayer color pattern. Most of functions in commercial software or open source library are copying a given image onto another slightly larger image and then automatically padding the boundary, such as Reflect, Reflect101 [5], Replicate [6], Wrap, Constant. But these boundary methods not always work well, especially when the blur filter window size is relatively larger with a high resolution of the backlight plate.

In this paper, we proposed a simple but effective method called center symmetric padding method. First of all, we analyze the padded region by calculating the gradient of the pixel values in 2 dimension. If the gradient is too sharp for traditional border handling method, we pad the boundary with our optimized center symmetric padding method. In case of causing unexpected noise, we blur the padded region. Hence, the information of the boundary will be kept, and the number of unexpected noise will be decreased as much as possible.

2 The Proposed Method

We define the center symmetric method as below: First of all, pad the boundary of the objective picture using the default padding method which is border reflect101 in OpenCV to make border in a given filter window. Then extract the pixel values and coordinate of each row or column near the boundary into a 2 dimensional plane, calculate the mean gradient of these discrete points. If the gradient of the line near the boundary is over than the threshold, using the symmetry method to process the padded boundary. Finally, we apply a local linear model to blur the padded boundary in case of sharp change pixels. So, it is necessary to collect some pixels next to the boundary to predict the gradient of the pixel values in a row, using a minimum mean-square error (MMSE) method.

Assume \hat{Y}_i is a vector of n predictions of the padded boundary, and Y_i is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the MSE of the predictor can be estimated by (1)

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \tag{1}$$

In this paper, the pixel values near the boundary in a row or column can be considered as a line in 2 dimension. Generally speaking, we can predict the pixel values outside the input image and padding the boundary, using the linear regression. For example, we choose to pick some pixels near the boundary in a row. We suppose the gradient of the fitted line which can be calculated as K in (2), and suppose the function of the fitted line is $f(x) = Kx + b$.

$$K = \frac{\sum_{i=1}^N (X_i - \frac{\sum X_i}{n})(Y_i - \frac{\sum Y_i}{n})}{\sum_{i=1}^N (X_i - \frac{\sum X_i}{n})^2} \tag{2}$$

After getting the parameter of the fitted line in linear regression, we can padding the boundary using the fitted line. Suppose we choose the 101*101 operator to filter the input image, the 50 pixels outside the boundary should be padded according to the rule of the center symmetric method: First, we can handle the boundary by reflect101, which can be used in OpenCV or MATLAB; then, reflect the padded boundary by a horizontal

line $f(x) = b$. Finally, we apply a local linear model to blur the padded boundary. We assume that $f(x, y)$ is the pixel values of the input boundary, $h(x, y)$ is the local operator which can be called kernel of the blurring, and $g(x, y)$ is the result pixel values in (3):

$$g(x, y) = \sum_{k,l} f(i + k, j + l)h(k, l) \tag{3}$$

In this paper, we take the horizontal padding as an example and illustrate it in detail. Figure 1 shows an example of a row pixel values in horizontal dimension. The region where the pixel values larger than 80 is the foreground, which is our object region. We can see that the pixel values decrease slowly around the boundary.

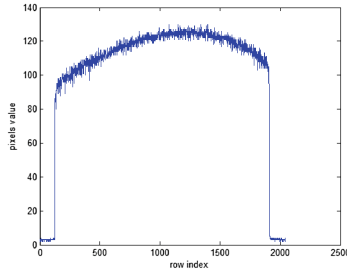


Fig. 1. Illustration of the pixel values distribution in horizontal dimension. The pixel values of the foreground has a sharp change adjacent to the boundary.

Our task is to detect flaws on the digital image of an illuminated screen, which is taken from a CCD camera. First of all, we need to find out the size and location of object region in the input image and crop it from the source image. After getting the object region, we start to do some image processing to locate the flaws of the backlight. According to the statistical results of the backlight, there are many kinds of flaws such as black or white dot, scratch, foreign objects, small or big shadow. In all of these flaws, big shadow is the most difficult problem on account of its ambiguous edges. In order to find out the big shadow flaw whose area is generally more than 20*20 pixels, the window of image filter operator must be set larger enough. Before the filtering operate, the appropriate padding method for image boundary handling is needed. In Table 1, it shows some typical methods for image padding on the boundary. As it mentioned in

Table 1. Image padding methods

	Border Types
Constant	iiiiilabcdefg hliiiii
Replicate	aaaaalabcdefg hhhhhh
Reflect	fedcbalabcdefg hlgfedc
Wrap	cdefghlabcdefg h l a b c d e f
Reflect101	gfedcblabcdefg hlgfedcb

introduction, different kinds of padding methods show different characters and lead to different results.

In order to visualize the differences of these kinds of padding and find out an appropriate padding method, we pad the source image and plot the pixel values of it on MATLAB. So the padded image results are in Fig. 2.

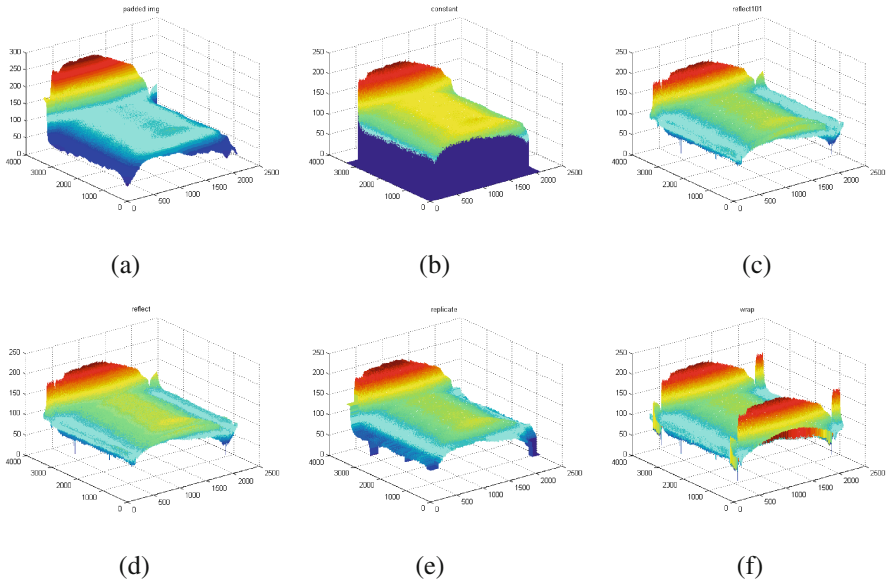


Fig. 2. Padded images in MATLAB with different border methods: (a) image deal with the proposed center symmetric method; (b) constant padded image; (c) reflect101 padded image; (d) reflect padded image; (e) replicate padded image; (f) wrap padded image

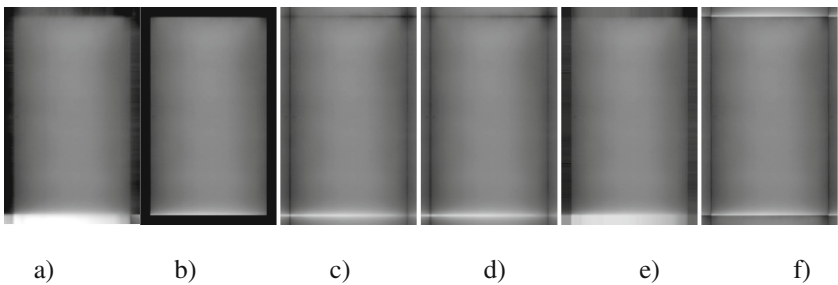


Fig. 3. (a) center symmetric padded image; (b) constant padded image; (c) reflect101 padded image; (d) reflect padded image; (e) replicate padded image; (f) wrap padded image

As the Fig. 2(a) and (e) shows, replicate border handling copies the boundary pixel values of an image or specified value and extends the values to the padded regions. From Table 1 and Fig. 2(b), (c), and (d), there are abrupt gradient changes around the boundary. Because of the abrupt gradient changes, the pixel values of the boundary after filtering

is much larger than the source pixel values. In this case, the boundary after padding becomes an inflection point in two dimension. Therefore, the boundary of the image are predicted as flaw by mistake, and the information of the flaw near the boundary are missing as well. When handling boundary artifacts, there are many algorithms are available such as fast image restoration without boundary artifacts [7, 8] and reducing boundary artifacts method [2]. But a significant majority of them are aimed at reducing the boundary artifacts when image filtering rather than the image padding.

As it shows above, we can draw a conclusion that most of the boundary padding methods can't solve the boundary problem efficiently. In addition, we find out a solution that can padding the boundary smoothly and preserve the boundary information, at the same time, without sacrificing the flaw information near the boundary.

3 Simulation

According to the algorithm presented in previous section, experiments are applied to verify the proposed padding method. Experiment platform is based on MATLAB and Visual Studio. In order to visualize the differences of these padding methods, we padded the border of the backlight plate which is cropped from the row image and applied canny edge detection. The subjective evaluation of the proposed method shows that it's effective comparing to traditional methods. In addition, objective evaluation of the effect of center symmetric padding method proves the algorithm validation.

As a concrete example of this center symmetric method, we try to solve the problem about a backlight plate flaw detection. Many approaches of padding has been taken to operate the source image about backlight plate, while the output boundary present different results respectively. After padding the boundary, it's time to extract the feature of the object, the flaws of the backlight plate. Edge correspond to abrupt changes or discontinuities in certain image properties between neighboring areas [9]. Canny edge detection in image processing is very sensitive to noise as well as the flaws in backlight plate. In this case, canny edge detection is an appropriate way to observe the differences of padding methods in the backlight plate flaw detection. We use canny edge detection on the padded image.

As the pictures show above, we can draw a conclusion that when the filter window size is too large for the traditional border methods to handle, and the gradient of the pixel values decreases around the boundary, our proposed optimized padding methods can fix the problem to some extent. In order to compare these method more intuitively, we apply a canny edge detection to evaluate these methods. After getting the canny results of these methods, we calculate the number of nonzero pixels as the statistical result. As shown in Table 2, our optimized padding method cause less artifacts comparing to others.

Table 2. Comparison on the number of Nonzero pixels in different canny images after padding.

	Center symmetric	Replicate	Reflect	Reflect101	Wrap	Constant
Nonzero	2190	9690	5076	5010	4066	3795

In order to compare the results of these padding images respectively, we choose the part which has distinct difference from the backlight plate image. In Fig. 4, there are the bottom left side of the backlight plate image after canny edge detection. In Fig. 4(b) and (f), the image applied constant and wrap border method remains an obvious line between the border and the boundary of the backlight plate owing to the sharp gradient. In Fig. 4(e), the replicate method makes the pixel values difference of each row or column more significant. From Fig. 4(a), (c) and (d), the boundary of center symmetric padding method is not so obviously than the reflect and reflect101, owing to the more gentle gradient on the boundary.

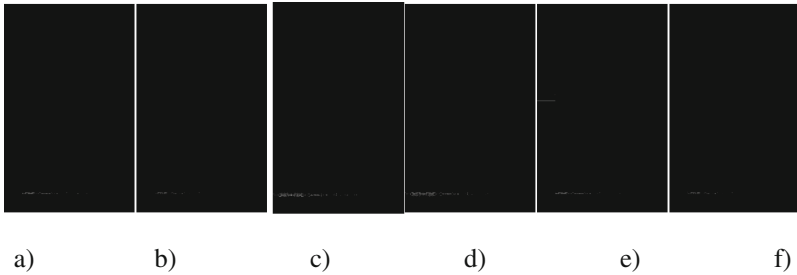


Fig. 4. The bottom left part of padded image after canny edge detection with different border methods: (a) center symmetric image; (b) constant padded image; (c) reflect101 padded image; (d) reflect padded image; (e) replicate padded image; (f) wrap padded image

Furthermore, in order to evaluate the effect of the proposed center symmetric, the PSNR (Peak Signal Noise Ratio) and MSE (Mean Square Error) are used to be objective evaluation parameters. PSNR is most commonly used to measure the quality of image reconstruction. A higher PSNR and lower MSE generally indicates that the reconstruction is of higher quality. Table 3 present the comparison on PSNR and MSE of center symmetric padded image and traditional padding methods respectively.

Table 3. Comparison on PSNR and MES among mentioned padded image

	Center symmetric	Replicate	Reflect	Reflect101	Wrap	Constant
PSNR	43.3643	43.0085	42.5242	42.5056	40.6914	33.6513
MSE	2.9968	3.2526	3.6363	3.6519	5.5455	28.0509

As the result shows above, the optimized center symmetric padding method has the biggest PSNR value, and the smallest MSE value in all the mentioned traditional methods. Therefore, we suspect our optimized algorithm, center symmetric padding method, is very effective. At the same time, it can preserve the edge detail information while not cause boundary artifacts.

4 Conclusion

In this paper, we present an optimized padding method which can be widely applied in image processing of computer vision. Different from the traditional border padding methods, we find out a center symmetric method which has a more gentle gradient on the boundary of the picture. And it can handle the abrupt changes near the boundary more smoothly. Since the center symmetric method take the pixels around the padded border into account and analyze the gradient of the border, the padded image might preserve more information of the image boundary.

The proposed center symmetric method shares a common limitation of the other border handling methods which is not directly applicable for all the raw images. However, we believe that the simplicity and effect of the border padding method still make it beneficial in many situations.

Acknowledgements. This research is financially supported by National Natural Science Foundation of China (No. 61203184).

References

1. He, K., Sun, J., Tang, X.: Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1397–1409 (2013)
2. Liu, R., Jia, J.: Reducing boundary artifacts in image deconvolution. In: 15th IEEE International Conference on Image Processing, San Diego Convention Center, San Diego, CA, USA (2008)
3. Leonard, G., Hamey, C.: A functional approach to border handling in image processing. In: *Digital Image Computing: Techniques and Applications (DICTA)*, Adelaide, Australia (2015)
4. Ragan-Kelley, J., Barnes, C., Adams, A., Paris, S., Durand, F., Amarasinghe, S.: Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. In: *Proceedings of the 2013 ACM Sigplan Conference on Programming Language Design and Implementation*, Seattle, WA, USA (2013)
5. Li, T., Zhang, X., Li, C.: An improved adaptive image filter for edge and detail information preservation. In: *International Conference on Systems and Informatics*, Yantai, China (2012)
6. Chen, W., Gu, C., Lee, M.C.: Repetitive and morphological padding for object-based video coding. In: *International Conference on Image Processing*, vol. 1, Santa Barbara, CA, USA (1997)
7. Reeves, S.J.: Fast image restoration without boundary artifacts. *IEEE Trans. Image Process.* **14**, 1448–1453 (2005)
8. Ng, P.-E., Ma, K.-K.: A switching median filter with boundary discriminative noise detection for extremely corrupted images. *IEEE Trans. Image Process.* **15**, 1506–1516 (2006)
9. Li, S.Z.: *Markov Random Field Modeling in Image Analysis*. Springer, London (2001)

Implementing a Stereo Image Processing for Medical 3D Microscopes with Wireless HMD

Cheolhwan Kim^{1(✉)}, Jiyoung Yoon², Yun-Jung Lee¹, Shihyun Ahn¹,
and Yongtaek Park¹

¹ Kyungpook National University, Daehakro 80, Bukgu, Daegu 41566, Korea
{kch1, knu35man, ytpark97}@knu.ac.kr, yjlee@ee.knu.ac.kr

² Daegu Gyeongbuk Medical Innovation Foundation,
Cheombokro 80, Donggu, Daegu 701-310, Korea
wngus235@gmail.com

Abstract. Recently 3D scenes are used in medical fields. We developed a medical 3D microscope with a HMD. We mounted two cameras on the 3D microscope. Incoming images through an object lens of an optical microscope are projected on sensors of mounted cameras by using reflector mirrors. These are processed in a computer and sent to HMD by wireless communications. However, there is the problem about 3D imaging display. Two camera images are easily distorted due to the defect in the 3D microscope and the color of the images are also a little changed because of the color characteristics of the cameras. In this paper, we suggested the method to correct the geometric and color distortion of the stereo camera image and real time implementation of these methods using GPU. We used a wireless HMD for ease of use by doctors. This system would be helpful for doctors to operate more comfortably.

1 Introduction

Recently 3D scenes are used in various industrial fields. 3D optical microscope [1] is being developed to see much precise detail in 3D. The 3D microscopy is widely used in industrial, medical and biological studies. In the industrial field, this system is used to identify errors of the electronic circuits or detects very small mechanical components. In the medical areas, it is used for actual medical procedures such as polyp removal. It can be used in hair transplant, ophthalmology and biology identifying such cells, chromosomes in research. A microscope with augmented reality technology is being researched to provide a guide to the surgical region and to show the necessary information [2] (Fig. 1).



Fig. 1. Various use cases of imaging devices in medical procedures

1.1 3D Microscope with HMD

Generally, doctors operated using a 3D microscope to see magnified images. In this case, a doctor's eyes should continuously contact with a microscope lens during operation. This is very unnatural and caused neck pain to doctors and other doctors could not see the surgery site. To solve this problem, camera and monitor system was added to a microscope. But operating with monitor is very difficult for doctors to concentrate on the operation. For these reasons, we developed an HMD 3D microscope system in which an HMD was used instead of a monitor. By using HMD, doctors could get their comfortable neck moving and focus on surgery for a long time with less fatigue. Our system consists of three parts, microscope with stereo cameras, image processing computer and HMD. In the microscope, two cameras are installed and the images are reflected at reflector and projected to the camera. The camera images sent to an image processing computer are reconstructed to side-by-side 3D image. Finally, this 3D image is sent to HMD.

1.2 Issues

When developing our system, we were faced with three problems. These are stereo image alignment error problem, color distortion problem and video output performance issue. Stereo image alignment error problem of our optical microscope is that the generated stereo images are not aligned and geometrically transformed. This problem occurred in the cause of the inaccurate reflection mirror alignment, which makes the microscope image could not be projected on the center of the camera's sensor. And, it makes users can have dizziness and nausea. Color distortion problem is that the color of the output image is not same with the original color of the object. This problem occurred in that the cameras could not generate exact color of the original scene. If display devices of microscope show incorrect colors, then users would feel very uncomfortable and sometimes cannot recognize medical treatment position. Video output

performance issue is that the display delay is not short enough that users cannot recognize it and the display frame rate is not fast enough that user cannot feel uncomfortable to work with the display device.

1.3 Our Solutions

In this paper, to solve above issues, we proposed the methods to correct the geometric and color distortion of the stereo camera image and real-time implementation of these methods using GPU. We used a wireless HMD for doctors use convenience.

2 Stereo Image Processing for 3D Microscope

In order to use 3D images for medical procedures, the following three problems must be solved. These are the stereo image alignment error, color distortion and high performance video output. We solved these problems by stereo image correction, colorimetric matching and real-time implementation with GPU.

2.1 Stereo Image Correction

To correct the geometric error of the image through the optical microscope, we used SURF(Speeded Up Robust Features) algorithm [3, 4]. SURF is a robust local feature detector that can be used in computer vision tasks like object recognition or 3D reconstruction. SURF is a detector and a high-performance descriptor for points of interest in an image where the image is transformed into coordinates, using a technique called multi-resolution. We used the SURF algorithm to extract key points and descriptors on left and right image. We could detect the almost same key points and descriptors on the left and right image even if the images were transformed. Then we got a homography [5] by key points and descriptors and used the homography to correct stereo image error. To correct stereo images for 3D optical microscope, the images are processed in the following order. First, shot the checker image by left and right camera. Second, extract key points and descriptors on left and right image using SURF algorithm. Third, transform left and right image to align center and perpendicular using key points. Fourth, clip missed image area by transforming the images. Finally, fit stereo images on a screen. Image correction is executed only once using first left and right image frame. Because the deformation mechanism of the camera connector and mirror part on the optical microscope does not change frequently. This method made a good stereo images. However, this method had two drawbacks. First, it occurs the loss of images on the clipping step. The loss of image due to clipping depends on a transformation of image using homography. But this is a very small area and we could get a sufficient image region. Second, the matrix operation to transform left and right images should be carried out at every frame. This caused an increase of computation and affected video output performance. We solved this issue using GPU processing. Figure 2 shows, correcting stereo images using checker.

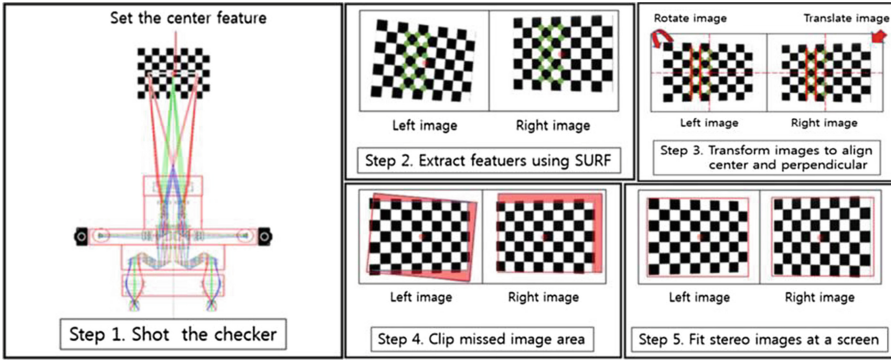


Fig. 2. Correcting stereo images using checker

2.2 Color Correction

Colorimetric matching for cameras is the relationship modeling of digital RGB values from cameras and XYZ values from color meter device, which is device independent color space. It is a characterization method for cameras [6, 7]. There is used the ColorChecker for the characterization of cameras. The ColorChecker is a color calibration target consisting of a cardboard-framed arrangement of many squares of painted samples. Camera characterization uses XYZ values of the ColorChecker from a color meter and RGB values from cameras. It makes the estimated coefficient through polynomial regression [8] using relationship modeling of RGB and XYZ values. We used the ColorChecker with 140 colors for Colorimetric matching. Additionally, we used

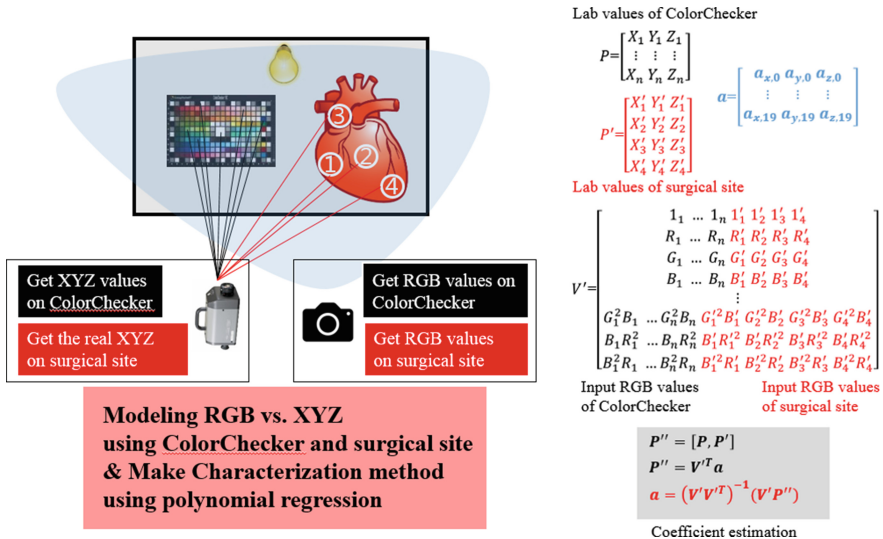


Fig. 3. Coefficient estimation method using ColorChecker and surgical area color values

color samples that we got it on the raw meat. Because doctors require an accurate representation of the red sequence of the surgical scene. After we got the coefficient of the camera characteristic, we made a lookup table for all RGB colors ($256 \times 256 \times 256$) and used this table for color correction every frame. Because the calculation of color correction for every pixel at every image frame significantly affects the video output performance. Figure 3 is the coefficient estimation method using ColorChecker and surgical area color values.

2.3 Real-Time Implementation with GPU

To display side-by-side 3D images at high frame rate, we used a high-performance graphics card with OpenGL. OpenGL is the most widely adopted 2D and 3D graphics API in the industry, bringing thousands of applications to a wide variety of computer platforms. Using the shader language of OpenGL, we could display high resolution stereo image with stereo image correction and color correction at high frame rate. Stereo image was displayed by two textures which were got by two cameras. To correct stereo image error, we transformed the texture location of the stereo image. This operation should be carried out at every frame, but GPU operations do not affect video output performance. To accelerate the colorimetric matching on GPU, we made the color lookup table on 2.2 sections to 3D texture. This 3D texture is used for color reference on shader language.

3 3D Surgical Microscope with Wireless HMD

We used a wireless HMD for doctors use convenience. If the HMD is wired to the system, this cable may interfere with doctor's operation. Therefore, it is necessary to transmit images through wireless communication. We used a sony HMZ-T3W for testing. This model uses WirelessHD 1.1 [9] specification. WirelessHD is a wireless high definition digital interface technology operating in the unlicensed and globally available 60 GHz frequency band and represents the first consumer application of 60 GHz technology. It supports data transmission rates at 10–28 Gbps, more than 20× faster than the highest 802.11n data rates. In the case of WirelessHD standard, high resolution image could be transmitted at a high speed and at a delay rate of less than 5 ms using a broadband of 60 GHz. However, when the wireless communication is interfered, the image is delayed or stopped. This can be a fatal problem for doctors' surgery. We tested the image transmission delay rate. As the result of the experiment, the transmission delay is about 2 ms, which does not affect the operation in real time. In order to use WirelessHD in medical treatment, it is necessary to ensure that no image delay or stop occurs. So it needs additional specification for safety critical. In the future, if a new WirelessHD standard for safety critical is established, it will be possible to commercialize HMD for medical fields.

4 Implementation and Result

4.1 System Configuration and Operation Explanation

The system consisted with a 3D optical microscope, two cameras, a wireless HMD, a monitor and a PC with a high performance graphic card. 3D optical microscope has a turret-type main tube and features of continuous zooming and auto focus. The camera is a CoaXPress type vision camera and operates at 60 Hz with 720p resolution. For the HMD, we tested two type. One is a wired version and the other is a wireless version. With the wired HMD, we used 2560×720 3D image ($2 \times 1280 \times 720$) and with the wireless HMD, we used 1280×720 3D image ($2 \times 640 \times 720$). Because the wireless HMD did not support 2560×720 resolution. The monitor is a normal 2D monitor for other doctors and nurses in the operating room.

The operation is as follows. At first, left and right images from two cameras are converted to side-by-side 3D image and it is sent to monitor and HMD. At the HMD a full image is displayed in 3D, and on the monitor side-by-side image is displayed in 2D with some GUI (Fig. 4). If a user clicks “Image Correction” button on the menu, with the images at that instance the homography calculation is executed and after that, geometrically corrected images are displayed at every frame. And if a user presses



Fig. 4. 3D surgical microscope system with wireless HMD

“Color Correction” button, after that time, the color of all pixels are converted using the color lookup table at every frame. Besides them, there are some functions of pause/play, still image saving, movie recording and so on.

4.2 Performance Testing and Result

We tested the performance of the geometric error correction, color distortion correction and the overall frame rate to display stereo image. To measure the performance of the geometric error correction, we measured three items, vertical error, rotational error, zoom size error. Table 1 is the result of geometric error correction performance.

Table 1. Geometric error correction performance

Vertical error	Rotational error	Zoom size error
0%	1.1°	2.0%

To measure the performance of the color correction, we measured CIELAB color distance. At CIELAB color space, the Euclidian color distance is proportional to the color difference human feels. The average CIELAB color difference between measured color and corrected color for 140 patches was 4.35.

To measure the frame rate of the system, we added frame count code at the display code. And the measurement result was 60fps, which is same with the camera frame rate. It means that by GPU implementation, we finished all image processing tasks within 16 ms and there is no delay from image processing task. Table 2 is the result of performance test.

Table 2. Result of performance test

Category	CPU	GPU
Stereo image display	32 fps	60 fps
Stereo image correction	24 fps	60 fps
Color correction	30 fps	60 fps

Figure 4 shows the 3D surgical microscope system with wireless HMD.

5 Conclusion

In the past, doctors used optical microscopes that show just a single image. Recently 3D scenes are used in medical field. In order to increase the precision of the surgery, doctors are trying to use 3D images for surgery. We developed 3D surgical microscope system with wireless HMD. In development, there were three problems, stereo image geometric error, color distortion and video output performance issue. We solved these problems by geometric correction by SURF algorithm and homography, and colorimetric matching and GPU implementation. Therefore, we developed wireless HMD 3D microscope system which is accurately aligned stereo image with realistic color and very fast

video output performance. We could watch 3D microscope scenes with 60fps on the wireless environment. However, wireless communication is necessary to ensure that no image delay or stop occurs in medical fields. So it needs additional specification for safety critical. In the future, if a new WirelessHD standard for safety critical is established, it will be possible to commercialize HMD for medical fields and many doctors will be able to make convenient surgery.

Acknowledgments. This research was financially supported by the Ministry of Trade, Industry & Energy (MOTIE), Korea Institute for Advancement of Technology (KIAT) and Gangwon Institute for Regional Program Evaluation (GWIRPE) through the Economic and Regional Cooperation.

This research was financially supported by the Ministry of Science, ICT and Future Planning (MSIP) and the Institute for Information & communication Technology Promotion (IITP).

References

1. Wikipedia contributors: Optical microscope, Wikipedia, The Free Encyclopedia. Wikipedia, Web (2015)
2. King, A.P., Edwards, P.J., Maurer, C.R., et al.: Stereo augmented reality in the surgical microscope. *Presence* **9**(4), 360–368 (2000)
3. Jiyoung, Y., HyunDeok, K., Cheolhwan, K.: Development of image error correction system for 3D optical microscope. In: *Computing Technology and Information Management*, pp. 105–108 (2015)
4. Bay, H., Tuytelaars, T., Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). doi:[10.1007/11744023_32](https://doi.org/10.1007/11744023_32)
5. Wikipedia contributors: Homography_(computer_vision), Wikipedia, The Free Encyclopedia. Wikipedia, Web (2016)
6. Hong, G., Luo, M.R., Ronnier, P.A.: A study of digital camera colorimetric characterization based on polynomial modeling. *Color Res. Appl.* **26**(1), 76–84 (2001)
7. Ilie, A., Welch, G.: Ensuring color consistency across multiple cameras. In: *IEEE International Conference on Computer Vision*, pp. 1268–1275 (2005)
8. Pointer, M.R., Attridge, G.G., Jacobson, R.E., Jacobson, R.E.: Practical camera characterization for colour measurement. *Imaging Sci. J.* **49**(2), 63–80 (2001)
9. World Wide Web: WirelessHD specification (2016). <http://www.wirelesshd.org/about/specification-summary>

Design of OpenGL SC 2.0 Rendering Pipeline

Nakhoon Baek^{1,2,3}✉

¹ School of Computer Science and Engineering, Kyungpook National University,
Daegu 41566, Republic of Korea
oceancru@gmail.com

² Software Technology Research Center, Kyungpook National University, Daegu 41566,
Republic of Korea

³ dassomey.com Inc., Daegu, Republic of Korea

Abstract. OpenGL SC 2.0 is a newly specified 3D graphics standard, derived from the OpenGL ES 2.0, as its safety-critical profile. In this paper, we represent the high-level design scheme of the OpenGL SC 2.0 context and rendering system. We also show the detailed implementation strategy, for its step-by-step implementation works. These implementation schemes are the fundamental and theoretical frameworks for the OpenGL SC 2.0 system implementation. In near future, we will implement all the OpenGL SC 2.0 API functions and its theoretical background, from the OpenGL SC 2.0 system implementations.

Keywords: OpenGL SC · Safety Critical Profile · Rendering pipeline · Design

1 Introduction

In the year of 2015, the Khronos Group, the fundamental standard management body of the famous OpenGL family, established a safety-critical working group. This “Safety Critical working group” is developing open graphics and compute acceleration standards for markets, including avionics and automotive displays.

As a result, the OpenGL SC 2.0 specification [1] defines a safety critical subset of OpenGL ES 2.0 [2]. Now the safety critical working group is working to adapt more recent Khronos standards including the new generation Vulkan API [3] for high-efficiency graphics and compute. The Safety Critical working group is also developing cross-API guidelines to aid in the development of open technology standards for safety critical systems.

The OpenGL SC 2.0, or equivalently, the Safety Critical Profile for OpenGL ES 2.0 is designed to be deterministic and testable to minimize implementation and safety certification costs, as shown in Fig. 1, while bringing GLSL shader programmability to safety critical graphics for enhanced graphics functionality with increased performance and reduced power.

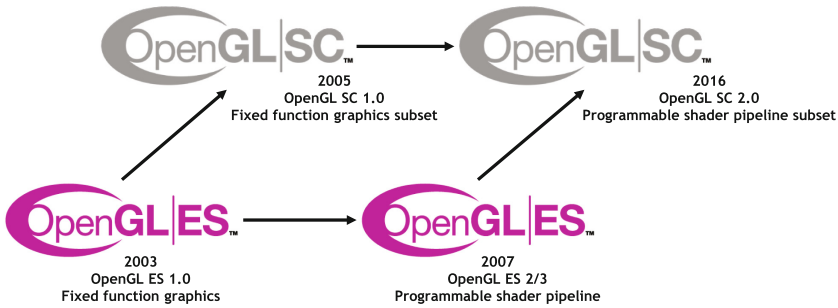


Fig. 1. Standard specification of OpenGL SC 2.0 derived from the OpenGL ES 2.0.

The OpenGL SC 2.0 API addresses the unique and stringent requirements of high reliability display system markets, including FAA DO-178C and EASA ED-12C Level A for avionics, and ISO 26262 safety standards for automotive systems.

Building on the large number of worldwide customer deployments and successful avionics certifications using OpenGL SC 1.0, OpenGL SC 2.0 enables high reliability system manufacturers to take advantage of modern graphics programmable shader engines while still achieving the highest levels of safety certification.

Although the OpenGL SC 2.0 was started from the subset of OpenGL ES 2.0, the details of the OpenGL SC 2.0 are a little bit different from its origin of OpenGL ES 2.0. Our final goal is to implement the whole system of OpenGL SC 2.0. As the first step, we show the high-level design scheme of the OpenGL SC 2.0 context and rendering system. We also show the detailed implementation strategy, for its step-by-step implementation works. These implementation schemes are the fundamental and theoretical frameworks for the OpenGL SC 2.0 system implementation. All the details are followed in the following systems.

2 Design of the Rendering System

Our ultimate goal is to provide all features and API functions in OpenGL SC 2.0. As the first step toward it, we here represent the overall design of our system in the following subsections.

2.1 Overall Design

The topmost design layer of our overall design consists of OpenGL SC 2.0 Context Part and OpenGL SC 2.0 Renderer Part, as shown in Fig. 2. Our focus is how to isolate the device-dependent features into the OpenGL SC 2.0 Renderer Part, as follows:

- **OpenGL SC 2.0 Context Part:** contains all OpenGL SC 2.0 state variables. These state variables are actually the core of OpenGL SC state machine. Upper level user interface of OpenGL SC 2.0 API function calls are also provided in this part.

- **OpenGL SC 2.0 Renderer Part:** contains hardware-specific features, depending on the underlying graphics devices. Our final goal is providing a set of Renderer Parts with a single Context Parts, for easy porting and immigration.

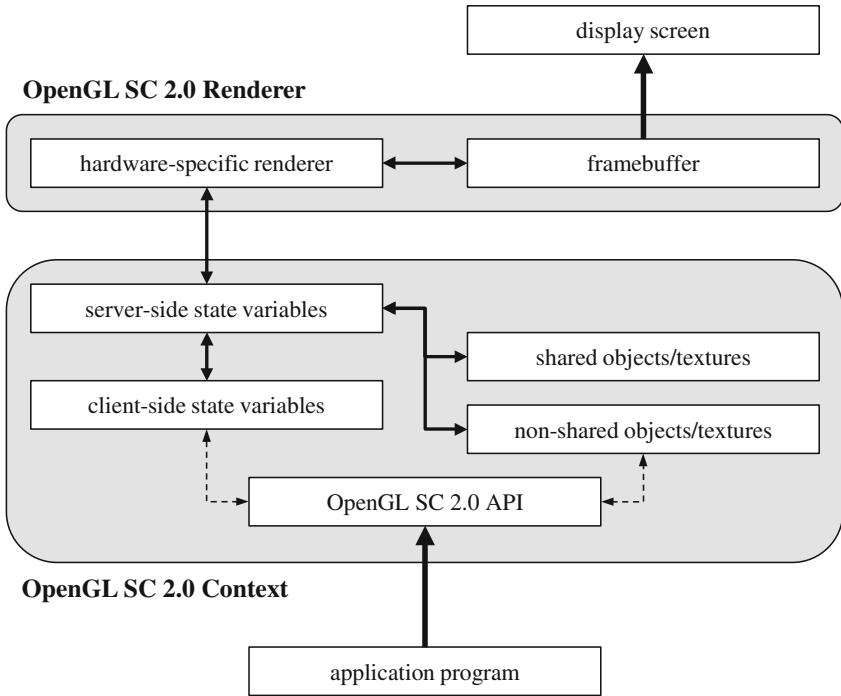


Fig. 2. Overall design of our OpenGL SC 2.0 system.

2.2 Context Part

Our OpenGL SC 2.0 Context Part naturally consists of several modules as follows:

- **OpenGL SC 2.0 API module:** provides the full API function bindings in C programming language, according to the OpenGL SC 2.0 standard specification. Higher-level user application programs will use these API function bindings to access any OpenGL SC 2.0 features.
- **client-side state variables:** contains the client-specific state variables for OpenGL SC 2.0 processing. It includes some vertex, color, normal vector, and texture coordinate arrays for user data.
- **server-side state variables:** contains the OpenGL SC 2.0 server specific state variables. It is the core of any OpenGL SC 2.0 implementation, as the OpenGL state machine.
- **non-shared objects/textures:** Any OpenGL SC implementation has a set of OpenGL objects and texture objects, as user provided graphics data. Most of them are non-shared, and available only for a specific OpenGL application program.

- **shared objects/textures:** In contrast, some OpenGL objects and texture objects can be shared across the applications, and can be shared by a set of OpenGL applications.

2.3 Renderer Part

Our OpenGL SC 2.0 Renderer Part consists of the following two modules:

- **hardware-specific renderer:** According to the underlying graphics hardware, we should implement its specific graphics primitive operations. This module is actually the hardware-dependent part of the whole OpenGL SC 2.0 implementation. Ideally, all the hardware-dependent parts should be isolated into this module, and we can easily immigrate to other hardware devices, through re-implementing this module.
- **framebuffer:** The final output of the OpenGL SC 2.0 implementation is a big two-dimensional array of pixel values, so called framebuffer. It will be directly connected to the graphics display devices, for screen output.

3 Implementation Strategy

Based on the high-level design of OpenGL SC 2.0 shown in the previous section, we have several implementation strategy, as shown in the following sections.

3.1 Implementing the Context Part

As the very first step to the OpenGL SC 2.0 implementation, we need to implement the OpenGL SC 2.0 Context Part. Since it provides all the OpenGL SC 2.0 API functions, any OpenGL application can call the OpenGL SC 2.0 functions and expect to execute their features. However, since this is actually the front-end of the whole OpenGL SC 2.0 system, we cannot get any graphics output from this Context Part. Conceptually, we can test its correctness through the query functions to the state variables.

3.2 Renderer Part: Emulator-like Implementation

The simplest implementation of the OpenGL SC 2.0 Renderer Part may be the OpenGL SC 2.0 Emulator over the existing OpenGL 2.x implementations. Especially, on the desktop PC's, we already have the similar implementation of OpenGL 2.x specification [4]. We can provide the whole OpenGL SC 2.0 features through emulating it with desktop OpenGL 2.x functions.

Since OpenGL SC 2.0 features are based on the subset of OpenGL 2.x features, we can implement this emulator in a straight-forward manner. A few distinct functions need some extra software implementations and it is the major characteristics for the emulator implementation. Anyway, this emulator implementation will show the graphics output on the screen, based on the OpenGL SC 2.0 function calls.

3.3 Renderer Part: GPU Command-Based Implementation

The full-fledged implementation of OpenGL SC 2.0 needs to use any specific GPU (graphics processing unit) as the underlying graphics hardware. We can select any one from commercial OpenGL-ready GPU's. Actually, Intel GPU's are more convenient for the developers, since they provide more sufficient documents on the GPU-based low-level development.

For a specific GPU, we can generate some GPU assembly instructions to control the fine details of the GPU behaviors. In our GPU command-based Renderer implementation, we will generate the proper GPU assembly instructions into the command stream, and the whole command stream will be sent to the GPU. Some GPU buffers and objects will be controlled by extra control commands to the GPU.

We can actually check the feasibility of the OpenGL SC 2.0 Context Part and this GPU command-based OpenGL SC 2.0 Renderer Part implementations, through sending OpenGL API function calls, and testing the output on the GPU framebuffer. This scheme may be used to check the original OpenGL SC Conformance Test Suits, if provided in near future.

3.4 Renderer Part: Vulkan-Based Implementation

Vulkan [3] is a recently-specified graphics and compute API that provides high-efficiency, cross-platform access to modern GPU's used in a wide variety of devices from PCs and consoles to mobile phones and embedded platforms. Since Vulkan aims to provide cross-platform access to commercial GPU's, we can use it as a low-level interface to the GPU's.

Our OpenGL SC 2.0 Renderer Part will generate Vulkan codes to control GPU's, and we get the final output from the GPU's. Vulkan can be a good way of achieving cross-platform features for the OpenGL SC 2.0 implementation. In contrast, for a specific GPU, the native GPU assembly instructions may show better performance than the Vulkan-based implementations.

3.5 Renderer Part: Massively-Parallel Implementation

OpenCL (Open Computing Language) [5] is the open, royalty-free standard for cross-platform, parallel programming of diverse processors found in personal computers, servers, mobile devices and embedded platforms. In these days, most commercial GPU's can also support this massively-parallel processing standard. Thus, OpenCL can also be used to provide cross-platform low-level massively-parallel framework for final rendering stage.

In this case, the renderer part can be executed in a massively-parallel manner. We think that this implementation can unify the graphics pipeline and the computing pipeline, to achieve a single stream. It can show a new unified paradigm for the GPU usage.

3.6 VLSI Chip-Based Implementation

Theoretically, OpenCL implementations can be converted into FPGA chips, through the OpenCL-to-VLSI logic compilers. Thus, if we got the OpenCL-based renderer implementation, we can convert it into physical VLSI chips. It is the fundamental low-level implementation for the OpenGL SC 2.0 Renderer Part.

4 Conclusions and Future Works

OpenGL SC 2.0 is a recently-released 3D graphics standard specification, with safety critical features. Although it is a member of OpenGL family, OpenGL SC 2.0 has some new and characteristic features. To implement this new 3D graphics standard, we present high level design of the overall OpenGL SC 2.0 system.

We also provide details of several implementation strategy. All these guidelines can be used to implement OpenGL SC 2.0-related systems. Our next steps are naturally the implementations with those guidelines. We plan to do them in a step-by-step manner. It will be a full-scale implementation of the OpenGL SC 2.0 system.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant 2016R1D1A3B03935488).

References

1. Fabius, A., Viggers, S.: OpenGL SC, Version 2.0.0 Full Specification (2016)
2. Munshi, A. Leech, J.: OpenGL ES Common Profile Specification, Version 2.0.25 Full Specification (2010)
3. Khronos Group: Vulkan 1.0.35 – A Specification (2016)
4. Segal, M., Akeley, K.: The OpenGL Graphics System: A Specification, Version 2.0 (2004)
5. Bourd, A.: The OpenCL Specification, Version 2.2 (2016)

Saliency Detection via Foreground and Background Seeds

Xiao Lin^{1,3}, Zhixun Yan¹, and Linhua Jiang^{1,2(✉)}

¹ Shanghai Key Lab of Modern Optical Systems,
University of Shanghai for Science and Technology, Shanghai, People's Republic of China
honorsir@yandex.com

² CCSR, Stanford University, 269 Campus Drive, Stanford, CA 94305, USA

³ The College of Information, Mechanical and Electrical Engineering,
Shanghai Normal University, Shanghai, People's Republic of China

Abstract. In this paper, we come up with a bottom-up saliency algorithm that both consider the background and foreground cues. First, we compute the coarse saliency map by manifold ranking on a graph using partly image boundaries which consider as background prior. In this step, we just select left and top sides as background seeds. Second, bi-segment the preliminary saliency map to extract foreground information. Third, we utilize Markov absorption probabilities to highlight objects against the background. Results on public datasets show that our proposed method achieve fabulous performance.

Keywords: Image saliency · Saliency detection · Foreground cues · Markov absorption probabilities

1 Introduction

Visual saliency detection attempt to extract the region which can attract human being the visual and cognitive system. Besides, visual attention mechanism plays a crucial role in the human visual system. It helps us deal with massive visual tasks and assign the perception for most valuable information (or saliency region). Therefore, there are many scholars in the nervous system and computer vision, and other fields do many types of research on the visual saliency. The research of computer vision can be described as given an image, to figure out the region which is significant for human, and quantified by gray. This process is saliency detection, and the gray scale map called saliency map. Visual saliency has been board applied to numerous computer vision tasks, such as image classification [1], object detecting [2], image resizing [3], image retrieval [4] and so on.

According to the perspective of saliency detection mechanisms, saliency algorithm can be summarized as bottom-up [5–9, 16–18, 20, 21, 23–28, 31, 32] (stimuli-driven) and top-down [13, 14, 19] (goal-oriented). A specific task drives top-down model, and its saliency map shows the location of the target. However, bottom-up models are data-driven, without specific task or prior knowledge and utilizing the low-level cues such as intensity, texture and so on to get the final saliency map. Considering top-down model associates with a specific task which is not universal. Therefore, mostly saliency detection algorithms employ the bottom-up model to acquire saliency map.

There are many saliency detection algorithms proposed by scholars. Koch et al. [15] proposed the first visual saliency detection model and given the definition of saliency map. Itti et al. [21] present a center-surround model, in which saliency map is acquiring by using Gaussian pyramid to integrating color, intensity and center-surround contrast at different scales. Also, some previous bottom-up model based on frequency domain analysis [22], transforming the image into the frequency domain and figure out the imparity between the original image with Gaussian smooth version. Moreover, some methods consider the hypergraph model for saliency detection, which obtains a great result in practice. Although most of the bottom-up models have mentioned achieved fabulous results. There are still existing some problem that marginal objects cannot be recognized well.

In this paper, we propose an effective algorithm to solve above problem. Firstly, we acquire a background-based saliency map using a part of boundary labels to decrease the influence of the situation which the object located in the image boundary. Secondly, bi-segment the coarse saliency map to get a series of foreground seeds. Moreover, then, we reconsider the Markov absorption probabilities and the relationship with saliency detection. Figure 1 shows some saliency map generated by our algorithm.

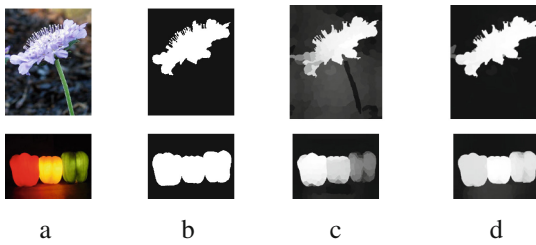


Fig. 1. From left to right: (a) Input image, (b) Ground Truth, (c) coarse saliency map, (d) final saliency map

2 Related Work

Recent years plentiful background-based model that based the background have come forth. Enormous saliency algorithm exploits boundary regions and other areas cues to detect saliency regions. In [5], Yang et al. considered the relationship of periphery region and the rest part of the image via manifold ranking. In [6], Gopalakrishnan et al. introduce hitting time into a fully connected graph and sparsely graph to seek the most salient seed. Besides, there is a difference of the hitting times applies to two kinds of seeds for calculating the saliency of each node. Although, the hitting time method can stress the rare global regions, does not restrain the background noise well. In [8], a more stable and robust method proposed, which exploit the probability of connection between each region and boundary region.

There are also effective approaches using object prior to finding salient region. In [10], Tong et al. employ a center bias model to acquire the week saliency map.

As we know, the model whether is based foreground or background has its limitation that cannot consider global information properly. Therefore, we propose a model to combine the foreground and background to avoid the problem that mentioned above (Fig. 2).

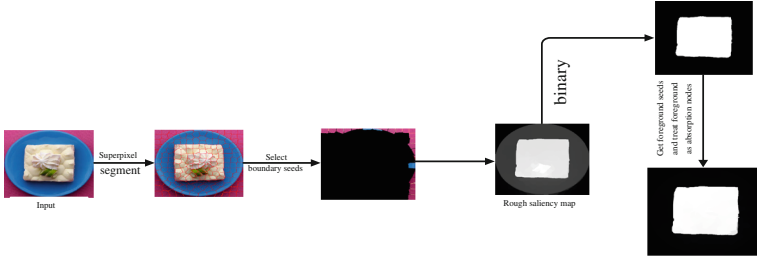


Fig. 2. The step of our algorithm

3 Algorithm

In this part, we propose an effective and efficient saliency detection model which combine foreground and background cues to detect the salient area in an image. Firstly, we construct a graph for each input image, and an optimal manifold ranking is used based on the image's two side background seeds. Then, the coarse saliency map is detected by manifold ranking. Finally, a more precious result is generated by random walk model.

3.1 Graph Construction

In this work, we should construct a graph $G = (V, E)$ where V is a series of nodes and E is a set of undirected edges. If using each pixel as the node is meaningless and will have an expensive calculation, so we adopted superpixels to instead. There are lots of superpixels segment algorithms, but we select SLIC algorithm [11]. It is widely used in the image segmentation, pose estimation, target tracking, target recognition and other computer vision tasks. After over-segmenting input image by SLIC, we get an image with n parts. So, we treat the superpixels as the vertices V rather than a pixel. And E is a series of undirected edges are between two connected superpixels. We construct a new graph model different from [12], which called sparse adjacency graph in this work, and it can better to present the local adjacent superpixels. First, each node concatenates to those nodes which nearby it. Besides, it also connected to the nodes that sharing the same boundaries with its neighboring nodes. Second, as we know there are lots of images present a certain symmetry. So we treat all boundary nodes connect, for reducing the geodesic distance between similar nodes.

We construct an affinity matrix W to represent the correlation between connected edges. If the edge between two nodes with a high weight which means they are tightly connected, vice versa. Then, the weight w_{ij} of two neighbor nodes present as

$$w_{ij} = -e^{-\frac{\|c_i - c_j\|^2}{2\sigma^2}}.$$

where c_i and c_j denote the feature of CIELab color space (because CIELab color space is better to conform to human visual system, so we translate the input image into CIELab space to instead of the origin RGB color space.) corresponding to two nodes i and j , and σ is a constant which controls the strength of the weight.

3.2 Coarse Saliency Detection Using Background Seeds

Manifold ranking applies in pattern classification. It aims to assign ranks to the elements in a dataset indicate their relationship in a certain group. After constructing a graph model for the input image, we also get its affinity matrix W (consist of the weight of edges). Besides, we can get a degree matrix, where $d_i = \sum_j w_{ij}$. A series of background seeds are selected from two sides as indicate vector $y = [y_1, y_2, \dots, y_n]$, where $y_i = 1$, if the background seeds is selected, otherwise $y_i = 0$. Let f be a ranking function to allocating rank values $f = [f_1, \dots, f_n]^T$ to each superpixels, which can acquired by solving the below minimization problem.

$$f^* = \arg \min_f \frac{1}{2} \left(\sum_{i,j} w_{ij} \left\| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right\|^2 + \mu \sum_{i=1}^n \|f_i - y_i\|^2 \right). \quad (1)$$

Then, a optimized solution is given as Eq. (2), where $\alpha = 1/(1 + \mu)$. (μ is a controlling parameter.)

$$f = (D - \alpha W)^{-1} \times y. \quad (2)$$

Finally, we exploit the Eq. (2) to detect the saliency of each superpixel. If those labels are background labels, the saliency region is presented by $1-f$, vice versa.

As we know the most area of images are the background for mostly images, so detect saliency region via boundary prior is quite an effective method [7, 8, 10]. Inspired by this theory, we choose a set of background seeds from both two sides.

3.3 Salient Detection via Foreground Seeds

Although using manifold ranking can estimate the foreground, and can detach it from the background. There still exists some situation that can't tell the foreground from complex background, and couldn't suppress background noise deeply. In this part, our main purpose is to select foreground cues for further saliency detection. After getting the coarse saliency map, we binary it using an adaptive threshold T to obtain foreground seeds. Then we introduce, the Markov chains [29] to purifier saliency map.

Here is some basic introduction of Markov chains. For any Markov absorbing chain corresponds to a transition probability matrix P . Given a set of $S = \{s_1, s_2, \dots, s_l\}$ is an arbitrary Markov absorbing chains and if rearrange all the states, so transient states will be listed before absorbing states. Consequently, we get a normalized form of transfer matrix P :

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}. \tag{3}$$

where I is the $k \times k$ identity matrix, Q is the probability transition matrix of transition states which belong to $[0,1]_{m \times m}$. 0 is a $k \times m$ zero matrix. The elements which belong to $R(R \in [0,1]_{m \times k})$ mean the probabilities between transition and absorbing states.

After got the absorption probability with Markov chain transfer standard form of a matrix P , we can get the basic Matrix N , the elements of the matrix have a special significance. And then we can get another matrix:

$$N = (I - Q)^{-1} = I + Q + Q^2 + \dots \tag{4}$$

Which called the basic form of matrix P . Each element n_{ij} of matrix N is described as the element starts in transient state s_i spend the expected number of times that the process has changed the transient state s_j . The absorption probability matrix:

$$B = N \times R. \tag{5}$$

demonstrate the probability which the process will be absorbed by given absorbing state.

Where N is the fundamental matrix, R has been introduced in (3), the item of B shows the probability that an absorbing chain will be absorbed by the absorbing state s_j .

Firstly, we use an adaptive threshold to segment the saliency map that acquired in section B, which easier to get the foreground. (i.e. the threshold is chosen as the average saliency across whole saliency map.) Then the nodes which separated by adaptive threshold are regarded as the absorbing nodes labeled as $n_{f1}, n_{f2}, \dots, n_{fl}$, and the rest of them are treated as transient nodes.

Similar to R and combined the Eq. (3), we can get an another matrix:

$$R' = \begin{bmatrix} r'_{11} & \dots & r'_{1l} \\ \vdots & \ddots & \vdots \\ r'_{m1} & \dots & r'_{ml} \end{bmatrix}. \tag{6}$$

Besides, we also treat all the boundary nodes as absorbing nodes. According to (5), a new matrix can be Q' constructed (matrix Q' contains the probability of transient nodes after normalized image). Using the formula (4), we can get $N' = I + Q' + (Q')^2 + (Q')^3 + \dots$.

It is obvious that the matrix N' and R' quite similar to matrix N and R , consider that

we multiply N' and R' defined as C' . Suppose that we have a matrix $N' = \begin{bmatrix} n'_{11} & \cdots & n'_{15} \\ \vdots & \ddots & \vdots \\ n'_{51} & \cdots & n'_{55} \end{bmatrix}$

$$\text{and } R' = \begin{bmatrix} r'_{11} & r'_{12} \\ \vdots & \vdots \\ r'_{51} & r'_{55} \end{bmatrix}.$$

The first two rows of matrix N' stand for foreground nodes and the rest rows represent the background nodes and matrix R' has the same structure as N' .

For example, we just consider the background nodes5. Multiplying the fifth row of matrix N' and the second column of matrix R' , and according to the (7), we get

$$c_{51} = n'_{51} \times r'_{12} + n'_{52} \times r'_{22} + \cdots + n'_{55} \times r'_{52}. \quad (7)$$

$n'_{ij} = i'_{ij} + q'_{ij} + (q'_{ij})^2 + \cdots$ and $(q'_{ij})^{(k)}$ is the i,j th element of $(Q')^k$. So we got $n'_{52} = i'_{52} + q'_{52} + (q'_{52})^2 + \cdots$. Because q'_{52} is normalized weight between foreground and background, so it is close to 0. Besides, we construct a graph model which is a sparse model. So the node 5 and node 2 may not connected, in this situation $q'_{52} = 0$. Consider that the value of r' tending to zero, which means c_{51} tend to 0, as well. If chosen foreground node1, $c_{11} = n'_{11} \times r'_{11} + n'_{12} \times r'_{21} + \cdots + n'_{15} \times r'_{51}$.

Obviously, r'_{11} and r'_{12} present the similarity of foreground nodes, their value will larger than (7) and tend to 1. Those observations can be an important approach to refine the coarse saliency maps. The elements of matrix C are considered as the similarity of the foreground. And then rearranging each row of matrix C (descending order), as follow, $c'_{i,1} \geq c'_{i,2} \geq \cdots \geq c'_{i,l}$, $(c'_{i,j} \in \bigcup_{j=1}^l c_{ij})$.

Selecting the first k $1 \leq d \leq 0.5l$ elements of each row in matrix C (in this paper $d = 0.4 * l$) and sum them up. The saliency of each node i is denoted as follow: $S_f(i) = \sum_{j=1}^d c'_{ij}$.

4 Experimental Results

In this part, we evaluate the proposed algorithm on five public available datasets: ASD [17], DUT-OMRON [5], THUS [25]. The first dataset provided by Achanta et al. which contained 1000 images selected from MSRA-5000 dataset, is broadly used in the almost algorithm. The second DUT-OMRON dataset is most challenging dataset, which contained 5168 images. And THUS include 10000 images which consist of multiple targets of different size.

We demonstrate the effectiveness of our algorithm to compare with the most classic or the state-of-art saliency detection models which contained IT98 [21], FT09 [17],

CA10 [16], SVO11 [32], RC11 [20], SF12 [31], GS12 [18], PCA13 [25], LMLC13 [9], HS13 [24], GC13 [27], GMR13 [5], DSR13 [28], MS14 [26], WCO14 [8], LPS15 [7].

4.1 Qualitative Results

We list some results of saliency maps generated by our saliency model and the other recent state-of-art models in Fig. 3. We obtain the saliency maps with highly correct results, and suppress background noises effectively, even though the images have clutter scenes or with multi-object. In a word, the proposed method reaches a quiet well result than previous algorithms.

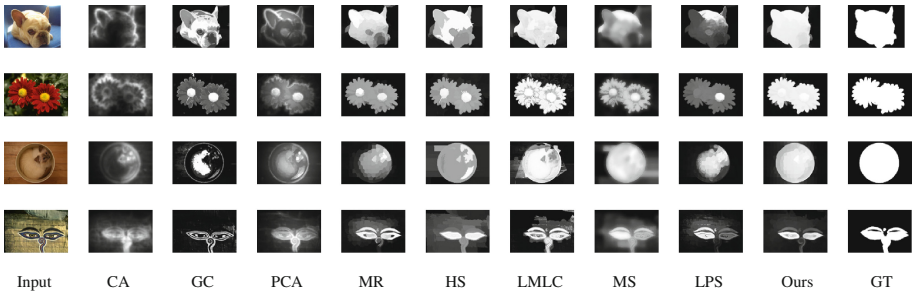


Fig. 3. Example results comparison of our method with nine recent state-of-the-art methods

4.2 Quantitative Results

We adopt the precision-recall curve, mean absolute error (MAE), and F-measure as quantitation rules to evaluate the precision of our method and the other state-of-art algorithm. P-R curve is using a threshold value to segment, which ranging the saliency map from 0 to 255. We compute the precision and recall and draw the P-R curve through comparing the pixel-level ground truth mask. The Mean Absolute Error shows the mean different between saliency and ground truth mask:

$$MAE = \frac{1}{H} \sum_{h=1}^H |S(h) - GT(h)|. \quad (8)$$

Besides, the F-Measure is the comprehensive evaluation metrics to evaluate the quality of saliency map. As the values of Precious and Recall is computed, the value of F-Measure is describing as $F = \frac{(1 + \eta^2) \times Precision \times Recall}{\eta^2 \times Precision + Recall}$ where η^2 is set to 0.3 to weigh precision more than recall as suggested in [17].

Figure 4 shows the result (P-R curve and MAE) and Table 1 present the F-Measure and MAE data on the three datasets. In general, the method that we proposed performs well compare with the previous method.

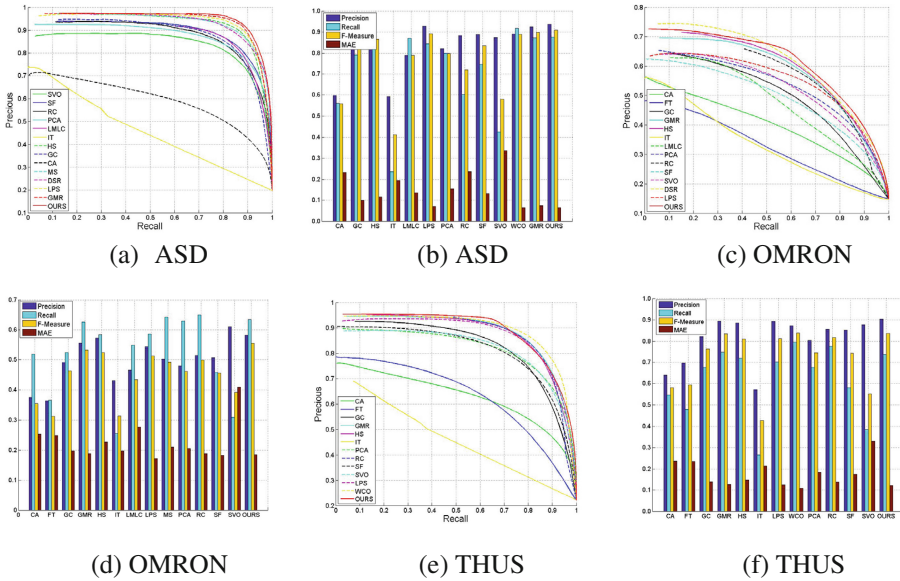


Fig. 4. The result are obtained on different datasets, including ASD, OMRON and THUS. The curves (e.g., (a)) refer to the P-R curves for different methods. The bar graphs (e.g., (b)) shows the precision, recall, F-measure and mean absolute error (MAE) for different methods.

Table 1. The numerical results: F-measure (F) and mean absolute error (MAE)

Metric	Dataset	Method											
		CA	FT	GC	GMR	HS	IT	LPS	PCA	RC	SF	SVO	OURS
F	ASD	0.558	0.640	0.820	0.898	0.865	0.412	0.891	0.796	0.719	0.834	0.579	0.909
	OMRON	0.355	0.313	0.464	0.533	0.525	0.313	0.172	0.461	0.499	0.456	0.392	0.556
	THUS	0.581	0.593	0.762	0.835	0.810	0.426	0.813	0.745	0.817	0.743	0.552	0.836
MAE	ASD	0.232	0.205	0.101	0.074	0.114	0.194	0.070	0.155	0.237	0.130	0.336	0.065
	OMRON	0.253	0.249	0.197	0.188	0.227	0.197	0.172	0.205	0.188	0.183	0.408	0.185
	THUS	0.237	0.234	0.139	0.126	0.148	0.213	0.124	0.185	0.137	0.175	0.331	0.120

5 Conclusion

In this paper, we propose a novel salient detection model via manifold ranking and Markov absorption probabilities, and take both background and foreground cues into consideration. Bi-segment the coarse saliency map from manifold ranking can get more accurate absorption nodes for purifying the final saliency map. Thus, we can handle the difficult image that the large targets connected with border or surrounded with complex scene. A large number of experimental results show that the proposed method performs favorably against almost recent state-of-the-art algorithm on benchmark datasets.

Acknowledgement. The research was partly supported by the program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, USST incubation project (15HJPY-MS02), National Natural Science Foundation of China (No. U1304616, No. 61502220).

References

1. Rabinovich, A., Vedaldi, A., Galleguillos, C., et al.: Objects in context. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE (2007)
2. Rutishauser, U., Walther, D., Koch, C., et al.: Is bottom-up attention useful for object recognition? In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, CVPR 2004, vol. 2, pp. II-37–II-44. IEEE (2004)
3. Fang, Y., Chen, Z., Lin, W., et al.: Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Trans. Image Process.* **21**(9), 3888–3901 (2012)
4. Wang, X.J., Ma, W.Y., Li, X.: Data-driven approach for bridging the cognitive gap in image retrieval. In: 2004 IEEE International Conference on Multimedia and Expo, 2004, ICME 2004, vol. 3, pp. 2231–2234. IEEE (2004)
5. Yang, C., Zhang, L., Lu, H., et al.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
6. Gopalakrishnan, V., Hu, Y., Rajan, D.: Random walks on graphs for salient object detection in images. *IEEE Trans. Image Process.* **19**(12), 3232–3242 (2010)
7. Li, H., Lu, H., Lin, Z., et al.: Inner and inter label propagation: salient object detection in the wild. *IEEE Trans. Image Process.* **24**(10), 3176–3186 (2015)
8. Zhu, W., Liang, S., Wei, Y., et al.: Saliency optimization from robust background detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2814–2821 (2014)
9. Xie, Y., Lu, H., Yang, M.H.: Bayesian saliency via low and mid level cues. *IEEE Trans. Image Process.* **22**(5), 1689–1698 (2013)
10. Tong, N., Lu, H., Ruan, X., et al.: Salient object detection via bootstrap learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1884–1892 (2015)
11. Achanta, R., Shaji, A., Smith, K., et al.: Slic superpixels (2010)
12. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2006)
13. Yang, J., Yang, M.H.: Top-down visual saliency via joint CRF and dictionary learning. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2296–2303. IEEE (2012)
14. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 73–80. IEEE (2010)
15. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of Intelligence, pp. 115–141. Springer, Netherlands (1987)
16. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(10), 1915–1926 (2012)
17. Achanta, R., Hemami, S., Estrada, F., et al.: Frequency-tuned salient region detection. In: CVPR 2009, IEEE Conference on Computer Vision and Pattern Recognition 2009, pp. 1597–1604. IEEE (2009)

18. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 29–42. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-33712-3_3](https://doi.org/10.1007/978-3-642-33712-3_3)
19. Zhang, L., Tong, M.H., Marks, T.K., et al.: SUN: a Bayesian framework for saliency using natural statistics. *J. Vis.* **8**(7), 32 (2008)
20. Cheng, M.M., Mitra, N.J., Huang, X., et al.: Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(3), 569–582 (2015)
21. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
22. Sun, J., Lu, H., Liu, X.: Saliency region detection based on Markov absorption probabilities. *IEEE Trans. Image Process.* **24**(5), 1639–1649 (2015)
23. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
24. Yan, Q., Xu, L., Shi, J., et al.: Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162 (2013)
25. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1139–1146 (2013)
26. Tong, N., Lu, H., Zhang, L., et al.: Saliency detection with multi-scale superpixels. *IEEE Sig. Process. Lett.* **21**(9), 1035–1039 (2014)
27. Cheng, M.M., Warrell, J., Lin, W.Y., et al.: Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1536 (2013)
28. Li, X., Lu, H., Zhang, L., et al.: Saliency detection via dense and sparse reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2976–2983 (2013)
29. Grinstead, C.M., Snell, J.L.: *Introduction to Probability*. American Mathematical Society, London (2012)
30. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
31. Perazzi, F., Krähenbühl, P., Pritch, Y., et al.: Saliency filters: contrast based filtering for salient region detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–740. IEEE (2012)
32. Chang, K.Y., Liu, T.L., Chen, H.T., et al.: Fusing generic objectness and visual saliency for salient object detection. In: 2011 International Conference on Computer Vision, pp. 914–921. IEEE (2011)

Identification and Annotation of Hidden Object in Human Terahertz Image

Guiyang Yue, Zhihao Yu, Cong Liu, Hui Huang, Yiming Zhu,
and Linhua Jiang^(✉)

Shanghai Key Lab of Modern Optical Systems,
University of Shanghai for Science and Technology, Shanghai 200093, China
lhjiang@usst.edu.cn

Abstract. Terahertz (THz) detection technology is a new security technology, it plays a significant role for social public security in the current situation. In this paper, we propose a new fast recognition algorithm for detection of suspicious objects according to the characteristics of human THz images. The algorithm consists of the following steps: (1), Smoothing and using gray stretch algorithm to enhance the terahertz images, (2), Distinguish the suspicious object connected and not connected to the background images. Here, THz images are classified by using our morphological classification algorithm, (3), Extracting a full human body contour by using our Bilateral Contour Tracking Comparison algorithm (BCTC). Finally, the computer can automatically identify and mark the hidden suspicious objects in Terahertz Image. Through a large number of experiments show that the new detection algorithm accuracy is reaching 92%. Our test results show that the new algorithm is quite effective for segmentation and extraction with human body contour and less time-cost.

Keywords: Recognition algorithm · Morphological classification · Contour tracking · BCTC algorithm · Suspicious object

1 Introduction

In recent years, criminals have hidden the knives, explosives through various means, undoubtedly they have brought a great of threat to the personal security in public places. Fast and accurate security technology is essential to protect the people's life and property security. Traditional security equipment, such as X-ray detector is playing a huge role in the security work, but X-ray technology has a strong ionization property will cause damage to the material.

As shown in Fig. 1, the frequency of THz wave is in the range of 0.1 THz to 10 THz. The penetration ability of THz is similar to X-ray, and its photon energy is small, which will not cause harm to human biological tissue [1, 2]. The principle is: The object to be detected by terahertz wave irradiation by using terahertz wave emitter, measurement the THz radiation signal of the objects, collecting relevant physical information carried by the signal (amplitude and phase information), then the two-dimensional distribution of the object is obtained. In 1995, Hu and Nuss established a terahertz imaging devices for the first time [3]. In 1996, real time imaging of terahertz

is realized by Zhang Xicheng et al. [4]. Then in 2003, Zhang Xicheng built a compact, portable continuous wave imaging system [5]. Currently the THz imaging system developed by the research institution is mainly divided into two system: Passive system and Active system. According to whether there is a THz emission source or not. As shown in Fig. 2, a active THz imaging system developed by our research group [6–9].

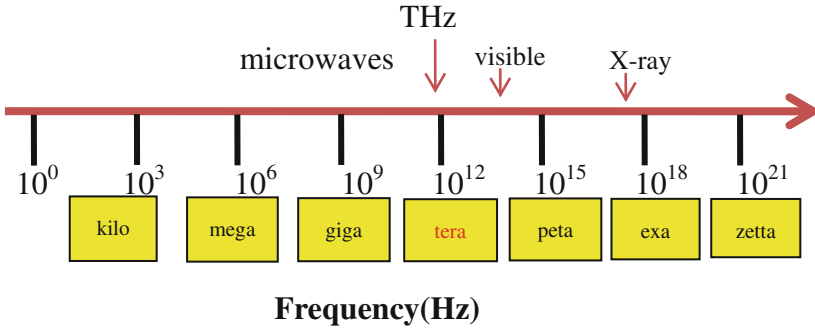


Fig. 1. The range of terahertz wave.

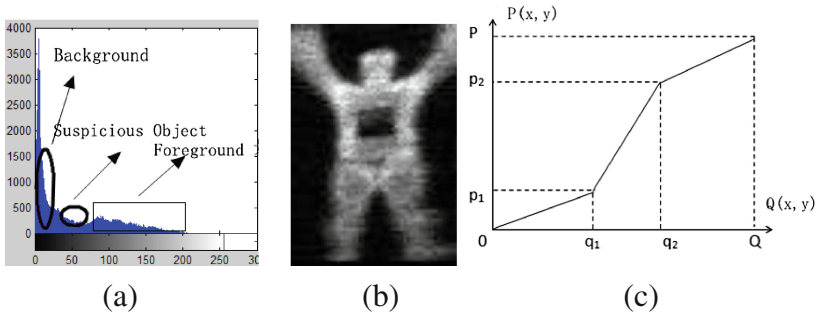


Fig. 2. (a): The histogram of THz image, (b): THz image, (c): Schematic diagram of gray stretch.

However, most of the terahertz imaging system is still on the stage of research and development. Mainly due to following reasons: (1), Measurement speed needs to be improved. (2), The power of the pulse type THz radiation source used by the existing device is generally low. (3) In some THz images, the suspicious object is connected to the background, which brings a great difficulty to the automatic recognition of the computer.

The algorithm mainly adopts the following steps: (1), Firstly, we use Wiener filter to smooth the terahertz image [10], then through gray stretch algorithm and genetic algorithm to enhance and segment images [11, 12]. (2), Secondly, using the morphological classification algorithm proposed in this paper to classify the human THz images. (3), At last, using our BCTC algorithm to extract a full human body contour.

2 Materials and Methods

2.1 Terahertz Imaging Acquisition

Terahertz imaging principle: firstly, we using terahertz emitter emits terahertz wave as an imaging ray. Through measuring objects or THz radiation signals reflected from the surface to obtain imaging information. Mainly including dimensional information of amplitude and phase. After further analysis and processing to obtain a two-dimensional map of object. The principle of our THz imaging system: Focusing element focus the THz wave to a point on the sample. After that focusing element collector collects THz wave that is through the sample or reflected from the sample. THz wave detector collects the THz wave that is reflected from the sample or through the sample. THz wave signal container collects the position information then convert it into a corresponding electrical signal by detection element. At last, the signal is converted to an image information by image processing unit.

Finally, we chose 50 Thz images as our test data.

2.2 Image Preprocessing

Usually, the SNR of the images produced by our THz imaging system is rather low (less than 30 db). Because the object has hindered the human THz emission, so the gray value of object region is different from other areas. According to this, after the image denoising, we enhance image by using gray stretch algorithm.

In (1), q is the gray value of gray image before stretching, and I is the gray value of gray image after stretching, q_1, q_2 is the peak of background radiation and human radiation respectively.

$$I = \begin{cases} \frac{p_1}{q_1}q & q < q_1 \\ \frac{p_2-p_1}{q_2-q_1}(q - q_1) & q_1 \leq q \leq q_2 \\ \frac{255-p_2}{255-q_2}(q - q_2) & q > q_2 \end{cases} \quad (1)$$

Genetic algorithm is an adaptive global optimization algorithm which simulates the genetic and evolutionary process of biological in the natural environment. Here, the length of the chromosome is 8. Population size is set as 10. The crossover probability and mutation probability are 0.7 and 0.4 respectively (Fig. 3).

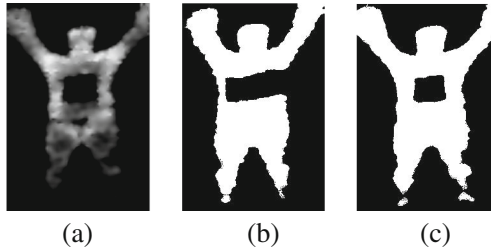


Fig. 3. (a) a gray stretch of THz image, (b) suspicious object is connected to the background after GA segment, (c) suspicious object and backgrounds do not connect after GA segment.

2.3 Classification of THz Image

Our THz image is roughly divided into three categories: (1): suspicious object is connected to the background, (2): suspicious object is not connected to the background, (3): THz images with no suspicious object. As shown in Fig. 4(A₁, B₁, C₁). Therefore, we propose a classification algorithm based on the principle of morphological algorithm.

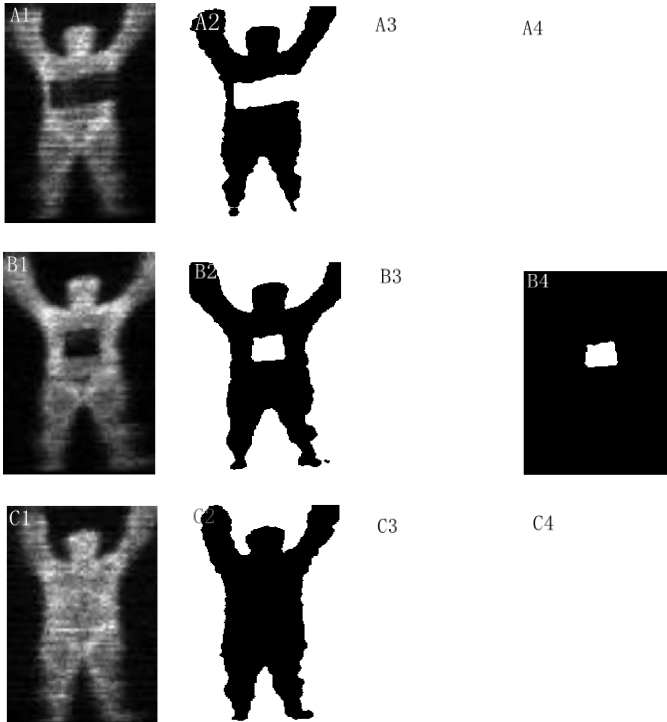


Fig. 4. Morphological classification results of three kinds of THz images. A₁, B₁, C₁: Original THz images. A₂, B₂, C₂: Images of after GA segment. A₃, B₃, C₃: Images of after morphological region filling. A₄, B₄, C₄: Extract closed region of images by using morphological algorithm.

Typically, a simple algorithm for region filling is presented as:

$$X_k = (X_{k-1} \oplus B) \cap A^c, k = 1, 2, 3, \dots \quad (2)$$

B is a structural element, assuming that all the background pixels are marked as 0, if $X_k = X_k - 1$, then the iteration stop, and end the fill step.

Through the comparison of the A₃, A₄, B₃, C₃, B₄ and C₄, when suspicious object is connected to the background or THz images with no suspicious object, such as Fig. 4 (A₁, C₁), the result of morphological region filling is the same as the result of extracting closed region of images by using morphological algorithm. Nevertheless, in the image suspicious object is connected to the background, the result of morphological region filling is different from the result of extracting closed region of images by using morphological algorithm.

2.4 Extract the Body Contour

In this paper, a Bilateral Contour Tracking Comparison algorithm (BCTC) is proposed by us. The algorithm is based on the eight neighbor contour tracking algorithm, through accumulate the length of the contour for bilateral, that is left direction and right direction. Then compare the contour length of two directions to synthesize a complete human body contour (Fig. 5).

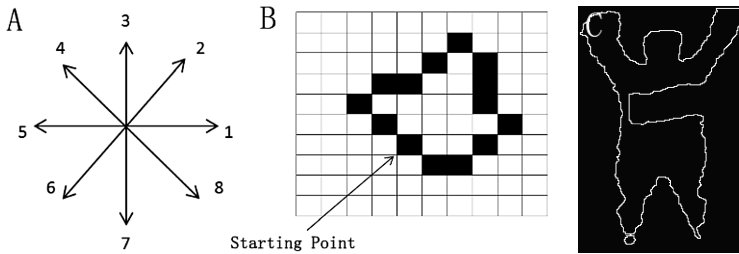


Fig. 5. A: Definition of chain code. B: Contour tracing. C: The contour of the human body.

In order to extract a full contour of the human body, our new algorithm consists of the following steps:

- (1): Using contour tracking algorithm for the THz image, search direction start from the left and right halves of the image, respectively. Then defined obtained contours as follows: left and right contour: L_1, L_2 ;
- (2): Calculate the number of pixels of L_1 and L_2 , at the same time set a threshold value S , the difference between the two contour pixel exceeds the threshold value S ;

$$|L_1 - L_2| > s \tag{3}$$

Then we abnegate the longer contour.

- (3): Then create a mirror image of short contour, in order to get a full contour of the human body (Fig. 6);



Fig. 6. The results of our contour extract algorithm. A: Left profile of human body. B: Right profile of human body. C: Full contour of human body.

According to our algorithm, we can distinguish the image without suspicious objects by calculating L_1 and L_2 and compare them. If the Eq. (3) is a true statement, we can determine the suspicious object is connected to the background in the THz image. On the contrary, the THz image doesn't has suspicious object. Here, we proposed a judgment algorithm according to neural networks [13, 14]. Firstly, we get a set of values of left and right contour profile according to THz images. As shown in Fig. 7, A represents the pixels difference between left and right profile when suspicious object is not connected to the background. B represents the pixels difference between left and right profile when suspicious object is connected to the background. In images of suspicious object is connected to the background, because there is the suspicious contour to the human body contour, a clear differences between left and right profile. The training vector is as follows:

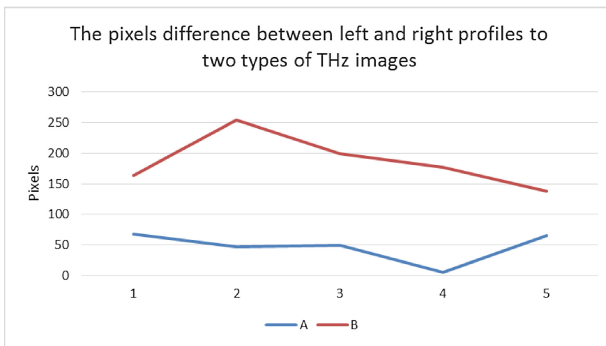


Fig. 7. The pixels difference between left and right profiles to two types of THz images.

$P = [869\ 858\ 851\ 867\ 787\ 823\ 937\ 854\ 861\ 827; 937\ 1022\ 1105\ 1066\ 834\ 1000\ 1075\ 903\ 866\ 892];$

Target vector is as follows,

$T = [-10\ 10\ 10\ 10\ -10\ 10\ 10\ -10\ -10\ -10];$

P is a matrix with 2 rows and 10 columns, the pixels of left and right contour are stored in each column. We assign 10 when suspicious object is connected to the background. On the contrary, we set -10.

Therefore, our algorithm is as follows:

- (1): Input a THz image, then use image denoising for the input image;
- (2): The gray stretch algorithm and GA algorithm is for image segmentation;
- (3): Then use our morphological classification algorithm for classification, so as to distinguish the suspicious object whether is connected to the background;
- (4): Using our proposed BCTC algorithm to extract a complete closed human body contour;
- (5): Annotating the suspicious objects.

Flowchart of our identification and annotation algorithm is as shown in Fig. 8.

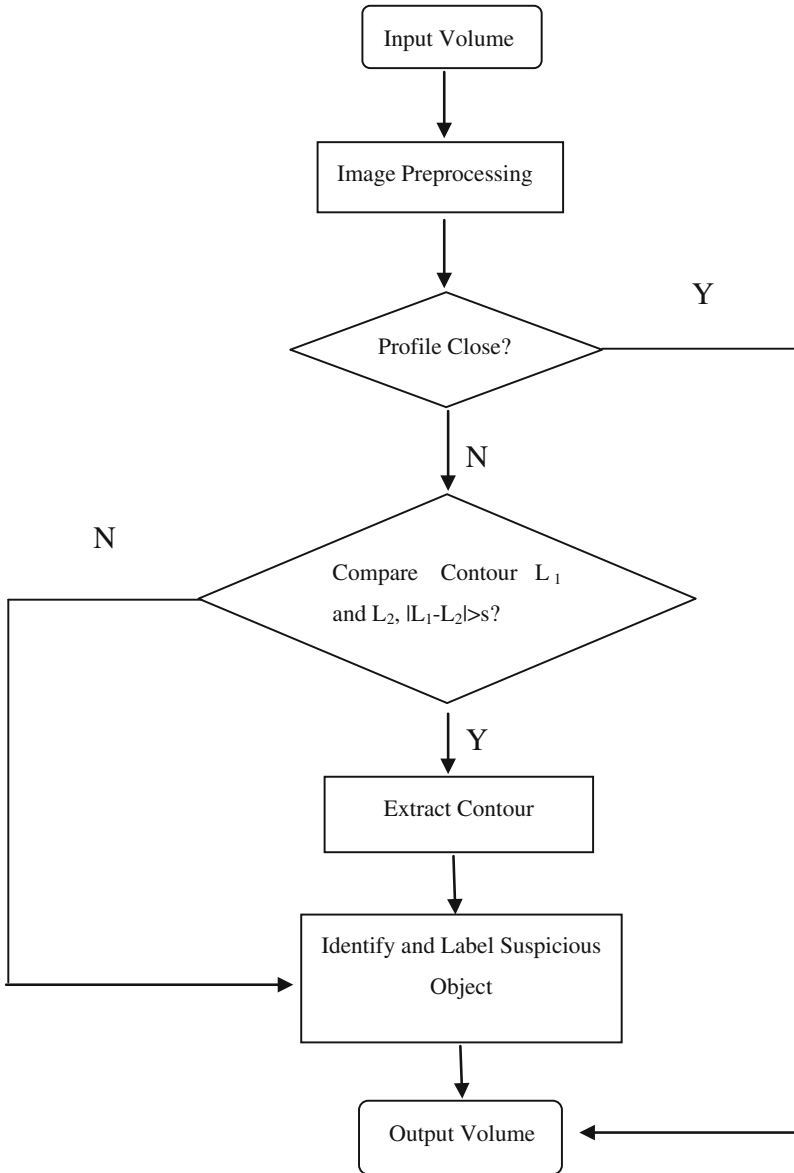


Fig. 8. Flowchart of our algorithm.

3 Results

Algorithm is based on the Windows7 platform with the MatlabR2014a environment, we test 50 images (220*320), there are 46 images are accurately identified, the accuracy rate reached 92%. At the same time, the average time of identification is short,

about 3.7 s, and recognition results are in Fig. 9. The background radiation and the peak value of the human body radiation are not accurate because of the image preprocessing, it affects the image quality after image segmentation, which affects the subsequent recognition process (Table 1).

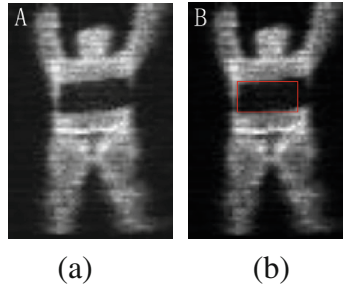


Fig. 9. (a): The original image, (b) Annotation of hidden object.

Table 1. Comparison of several security apparatus.

Detecting instrument	SONO-8065 X-ray	L3 system(USA)	Our instrument
Detecting time	>5 s	3 s	3.7 s
Frequency	–	35 GHz	94 GHz
Recognition method	Human recognition	Computer recognition	Computer recognition

The active millimeter wave scanning security system developed by the L3 communications company in the United States is composed of the operator control panel and the analysis work station [15]. However, the new generation of security systems such as L3 systems and our research group developed active THz imaging system can completely overcome these shortcomings, the imaging system of the time and the computer to identify suspicious objects is very short, usually take about 3.7 s. Also the ray detection is harmless to the human body.

4 Conclusion

In this paper, we proposed a new algorithm for fast detection of THz images. The algorithm is as follows: (1), Firstly, the terahertz image is preprocessing to smooth the image, and the human body is enhanced by gray stretch. (2), Secondly, by using a morphological classification algorithm proposed in this paper to distinguish the discontinuous contour of the human body and the continuous contour of the human body, (3), By using BCTC algorithm proposed in this paper to extract a complete closed

human body contour, (4), At last, mainly focus on annotation of suspicious objects. A large number of tests show that our new method has a high accuracy rate, about 92%. The average total time of terahertz image recognition is less than 4 s.

Acknowledgements. The research was partly supported by the program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, USST incubation project (15HJPY-MS02), Construction project of pilot project for terahertz technology products (ZJ2014-ZD-004), the new terahertz source, National 973 Project (2014CB339800), National Natural Science Foundation of China (61502220; U1304616) and Shanghai Engineering Research Center Project (GCZX14014).

References

1. Tribe, W.R., Newnham, D.A., Taday, P.F., et al.: Hidden object detection: security applications of terahertz technology. In: *Integrated Optoelectronic Devices*, pp. 169–176 (2004)
2. Yao, J.Q.: Introduction of THz-wave and its applications. *J. Chongqing Univ. Posts Telecommun.* **22**, 703–707 (2010)
3. Hu, B.B., Nness, M.: Imaging with THz waves. *Opt. Lett.* **20**(16), 1716–1719 (1995)
4. Zhang, L.L., Nick, K., Zhang, C.L., Zhao, Y.J., Zhang, X.C.: Real-time nondestructive imaging with THz waves. *Opt. Comm.* **281**(6), 1473–1475 (2008). Harvard.edu/abs
5. Nicholas, K.P., Zhong, H., Xu, J.Z., Zhang, X.-C.: Non-destructive sub-THz CW imaging. In: *Proceedings of SPIE*, vol. 5727 (2005)
6. Zhao, R., Zhu, Y.M., Zhang, C.L.: Target aided identification in passive human THz-image. *High Power Laser Particle Beams* **26**, 126–130 (2014)
7. Qiao, L.B., Wang, Y.X., Zhao, Z.R., Niu, Y.J., Chen, Z.Q.: Analysis of active near-field terahertz imaging for personnel surveillance. *J. Microwaves* **31** (2015)
8. Zang, X.F., Li, Z., Shi, C., Chen, L., Cai, B., Zhu, Y.M., Li, L., Wang, X.B.: Rotatable illusion media for manipulating terahertz electromagnetic waves. *Opt. Express* **21**, 25565 (2013)
9. Chen, L., Gao, C.M., Xu, J.M., Zang, X.F., Cai, B., Zhu, Y.M.: Observation of electromagnetically induced transparency-like transmission in terahertz asymmetric waveguide-cavities systems. *Opt. Lett.* **38**, 1379 (2013)
10. Gonzalez, R., Wood, R.: *Digital Image Processing*. Addison-Wesley, New York (1992)
11. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
12. Sahiner, B., Chan, H.P., Wei, D., et al.: Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue. *Med. Phys.* **23**(10), 1671–1684 (1996)
13. Vincent, L., Soille, P.: Watersheds in digital space: an efficient algorithms based on immersion simulation. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(6), 583–598 (1991)
14. Hoskins, J.C., Himmelblau, D.M.: Process control via artificial neural networks and reinforcement learning. *Comput. Chem. Eng.* **16**(4), 241–325 (1992)
15. L-3 Communications, New York. <http://www.l-3com.com/>

Information Visualization for Mobile-Based Disability Test Applications

Jongmun Jeong¹, Seungho Kim¹, Changsoon Kang², and Mintae Hwang²(✉)

¹ Department of Eco-friendly Offshore Plant FEED Engineering, Changwon National University, Changwon, Republic of Korea

jhs7986@gmail.com, shk0529@changwon.ac.kr

² Department of Information and Communication Engineering, Changwon National University, Changwon, Republic of Korea

cskang@changwon.ac.kr, professorhwang@gmail.com

Abstract. As part of the development of ICT convergence technologies for special education, we implemented a mobile-based disability test application. In this paper, we studied an information visualization method that uses a graph to optimize the display on the screen of mobile devices. The implemented application properly divides the screen of the mobile device, and the test scores obtained from the results of the disability test are displayed in a graph that can be moved left and right using the GraphView library. Test comments are also displayed in a table format that can be simultaneously moved up and down. This provides efficiency to users due to the large amount of data that can be represented.

Keywords: Disability test · Information visualization · Mobile-based system

1 Introduction

The current digital age exposes individuals to a large quantity of information, and it is thus necessary to classify information so that necessary information can be obtained. However, even when classifying a large quantity of information, it is impossible to check all information even if it is restricted to a single category. Therefore, it is necessary to efficiently process information.

Information visualization refers to presenting information in a visual manner to communicate it more efficiently. In other words, information expressed in tables, figures, graphs, etc. will improve readers' understanding [1].

The following guidelines are applicable to information visualization [2]. First, try to utilize graphic expressions, such as icons and images, to display intricate information. Second, consider the impact of the position when arranging information. Third, express relationship and differences between variables clearly. Finally, avoid listing a large volume of information on one page.

This paper contributes to research on information visualization for mobile based disability test applications, and we studied the information visualization method that can optimize the test results on a mobile screen. We eliminate the inconvenience of providing

disability test results through text and tables by using a graph and table together to efficiently provide optimized results on a mobile devices screen.

This paper is organized as follows. Section 2 introduces a case study on information visualization for mobile applications, Sect. 3 briefly introduces a mobile-based disability test system developed in this paper as part of the research on information visualization, and Sect. 4 presents a method to apply information visualization to a mobile disability test application as well as its result. Finally, Sect. 5 presents the conclusion and future work.

2 Related Research

Information visualization is intended to more efficiently convey information to users by using graphic elements to shape data so that meaning is generated as information and to organize and express information in forms and structures that are easy for viewers to understand. When discussing information visualization, we need to focus on how information is understood by users. It is thus necessary to understand human perception and cognitive processes [3].

A study analyzing the GUI design of applications used in smartphones [4] showed that the usage rate of a web browser on mobile device has been rapidly increasing. Nevertheless, websites that are optimized for a desktop environment can have reduced accessibility and usability when viewed on mobile devices. With this in mind, web design targeting mobile devices can improve user satisfaction.

Another study analyzing the GUI design of the smartphone applications [5] showed that the visual satisfaction of smartphone users can be further improved by recognizing the importance of GUI design as a key part of application design. Going forward, an application should be developed to help users accurately grasp useful information while maintaining professional use, functionality, and practicality of the application in response to diverse user needs.

An study on user interface design for information visualization in mobile augmented reality [6] summarized and analyzed the trends of mobile augmented reality applications. Considering the necessity and importance of information visualization in a mobile environment, user UI elements and functions and design changes are analyzed in terms of information visualization through interface design examples. An analysis was also conducted of the correlation between visualization elements of this information and the user interface when applying it to application development.

In addition to theoretical research on information visualization in mobile applications, a search on Google Play reveals many mobile applications in which information visualization is applied [7]. These applications are shown in Fig. 1, instead of showing text-based data, these display information visually by using suitable shapes, icons, and graphs.

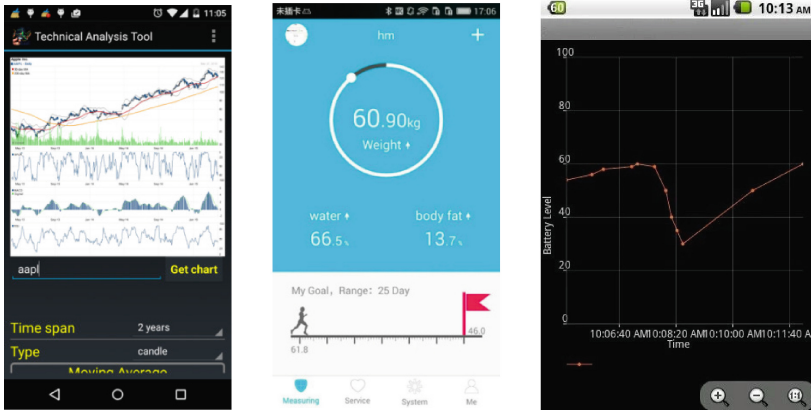


Fig. 1. Examples of information visualization in mobile applications

The mobile applications related to special education covered in this paper are limited to several parts, mainly including AAC (Augmentative and Alternative Communication) technology and development of special educational contents. Therefore, we could not find examples of application development applying information visualization technology for disability tests [7].

3 A Study of Mobile Applications for Disability Tests

Research on mobile-based disability test application development is currently underway as part of special education and ICT convergence technology research in order to resolve the inefficiency and inaccuracy of the disability test methods using a test sheet or a checklist. A simple test to score the level of disability with an autism behavior test and learning disorder test can be easily developed for a mobile basis. As shown in Fig. 2, it can be utilized at any time and place in special education.

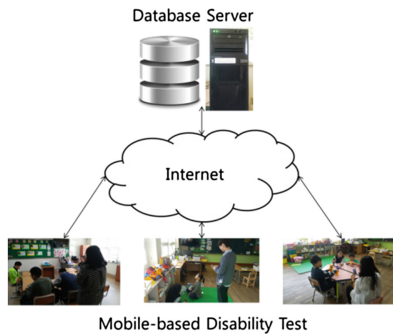


Fig. 2. Operating environment of mobile based disability test system

The disability test result is stored and managed in the server database, and it is thus possible to search anytime and anywhere and guarantee the accuracy of the test results.

In this paper, a mobile-based disability test applications is developed for the Android operation system targeting autism behavior and learning disorder tests. These were developed using the Java and PHP programming languages in the Android Studio integrated development environment, and the Apache web server and MySQL database management system are also used.

Figure 3 shows an example of the implementation of a mobile-based disability test application.

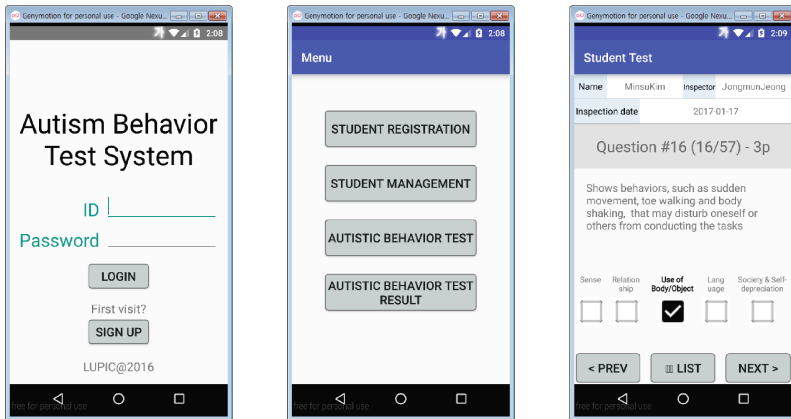


Fig. 3. Implementation of an example of a mobile-based disability test application

The disability test mobile application developed in this paper not only offers a member registration function for the inspector (e.g., special school teacher), but the registered inspector can also register and manage disabled students as the test subjects. And application includes the function to directly test the disability behavior and to refer and manage the test results. The disability test is based on questions developed by the National Institute of Special Education. When a student observes the behavior described in a given question, the response is marked with a checkmark.

The disability test mobile application developed in this paper allows the parents of a disabled student to view the results for their child, and it provides functions to directly test disability-related behavior. These functions allow for parents and teachers to share test information for disabled students to effectively support and manage disabled students at school and at home.

4 Information Visualization for Mobile-Based Disability Test Applications

Figure 4 shows the result of applying the information visualization technique to the disability test mobile application developed in this paper.

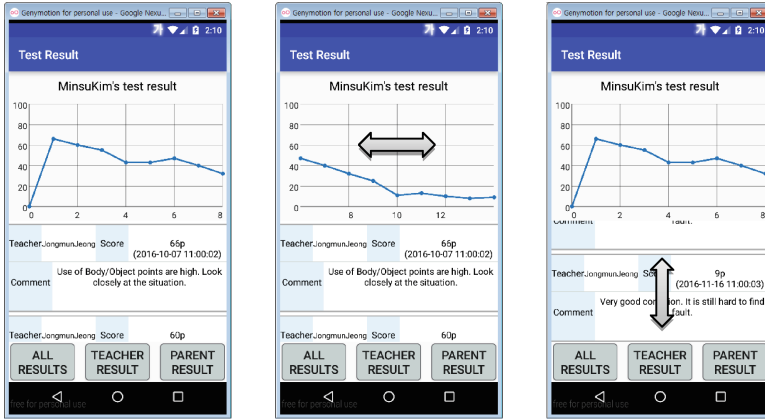


Fig. 4. Disability test application with information visualization

The implemented application divides the screen of the mobile device into two areas. In the upper part, the test score obtained from the test result is shown as a graph, and in the lower part, the test comment of the inspector is shown in tabular form.

By allowing the upper graph to move to the left and right, the lower table to move up and down, we implemented a function that can express a large amount of data optimized for the small screen of a mobile device.

In this paper, we used the GraphView library to visualize information for the mobile application [8], and the core code for the information visualization is as follows.

```

DataPoint[] dataPoints;
    LineGraphSeries<DataPoint> series;
dataPoints = new DataPoint[array.length + 1];
dataPoints[0] = new DataPoint(0, 0);
    for (int i = 0; i < array.length; i++) {
        dataPoints[i + 1] =
            new DataPoint(i + 1, array[i]);
    }
series = new LineGraphSeries<DataPoint>(dataPoints);
graphView.getViewPort().setScrollable(true);
graphView.getViewPort().setMinY(0.0);
graphView.getViewPort().setMinX(0.0);
graphView.getViewPort().setMaxY(100.0);
graphView.getViewPort().setMaxX(8);
series.setDrawDataPoints(true);
series.setDataPointsRadius(10);
series.setThickness(8);
graphView.addSeries(series);

```

This source code stores the disability scores in an Array, converting it into a Data-Point format and adding it to GraphView. The minimum value and maximum value of the graph is set and then a line graph is added.

The GraphView library can be used to visualize the test results in various forms. Figure 5 shows the results of the disability test implemented in this paper, shown with a number of graphs.

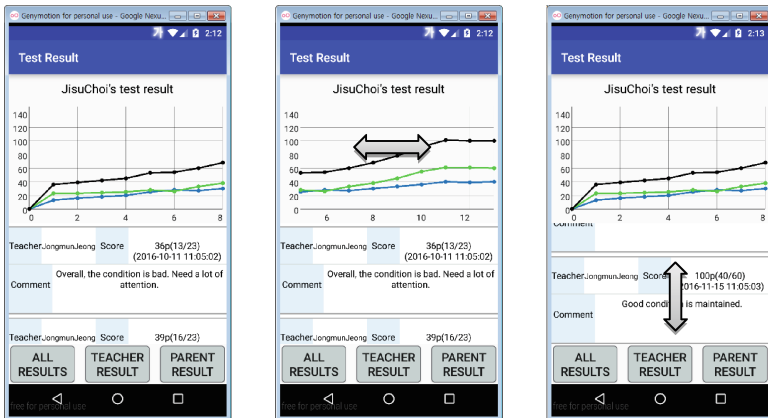


Fig. 5. Application examples of information visualization

5 Conclusion

This paper implemented a disability test mobile application, and we investigated how the information visualization method can optimize the test results on the screen of the mobile device by using suitable graphs and tables.

The implemented disability test application divides the screen of the mobile device into two areas. In the upper part, the test scores obtained from the disability test are displayed in the graph and can be moved using the GraphView library. In the lower part, the test comments are presented in a table that can be moved up and down, so the a large amount of data can be efficiently presented.

In the future, we will attempt to access various graph types for disability test mobile applications, and we will also try to investigate visualizations for integrated information by improving the GUI function of the table.

References

1. Cho, S.-N., Seo, T.-S.: Study on development of journal and article visualization services. J. Korean Soc. Libr. Inf. Sci. **50**(2), 183–196 (2016)
2. Zuo, Y.X., Kim, H.H.: A study on the effective communication of information visualization - focusing on the first page of weather forecasting application. J. Digit. Des. **14**(3), 83–94 (2014)

3. Byeongeun, O., Seongjung, G.: Textbook of Information Design, Ahn Graphics, pp. 99–100 (2008)
4. Kim, H., Park, J.: A study on the designing user interfaces for mobile web. *J. Digital Des.* **10**(2), 65–74 (2010)
5. Yang, H.-J., Lee, J.R.: A study on the analysis of the GUI design of the applications used in the smartphone base. *J. Soc. Korea Illusart* **14**(2), 107–116 (2011)
6. Kim, B., Kim, Y.: A study on user interface designs through analysis of information visualization in a mobile-augmented reality. *J. Korea Des. Knowl.* **16**, 72–81 (2010)
7. Google Play (2016). <https://play.google.com/>
8. Github. <https://github.com/jjoe64/GraphView/>

Deep Convolutional Neural Networks for All-Day Pedestrian Detection

Xingguo Zhang^(✉), Guoyue Chen, Kazuki Saruta, and Yuki Terata

Faculty of Systems Science and Technology, Akita Prefectural University, Yurihonjo, Japan
{xingguozhang, chen, saruta, terata}@akita-pu.ac.jp

Abstract. Pedestrian detection is a special topic in computer vision and plays a key role in intelligent vehicles and unmanned drive. Although recent pedestrian detect methods such as RPN_BF [1] have shown good performance from visible spectrum images at daytime, they have limited study for near-infrared image at nighttime. Unfortunately, when the traffic accident happened at night, the pedestrian is one of the most serious victims. Recently deep convolutional neural networks such as R-CNN/Faster R-CNN [2, 3] have shown excellent performance for object detection. In this paper, we investigate issues involving Faster R-CNN for construction of end-to-end all-day pedestrian detection system. We propose an effective baseline for pedestrian detection both on visible spectrum images and infrared images, using a same pre-train Faster R-CNN model. We comprehensively evaluate this method, the experiment results presenting competitive accuracy and acceptable running time.

1 Introduction

Automobiles bring people great convenience and changed societies in many aspects. However, such a technology has also carried a dark side: traffic accidents. Every year almost 1.2 million people are killed in traffic crashes while the number of injured rises to 50 million. Meanwhile, fatality rates of traffic accidents occur at night are more than double compared to daytime [4].

In order to improve safety, Pedestrian Protection Systems (PPSs) have attracted an extensive amount of interest from the computer vision community over the past few years [5, 6]. A PPS is defined as a system that detects both static and moving people in the surroundings of the vehicle (typically in the front area) in order to provide information to the driver and perform evasive or braking actions on the host vehicle if needed. So more traffic accidents are expected to be avoided in future.

However, pedestrian detection also is an extremely challenging task due to the large intra-class variability caused by different articulated poses and clothing, cluttered backgrounds, abundant partial occlusions and frequent changes in illumination.

Almost all current leading pedestrian detectors have being evaluated by visible spectrum images at daytime. The detection performance of the images at nighttime is still unexplored. They seem to overlook the fact: the pedestrian is one of the most serious victims in traffic accident occur at night.

As mentioned at [7], normal cameras based on visible spectrum images (hereafter called VS images) are not very satisfactory in the absence of plenty of illumination. In the day time, this illumination can come from the sun, but at night, artificial illumination is required. Important areas of interest could be lit with bright lights but undesirable activities are more likely to occur in darker areas. Infrared (IR) cameras are ideally suited to imaging under these conditions, as they sense emitted radiation from the objects of interest, such as pedestrians. However, IR cameras are still expensive to deploy on a large scale. Therefore, we expect to detect pedestrian at night using near-infrared (NIR) cameras which are cheaper.

Recently deep convolutional neural networks such as R-CNN/Faster R-CNN [2, 3] have shown excellent performance for object detection, but they have limited study for pedestrian detection at nighttime. In this paper, we propose an effective baseline for pedestrian detection not only on VS images but also on NIR images, using a same pre-train Faster R-CNN model.

The rest of the paper is organized as follows: Sect. 2 introduce the related works. Section 3 presents the pipeline that we use for detecting pedestrians. Section 4 is devoted to experimental evaluation, whereas conclusions are drawn in Sect. 5.

2 Related Work

Most pedestrian detection algorithms share similar computation pipelines. Such pipeline comprises two main stages: (i) region proposal, (ii) region classification.

Region Proposals. As regards the first stage, the entire frame is analyzed so as to extract a set of candidate regions.

The simplest technique to obtain the candidate regions is the sliding window approach, where detector windows at multiple scales and locations are shifted over the image. But the computational costs are often too high to allow for real-time processing [4]. More complex approaches analyze the visual content to filter out regions that are believed not to contain objects or salient content. Selective Search [8] is instance of such class of algorithm.

Region Classification. Current methods for pedestrian detection can be generally grouped into two categories, the models based on hand-crafted features [9–12] and deep convolutional features [13–17]. As the first category, HOG [18] and Deformable Part Models (DPM) [9] are very popular methods. Although they are sufficient to certain pose changes, the pedestrian detection accuracy bottlenecks are already appearing.

Recently the prevalent success of deep learning approach in computer vision, such as R-CNN [2], RPN_BF [1] have shown good performance for pedestrian detection.

The R-CNN method [2] trains CNNs end-to-end to classify the proposal regions into object categories or background.

Fast R-CNN [19] enables end-to-end detector training on shared convolutional features and shows compelling accuracy and speed. In [1], an RPN_BF approach have been proposed. An RPN that generates candidate boxes as well as convolutional feature

maps, and a Boosted Forest that classifies these proposals using these convolutional features.

3 All-Day Pedestrian Detection Based on Faster-RCNN

In this section, we describe our all-day pedestrian detection pipeline by Faster R-CNN [3].

Faster R-CNN consists of two components, as show in Fig. 1: an RPN that generates candidate boxes as well as convolutional feature maps, and a Fast R-CNN used for object detection. In here, we using two datasets (VS images and NIR images) to training the RPN and Fast R-CNN network, and created a single detectors. As such, we expect to use it to detect the all-day pedestrian.

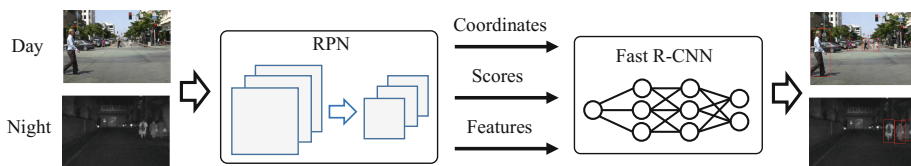


Fig. 1. Our all-day pedestrian detection pipeline. RPN is used to compute candidate bounding boxes, scores, and convolutional feature maps. The candidate boxes are fed into Fast R-CNN for further classification, using the features pooled from the convolutional feature maps computed by RPN. Finally, NMS is used to merge the similar results and get the output.

3.1 RPN Network

The RPN network shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection.

We fixed the aspect ratio of anchors (Region Proposal Boxes) [3] as 0.45 (width/height). This is the average aspect ratio of pedestrians as indicated in [6]. This is unlike the original RPN [3] for detect general object that has anchors of multiple aspect ratios. In order to detect multi-scale pedestrians, we use anchors of 9 different scales, starting from 30 pixels height with a scaling stride of 1.2.

Following [3], we adopt the VGG-16 net [20] pre-trained on the ImageNet dataset [21] as the initial network. The RPN is built on top of the Conv5_3 layer, which is followed by an intermediate 3×3 convolutional layer and two sibling 1×1 convolutional layers for classification and bounding box regression (more details in [3]).

The output layer of the RPN net provides confidence scores and regression coordinate of the predicted boxes, which can be used as the input for Fast R-CNN network.

3.2 Fast R-CNN Network

For the detection process, we adopt Fast R-CNN network as mentioned at [3].

To speed up the process, [3] developed a technique that allows for sharing convolutional layers between the two networks, rather than learning two separate networks. For the convenience of the reader, we briefly recap such approach.

A 4-step training algorithm has been adopted to learn shared features via alternating optimization. In the 1st step, the RPN network has been trained. And this network is initialized with an ImageNet pre-trained model and fine-tuned end-to-end for the region proposal task. In the 2nd step, they train a separate detection network by Fast R-CNN using the proposals generated by the 1st step. In the 3rd step, they use the detector network to initialize RPN training, but fix the shared convolutional layers and only fine-tune the layers unique to RPN. Finally, keeping the shared convolutional layers fixed, and fine-tune the unique layers of Fast R-CNN. As such, both networks share the same convolutional layers and form the Faster R-CNN network.

In this solution, the RPN and Fast R-CNN networks are merged into one network during training. Two examples of convolutional feature map are shown in Fig. 2.

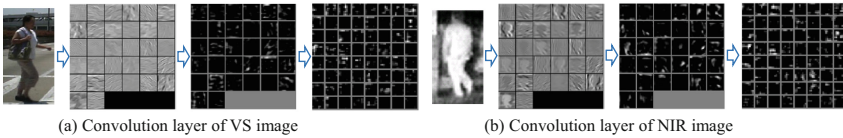


Fig. 2. Two examples of feature maps for VS and NIR image obtained at each layer of a convolutional neural network, respectively. Former layers identify simple structures such as edges, corner and details, whereas deeper layers identify more complex abstract concepts.

3.3 Implementation Details

We adopt a single scale training and testing as in [1]. We do not use feature pyramids, because multi-scale feature extraction may improve accuracy but does not suit the task for real-time.

For RPN and Fast R-CNN training, an anchor is considered a positive example if it has an Intersection-over-Union (IoU) ratio greater than 0.7 with one ground truth box, and otherwise consider as negative. For Fast R-CNN training, we construct the training set by selecting the top-ranked 100 proposals of each image by RPN network. At test process, we only use the top 100 proposals in an image, which are classified by the Fast R-CNN. We adopt non-maximum suppression (NMS) to output the detect results.

4 Experiment Result

In this section, we describe our experimental setup details and the results.

4.1 Datasets

We used two datasets to training the Faster R-CNN model, and created a single detectors. One for the NIR images and the other for VS images, as shown in Fig. 3.

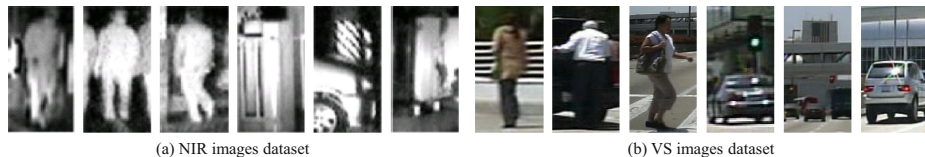


Fig. 3. Example images of NIR images dataset and VS images dataset used in this experiment.

NIR Image Data: For NIR images, we collected a set of video sequences containing pedestrians from multiple view points and of multiple sizes, using a monochrome board camera KPC-EX500BA and a NIR lamp RM-240 (spectral wavelength in 0.7–2.5 microns). Images were captured at night and the height of the persons in the images ranged from 50 to 300 pixels.

The annotations include temporal correspondence of the same pedestrian across different frames. Nearly half of the frames do not contain any pedestrian, whereas 30% of the frames contain two or more.

On average, a pedestrian is visible for about 4 s. Some of the training and testing images are shown in Fig. 3(a).

VS Image Data: For VS images, we used the Caltech which is a publicly available dataset [6]. The dataset is collected by an on-board camera within a vehicle that includes pedestrians from different viewpoints under good weather conditions (Fig. 3(b)). The resolution and the sampling rate of the video content are 640×480 and 30fps, respectively.

We selected 5000 frames NIR images and 20000 frames VS images as the training samples respectively. The test samples also included 500 frames NIR images and 500 frames VS images selected from the rest of the dataset.

4.2 Evaluation Metrics

We introduce the evaluation metrics defined by the Caltech Pedestrian Detection [6], which also is the most common evaluation protocol.

In particular, the performance of an algorithm is evaluated in terms of the tradeoff between the miss rate (MR) and the number of false positives per image (FPPI). First, a detected bounding box and a ground truth bounding box (GT) are considered a true positive (TP) if the area covered by their intersection $\geq 70\%$. A GT that does not have

a match is considered a False Negative (FN), or a Miss. A detected bounding box that does not have a matching GT is considered as a False Positive (FP). Then, it is straightforward to define the average number of false positives per image (FPPI), that is, the average number of regions of each image that are erroneously detected as a pedestrian. The Miss Rate is computed by the following equation:

$$MissRate = FN/C_{GT} \quad (1)$$

where C_{GT} is the number of GT in images. In here, we count the miss rate at $FPPI = 0.1$, that has been identified as a reasonable working condition for a real-world system.

4.3 Comparison with the State-of-the-Art

In this section, our final detectors were evaluated with other state-of-the-art methods using our NIR dataset and Caltech.

We performed the standard per-image evaluation used in pedestrian detection [6]. Results are shown in Table 1. We found that every detector has higher accuracy for VS images than NIR images. This is probably due to the VS images have plenty of details. These details might be important for pedestrian detection. But in NIR images, the details of the textures are lost as the human body temperature is relatively constant over the entire body.



(a) Our approach for VS image



(b) RPN_BF for VS image



(c) Our approach for NIR image



(d) RPN_BF for NIR image

Fig. 4. Examples of some results using different detector. Our detector show better performance for the both of two datasets.

Table 1. Comparisons the *MissRate* of different classifiers on the VS dataset and NIR dataset.

	Our (using Faster R-CNN)		RPN_BF [1]		HOG [18]	
	VS	NIR	VS	NIR	VS	NIR
<i>MissRate</i> (FPPI = 0.1)	0.19	0.24	0.26	0.56	0.45	0.62

For VS images dataset, our detector has an MR of 19%, which is better than the RPN_BF's 26% and HOG's 45%. For NIR images dataset, our detector also obtain the best accuracy. Furthermore, the accuracy have a significant decline when using RPN_BF.

Figure 4 show the results of the NIR images and VS images. From the results of the images it can be seen that for the both of two datasets, our detector is competitive in terms of the detection quality with respect to RPN_BF and provides significant improvement over HOG + SVM.

4.4 Analysis of Computational Time

We profiled the execution of our system on a desktop architecture which features a 2.4 GHz Intel i7 CPU, a NVIDIA GTX750 GPU and 32 GB of RAM. The systems requires, on average, 103 ms to process a frame at a resolution of 640×480 pixels. It can be consider as an acceptable running time.

5 Conclusions

In this paper, we investigate issues involving Faster R-CNN for construction of end-to-end all-day pedestrian detection system. We proposed an effective baseline for pedestrian detection both on visible spectrum images and infrared images, using a same pre-train Faster R-CNN model. We comprehensively evaluate this method, the experiment results presenting competitive accuracy and acceptable speed. In future work, more theoretical and experimental studies will be conducted to analyze the performance.

References

1. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: ECCV 2016, 1–15 (2016)
2. Girshick, R., Donahue, J., Darrell, T., Berkeley, U.C., Malik, J.: R-CNN: region-based convolutional neural networks. In: CVPR 2014, 2–9 (2014)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, 1–10 (2015)
4. World Health Organization: Global Status Report on Road Safety 2015 (2015)
5. Yan, J., Zhang, X., Lei, Z.: Robust multi-resolution pedestrian detection in traffic scenes. In: IEEE Conference Computer Vision Pattern Recognition, 3033–3040 (2013)
6. Dollár, P., Wojek, C.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 743–761 (2012)

7. Zhang, X., Chen, G., Saruta, K., Terata, Y.: Discriminative feature points distribution in near-infrared pedestrian images. *IEEJ Trans. Electron. Inf. Syst.* **135**, 1222–1228 (2015)
8. Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. *Int. J. Comput. Vis.* **104**, 154–171 (2013)
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1627–1645 (2010)
10. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed Haar-like features improve pedestrian detection. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 947–954 (2014)
11. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1532–1545 (2014)
12. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: *BMVC 2009, London, England*, 1–11 (2009)
13. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3258–3265 (2012)
14. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship in pedestrian detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3222–3229 (2013)
15. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 899–905 (2014)
16. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: *Proceedings of IEEE Computer Society Conference Computer Vision Pattern Recognition*, 7–12 June 2015, 5079–5087 (2015)
17. Tomè, D., Monti, F., Baroffio, L., Bondi, L., Tagliasacchi, M., Tubaro, S.: Deep convolutional neural networks for pedestrian detection. *Arxiv* (2015)
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **1**, 886–893 (2005)
19. Girshick, R.: Fast R-CNN: fast region-based convolutional networks for object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448 (2016)
20. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: *ICLR*, 1–14 (2014)
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015)

An Augmented Reality Learning System for Programming Concepts

Kelwin Seen Tiong Tan^(✉) and Yunli Lee

Department of Computing and Information Systems, Sunway University Selangor,
Subang Jaya, Malaysia

13014865@imail.sunway.edu.my, yunlil@sunway.edu.my

Abstract. Learning programming concepts through traditional learning method is challenging for novice programming students due to it lacks the characteristics to motivate students to further study on the particular content. This paper analyses the results of survey given to students enrolled in computing programmes at Sunway University in 2016. The survey focused on student involvement in learning using student engagement matrix. The overall findings states that students are only moderately engaged, hence we aim to enhance the interactivity and motivation of learning through the use of Augmented Reality (AR) technology in designing the learning system for programming concepts. Two basic of programming structures are implemented and evaluated by 20 novice programming students. 80% of students agreed that AR learning method is an effective method as it provides more fun, interest, provide basic understanding and good effect for students to learn programming.

1 Introduction

The current learning approach which was adopted by universities are still the traditional learning method including lectures, tutorials, practical and examinations [1]. These traditional learning such as power point slides and images are one-way interaction. Hence, it lacks the characteristics to motivate students to further study on the particular subject or content. Motivation plays a crucial role in the teaching style because student may feel isolated and they will stop learning [2]. In order to enhance a student's learning behaviour, a better content presentation method is required. Currently, there are various solutions have been proposed to overcome this issue like Second Life (A virtual reality classroom) [3], Gamification approach [4], Short Video Lecture [5], Twitter (social media) learning [6], Augmented Reality (AR) with interactive content [7] and many more. Among these solutions, AR is one of latest advancement which is predicted to become a technical trend in higher education in two to three years' time [8]. It has been implemented in several education scope to enhance student's interactivity such as Mathematics, Nursing, Science and robotics programming logic and morphology.

AR is a technology that provides composite view by combining virtual objects with real world environment [9]. It is an interactive technology that contains real time processing and it is designed in a three dimensional structure. Unlike Virtual Reality

(VR), AR contains less virtual content but more of realness to the user as it uses the real environment as a background to augmented the targeted image. Through the implementation of AR into current learning system, various studies states that it can significantly contribute to student's perception, interaction and motivation. In addition, the motivation level increases as the learning environment becomes more collaborative and interactive. The proposed learning system aims to aid students with the basic concepts of programming. These concepts will be taught through an AR environment within a certain scene which is generated by the system using marker based technique. Hence in this project, AR approach is proposed to enhance the interactivity of students in class. This project proposes a new AR system that uses interactive marker to learn basic programming concepts. The evaluation result shows that 80% of students agreed that AR learning method is an effective method as it provides more fun, interest, provide basic understanding and good effect for students to learn programming.

2 Literature Review

2.1 Current State of AR in Education Field

Even though AR has only been implement in educational fields in recent years, a considerable amount of researches has been conducted within the educational context in order to achieve a wide variety of learning domains. However, the current state of AR for education is still just a start [10]. Researchers started implementing AR in many different ways in different educational subjects such as mathematics, science, arts, business and many more. In the field for instance, different implementation methods are being used for different educational context. For example, researchers can use either different devices, different mode of interaction, different method for sensory feedback, fixed/portable experience and many more [11]. In addition, researchers might also use different technique to implement AR in education, for example, using marker to create movement of an object to learn its function, integrating virtual class materials with real environment, AR with mobile GPS and many more. AR implementation in education can also be implemented in different environment, such as AR in classroom, AR laboratory and more.

2.2 AR Development Tools – Unity 3D

Unity 3D is a software package that can be fully integrated into development engines such as Vuforia and Visual Studio which provides stunning and efficient functionality for the creation of effective and interactive 3D content. The perceived ease of usefulness using this software package in the development tool makes it easier for user to venture into AR projects since it is a platform with pre-filled AR programs, databases, animation are well as an interactive design platform. By implementing Unity software, publication of the developed application can be done on vast platform such as on Windows PC, Mac PC, IOS, Android, Xbox, Web, Linux and many more [12]. On the side note, the platform also provides intuitive workspace which allows the user to control the marker size, augmented animation as well as other programming related acts on the application.

A real-time testing can be done as long as the program is being compiled properly by clicking the play button on the workspace. Thence, this makes the editing and functionality testing of the AR application to be faster and more efficient. The Vuforia AR extension allow an additional detection of vision and tracking functionality which can be alter at ease within Unity, thus allowing developers to create AR application in a fast manner as long as internet is available [12].

3 Design and Implementation

3.1 Pre-evaluation Survey

The sample method used in this research is random sampling. The target sample population is 82 computing students taking programming principles under Department of Computing and Information Systems. The survey questionnaire is developed based on the Student Engagement Matrix's dimension 3 – involvement in learning as shown in Fig. 1 [13]. The survey's quantitative data on the effectiveness and student respond towards the current learning system (traditional learning method) has been recorded to understand their perception towards the current learning approach. The attention and memory, class participation, participation in learning and literacy/numeracy levels is used to understand the current learning method, whereas dealing with feedback and resilience provides contributes to the design of this AR system.

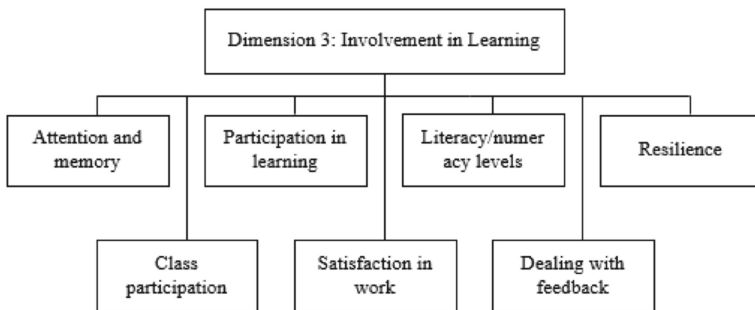


Fig. 1. Involvement in learning in Student Engagement Matrix.

From the findings obtained from pre evaluation, students are being categorized into different level of engagement within the matrix. The findings show that majority of students falls in moderately engaged for attention and memory, participation in learning, resilience, and class participation. This instated that the students, with accordance to the student engaged matrix, usually remembers instructions and concepts, completes work in most field but may need extra time, can use literacy to achieve most age-appropriate task, can manage challenges with support, answers may pose some relevant questions, gains some satisfaction from own work and shows some capacity for accepting feedback. This overall states that students are only moderately engaged, hence this system aims to enhance interactivity and motivation of students in learning basic programming concepts.

3.2 Marker Development

The marker of this project is developed by referring to the design of black and white fiducial marker. As opposed to the traditional marker, this new marker is done with the purpose of allowing year one computing students to learn about the basic programming concepts. The creation of marker is based on three basic concepts of programming, which is the IF condition, WHILE loop and COUNTER loop. Each loop provides two case scenario using Boolean format, if the pattern matches the trained marker, the augmented animation will perform a set of actions which is pre-set in the program. In addition, the markers are designed on transparent paper where each marker is represented as a specific programming instruction that allow student to stack few markers to form a complete marker of respective desired solution. Figure 2 shown the various type of marker with each instruction, by stacking these markers could enhance the interactivity of students, allowing the students to play around with different loop. The interactive markers are being printed in a separated form of arrow, loop, number, and action. The transparent type of markers allowed students to overlay each layer on top of one another to form a loop pattern that is being set. The marker based is white and black, due to the printing on a transparent paper, a white paper as background is attached.

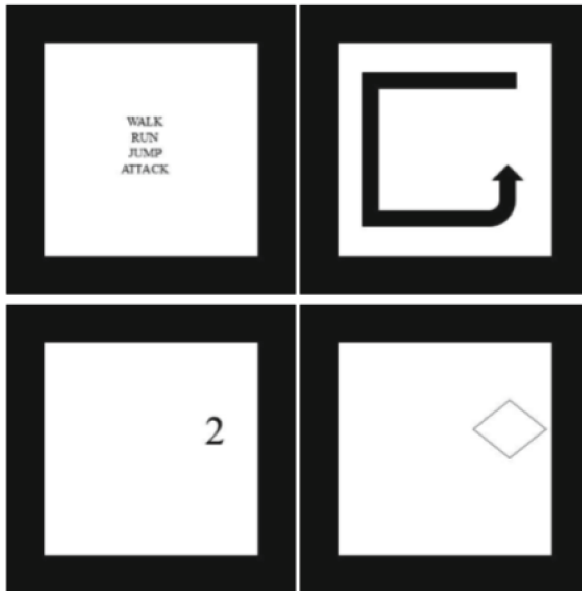


Fig. 2. Transparent markers with four different simple instruction

3.3 Identify and Altering Animation

A suitable animation, which is the Red Samurai shown in Fig. 3 is being used in this project as the animation itself covers a few actions such as attack, jump, walk, run and idle. In addition to that, the samurai animation freely available on the unity3d online

animation store which is ready to be altered whenever necessary to suit the purpose of the project. The action done by the chosen augmented object can be looped forever, ping pong, susceptible to physics concepts as well as played only once, these feature can be implemented in teaching programming since it will provide the loop understanding to the users using the system. For instance, while the loop is 2 shown in Fig. 4, the samurai will jump twice. Upon identifying the correct animation and feature it contains, animation is being applied onto the marker with the desired size as well as suitable projection to prevent it from oversizing when augmentation is done.



Fig. 3. Red Samurai animation

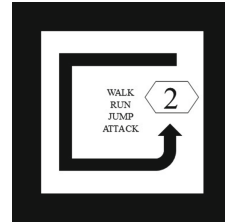


Fig. 4. While loop with value of 2

3.4 Marker Training

The marker is first being created using the Microsoft Power Point as the program provides feature to alter the slides size to a desired, squared marker size for the ease of creating a square marker. The marker size is set to 10 × 10 centimetre. We use black casing as the base to design each programming instruction. Figure 5 below shows an example of full marker for the project with description and its measurement in centimetre.

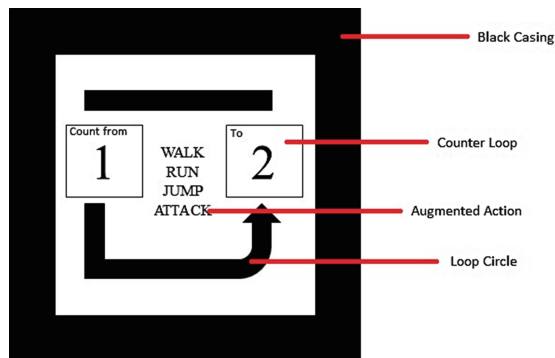


Fig. 5. Counter loop design example with scales and description

The full marker is exported as JPG from Power Point upon completion. The JPG files of the marker is imported into the Vuforia cloud server, which is an online cloud database which allows developers to create database to store AR target or base image to be used in the program that will developed. An automatic feature extraction is done by the Vuforia cloud on the marker by choosing the desired feature it requires to use. These markers be downloaded and used in the Unity3D format for the system development stage. The database itself will be saved in Vuforia pre-prepared format to suit the development tools. A total of 6 markers is being imported, including two If conditions, two while loop and two counter loop into the Unity3d which is stored in the Vuforia developer cloud server as per this project. A width with scale of 0.5 for both the augmented animation as well as the marker is set since the projection should fit the screen size of any android phone as the system testing will be done in application base. The full functional and trained system is being built on an android system which suits the current Android SDK of majority android OS phones. The outcome of the android application of programming concept is shown in Fig. 6.

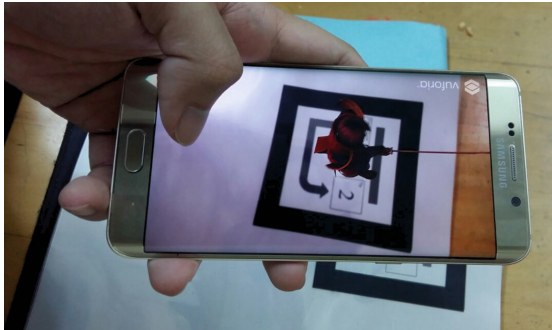


Fig. 6. AR based application for programming concept

3.5 Post-evaluation Survey

An evaluation is being done on a group of 20 Sunway University year 1 computing students to test and answer a post evaluation survey on the system. Upon using the designed system, 45% of the 20 student respondent in Fig. 7 states that the learning method will be able to improve understanding on programming concepts. 35% of students in Fig. 8 also agree that this method can help students in managing any programming related challenges. Out of the total of 20 responses shown in Fig. 9, 80% of students actually states that AR learning method is an effective method as it provides more fun, interest, provide basic understanding and good effect for students to learn programming. As opposed to the 80%, the remaining states that this method is ineffective for learning programming since it only provides understanding on the theory and the detection of the marker is not easy due to human error.

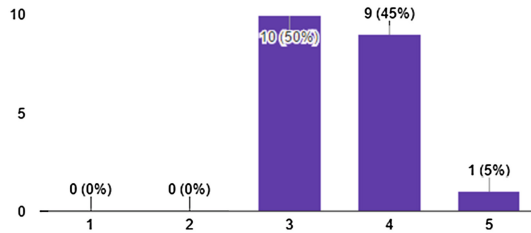


Fig. 7. This learning method would be able to improve my understanding on the programming concept

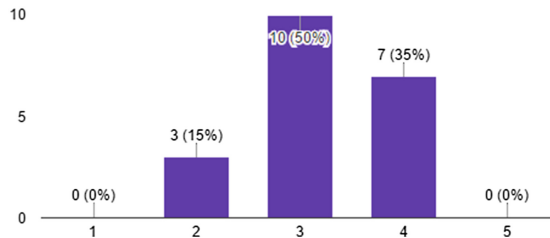


Fig. 8. I think I will be able to manage any programming concept related challenges if this method of learning is used

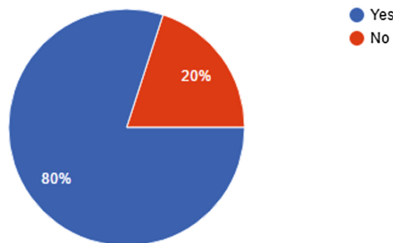


Fig. 9. Do you think Augmented Reality learning method is an effective and interactive method to learn basic programming concept?

4 AR Learning System Performance Evaluation

This method requires enhancement in many different aspects since problems such as reflection, blurriness of ink printing on the marker, feature detection and limited animation of the augmented object occurs. As the marker is being printed on a smooth plastic surface A4 paper, the ink of the printing fades as the paper is incompatible with the printing machine. Due to limited resources, the printing of the paper can only be done using the University facility, which is a normal photocopy machine, which result in ink blurriness. In addition, the paper is susceptible to reflection which causes light reflection when light shown on it. This problem distorts the ability of the system to identify the marker which eventually leads to either the system couldn't recognize the marker pattern

or the system recognize the wrong marker. Other than that, due to the limited feature detection, it makes the marker detection harder since the feature detection is set default by Vuforia server itself.

The limited feature detection is also caused by the marker developed since unlike traditional marker, this marker aims to teach basic programming languages, where instead of QR code being printed, shapes are being printed on the marker, hence, only limited feature can be detected by the system. Lastly, the limited access on the action performed by the augmented animation creates problem for the system to show and effective portrayal of loop. Advanced programming is required to recode the animation if needed which falls out of the scope of this project. Thence, the limited animation causes the system to only be able to perform four types of action on the augmented object.

5 Discussion and Conclusion

In a nutshell, it is likely to find that this method according to the result of post-evaluation of survey will be able to help students to learn programming as it provides a better interaction for students to learn. This can be seen as 80% of control group students suggest that this is a good method to learn programming. This instate that the method effective in providing a better interactive for students to learn about basic programming concepts before venturing into harder ranges of programming.

References

1. Phillips, R.: Challenging the primacy of lectures: the dissonance between theory and practice in university teaching. *J. Univ. Teach. Learn. Pract.* **2**(1), 1–12 (2005)
2. Tiago, D.G., Antonio, C.S., Fábio, A.N., João, F.M., Humberto, F.F.: An Augmented Reality E-learning System for the Teaching of Robotic Morphology and Programming Logic, *Virtual and Augmented Reality Workshop*, Bauru (2012)
3. Wade, H., Victoria, L.C., Leyland, P.: Teaching cases in a virtual environment: when the traditional case classroom is problematic. *Decis. Sci. J. Innovative Educ.* **9**(3), 485–492 (2011)
4. Jorge, F.F.F.: Using gamification to enhance second language learning. *Digital Educ.* **27**, 32–54 (2015)
5. Hanna, K.K.: Using short video lectures to enhance mathematics learning - experiences on differential and integral calculus course for engineering students. *Inform. Educ.* **14**(1), 67–81 (2015)
6. Bettina, W., Helene, M., Ben, B.: Beyond the tweet, using twitter to enhance engagement, learning, and success among first-year students. *J. Mark. Educ.* **37**(3), 160–170 (2015)
7. Fotis, L., Panagiotis, P., Martin, W.: Multimedia augmented reality interface for e-learning (MARIE). *World Trans. Eng. Technol. Educ.* **1**(2), 173–176 (2002)
8. Jorge, M.G., Maria, D.M.F.: Augmented reality environments in learning, communicational and professional contexts in higher education. *Digital Educ. Rev.* **26**, 22–35 (2014)
9. Azuma, R.T.: A survey of augmented reality. *Massachusetts Inst. Technol.* **6**(4), 355–385 (1997)
10. Bacca, J., Baldiris, S., Fabregat, R., Graf, S., Kinshuk: Augmented reality trends in education: a systematic review of research and applications. *Educ. Technol. Soc.* **4**(17), 133–149 (2014)

11. Elizabeth, F., Rebecca, F., Anne, A., Mark, G., Yishay, M., Rhodri, T.: Augmented reality and mobile learning: the state of the art. *Int. J. Mob. Blended Learn.* **5**(4), 19 (2013)
12. Dipti, R.D.: Research on Object based augmented reality using unity3d in education system. In: *IJARIE*, Mumbai (2016)
13. G. o. S. Australia. ICAN, Student Mentoring and Youth Development: Student Engagement Matrix Guidelines. Youth Engagement and Inclusion

Middleware and Operating Systems

Efficient vCore Based Container Deployment Algorithm for Improving Heterogeneous Hadoop YARN Performance

SooKyung Lee, Min-Ho Bae, Jun-Ho Eum, and Sangyoon Oh^(✉)

Department of Computer Engineering, Ajou University, Suwon, Korea
syoh@ajou.ac.kr

Abstract. Hadoop has been widely utilized in processing a large-scale of data. Even though Hadoop-YARN has highly advanced processing performance, it still has performance limitations on heterogeneous environment. Its processing performance can be degraded when it is utilized in servers with different capabilities as well as the processes are scheduled for multiple users. To address this performance degradation problem caused by imbalanced load on heterogeneous environment, we propose an efficient vCore based container deployment algorithm that allocates the processing load for each container equally in order to minimize the deviation among processing task loads by controlling the number of vCores. The experiments show that our proposed method improved the performance on completion time by 18% on average.

Keywords: Container · Hadoop · Heterogeneous · vCore · YARN

1 Introduction

The MapReduce framework [1] has been widely accepted as the de-facto standard of the distributed parallel data processing paradigm for large-scale data. Hadoop, the most popular MapReduce framework, achieves a high-speed processing by dividing one large job into multiple small tasks and processing each task in an environment where the cluster is established with multiple nodes. With the new resource management platform YARN (Yet Another Resource Negotiator), Hadoop achieves a separation between the programming model and resource management by resource and node managers and application master. With this approach, various functions of the job tracker including resource management, monitoring and scheduling can be separated since the job tracker is the main origin of single-point failure and bottleneck for the previous Hadoop, and it can handle fault handling and resource management more effectively. In addition, Hadoop-YARN can dynamically allocate Map and Reduce processes by using the container whereas the previous Hadoop allocates Map/Reduce slots statically, which occurs a dissipation of resources. In Hadoop-YARN, the container is allocated and deployed by the number of vCores and the size of the memory per user's request. Thus, YARN can utilize resources in a more flexible and efficient manner than previous Hadoop by scheduling vCores and the memory [2].

The degradation of the performance in a heterogeneous environment is an inevitable problem even for Hadoop-YARN because Hadoop assumes its runtime environment as homogeneous. How to overcome the degradation due to this heterogeneity has become a critical challenge to the Hadoop runtime management [3]. Since the MapReduce follows Bulk Synchronous Parallel (BSP) paradigm, the overall performance of its cluster can be determined by the slowest process (called straggler). Thus, the difference of the performance among the physical nodes may cause poor load balancing, performance and utilization of runtime platform. In addition, the deviation of the performance worsens to guarantee the fairness among tenants in a multi-tenant environment either.

There have been various research approaches to resolve this task/cluster scheduling and tuning problem. However, these approaches require to calculate a complex formulas to get the difference of the capabilities of nodes when allocating tasks to each node to guarantee load balancing, the best-fitted node selection and the fairness among tenants. The implication of the problem becomes more serious as the number of nodes and the degree of the heterogeneity increases.

In this paper, we propose a novel approach to unify the performance of each containers, the unit for task processing in Hadoop-YARN, by determining the number of vCores. We aim to equalize the performance among containers in order to simplify the calculation that is required to guarantee the fairness and allocate tasks to its nodes in heterogeneous environment. It improves the performance of the task processing by enhancing the utilization of resources and minimizing the occurrence of the stragglers during processing. Usually, the accurate measurement on the performance is required to accomplish the equalization of the performance. However, the quantitative measurement on the performance of Hadoop is a challenging task because of various factors (e.g., various types of workload, processing phases of Hadoop and the implication of the setting parameters for Hadoop processing). The previous researches have determined the performance of Hadoop by measuring actual processing time when processing its tasks (i.e., performance profiling) [4, 8]. However, they require excessive resources to actually process the tasks and measure the performance, and it is limited to estimate the expected performance when adjusting the unit of task processing. Thus, we consider that it is not adequate to apply their approaches on our research. On the other hand, the approaches used for benchmarking virtualized resource generally determine the performance based on the architectural metrics, and it can simplify the estimation of the completion time when the unit changes. Therefore, we aim to get a simple but effective indicator by adopting the approaches to benchmark virtualized resource. In this paper, we determine the number of vCores for each node based on the indicator.

This paper is organized as follows; related works are presented as well as describing the container and vCore in YARN in Sect. 2. We present our proposed method to allocate containers to guarantee the fairness as well as to enhance the utilization in the heterogeneous environment in Sect. 3. In Sect. 4, we present our evaluation results, and we conclude this paper with future works in the Sect. 5.

2 Background

There have been various researches to utilize Hadoop in the heterogeneous environment because it is originally assumed that Hadoop works in the homogeneous. In this section, we introduce related works regarding previous researches on Hadoop and its benchmark methods. We also cover the allocating method of the containers and the working mechanism of the vCore in YARN.

2.1 Handling Heterogeneity Through Scheduling

Task Scheduling. In Hadoop, task scheduling with speculative execution is essential to minimize processing time. The speculative execution aims to migrate the task from the slower node to the faster node. The LATE [3] introduced a method to determine a straggler by estimating the completion time because the previous Hadoop determines a straggler based on the progress score which is an inadequate indicator in the heterogeneous environment. There are researches on marking the straggler or determining the node to process the task with more subdivided criteria [4], and applying data placement for processing [5].

Cluster Scheduling with Heterogeneous Aware. The consideration on the difference of the capability of each node is important to enhance the performance. In Tarazu [6], the degradation of the performance due to the heavy shuffle is addressed. However, it does not take a multi-tenant environment into account for its solution. Other researches such as the method that improves the performance by determining adequate nodes and scheduling to allocate tasks based on the characteristics of the task (i.e., CPU or I/O intensive tasks) [7] have been introduced, and the method to establish a model that acquires the level of capability for each node and then groups nodes into sub-clusters for scheduling [8]. However, they still require computation to determine optimum method to allocate tasks.

Tuning Approach. Tuning is a method that adjusts Hadoop parameters to achieve the best performance with optimized ones. The ANT [9] introduced a method using a self-adaptive tuning. However, this approach still determines optimized setting parameters by actual job profiling. Since there are some setting parameters that require re-run of Hadoop, this approach is limited in dynamic configuration setting.

2.2 Benchmark on Virtualized Resource

The performance measure of virtualized resources are frequently used as the metric for VM consolidation. The vConsolidate [10] is a popular benchmark tool that measures the performance of CPU by using IPS (Instruction per Second) and MPI (Misses per Instruction). IPS and MPI measures are very effective to measure the performance of multi-core and multiple nodes [11]. In this regard, the disk I/O performance can be effectively measured from the IOPS benchmark [13].

2.3 Container and vCore in YARN

In Hadoop-YARN, the size of the container is determined by the number of vCores and required memory size. For container allocation, YARN tries to avoid the case where the size of the container is exceeding the size of available resource. It is generally recommended that the size of the vCore should be determined according to the number of physical cores or disks. Even though the task may be slowed down or become faster if an arbitrary number is applied, the task is never be failed (i.e., no failure). However, if the number of vCore is not set adequately, it can cause negative impacts on the CPU utilization. For instance, if less vCores than available physical cores is set by the user, there could remain idle physical cores that are not occupied by any vCore.

3 vCore Based YARN Container Deployment for Heterogeneous Environment

Our proposed method determines the adequate number of vCores on each node in order to equalize the processing unit on each node. It can enhance the utilization and the fairness among tenants. In this section, we present the performance factors that impact on the performance of the container and the equalizing unit, and we cover the method for equalization of the containers. We also present how our proposed method can satisfy the fairness.

3.1 Modeling the Performance Impact Factors

To equalize the performance of each container based on the vCore, we first need to calculate the performance difference between nodes. We measure the IPS of a single core to measure the performance of the CPU, and then we multiply the number of physical cores and our measured IPS to get the capacity. The MPI which is another critical metric for the measurement includes the weight on the cache hit rate. We exclude the cache hit rate from our measurement because the cache hit rate is no importance that the MapReduce clears the memory before processing another task even for repetitive tasks. The previous researches on the performance of virtual machines (VM) [12, 14] include the virtualization overhead, core contention and cache contention as main metrics for their measurement whereas every established container in YARN has identical overhead. In addition, we found that the deviation of the completion time on the change of the container remains less than 1% when the same amount of task is processed during our experiment which is presented in Appendix A. According to these observation, we assume that the contention among the Map/Reduce tasks are not significantly large.

Several researches to enhance the performance of the MapReduce framework have presented that the completion time for the entire job depends on not only the performance of the CPU but also that of I/O. They indicated that the performance of I/O results in the difference of the performance in the heterogeneous environment [9]. Therefore, it is

required to consider the performance of both CPU and I/O at the same time to realize the performance equalization of the containers. We use the IOPS, the metric that has been widely utilized to analyze the I/O performance on the virtualized environment as well as on the desktop platform. The capacity of one node can be represented as the Eq. (1) below:

$$(IOPS_k \times No. \text{ of } cores_k) \times W_{CPU} + IOPS_k \times W_{I/O} = capacity_k \quad (1)$$

It is much more challenging to represent the implication of the performance of CPU and I/O on Hadoop in one equation, and we calculate its overall performance with W_{CPU} and $W_{I/O}$. We observed that most of the jobs on Hadoop require disk I/O processing in general. Therefore, we present the performance by adding the value for I/O with the value for the CPU performance. We conducted an experiment to acquire the distribution of processing time on the variation of the weight value, and we observed that the weight value on the Eq. (2) results in the minimum deviation of the performance in our experimental environment. We then measure the performance of each node with this equation. Appendix B contains detailed information on this calculation.

$$(IOPS_k \times No. \text{ of } cores_k) \times 0.8 + IOPS_k \times (0.2 \times 1 \text{ Billion}) = capacity_k \quad (2)$$

3.2 Container Equalization

The experiments that we conducted indicate that the sum of the tasks on a node is not changed significantly if we allocate more vCores than the number of physical cores. Also, the overall performance of all containers can be equalized if the number of vCores is adjusted to produce equal or similar performance. The Appendix B contains the detail information on this experiment and our assertion.

Algorithm 1. The algorithm to select the number of vCores

- 1: $capa_{max} = \max[capa_i], i \in \{1, \dots, k\} : nodes$
 - 2: $vCore_{num} = No. \text{ of } physical \text{ cores in } capa_{max}$
 - 3: **while** No. of vCore is not selected **do**
 - 4: $capa_{criteria} = capa_{max} \div vCore_{num}$
 - 5: **for** $i = 0$ **to** k
 - 6: $vCore_i = round(capa_i \div capa_{criteria})$
 - 7: $rest = abs[vCore_i - (capa_i \div capa_{criteria})]$
 - 8: **if** No. of physical cores _{i} > $vCore_i$ **or** $rest > threshold \times No. \text{ of } physical \text{ cores}_i$
 - 9: $vCore_{num} = vCore_{num} + 1$
 - 10: No. of vCore is not selected
-

We present our proposed algorithm that determines the number of vCores for the container in the heterogeneous environment. First, it selects the node with the maximum capacity. It then selects the criteria of the capacity for each vCore. Finally, it searches the number of vCores that meets the following criteria from all nodes; (1) the number

of vCores should be larger than the number of physical cores, and (2) the difference of the performance among vCores should not exceed the threshold.

We apply criteria (1) in order to avoid the degradation of the utilization when there are less vCores than the number of physical cores whereas the completion time of the task does not change significantly when there are more vCores than the number of physical cores. We then apply the threshold on criteria (2) because the bottleneck effect on the I/O occurs and the requirement of the memory becomes larger as there are more vCores even it is ideal that every vCore has exactly same capacity. To mitigate these problems, we attempt to find the minimum value with the threshold by partly accepting the deviation of the performance as we consider the difference within 10% as identical.

3.3 Fairness Guarantee

The fair-scheduler on Hadoop satisfies the fairness by allocating the same number of containers for each job (i.e., to allocate the same duration of CPU usage for multiple users). However, this approach does not guarantee the fairness in the heterogeneous environment because it is based on the assumption that the performance of each container is identical. Our proposed method can partly guarantee the fairness by simply using the number of allocated containers as the metric (i.e., it equalizes the performance of each container) even though a complete fairness cannot be achieved.

The DominentResourceCalculator of the capacity-scheduler uses the number of vCores as an essential metric to measure the fairness. Since the performance of the vCore in the heterogeneous environment, it also cannot guarantee the fairness. However, our proposed method can partly achieve the fairness because it equalizes the performance of each container. The DefaultResourceCalculator allocates the resource based on the size of RAM itself, we do not address this factor in this paper because we focus on equalizing computing resource including CPU and I/O.

As we mentioned above, there can be an occasion that the performance of each vCore is not completely identical. In this case, we estimate that the gap in the fairness among tenants can be significantly reduced even though the fairness cannot be completely guaranteed. However, we did not address this topic in this paper as well in order to focus on our approach.

4 Evaluation

In this section, we present our evaluation on the degree of the equalization and the measurement of the performance. Table 1 shows the experimental environment for our evaluation.

Table 1. Experimental environment

Node	CPU model	Frequency (GHz)	No. of physical cores	IPS (billion)	IOPS	Capacity (billion)
Node A	Intel Core i7-5930	3.50	6	625	65	4,301
Node B	Intel Xeon E3-1231	3.40	4	647	64	3,353
Node C	Intel Xeon E3-1231	3.40	4	639	57	3,184
Node D	Intel Core i5-3450	3.10	4	500	67	2,942
Node E	Intel Xeon E3-1220	3.10	4	394	56	2,380
Node F	Intel Xeon X3430	2.40	4	348	58	2,274

4.1 The Degree of Equalization

We conducted our experiment both on Hadoop with default recommended setting and the one by our proposed method under our experimental environment. The results of the experiment is presented on Table 2.

Table 2. Evaluation on equalization

Job name	Default		Our proposed method	
	Aver. map time (sec)	Stdev ($\sqrt{\sigma^2}$)	Aver. map time (sec)	Stdev ($\sqrt{\sigma^2}$)
PI/TG	174.7/281.8	56.2/111.6	259.5/328.5	33.7/75.2
TS/WC	11.5/10.6	9.4/10.8	14.9/13.7	3.8/5.9

We observed that the processing speed of each container becomes similar even the task processing speed for each Map becomes slower since the amount of the task for the container is determined by that of the slowest node. However, the effectiveness of our proposed method is not significant when the given job is excessively CPU or I/O intensive such as PI or TG (Teragen). We estimate that it results in our proposed method considers that both CPU and I/O are equally important through the weight values on CPU and I/O for the measurement of the performance. On the contrast, the results of TS (Terasort) and WC (Wordcount) show that our method is effective to minimize the deviation of performance among containers.

4.2 Utilization and Fairness

We measured the utilization of the CPU and the fairness by measuring the result based on the fair scheduler when five tenants submit jobs at the same time. We classified the result that the job from one or two tenants occupies entire cluster as outliers and excluded them from our measurement because the cluster becomes idle as there are no more jobs

for the cluster. The result on the utilization is calculated based on the peak time. Table 3 shows the results of the experiment.

Table 3. Utilization and fairness evaluation

Job name	Default			Our proposed method		
	Avg. Cmpl. time (sec)	Stdev ($\sqrt{\sigma^2}$)	Avg. CPU Util.	Avg. Cmpl time (sec)	Stdev ($\sqrt{\sigma^2}$)	Avg. CPU Util.
PI/TG	962/752	146/291	72/55	741/548	64/142	88/68
TS/WC	831/778	238/201	44/45	726/643	130/97	66/57

We observed that the overall CPU utilization and the return time have improved by 28% and 21% on average, respectively. It shows that our proposed method reduces the occurrence of the straggler and improves the utilization since the powerful node can have more tasks to be processed simultaneously while the occurrence of the straggler strongly impacts on the entire job when processing multi-jobs. In addition, Hadoop with default recommended setting shows a high deviation on the completion time while our proposed method reduced such deviation significantly. Table 4 shows the completion time of the entire job when we assigned each job ten times. We evaluate that our proposed method improved the performance on the completion time by 18% on average.

Table 4. Average completion time for the entire job

Job name	Default (sec)	Our proposed method (sec)	Improvement (%)
PI/TG	3,142/2,457	2,424/1,878	22.8/23.5
TS/WC	569/2,499	499/2,201	12.3/11.9

5 Conclusion

The homogeneous environment that Hadoop assumes can occasionally be broken down, which can result in the degradation of its performance. In this paper, we aim to guarantee the fairness and improve the utilization by equalizing each container by determining the number of vCores. Our proposed method shows that it reduces the deviation of the fairness significantly even though it does not guarantee the complete fairness. In addition, it improves the performance on the completion time when processing multiple jobs by 18% on average. In addition, we represent the correlation between CPU and I/O with weight values since it is limited to establish a comprehensive model to represent the implication of the CPU and I/O performance on Hadoop. We estimate that the allocation of the container can be more sophisticated if we can establish a more detailed mathematical model for this correlation. In addition, we note that the effectiveness on the improvement of the utilization could be reduced if there are more resources available for the processing. However, it still can contribute to not only the guarantee of the fairness and but also the simplification of task processing and cluster scheduling.

We will follow up on our research in this paper to improve data locality with variation of the number of containers. We will focus on improving data locality since we estimate

that the powerful nodes can have more containers in general, and the HDFS still manage the blocks in a balanced manner across the cluster. The distribution and storage of the blocks according to the number of containers can improve the performance with enhanced data locality.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01059557).

This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & Communications Technology Promotion) (R22151610020001002).

Appendix A: The Variation on the No. of vCores and Job Completion Time

We conducted the experiment to evaluate the implication on the performance on Hadoop when we set the vCore as both default recommended setting and abnormal setting. We changed the number of containers by changing the number of vCores on the node A on Table 1 in Sect. 4. Table 5 shows the jobs utilized for the experiment, and we processed each job 30 times in our experimental environment. The Map wave indicates that how many reallocation of the container is required to complete the assigned job. We observed that the CPU utilization was degraded when the number of containers went below the number of physical cores. In addition, the job completion time was delayed when there are more Map waves, which occurs because the time to reallocate the Map task to the container after a single map task is completed becomes an overhead for the job processing. However, it indicates that there was no significant change on the overall job completion time. We estimated that it occurs because the MapReduce does not require any synchronization or communication among tasks and there is little possibility to occur any bottleneck effect except the I/O. In addition, the vCore scheduling follows the policy of the JVM, and its policy is optimized with threading even it is subdivided more than physical core level with its multi-core and multi-threading technologies. Therefore, the throughput for one node does not change significantly, and it implies that the performance among containers can be equalized by adjusting the number of vCores so that all the containers have identical or similar level of performance.

Table 5. Job configuration and result

Job name	Map tasks	No. of containers	Map wave	Avg. map time (sec)	Avg. CMPL. time (sec)	Avg. CPU Util. (%)
PI 24/12	24/24	24/12	1/2	610/313	649/640	93/97
PI 6/3	24/24	6/3	4/8	159/160	666/1,298	84/67
TS_24/12	744/744	24/12	31/62	111/52	3,457/3,255	71/62
TS 6	744	6	124	27	3,410	65

Appendix B: The Result of the Variation on the Weight

We also conducted the experiment with weight variation in order to evaluate how the weights on the CPU and I/O impact on the allocation of the container. The experimental environment is identical to that in the Table 1 in Sect. 4, and Table 6 shows the results of the experiment. The weight value with no experimental result implies that there was no change on the allocation of the containers. The results of the experiment are presented in Fig. 1. We identified that the weight labels #1–3 on Table 6 minimize the distribution. We conducted the experiment with the weight label #2 since it shows the least standard deviation among the weight values.

Table 6. The changes on the number of vCores on the adjustment of the weight

Weight label	CPU weight	I/O weight	The number of vCores					
			Node A	Node B	Node C	Node D	Node E	Node F
1	90	10	8	6	6	6	4	4
2	80	20	8	6	6	5	4	4
3	70	30	7	6	5	5	4	4
4	60	40	6	5	5	5	4	4
5	40	60	6	5	5	5	4	5
6	30	70	6	6	5	6	5	5
Default	-	-	6	4	4	4	4	4

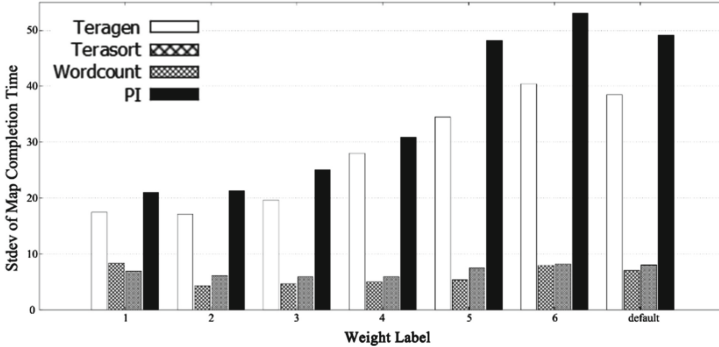


Fig. 1. The distribution of the weight values and completion time

References

1. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**, 107–113 (2008)
2. Vavilapalli, V.K., Murthy, A.C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S.: Apache Hadoop yarn: yet another resource negotiator. In: 4th Annual Symposium on Cloud Computing, p. 5. ACM, Santa Clara (2013)

3. Zaharia, M., Konwinski, A., Joseph, A.D., Katz, R.H., Stoica, I.: Improving MapReduce performance in heterogeneous environments. *OSDI* **8**, 7 (2008)
4. Chen, Q., Liu, C., Xiao, Z.: Improving MapReduce performance using smart speculative execution strategy. *IEEE Trans. Comput.* **63**, 954–967 (2014)
5. Xie, J., Yin, S., Ruan, X., Ding, Z., Tian, Y., Majors, J., Manzanares, A., Qin, X.: Improving MapReduce performance through data placement in heterogeneous Hadoop clusters. In: 2010 IEEE International Symposium on Parallel and Distributed Processing, pp. 1–9. IEEE, Atlanta (2010)
6. Ahmad, F., Chakradhar, S.T., Raghunathan, A., Vijaykumar, T.: Tarazu: optimizing MapReduce on heterogeneous clusters. *ACM SIGARCH Comput. Archit. News* **40**, 61–74 (2012)
7. Gupta, S., Fritz, C., Price, B., Hoover, R., Dekleer, J., Witteveen, C.: ThroughputScheduler: learning to schedule on heterogeneous Hadoop clusters. In: 10th International Conference on Autonomic Computing, pp. 159–165. ICAC, San Jose (2013)
8. Xiong, R., Luo, J., Dong, F.: Optimizing data placement in heterogeneous Hadoop clusters. *Cluster Comput.* **18**, 1465–1480 (2015)
9. Cheng, D., Rao, J., Guo, Y., Zhou, X.: Improving MapReduce performance in heterogeneous environments with adaptive task tuning. In: 15th International Middleware Conference, pp. 97–108. ACM, Bordeaux (2014)
10. Casazza, J.P., Greenfield, M., Shi, K.: Redefining server performance characterization for virtualization benchmarking. *Intel Technol. J.* **10**, 243–251 (2006)
11. Tickoo, O., Iyer, R., Illikkal, R., Newell, D.: Modeling virtual machine performance: challenges and approaches. *ACM SIGMETRICS Perform. Eval. Rev.* **37**, 55–60 (2010)
12. Wood, T., Shenoy, P., Venkataramani, A., Yousif, M.: Sandpiper: black-box and gray-box resource management for virtual machines. *Comput. Netw.* **53**, 2923–2938 (2009)
13. Gulati, A., Ahmad, I., Waldspurger, C.A.: PARDA: proportional allocation of resources for distributed storage access. *FAST* **9**, 85–98 (2009)
14. Iyer, R., Illikkal, R., Tickoo, O., Zhao, L., Apparao, P., Newell, D.: VM 3: measuring, modeling and managing VM shared resources. *Comput. Netw.* **53**, 2873–2887 (2009)

A Real-Time Operating System Supporting Distributed Shared Memory for Embedded Control Systems

Yuji Tamura, Doan Truong Thi, Takahiro Chiba,
Myungryun Yoo, and Takanori Yokoyama^(✉)

Tokyo City University, 1-28-1, Tamazutsumi, Setagaya-ku, Tokyo 158-8557, Japan,
{myoo, tyoko}@tcu.ac.jp

Abstract. The paper presents a real-time operating system (RTOS) that supports distributed shared memory (DSM) for distributed embedded control systems. The RTOS provides a location-transparent environment, in which distributed software modules can exchange input and output values through the DSM. The RTOS is an extension to OSEK OS and it utilizes a real-time network called FlexRay. The consistency of the DSM is maintained according to the order of data transfer through FlexRay, not using inter-node synchronization. The worst case response time of the DSM is predictable if the FlexRay communication is well configured.

Keywords: Operating systems · Real-time systems · Embedded systems · Distributed shared memory · Distributed control systems

1 Introduction

An application program of an embedded control system such as an automotive control system is designed as a set of software modules, which are executed by tasks on a real-time operating system (RTOS) such as OSEK OS [1]. Model-based design is also widely adopted in embedded control software design. A controller model is designed and verified using a model-based design tool such as MATLAB/Simulink [2] in model-based design. The source code of software modules can be generated from the controller model by a code generator such as Real-Time Workshop/Embedded Coder [2]. The generated software modules exchange their input and output values through global variables. However, if we build a distributed control system with the software modules on a message-based communication environment such as OSEK COM [3], we have to rewrite the source code to exchange input and output values by messages, not global variables.

Distributed shared memory (DSM) provides location-transparent shared variables, through which distributed software modules can exchange their input and output values. Existing DSM systems are, however, not suitable for embedded control systems. Most DSM systems are based on page-based DSM [4, 5], the response time of which is difficult

Y. Tamura—Presently with Fujitsu Public Solutions Limited.

T. Chiba—Presently with Systems Engineering Consultants Co., LTD.

to predict. It is also difficult to implement a page-based based DSM mechanism in a small RTOS with no virtual memory on a microcontroller without MMU (Memory Management Unit), which is widely used in embedded control systems.

The goal of the research is to develop a RTOS with DSM for embedded distributed control systems. We have already presented the previous version of the RTOS with DSM as an extension to OSEK OS, which supports a consistency model called partially-sequential consistency [6]. This paper presents a stronger consistency model called equivalently-sequential consistency, which is equivalent to sequential consistency in cyclically executed distributed control software.

The rest of the paper is organized as follows. Section 2 describes a DSM model and Sect. 3 describes the specification and the implementation of the DSM. Section 4 describes experimental evaluation and Sect. 5 concludes the paper.

2 Distributed Shared Memory Model

2.1 Design Policies

A controller model is usually built as a set of subsystem blocks using MATLAB/Simulink in model-based design. Figure 1 illustrates a Simulink model, which consists of *SubsystemX*, *SubsystemY*, *SubsystemZ* and *SubsystemW*.

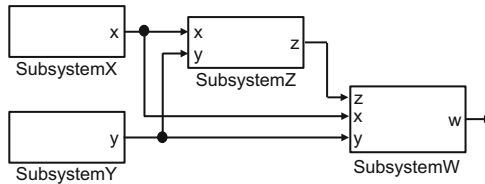


Fig. 1. Example Simulink model

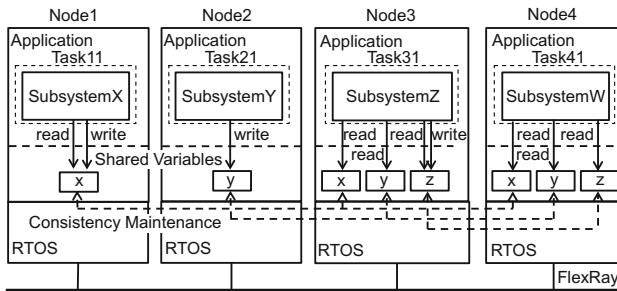


Fig. 2. Example distributed control software

Figure 2 illustrates the structure of example distributed control software with DSM. Software modules generated from the Simulink model are distributed to four nodes. *SubsystemX*, *SubsystemY*, *SubsystemZ* and *SubsystemW* are respectively executed by

Task11 on *Node1*, *Task21* on *Node2*, *Task31* on *Node3* and *Task41* on *Node4*. The copies of shared variables are located on each node. For example, the copies of x are located on *Node1*, *Node3* and *Node4*.

The DSM mechanism is based on a shared-variable DSM architecture [5], not a page-based DSM architecture, because only certain variables are shared in distributed control software developed with MATLAB/Simulink.

The design policies of the DSM are shown below [6].

- A MMU should not be used because most microcontrollers used in embedded control systems have no MMU.
- Inter-node synchronization should not be used because it may cause a performance problem (Intra-node synchronization (inter-task synchronization) is acceptable).
- No new API of RTOS for DSM is required because new API may violate the compatibility (an extension of the semantics of an API is acceptable).
- Consistency sufficient for the control software generated from Simulink models should be provided.

The RTOS utilizes a TDMA-based real-time network called FlexRay [7] to predict the worst case response time. The maximum communication delay time is predictable if FlexRay communication is well configured.

2.2 Consistency Model

Some consistency models for DSM have been presented [4, 5, 8]. Sequential consistency [9] is desirable because the same write operation sequence is observed by every node. However, strict sequential consistency is not needed for cyclically executed software generated from Simulink models, because input and output values of software modules are not events but states. A written value may be overwritten before being read. The semantics of the DSM consistency is similar to the semantics of the state message [10], which is called unqueued message in OSEK COM [3].

We present two consistency models of the DSM: partially-sequential consistency and equivalently-sequential consistency. The former maintains the write operation sequence for each shared variable and the latter maintains the write operation sequence for all shared variables. Partially-sequential consistency and equivalently-sequential consistency are selectable.

The consistency is maintained according to the sequence of data transfer through FlexRay. A read operation is inhibited for a certain interval after the write operation as shown later. The inhibition is needed just for tasks that perform both read and write operations, not for tasks that perform just write operations or read operations.

Figure 3 illustrates an example DSM access sequence in the case of Fig. 2. The operations performed by a task on each node are shown horizontally, with time increasing to the right. The notation $r(x,a)$ means that a task reads the value a from the variable x . The notation $w(x,b)$ means that a task writes the value b into the variable x . The value of a shared variable written by a task is transferred to other nodes through FlexRay. FlexRay communication is periodically performed with a communication cycle. The notation $t(x,b)$ means that the value b of the variable x is transferred.

Transferred values are received at the beginning of the communication cycle. For example, the value b of the variable x is transferred during the n th cycle and is received by *Node3* and *Node4* at the beginning of the $n + 1$ th cycle.

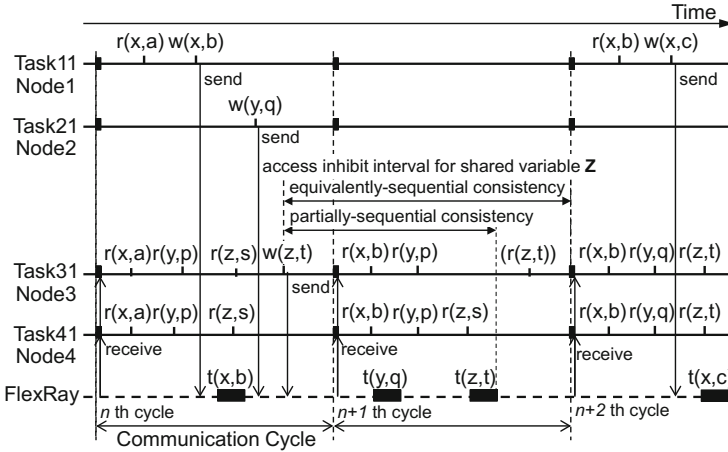


Fig. 3. Example DSM access sequence

Partially-sequential consistency is realized by inhibiting a read operation until the data transfer completes. After writing t into z , *Task31* cannot read z until the data transfer is completed in this case. The same write operation sequence for the variable z is observed by every node. However, the different write operation sequences for all shared variables may be observed by nodes.

Equivalently-sequential consistency is realized by inhibiting a read operation until the next communication cycle. After writing t into z , *Task31* cannot read z until the data is received by other nodes in this case. The same write operation sequence is observed by every node as follows: $w(x,b)$, $w(y,q)$, $w(z,t)$, $w(x,c)$.

We assume the FlexRay communication cycle period is sufficiently shorter than the periods of application tasks. The typical FlexRay communication cycle is 1 ms and the typical period of automotive application tasks is 10 ms or longer. So the access inhibit intervals are also sufficiently shorter than the interval time from the write operation to the next cycle's read operation performed by periodic tasks. The inhibition does not influence the behavior of the application in most cases.

3 Specification and Implementation

3.1 API and OIL for Distributed Shared Memory

OSEK OS provides resource access system calls for mutual exclusion: *GetResource()* and *ReleaseResource()*. We extend the semantics of the system calls for DSM. A shared variable or a set of shared variables is dealt with as a distributed shared resource. A task calls *GetResource()* and *ReleaseResource()* to access a distributed shared variable.

Figure 4 shows a fragment of example source code of application program. The name of the shared variable is *sharedData0* and the identifier of the resource for *sharedData0* is *ResourceSharedData0*.

```

. . . . .
/* get the resource for the shared variable */
GetResource(Resource_sharedData0);
/* update the shared variable */
sharedData0 = a * sharedData0 + b;
/* release the resource for the shared variable */
ReleaseResource(Resource_sharedData0);
. . . . .

```

Fig. 4. Example source code

The configuration of an OSEK application is described in OIL (OSEK Implementation Language) [11]. We extend OIL to declare distributed shared variables. Figure 5 shows an example OIL description, in which the shared variable *sharedData0* is declared. Its data type is *long* in C language, its initial value is zero, and equivalently-sequential consistency is selected. The task *Task11* on *CPU1* shares the shared variable *sharedData0*. We also extend the system generator (SG) to generate DSM configuration data.

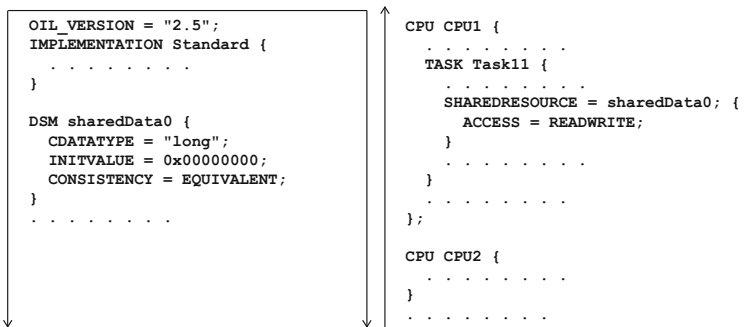


Fig. 5. Example OIL description

3.2 Distributed Shared Memory Mechanism

We developed a distributed RTOS with location-transparent system calls [12] as an extension to TOPPERS/ATK1 [13], an OSEK-compliant operating system. We extend the RTOS to support DSM by adding a distributed shared memory module, which manages the copies of shared variables and maintains the consistency. The copies of shared variables are allocated in the data section of application program on each node. The RTOS has shared data buffers and received data buffers. The messages for DSM are transmitted in the dynamic segment of FlexRay communication.

Figure 6 shows a time chart of the DSM processing on the writer node. When an application task calls *GetResource()*, the RTOS executes the processing of the original

GetResource() of OSEK, and then executes the DSM access preprocessing, which copies the value of the shared data buffer to the shared variable.

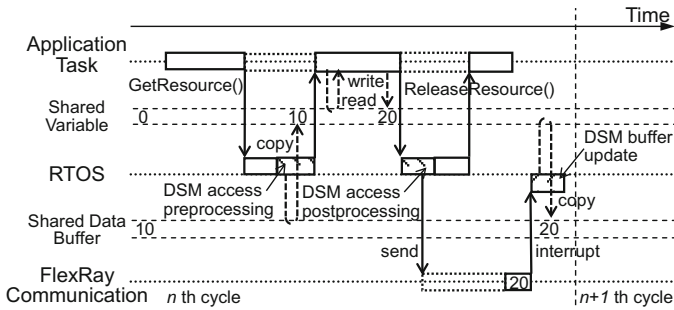


Fig. 6. Time chart of writer node

When the application task calls *ReleaseResource()*, the RTOS executes the DSM access postprocessing, which compares the value of the shared variable and the value of the shared data buffer, and calls the FlexRay driver to send the former value if the values are different. Then the RTOS executes the processing of the original *ReleaseResource()* of OSEK OS. An interrupt is activated when the data transfer is completed. The interrupt executes DSM buffer update, which copies the value of the shared variable to the shared data buffer.

Figure 7 shows a time chart with an access inhibit interval for equivalently-sequential consistency. When the application task calls *GetResource()* before completing the communication cycle during which the data transfer is performed, the RTOS changes the state of the task to waiting. The cycle start processing of the next communication cycle changes the state of the task to ready. In case of partially-sequential consistency, the task state change is performed by the data transfer completion interrupt, not by the cycle start processing.

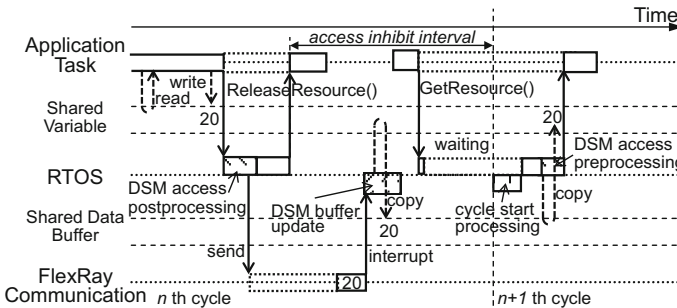


Fig. 7. Time chart with access inhibit interval

Figure 8 shows a time chart of the DSM processing on the reader node. The cycle start processing executes received data update, which interprets the received data and

writes the received value into both the received data buffer and the shared data buffer if no task holds the DSM resource as shown in Fig. 8. If the resource is held by a task, the RTOS writes the value into just the received data buffer.

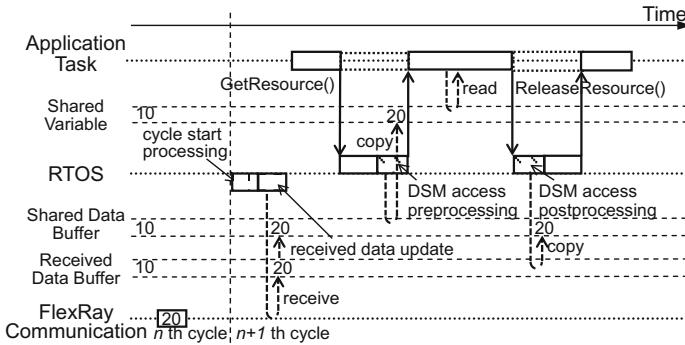


Fig. 8. Time chart of reader node

If FlexRay communication is well configured, the maximum communication delay time is predictable. The maximum total time of the DSM access postprocessing execution time and the FlexRay communication delay time is twice the communication cycle period in the worst case, in which data transfer is postponed to the next communication cycle. So the worst case response time is about twice the communication cycle period because the cycle start processing execution time and the received data update execution time are sufficiently less than the communication cycle period.

4 Experimental Evaluation

We have developed the RTOS with DSM on an evaluation board called GT200N10, the CPU of which is V850E/PH03 with an on-chip E-Ray FlexRay controller. The clock rate of the CPU is 128 MHz. The data transfer rate of FlexRay is 10 MHz and the communication cycle period is 1 ms.

We have measured the CPU execution times of DSM functions: DSM access preprocessing, DSM access postprocessing, cycle start processing, received data update, and DSM buffer update. Table 1 shows their average values.

We think each execution time is practically small for automotive control systems. The typical period of the automotive control application periodic tasks is 10 ms or more. The dominant factor of the response time of the DSM is the FlexRay communication delay time, so the worst case response time is about twice the communication cycle. We think the response time is practically small.

Table 1. Execution time of DSM mechanism

Processing		Execution time [μ sec]				
		1Byte	2Byte	4Byte	8Byte	16Byte
DSM Access preprocessing		1.2	1.2	1.4	1.7	2.3
DSM access postprocessing	With data transfer	10.7	10.8	11.0	11.3	11.9
	Without data transfer	1.3	1.4	1.6	2.1	2.9
Cycle start processing		31.9	32.0	32.2	32.5	33.1
Received data update		1.5	1.5	1.7	2.0	2.6
DSM buffer update		1.6	1.7	1.8	2.2	2.8

5 Conclusion

We have presented a RTOS that supports DSM for distributed embedded control systems. The consistency of the DSM is maintained according to the order of data transfer through FlexRay, not using inter-node synchronization. The worst case response time is predictable if the FlexRay communication is well configured. We have also evaluated the performance of the DSM. According to the evaluation results, we think the performance is practically sufficient for automotive control applications.

Acknowledgment. We would like to thank the developers of TOPPERS/ATK1. This work was supported in part by JSPS KAKENHI Grant Number JP24500046 and JP15K00084.

References

1. OSEK/VDX, Operating System, Version 2.2.3 (2005)
2. The MathWorks Inc. <http://www.mathworks.com/>
3. OSEK/VDX, Communication, Version 3.0.3 (2004)
4. Tanenbaum, A.S.: Distributed Operating Systems. Prentice Hall, New Jersey (1995)
5. Protic, J., Tomasevic, M., Milutinovic, V.: Distributed shared memory: concepts and systems. *IEEE Parallel Distrib. Technol.: Syst. Appl.* **4**(4), 63–71 (1996)
6. Chiba, T., Yoo M., Yokoyama, T.: A distributed real-time operating system with distributed shared memory for embedded control systems. In: *Proceedings of IEEE 11th International Conference on Dependable, Autonomic and Secure Computing*, pp. 248–255 (2013)
7. Makowitz, R., Temple, C.: FlexRay - a communication network for automotive control systems. In: *Proceedings of 2006 IEEE International Workshop on Factory Communication Systems*, pp. 207–212 (2006)
8. Adve, S.V., Gharachorloo, K.: Shared memory consistency models: a tutorial. *IEEE Comput.* **29**(12), 66–76 (1996)
9. Lamport, L.: How to make a multiprocessor computer that correctly executes multiprocess programs. *IEEE Trans. Comput.* **C-28**(9), 690–691 (1979)
10. Kopetz, H., Merker, W.: The architecture of MARS. In: *Proceedings of 24th International Symposium on Fault-Tolerant Computing*, pp. 50–55 (1995)
11. OSEK VDX, OSEK/VDX System Generation OIL: OSEK Implementation Language Version 2.5 (2004)

12. Chiba, T., Itami, Y., Yoo, M., Yokoyama, T.: A distributed real-time operating system with location-transparent system calls for task management and inter-task synchronization. In: Proceedings of IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1133–1138 (2011)
13. TOPPERS Project. <http://www.toppers.jp/en/>

Security and Privacy

MBR Image Automation Analysis Techniques Utilizing Emulab

Gibeom Song and Manhee Lee^(✉)

Department of Computer Engineering, Hannam University, Daejeon, Korea
ssonggibum@gmail.com, manheelee@gmail.com

Abstract. Virtual environment is frequently used for malware analysis. To hide their behavior, malware began to adopt virtual environment detection techniques. One of trickiest things when analyzing malware on real systems is that the operating system became unbootable due to the crash of partition and boot loader stored in the first sector of hard disk called the master boot record (MBR). It is quite time consuming to extract its MBR image from the crashed hard disk, so running malware on real system is usually considered as the last resort. In this research, we proposed a malware analysis system utilizing Emulab to extract crashed MBR images very easily.

Keywords: Emulab · Virtualization · Malware · Analysis automation

1 Introduction

While the number of malware keeps increasing very quickly, their behaviors are also evolving. According to [1], the number of unique malware is more than 430 million. Since nobody knows how many malware is undetected, it is not easy to estimate the total number of malware.

To cope with this situation, the automatic malware analysis becomes necessary. Generally, malware analysis is categorized into static or dynamic analysis. In static analysis, a malware under test is not run, but its binary code is analyzed by malware analysts. In dynamic analysis, in contrast, analysts make malwares run in a test environment and collect behavior information for further analysis. Owing to advances in hardware virtualization techniques, it becomes easier to utilize virtual machines for automatic malware analysis.

However, to evade this analysis trend, hackers started to use virtual environment detection techniques. It was reported that 28% of malwares found in 2014 are reported to have the virtual environment detection function [2]. When a malware detects any virtual environment, it quits or disguises as normal applications by performing naïve operations. Many researchers are trying to solve this problem in many ways, but there is not a practical solution yet.

In our previous study, we used a completely different approach; real machines are used, not virtual machines. We are not the first to use this approach. Some authors already proposed to use real machines which are called bare

metal systems in [3,4]. Different from previous research, we utilized the existing research facility, Emulab, developed by Utah University [5]. Its main benefit is to dynamically assign real machines running on various OSES with any network topologies. Many researchers found Emulab very useful for network and security research [6]. Our previous study showed that Emulab is almost like bare metal systems by showing that many virtual environment detection functions available to us could not detect Emulab as virtual environments [7]. In addition, we showed how to extract the MicroSoft Windows MBR image from a hanged Emulab system.

Although our idea could save researchers' time and efforts, there are two main disadvantages. First, our idea is not scalable because every command is executed one by one. So, it is not adequate for automatic analysis. Second, users need to check the system's status often and they had to extract the system's MBR image for testing whether the malware destroyed or modified that part.

This research focused on automating the above operations. When a large number of nodes are used, the automatic analysis will be very essential. In addition, we tried to provide useful information about the automatically extracted MBR images that will help analysts by saving time for basic test. We use email to send the information to users so that they do not have to access Emulab for getting the information, which will be very convenient to them.

2 Related Work

2.1 Malware Testing Environment Validation for Emulab

The most important prerequisite to use Emulab for malware analysis is that virtual environment detection techniques should not differentiate Emulab from real machines. We use Paranoid fish (Pafish) [8]. It implemented 46 detection techniques commonly adopted by many malwares. The tests are categorized into ten groups as shown in Table 1. The first four categories are general tests checking normal differences between virtual environments and real systems. For example, the CPU performance is to measure the number clocks between two consecutive instructions. Compared to real systems, the number in virtual environment is much bigger or shows a bigger variance in multiple tests. The remaining six categories detect specific virtual systems. Each virtual environment used to have its special values in some files or registry information. For example, in VMware [9], there are special processes named as *vmusrvc.exe* and *vmachhlp.exe*. The value of HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Disk\Enum registry information is DiskVMware.

Table 1 shows the Pafish results of Emulab and VMware. VMware was detected by Pafish in three categories: CPU performance, generic sandbox, and VMware. In total, 15 tests successfully detected VMware. In contrast, Emulab is detected by none of these tests, thus meaning that Emulab is good for malware analysis like bare metal systems [7].

Table 1. Pafish analysis for Emulab

Test category	No of tests	Vmware detections	Emulab detections
Debugger	1	0	0
CPU performance	4	3	0
Generic sandbox	10	5	0
Hook	1	0	0
Sandboxie	1	0	0
Wine	2	0	0
VirtualBox	17	0	0
Vmware	7	7	0
QEMU	2	0	0
Cuckoo	1	0	0

2.2 MBR Recovery from Crashed Operating System in Emulab

We found Emulab very useful when malwares under test destroyed operating systems. Especially, if a malware makes a real system's operating system unbootable by destroying MBR, it would take lots of time and effort to investigate the situation because it is necessary to use other bootable device to access the crashed hard drive. In [7], we provided a processing flow to extract a damaged disk image from a remote system as follows. First, when an Emulab node is not accessible after running malware, it accesses a server called as OPS server that is a file server for managing Emulab nodes. The second step is to initiate the administration mode for the damaged system. Under the mode, a very light FreeBSD operating system is booted and various maintenance operations can be done. After the mode begins, an access to the damaged system is possible. The third step is to extract an MBR image. We found that, even after a new operating system is booted, the device file, */dev/mfid0*, is still connected to the hard disk and it contains previously destroyed operating system. We used *dd* command to make a disk image by extracting information from the disk [10].

3 System Architecture

Our first design goal is to automate the above operations and generate useful information for analysts. The second design goal is that our system can be easily used by as many Emulab users as possible. For this, we implemented our system with a normal user privilege, not with root privilege on Emulab. Therefore, we avoided any ideas that need to change Emulab configuration including package installation so that our system architecture can be used easily by other Emulab users. Due to this, our architecture may look a little redundant, but we believe this trade-off will be paid off.

Figure 1 shows the overall architecture of the proposed system. Node 1 and OPS server belong to Emulab and the external server is located outside of Emulab. We first explain why we chose this design and then how we automated all the numbered processes. Once a malware test starts on Node 1, we keep sending Ping packets to decide when Node 1 hangs (step 1). At the event of no reply (step 2), we determine that the system is no longer alive, so rebooting is necessary. From the OPS server, we activate the admin mode of the Node 1 to reboot with FreeBSD (step 3). While waiting for rebooting, we resume to send Ping packets to know when the system's rebooting is completed. When a reply packet returns again (step 4), it is good to access the node for the next step. In order to automate accessing the node and running some commands, we needed *sshpass* for non-interactive ssh password authentication [11]. Unfortunately, we only had a normal user privilege, so we cannot install the package. Therefore, we prepared an external Linux server where we can install any packages as the root administrator.

To use the external server, the OPS server sends a disk dump request via the HTTP protocol (step 5). When the web server on the external server receives the request message, it accesses Node 1 and executes some commands to dump a disk image (step 6). Please remember that it is not necessary to send the image to the OPS server. Emulab uses the network file system (NFS), so the extracted image is stored in the NFS so that it is possible for the OPS server to access the image directly.

When the disk dump is done, the external server sends a complete message to the OPS server by using a HTTP response packet with the successful code, 200 (step 7). Once the OPS server receives the response, it means that a new MBR image is created. Instead of stopping here, we go one step further by performing several simple analysis on the image because we would like to save analysts' time and also give some hints. We compare the newly created disk image with the previously created one to see if there is something different. If so, it is highly probable that the system shutdown was caused by MBR modification. In addition, we translate the image into three formats: hexadecimal, ASCII, and assembly. The ASCII format may reveal any readable strings that the malware

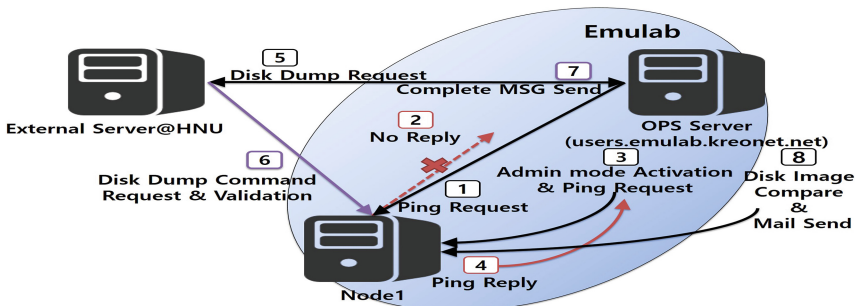


Fig. 1. System architecture

left on MBR. Analysts may be able to identify how instructions are changed. These operations are aggregated into the step 8 in Fig. 1.

Figure 2 shows the flow how all the above operations occur among three systems. *Automation.py* on the OPS server is the main program that communicates with the testing node and the external server. After completing reboot of the node, *Automation.py* generates a HTTP POST request and sends it to the external server. *Emulab.php* is the main web server script that receives the requests from *Automation.py* and accesses nodes by using *sshpass* package.

Again since we decided not to have the administrator privilege, we installed the Netwide Assembler that contains a disassembler, *ndisasm*, at the external server [12]. In order to use the disassembler, we send the extracted images to the external server, which is depicted as the HTTP POST requests to the external server's *Disassembly.php* in Fig. 2. The first one is for the original MBR backup image, and the second is for the newly extracted one. *Disassembly.php* receives these images, runs *ndisasm*, and sends results back to *Automation.py*. Then, *Automation.py* generates hexadecimal and readable strings from the two images. Now it is ready to send an email to the user. The biggest benefit of using email for an alerting method is that it is not necessary to install any database or additional servers. We can simply construct an email with an analysis report and send it. That's all. Another benefit is that users do not have to check the Emulab to see if all the analysis is done. They can just wait for emails. In addition, since the email contains meaningful hints for deciding what to do next, the users even do not have to access Emulab when there is not anything wrong. For this purpose, we programmed *MailingService.py*.

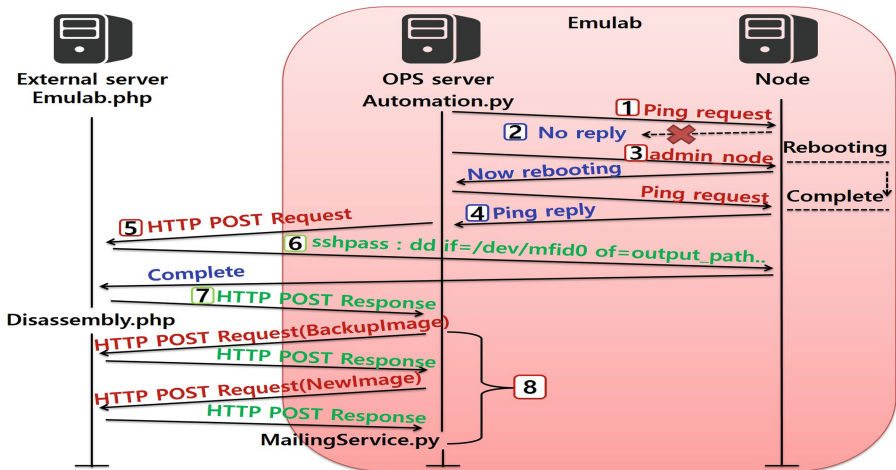


Fig. 2. Automation.py Flow

The last program we made is *DiskImageBackup.py*. This script should be run once before running malwares on the node. Figure 3 depicts its flow.

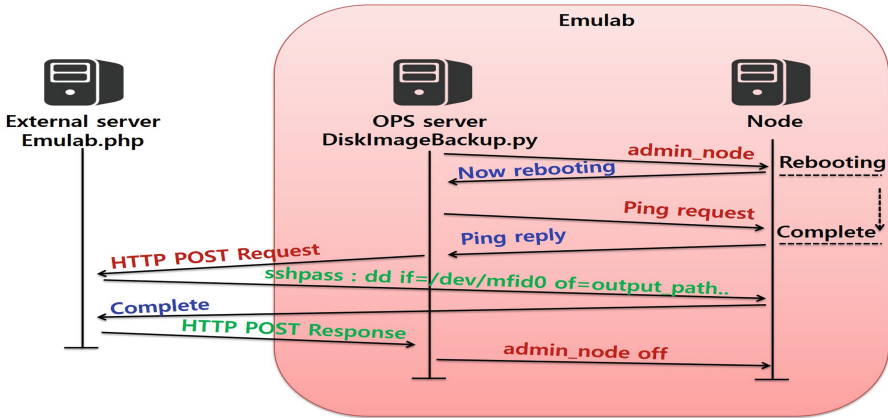


Fig. 3. DiskImageBackup.py flow

4 Experiment

We implemented our system and tested on KREONet Emulab [13]. We picked a malware known as DarkSeoul that staged a serious cyber attack on South Korea in March 2013 [14]. According to a report from McAfee [15], it wiped out the MBR image. After running the malware, then Emulab node status changed to “possible down (BOOTING)” shown in Fig. 4, but the node was not rebooting actually. Rather it simply hanged, not accessible from the outside. When we ran *Automation.py* on the OPS server, it detected this situation and performed automatic analysis on pc41.

Reserved Nodes

Node ID	Name	Type	Default OSID	Node Status	Hours Idle[1]	Startup Status[2]	SSH URL	SSH mime	Console	Log	RDP
pc41	node2	dellR710	Windows_7	possibly down (BOOTING)	210.95?	none					
pc42	node3	dellR710	Windows_7	possibly down (BOOTING)	210.18?	none					

Fig. 4. Emulab node status

All figures from Figs. 5 to 8 are included in one email sent after the analysis of pc41. The first line of the email body is “Result: Different”, meaning that there were some changes on the MBR. Figures 5 and 6 show its binary dump in the hexadecimal and ASCII formats. The reason why strings in ASCII format are broken is that the binary part is x86 instructions as shown in Fig. 7. Figure 8 shows ASCII values of the changed MBR. Interestingly, McAfee reported that DarkSeoul wipes MBR with “principes” and “hastati”, but in our

5 Conclusion

This research utilized Emulab for running malwares and automated all the processes to extract MBR and generate useful information for analysts. This automation will help many Emulab users do their research on malware analysis by saving time and effort.

For future research, we are currently investigating how to use Emulab as on-demand malware analysis web resources so that normal users who is not familiar with Emulab can utilize it easily. Another topic is the malware traffic containment for KREONet Emulab. Current KREONet Emulab allows all traffic generated by test nodes to access the Internet without any restriction.

Acknowledgments. This research is financially supported by 2016 Hannam University Research Fund.

References

1. Internet Security Threat Report, Vol. 21, Symantec Corp., April 2016. <https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf>
2. Internet Security Threat Report, Vol. 20, Symantec Corp., April 2015. <https://www.symantec.com/content/dam/symantec/docs/reports/istr-20-2015-en.pdf>
3. Kirat, D., Vigna, G., Kruegel, C.: Barecloud: bare-metal analysis-based evasive malware detection. In: Proceedings of the 23rd USENIX Security Symposium, pp. 287–301, August 2014
4. Kirat, D., Vigna, G.: MalGene: automation extraction of malware analysis evasion signature. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 769–780, October 2015
5. Utah Emulab, Network Emulation Testbed Home. <https://www.emulab.net/>
6. Lee, M., Seok, W.: Research on the trend of utilizing emulab as cyber security research framework. J. Korea Inst. Inform. Secur. Cryptology **23**, 1169–1180 (2013)
7. Lee, M., Seok, W.: Research on utilizing emulab for malware analysis. J. Korea Inst. Inform. Secur. Cryptology **26**, 117–124 (2016)
8. Pafish. <https://github.com/a0rtega/pafish>
9. VMware Virtualization, VMware. <https://www.vmware.com/>
10. dd. [https://en.wikipedia.org/wiki/Dd_\(Unix\)](https://en.wikipedia.org/wiki/Dd_(Unix))
11. sshpass. <http://manpages.ubuntu.com/manpages/trusty/man1/sshpass.1.html>
12. Netwide Assembler. <http://www.nasm.us/>
13. KISTI Emulab, Network Emulation Testbed Home. <https://www.emulab.kreonet.net/>
14. South Korea cyberattack. https://en.wikipedia.org/wiki/2013_South_Korea_cyberattack
15. Dissecting Operation Troy: Cyberespionage in South Korea, McAfee. <http://www.mcafee.com/kr/resources/white-papers/wp-dissecting-operation-troy.pdf>

Detection of DNS Tunneling in Mobile Networks Using Machine Learning

Van Thuan Do¹, Paal Engelstad², Boning Feng², and Thanh van Do^{2,3(✉)}

¹ Wollfia AS, Martin Linges vei 15, 1364 Fornebu, Norway
vt.do@wollfia.no

² Oslo and Akershus University College of Applied Sciences, Pilestredet 46, 0167 Oslo, Norway
{paal.engelstad,boning.feng}@hioa.no

³ Telenor ASA, Snarøyveien 30, 1331 Fornebu, Norway
than-van.do@telenor.com

Abstract. Lately, costly and threatening DNS tunnels on the mobile networks bypassing the mobile operator's Policy and Charging Enforcement Function (PCEF), has shown the vulnerability of the mobile networks caused by the Domain Name System (DNS) which calls for protection solutions. Unfortunately there is currently no really adequate solution. This paper proposes to use machine learning techniques in the detection and mitigation of a DNS tunneling in mobile networks. Two machine learning techniques, namely One Class Support Vector Machine (OCSVM) and K-Means are experimented and the results prove that machine learning techniques could yield quite efficient detection solutions. The paper starts with a comprehensive introduction to DNS tunneling in mobile networks. Next the challenges in DNS tunneling detections are reviewed. The main part of the paper is the description of proposed DNS tunneling detection using machine learning.

Keywords: Mobile network security · Mobile fraud · Mobile privacy · Cyber security · Cyber attacks · Mobile vulnerability · Machine learning

1 Introduction

The emergence of fancy, powerful but user-friendly mobile devices such as smartphones, tablets, etc. combined with the deployment of mobile wireless broadband access like 3G/4G have made mobile wireless Internet access the most popular Internet usage form surpassing by far the fixed Internet access. However, although affordable for the majority, the mobile wireless access to the Internet is not free of charge and the user is charged based on the used data volume. Indeed, mobile operators usually offer flat fees for different data volumes per month. Further, mobile data usage while roaming on foreign mobile network is very expensive and unaffordable for most of regular non-business users. This explains some individual's motivations and efforts to bypass the mobile operator's charging function and to get free Internet access. But, most difficult to accept is the behavior of a few dishonest mobile operators who equip their customer's smartphones with apps that enable the evasion of the visited operator's charging function

when roaming. These apps use quite often DNS (Domain Name System) [1, 2] tunneling, a method which is motivated by the need of Internet access at Wifi hotspots while evading the fees. DNS tunneling causes obviously revenue losses to mobile operators. But, most importantly, it could be used by any attack that requires firewall evasion i.e. an attacker can send and receive commands and data bypassing the firewall. The simplest but not less serious attack is the theft of confidential information such as personal data, health care data, credit card numbers, payroll, etc. that has financial value. For mobile operators DNS tunneling is not only causing loss of revenues but also deteriorating the quality of service of the overall wireless access and damaging their reputation. A solution preventing DNS is urgently needed.

One obvious solution to prevent DNS tunneling abuses is to block all malicious DNS queries and responses. Unfortunately, this is not a trivial task because differentiating malicious traffic from legitimate one is very challenging if not impossible while blocking the entire DNS traffic is not an inadmissible option. Actually, there is currently no really efficient prevention solution that is mobile operators can use.

In this paper we propose to use machine learning techniques to detect and mitigate DNS tunneling. The paper starts with a state-of-the-art detection and prevention of DNS tunneling, which is followed by a comprehensive introduction to DNS tunneling in the mobile network. Next the challenges of DNS tunneling detection are analyzed. A brief introduction of machine learning and clarifications on how it can be useful in the DNS tunneling detection are then given. The main part of the paper is the description of the proposed DNS tunneling detection using machine learning. The paper concludes with some suggestion for further works.

2 State-of-the-Art Detection and Prevention of DNS Tunneling

Actually DNS tunneling is a known vulnerability that has been known for many years now [3, 4] and there were a lot of works on detection and prevention of DNS tunneling both in academia and in industry. The prevention tools can be classified as following:

Firewalls. All firewalls allows the definition of rules to prevent IP spoofing and to deny DNS queries from IP addresses outside the defined numbers space to prevent the name resolver from being exploited as an open reflector in DDoS attacks. They also enable inspection of DNS traffic for suspicious byte patterns or anomalous DNS traffic to block DNS tunneling. Popular firewalls such as Palo Alto Networks, Cisco Systems, WatchGuard, etc. can detect and block certain DNS tunneling traffic. Unfortunately, they are only efficient against known DNS tunneling methods but are not usable when it comes to the unknown ones.

Intrusion Detection Systems. Intrusion detection systems (IDS) like Snort, Suricata or OSSEC allow the composition of rules to report DNS request from unauthorized clients, to count DNS queries and responses, DNS queries made using TCP, DNS queries to nonstandard ports, suspiciously large DNS queries, any value in any field of the DNS query, etc. However, these IDS can only detect the known attacks.

Traffic Analyzers. Passive traffic analyzers i.e. analyzers that monitor traffic without injecting traffic into the network or modify the traffic that is already on the network, can be used in the identification of DNS tunneling. Unfortunately, these analyzers rely on the knowledge of the traffic amount and patterns.

Passive DNS Replication. Replication of every DNS queries and responses enables analysis that could identify malware using domain name generated by Domain Generation Algorithm (DGA). Passive DNS replication can be used together with IDS to block known malicious domains but again are not usable for the unknown ones.

The DNS tunneling in the mobile network poses additional challenges in terms of processing capability and real time response because of the much larger number of users and considerable number of unknown visiting users but so far according to our knowledge there is not yet any detection work dedicated especially for mobile DNS tunneling.

3 Brief Introduction to DNS Tunneling in the Mobile Network

To introduce DNS tunneling in the mobile network, it is necessary to explain the Internet access from the mobile network. The mobile network is actually a complex network consisting of several mobile networks e.g. 2G, 3G and 4G with a multitude of network elements having different functions. However, since our focus is on the access to the Internet, it is sufficient to consider a simplified representation of the mobile network as shown in Fig. 1.

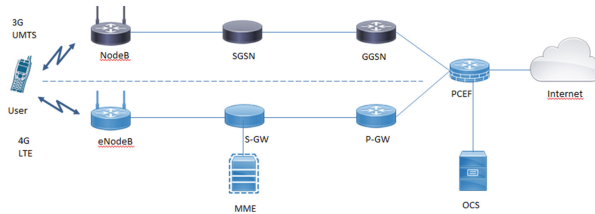


Fig. 1. Mobile wireless access to the Internet

As the user attempts to browse and visit a certain web site <http://www.website.com> a Packet Data service is started. A Packet Data Protocol (PDP) Context Activation is initiated. It establishes a bearer between the mobile phone through the SGSN (Serving GPRS Support Node) and the GGSN (Gateway GPRS Support Node) to the Internet. A Packet Data Protocol (PDP) context is established with the GGSN. The PDP context is a data structure that contains the subscriber’s session information, such as IP address, International Mobile Subscriber Identity (IMSI), and Mobile Station International Subscriber Directory Number (MSISDN). A tunnel is established between the SGSN and the GGSN. User traffic is encapsulated using GTP-U protocol. At the GGSN it is decapsulated and sent to the PDN through the Gi interface (or the Gp in case of roaming). PDP establishment and termination occurs through the GPRS Tunneling Protocol GTP-C protocol.

On 4G/LTE networks the functionality of the GGSN has been replaced by the Serving Gateway (SGW) and the PDN Gateway (PGW).

Before reaching the Internet the data traffic passes through the Policy and Charging Enforcement Function (PCEF), which is quite often integrated within the GGSN or the PGW. It can block the traffic when the user has exceeded the data quota and may also be equipped with Deep Packet Inspection (DPI) functionality.

The Domain Name System (DNS) is a hierarchical decentralized naming system for computers, services, or any resource connected to the Internet or a private network, which allows the translation of human recognizable domain names into numeric IP addresses and enables servers and computers to look up and communicate with each other. The DNS is defined by the IETF (Internet Engineering Task Force) RFC (Request for Comments) 1034 [1] and RFC 1035 [2].

DNS tunneling in the mobile network in the same way generic DNS tunneling in IP network exploits the fact that most operators allows all DNS traffic out and also through port 53 without charging to set up a tunnel for IP traffic bypassing firewall and charging functions.

In fact, it is quite simple to bypass the PCEF by establishing a modified name server on the internet and by creating a special client that is capable of encoding information in the DNS packets. As shown in Fig. 2 the client might send a chunk of data as an “A” or “AAAA” record which may look something like this “nslookup VGhIIHgbWFrZSB1cCB0aGUgNjQgY2hhcmFjdGVycyByZXFlaXJlZCBmb3JgYmFzZ.myDNSTunnel.com” which may encapsulate personal information about John Doe as shown in Fig. 2.

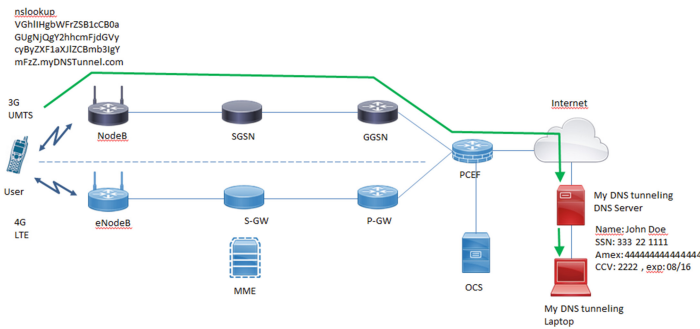


Fig. 2. A simplified DNS tunneling in mobile network

Once this query arrives at the modified DNS server, the server can send any data that is waiting for the client by responding to the A query with a CNAME record - “CNAME:JIIHRyYWRpdGlVbmFsbHkgbm9NsZWFuLiBGb3JgYmFzZ.myDNSTunnel.com”.

Another way to do this involves using DNS TXT or EDNS type records, which allow large unstructured strings to be sent. Reverse lookups can also be used to fetch the data responses.

There are currently a variety of DNS tunneling tools for both PC and Android phones using different forms for data encoding such as OzymanDNS [5], Dns2tcp [6], Iodine [7], Heyoka [8], DNSCat [9], MagicTunnel [10], Element53 [11], VPN-over-DNS [12], etc.

4 Challenges in Detection of DNS Tunneling

In this section, the different DNS tunneling detection techniques are reviewed and their limitations are identified and analyzed. As stated in [13, 14] the detection techniques can be classified into two categories:

Payload analysis: This analysis category can again be divided into sub-categories as follows:

- ***Size of request and response:*** This technique focuses on the size of the request and response, e.g. length of DNS queries and responses [16], ratio of the source and destination bytes [15], size of host name request, etc. The difficulty of this technique type is to find size thresholds that are optimal against all the tunneling methods.
- ***Hostnames entropy:*** This technique is based on the assumption that encoded names have higher entropy than legitimate DNS names. Unfortunately, some tunneling methods do not create high entropy hostnames and some content delivery networks do use hostnames with high entropy to represent some type of information.
- ***Statistical analysis:*** Tunneling can be detected by looking at specific character makeup of DNS names, e.g. percentage of numerical characters in domain names, number of unique characters, percentage of the length of the Longest Meaningful Substring (LMS), number of repeated consonants, etc. The challenge is to determine the threshold value for these specific character makeups. Further, an intelligent DNS tunneling tool will be able to abandon these makeups when their traffic is blocked.
- ***Uncommon record types:*** Not commonly used record types e.g. “TXT” could be used in the detection. Unfortunately, this technique is not decisive.
- ***Specific signatures:*** Each DNS tunneling tool does have specific way of using the attributes in a DNS header, which can be used as signature in the detection. This technique can only be used for known DNS tunneling.

Traffic analysis: This type of analysis considers multiple queries and response pairs over time to detect tunneling:

- ***Volume of DNS traffic per IP address:*** The amount of DNS traffic generated by a specific client IP address [16] can be used in the detection because tunneled data is typically limited to 512 bytes per request and a large number of requests are need for communication. Unfortunately, advanced DNS tunneling tools can spoof the source IP address and spread the requests within a larger range of IP addresses to avoid detection.
- ***Volume of DNS traffic per domain:*** Large amounts of traffic to a specific domain can be used to detect tunneling because DNS tunnel utilities are quite often set up to tunnel the data using a specific domain name. Unfortunately, sophisticated tunneling tools can make use of multiple domain names and hence avoid detection.

- **Number of hostnames per domain:** The number of hostnames for a given domain can be an indicator for tunneling. However, it is not trivial to determine the optimal threshold because different tunneling methods have different numbers of hostnames.
- **Geographic location of DNS server:** Large amounts of DNS traffic to different parts of the world may be an indicator for tunneling. Unfortunately, in the mobile network there will be users coming from all over the world that may have request to DNS resolvers from their country of origin.
- **Domain history:** Domain history can also be an indicator for detection of DNS tunneling. By checking when an A record or NS record is added because a domain could be acquired only recently for DNS tunneling. However, not all newly acquired domains are used in tunneling and one cannot trust every old domain.
- **Orphan DNS requests:** Orphan DNS requests can be also used to detect DNS tunneling because they are the ones that do not have a corresponding request by another application such as http. However, orphan DNS requests may be legitimately used by security devices and program for IP address lookups.

In brief, there is so far no DNS tunneling detection that is really satisfactory for mobile network [17].

5 How Can Machine Learning Help

As in [18] which proposes the usage of machine learning in the protection of mobile networks, Tom Mitchell's definition of machine learning [19] is adopted in this paper as follows:

“The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience”

He provides also a short formalism as follows:

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .”

The machine's ability to learn and improve its solutions to problems is hence central in machine learning.

As shown in Sect. 4, the reviewed DNS tunneling methods do not provide firm criteria but rather non-conclusive indicators. In order to build an efficient DNS tunneling detection it is necessary to use a combination of a large number of both payload and traffic analysis method. Such an approach is not suitable for mobile network because with the huge number of users coming from all over the world it is time consuming and requires huge processing capability.

In this challenging situation, Machine Learning can come to rescue by providing a sound way to define normal behavior of the mobile network when there is no tunneling. Upon the emergence of a DNS tunnel, machine learning techniques will detect anomalies which indicate the presence of the DNS tunnel. At the beginning there will be false positives, i.e. anomalies that are not due the presence of a DNS tunnel but the machine will receive the feedback, learn it and get better for the next time.

For the detection of DNS tunneling, we experiment 2 machine learning methods, namely One Class Support Vector Machine (OCSVM) and K-Means as follows:

One Class Support Vector Machine (OCSVM). OCSVM [20] belongs to a supervised learning class called Support Vector Machine (SVM) that divides the input spaces into two regions, separated by a linear boundary and classifies input either inlier, i.e. falling into one category such as normal or outlier, i.e. falling outside such as abnormal. Although the training of SVM is performed on both positive and negative data the OCSVM extension makes it possible to use only positive data in the training process.

K-Means. k-means clustering, an unsupervised method [21] aims at partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

6 DNS Tunneling Detection Using Machine Learning

In order to test and verify the proposed DNS tunneling detection using machine learning we need DNS traffic data which are both benign and malicious. A testbed is created with four clients: one malicious DNS tunneling client on mobile phone and 3 regular browsers.

- The mobile phone client is realized by an Android 4.4 running on a Virtual Machine (VM) and hosting Slow DNS, a DNS tunneling app.
- The other three clients are realized by a Ubuntu 14.04 running on virtual machines and hosting regular browser.

The 3 browser are regularly accessing the World Wide Web and generating both DNS traffic and http traffic. Before initiating browsing from the mobile phone client a Slow DNS tunnel is established.

All the DNS traffic are then gathered and captured by Wireshark. The DNS packets were filtered out in Wireshark and saved as a comma-separated values (csv) file. Each line in the csv file contained meta data for one packet, with the features No., Time, Source, Destination, Protocol, Length, Info. The raw data had to be reformatted to be used for the machine learning models. To do this a python script is developed and used, which went through the csv file to find the response to each request and creating a new csv file.

The features in the new file were Time, Source, Destination, Protocol, LengthUp, LengthDown, Info, Label.

The Time feature now was the time between the request and response, not the time since the capture started. Table 1 shows how two lines of the new csv file looks like, one line with regular DNS traffic and one with malicious traffic.

Table 1. A line with regular DNS traffic and one with malicious traffic

Time	Source	Destination	Protocol	LengthUp	LengthDown	Info	Label
0.021829999999997796	192.168.1.60	192.168.1.1	DNS	89	276	Standard query 0x0cf7 A safebrowsing-cache.google.com	1
0.51761099999999987	192.168.1.14	192.168.1.1	DNS	209	274	Standard query 0x3a8c NULL 149N2546851188122-246-109-MHoQF3dk88nOvvCbaPdyeLvksPKAAAAAdgAbh.u4KADQAMgAOAA0AGQALA AwAGAAJAAoAFgAXAAgABgAHABQAFQAEAAUA EAATA.AEAAGADAASie.tg16.m7q.in	-1

To test and verify the two proposed machine learning methods for DNS detection, SciKit-Learn, a library for Python which contains functions to create machine learning classifiers and support for training and testing is used. The metrics class in SciKit-Learn contains many functions to evaluate classifiers. For an outlier and categorization classification the normal way to determine the success is by measuring precision, recall and F-score. The precision of a classifier determines the percentage of the elements selected that are true positives. The recall determines the percentage of the relevant elements was selected. The F-score measurement is derived from both precision and recall and gives a result which better represents the overall character of the classifier. The closer to 1 the higher are both the precision and the recall; and classifier is working well.

DNS Tunneling Detection Using OCSVM. Since the SciKit-Learn OCSVM has four kernels linear, polynomial, Radial Basis Function (RBF) and sigmoid, experiments are carried out with each kernel.

As shown in Table 2, the poly kernel has best result. However, after some adjustment of the nu and gamma parameters, the RBF kernel obtains an f1 of 96% which is higher than the poly kernel.

Table 2. Classification report for OCSVM with different kernels

		precision	recall	f1-score	support
Kernel = rbf	Outlier	0.40	1.00	0.57	1124
	Inlier	1.00	0.51	0.67	3394
	avg / total	0.85	0.63	0.65	4518
Kernel = sigmoid	Outlier	0.25	1.00	0.40	1124
	Inlier	0.00	0.00	0.00	3394
	avg / total	0.06	0.25	0.10	4518
Kernel = linear	Outlier	0.24	0.94	0.38	1124
	Inlier	0.57	0.03	0.05	3394
	avg / total	0.49	0.25	0.14	4518
Kernel = poly	Outlier	0.86	0.94	0.90	1124
	Inlier	0.98	0.95	0.96	3394
	avg / total	0.95	0.95	0.95	4518

DNS Tunneling Detection Using K-Means. The K-means classifier is tested with three different initiation methods, namely k-means++, random and ndarray. The results of the experiment are shown in Table 3. All the initiation methods have a quite even total f1-score, but a closer look of each line may identify a weakness. With init set to random the model is almost not able to predict any outlier, with recall at 1% and both precision and f1-score at 0%.

Table 3. Classification report for K-means models with different initiation

		precision	recall	f1-score	support
init = ndarray	Outlier	0.14	0.99	0.24	286
	Inlier	1.00	0.48	0.65	3441
	avg / total	0.93	0.52	0.62	3727
init = k-means++	Outlier	0.14	0.99	0.25	286
	Inlier	1.00	0.50	0.66	3441
	avg / total	0.93	0.53	0.63	3727
init = random	Outlier	0.00	0.01	0.00	286
	Inlier	0.86	0.50	0.64	3441
	avg / total	0.79	0.47	0.59	3727

Evaluation. The experiments show that the DNS detection using OCSVM is superior to the one using K-means. This is not surprising since K-means is a cluster classifier and work best when the clusters are even, which is not the case of DNS tunneling where only a minor part of the traffic data is malicious. Further, the experimented data set is too small for K-means.

OCSVM gave great results with the poly kernel with default parameters, and with the RBF kernel when the gamma and nu parameters are tuned. As the poly kernel only seemed to work with the default parameters and with two features from the dataset, it seems to be quite unstable and might not be the best to use in a real implementation. The RBF kernel had a recall of close to 100% on the outliers in most of the tests, which means it was able to categorize all the outliers correctly. This is important for a DNS tunneling detection. The weakness of the method is the precision of outliers and recall of inliers, which means it produces some false positives. By working with the initiation parameter of the model it is possible to reduce the number of false positives down. The OCSVM with RBF kernel is a good method to implement DNS tunneling detection.

7 Conclusion

In this paper, we propose to use machine learning techniques in the detection of DNS tunneling, which so far does not have any really efficient solutions. Two machine learning methods, namely OCSVM and K-means have been selected for the experiments. A testbed able to generate and collect both regular and malicious DNS traffic is established. Experiments have been carried out and the results prove that machine learning is a feasible technique that could be used in the detection of DNS tunneling. However, the efficiency depends heavily on the machine learning method in use and on some degree the fine-tuning of their respective parameters. The experiments have many limitations. First, the dataset used in the experiment is too small and is not representative for the huge DNS traffic in the mobile network. Next, the dataset is generated only by 4 clients, one malicious and three benign and hence does not contain sufficient variations in terms of IP addresses, domains, DNS tunneling methods, etc. Third, the performance and the scalability of the detection have not been evaluated due to the small size of the dataset. As further work, it is quite interesting to carry out the experiments using real DNS traffic data both benign and malicious collected from the Telenor mobile networks. It might be also quite relevant to make use of a machine learning technique called the Deep Learning Auto-Encoder (DL-AE) [23].

References

1. IETF: RFC 1034 Domain names – concepts and facilities, Internet standard, November 1987
2. IETF: RFC 1035 Domain names - Implementation and specification - Internet standard, November 1987
3. Pure Hacking: Reverse DNS Tunneling – Staged Loading Shellcode, Ty Miller, Blackhat (2008)
4. Ayaya: Black Ops of DNS, Dan Kaminsky, Blackhat (2004)
5. OzymanDNS – Dan Kaminsky (2004). <https://dankaminsky.com/2004/07/29/51/>
6. Dns2tcp - Hervé Schauer Consultants. <http://www.hsc.fr/ressources/outils/dns2tcp/>
7. Iodine. <http://code.kryo.se/iodine/>
8. Heyoka. <http://heyoka.sourceforge.net/>
9. DNScat. <http://tadek.pietraszek.org/projects/DNScat/>
10. MagicTunnel. <http://www.magicunnel.net/>
11. Element53 – Sander Nijhof. <https://nijhof.biz/element53/>
12. VPN over DNS. <https://www.vpnoverdns.com/>
13. SANS Institute: Data Charging Bypass - How your IDS can help, Hassan Mourad, September 2014
14. SANS Institute: Detecting DNS Tunneling, Greg Farnham, February 2013
15. Bianco, D.: A traffic-analysis approach to detecting DNS tunnels. <http://blog.vorant.com/2006/05/traffic-analysis-approach-to-detecting.html>. Accessed 3 May 2006
16. Pietraszek, T.: Dnscat. <http://tadek.pietraszek.org/projects/DNScat/>. Accessed 31 Oct 2004
17. Heavy Reading: DNS Security for Service Providers: An Active Approach at L7 – White Paper – Patrick Donegan, October 2015
18. Do, V.T., Engelstad, P., Feng, B., van Do, T.: Strengthening mobile network security using machine learning. In: Younas, M., Awan, I., Kryvinska, N., Strauss, C., van Thanh, D. (eds.) *MobiWIS 2016*. LNCS, vol. 9847, pp. 173–183. Springer, Heidelberg (2016). doi: [10.1007/978-3-319-44215-0_14](https://doi.org/10.1007/978-3-319-44215-0_14)
19. Mitchell, T.M.: *Machine Learning*. McGraw-Hill Companies Inc, New York (1997). ISBN 0-47-042807-7
20. Manevitz, L.M., Yousef, M.: One-class SVMs for document classification. *J. Mach. Learn. Res.* **2**, 139–154 (2002)
21. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297. University of California Press, Berkeley (1967). MR 0214227, Zbl 0214.46201, Accessed 07 Apr 2009
22. SlowDNS: A free VPN over DNS Tunneling Tool. <http://slowdns.com/>
23. Bengio, Y.: Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2**, 1–127 (2009). doi: [10.1561/2200000006](https://doi.org/10.1561/2200000006)

On the Security Analysis of Weak Cryptographic Primitive Based Key Derivation Function

Chai Wen Chuah¹(✉), Mustafa Mat Deris¹, and Edward Dawson²

¹ University Tun Hussein Onn Malaysia, Parit Raja, Malaysia
{cwchuah,mmustafa}@uthm.edu.my

² Queensland University of Technology, Brisbane, Australia
e.dawson@qut.edu.au

Abstract. A key derivation function is a function that generate one or more cryptographic keys from a private string together with some public information. The generated cryptographic key(s) must be indistinguishable from random binary strings of the same length. To date, there are designed of key derivation function proposals using cryptographic primitives such as hash functions, block ciphers and stream ciphers. The security of key derivation functions are based on the assumption that the underlying cryptographic primitives are secure from attacks. Unfortunately, the current works do not investigate the consequences for key derivation functions if the cryptographic primitives that are used to build the key derivation functions are broken. In this paper, we are confirmed by results of having the cryptographic primitives that are used to build the key derivation functions are broken, it allows the adversaries to distinguish the cryptographic key from the random binary string of the same length.

Keywords: Key derivation function · Cryptographic key · Hash function · Block cipher · Stream cipher

1 Introduction

Many cryptographic systems such as Host Identity Protocol (HIPv2) [8], PKINIT algorithm agility [17] are require cryptographic algorithm and pseudorandom cryptographic keys to protect electronic data transmit at insecure channel. Key derivation functions (KDFs) are used to generate these pseudorandom cryptographic keys. KDFs transforms private string together with public strings into one or more pseudorandom cryptographic keys. It is critical to ensure that the KDFs are secure to be used in these cryptographic systems.

To date, many of the existing KDF proposals have been designed using three major cryptographic primitives. There are hash functions [3, 11], block ciphers [3] and stream ciphers [5]. In this paper, we investigate the impacts of weak cryptographic primitives are used to construct these KDFs.

2 Key Derivation Function

KDFs take two inputs, namely private string and public string to generate n -bit cryptographic key. The private string p contains certain entropy. The public strings are salt s and context information c . The derived cryptographic key is said to be computationally indistinguishable from a binary random string, if no polynomial time algorithm can distinguish between the cryptographic key and a binary random string of the same length. The length, n , of the derived cryptographic key is an application specific security parameter.

Definition 1 [5] (*Key derivation function*). A key derivation function is defined as: $K \leftarrow KDF(p, s, c, n)$, where

- p is a private string, which is chosen from the space of all possible private strings $PSPACE$. We denote the length of p as pl and the probability distribution of p as \mathcal{P} .
- s is a salt, a public random string chosen from the salt space $SSPACE$. We denote the length of s as sl and the probability distribution of s as \mathcal{S} .
- c is a public context string chosen from a context space $CSPACE$. We denote the length of c as cl and the probability distribution of c as \mathcal{C} .
- n is a positive integer that indicates the number of bits to be produced by the KDF;
- K is the derived n -bit cryptographic key.

The basic operation of a KDF is to transform the secret p and the public inputs (s and/or c) into an n bit string which can be used as a cryptographic key.

2.1 Collision Analysis

Assume a message m has length of ml and a random function H maps m to an output with length of n . Collision will happen when $ml > n$. For the function H on a random message m , we have message collision when $H(m_1) = H(m_2)$, where $m_1 \neq m_2$. When the length of the output is n -bit then by birthday paradox [13] after calculating H for $2^{\frac{n}{2}}$ distinct messages, there is a 50% chance of message collision. It may be possible to construct message collisions for algorithms like MD5 and SHA1 in substantially less than $2^{\frac{n}{2}}$. For MD5 [16] and SHA1 [15], Wang *et. al* found a message collision in less than 2^{64} calculations and 2^{69} calculations respectively which are relatively faster than the birthday paradox.

2.2 Provable Security - Random Oracle Model (ROM)

In 1993, Bellare and Rogaway made proving cryptographic protocols easier and more efficient by introducing the idea of ROM that allows all parties to access the public random oracle [1]. In the ROM, in order to obtain the value $H(x)$, the adversary needs to query the random oracle with input x , where H can be a hash function, block cipher or stream cipher. The random oracle queries are simulated by the challenger as follows. On input a string x , if x has not been

queried before, then output $H(x) \in_R \{0, 1\}^n$, where n is the output length of the H function. If x has been queried before, output the same value $H(x)$ as before.

2.3 Formal Security Analysis on KDF Proposals

KDFs are multipurpose functions that used to generate cryptographic keys. It is widely use in many application, thus there are some little formal work to proof the security of KDFs. The first work to analyze the KDF is done by studying the suitability of common pseudorandomness modes associated with cryptographic hash functions and block ciphers (CBC-MAC, Cascade and HMAC) [6]. This paper examined the formal foundations for the HMAC based KDF where the HMAC is using either SHA1 or MD5. This scheme then being used in IKE protocols [7]. However, the paper did not provide the formal security model for the KDF, this, motivated Krawczyk to propose cryptographic extraction and key derivation, the HKDF scheme [10]. The proposed HKDF scheme consists of extractor and expander. The extractor and expander are built using HMAC-SHA256 or HMAC-SHA512 or combination of both. This scheme is then being used in HIPv2 [8]. Krawczyk proposed security model restricts the capability of the strong active adversary. The adversary is only able to change the context information. Multiple derived cryptographic keys were generated by the KDF by using a fixed private string, a single salt and different context information. The salt was chosen by the challenger and known by the adversary, the context information was chosen by the adversary. Hence, a robust security model called CPM-secure is proposed [4] that captures the strong active adversary extending Krawczyk result. Multiple derived cryptographic keys were generated by the KDF by using a fixed private string, multiple salts and different context information. Both salt and context information was chosen by the adversary. Both security proof [4, 10] are based on the ROM.

2.4 The Security of KDF - Adaptive Chosen Public Inputs Model With Multiple Salts (CPM)

A robust security model called CPM-secure is proposed by Chuah *et. al* [4]. For a KDF to be CPM-secure as shown in Table 1, an adversary A is assumed to be allowed to choose both public inputs, salt and context information. For instance, the adversary may choose a null or non-random salt value. The adversary's chosen salt value and different chosen context information is used to generate the cryptographic keys. The adversaries are also able to choose whether to respond to the challenger immediately or to progress to the adaptive stage. The adversaries are allowed to make no more than q queries.

Due to birthday paradox, a restriction to the adversary to make queries q not greater than σ where collision is found for the PRK , such that $\sigma < |\mathit{SSPACE}|$. However, if $\sigma > 2^m$, where m is the min-entropy of the private string, then the adversary is allow to make queries q not greater than 2^m . If $q < \sigma$ or $q < 2^m$, all two-phase KDF proposals are secure in CPM model [4].

Table 1. CPM-secure.

Learning stage	1. C chooses $p \leftarrow PSPACE$	
	2. For $i = 1, \dots, q' \leq q$,	(2.1) A chooses $s_i \leftarrow SSPACE$
		(2.2) A chooses $c_i \leftarrow CSPACE$
		(2.3) C computes $K_i = F(p, s_i, c_i, n)$
	(2.4) A is provided K_i	
Challenge stage	1. A chooses $s_i \leftarrow SSPACE$ and $c \leftarrow CSPACE$ (subject to restriction $s \notin s_i, \dots, s'_q$ and $c \notin c_i, \dots, c'_q$)	
	2. C chooses $b \xleftarrow{R} \{0, 1\}$	(2.1) If $b = 0$, C outputs $K' = F(p, s, c, n)$,
		(2.2) else C outputs $K' \xleftarrow{R} \{0, 1\}^n$
	5. C sends K' to A	
Adaptive stage	1. Step 4 in <u>Learning stage</u> is repeated for up to $q - q'$ queries (subject to restriction $s_i \neq s$ and $c_i \neq c$)	
	2. A outputs $b' = 0$, if \mathcal{A} believes that K' is cryptographic key, else outputs $b' = 1$	
<p>A wins the game if $b' = b$</p>		

Definition 2 {CPM-secure [4]}. *CPM-secure. The KDF is (t, q, ϵ) CPM-secure if for all probabilistic polynomial-time t adversaries A can make at most $q < |SSPACE| \times |CSPACE|$ queries to the KDF who can win the following indistinguishability game with probability not larger than $(\frac{1}{2} + \epsilon)$.*

2.5 Two-Phase Key Derivation Function

The first phase is an extractor process, denoted as Ext , which takes a private string p and a salt s as the inputs, and generates an output, which denoted as PRK . The PRK is an intermediate value derived from the secret p , so PRK is also secret. The second phase is an expander process, denoted as Exp , that takes the secret intermediate value PRK and public string namely context information c as the inputs and produces an n -bit cryptographic key. This basic operation is $KDF(p, s, c, n) = Exp(\{Ext(p, s)\}, c, n)$.

a. First Phase: Extractor

Extractor is a function that takes the private string p which contains randomly generated secret information and the salt s which is random string which is not kept secret. The aim of the extractor is to extract all the entropy from p and to transform the entropy to the value of PRK which is computationally indistinguishable from a random binary string of the same length.

Definition 3 (*Computational Extractor Multiple Salts*). Let $PSPACE$ and $SSPACE$ be set spaces of $\{0, 1\}^{pl}$ and $\{0, 1\}^{sl}$ respectively. A function $Ext : \{0, 1\}^{pl} \times \{0, 1\}^{sl} \rightarrow \{0, 1\}^{kl}$ is called a (t_X, q_X, ϵ_X) -computational extractor if an adversary A running in a polynomial number of time steps t_X and making at most q_X queries to the extractor, A can distinguish between PRK (derived from p) or a random string of the same length, with probability not larger than $(\frac{1}{2} + \epsilon_X)$ where p is chosen from $\{0, 1\}^{pl}$ and s chosen from $\{0, 1\}^{sl}$. If Ext is a (t_X, q_X, ϵ_X) -computational extractor with min-entropy m then we call it a $(m, t_X, q_X, \epsilon_X)$ -computational extractor, where ϵ_X is negligible.

- b. **Second Phase: Expander.** Expansion is a function that takes as input the PRK and the context information c , then transforms these inputs into one or more arbitrary length cryptographic key(s). The aim of the expander is to form a cryptographic key(s) which is computationally indistinguishable from a random binary string of the same length.

Definition 4 (*Expander*) [10]. An expander is a (t_Y, q_Y, ϵ_Y) -secure variable-length-output pseudorandom function family if an adversary A running in a polynomial number of time steps t_Y and making at most q_Y queries to the expander, A can distinguish the cryptographic key generated by the expander from a random string of the same length with probability not larger than $(\frac{1}{2} + \epsilon_Y)$, where ϵ_Y is negligible.

3 KDF Security Analysis

HIPv2 is based on DH shared secret key exchange protocol which provides secure communications and maintains shared IP-layer state between two separate parties. HIPv2 provides protection against attacks on the confidentiality and integrity of the communication between these two parties. These protections require cryptographic keys. KDFs are used to generate these cryptographic keys, the inputs are DH shared secret key (private string) together with some public informations stated in [8]. The length of the DH shared secret key indicating that the security strength provided by the cryptographic keys. For example, if the DH share secret key is 1536-bits, the estimated security provided from this DH share secret key protocol is approximately 90-bits. However, if a KDF based on HMAC-MD5 is CPM-secure (Definition 2) which is used to transform the 1536-bits DH share secret key together with some public strings into cryptographic key, the practical security provided by this KDF is only 64-bits. This scenario is happened due to the birthday paradox paradigm.

In this section we present the security analysis of KDF based on eight cryptographic primitives which are shown in Table 2. The KDFs are two-phase KDFs which consist of extractor and expander. Both extractor and expander are build using the same cryptographic primitive. The cryptographic primitives are either theoretically broken or practically broken. The assumptions that we made in the proof are the adversary can play the challenge game with the KDFs and at the

Table 2. Existing attacks on cryptographic primitives.

Cryptographic primitive		Birthday paradox	Attack	Complexity	Data require	Reference
HMAC	HMAC-MD5	2^{64}	Collision Attack	2^{41}	-	[16]
	HMAC-SHA0	2^{80}	Collision Attack	2^{39}	-	[14]
	HMAC-SHA1	2^{80}	Collision Attack	2^{69}	-	[15]
Block cipher	DES	2^{64}	Birthday Attack	2^{48}	2^{16}	[2]
Stream cipher	Sosemanuk	2^{64}	Differential Fault Analysis	$2^{23.46}$	$2^{35.16}$	[12]
	LILI-128	2^{64}	Fault Attack	2^{25}	1M	[9]
	SOBER-t32	2^{128}	Fault Attack	2^{30}	100K	[9]
	RC4	2^{1024}	Fault Attack	2^{26}	2^{26}	[9]

same time the adversary can exploit the existing attacks towards the cryptographic primitives that are used to build the KDFs. An interesting point is, if the adversary aim to generate these cryptographic keys, they could just find the intermediate value PRK (derived from the private string and salt), where PRK is the input to the expander phase. With the PRK together multiple known context information, the adversary could generate all the cryptographic keys. The detail attack is shown in Theorems 1, 2 and 3.

Theorem 1. *When H is HMAC-MD5 or HMAC-SHA0 or HMAC-SHA1, the KDFs are not CPM secure.*

Proof: Here we give that the KDFs are not CPM-secure. Recall from the Definition 2, a KDF is CPM-secure, if A is unable to determine whether the challenge output K' is the cryptographic key K generated from the KDF or a random binary string of the same length after making q queries in the learning stage and adaptive stage, where $q < |SSPACE| \times |CSPACE|$ and the probability that A can distinguish K' is not greater than $\frac{1}{2} + \epsilon$. Assume the KDFs are hash based KDFs which follows the two-phase model and which are based on the hash function HMAC-MD5, HMAC-SHA0 and HMAC-SHA1. The security for this KDF is based on the underlying security of these ciphers (HMAC-MD5, HMAC-SHA0 and HMAC-SHA1) which are used to construct the KDF. This means, if the security of underlying ciphers are compromised, it will affect the security strength of KDFs itself. For example, MD5, SHA0 and SHA1 are broken; we can find the collision using about 2^{41} MD5 operations [16], 2^{39} SHA0 operations [14] and 2^{69} SHA1 operations [15] respectively. We denote the collision operations as ϕ . As discussed in Sect. 2.4, the KDF is not considered CPM-secure if $q \geq \sigma$, where $sl > \sigma$. For MD5, if using collisions based on birthday paradox, the adversary can make queries at least σ queries, where σ for MD5 is 2^{64} . For SHA0 and SHA1, if using birthday paradox the adversary can make $\sigma \approx 2^{80}$ queries.

Firstly, we show the HMAC-MD5 based KDFs are not CPM secure. The adversary A in CPM is an active adversary who can make queries at most $q < \sigma$. For each query, the A queries different s_i and with a single c , such that

$KDF(p, s, c, n) = (H\{p, s_i\}, c)$, where $i = 1, 2, \dots, \sigma$ and H is the KDF function which is HMAC-MD5. As MD5 is broken, with around $q \approx \phi$ queries to HMAC-MD5 based KDF, A will be able to find collision to the intermediate value PRK . For example, $H(H\{p, s_i\}, c) = H(H\{p, s_\phi\}, c)$, where $i = \tau$ and, $\tau < \phi$. The A still can make more queries $(\sigma - \phi)$. Next, A queries pair (s_τ, c') . The challenger computes $K_{\phi+1} = H(H\{p, s_\tau\}, c')$ and sends $K_{\phi+1}$ to A . During the challenge stage, A queries (s_ϕ, c') and A received the challenge output K' . The A wins the game as A will be able to distinguish the challenge output by simply verifying $K' \stackrel{?}{=} K_{\phi+1}$. Hence, HMAC-MD5 based KDFs are not CPM secure.

By the same token, we may show that HMAC-SHA0 and HMAC-SHA1 are used to build the KDFs also are not CPM secure. The A may find the collision of the intermediate value PRK with approximately $q \approx \phi$ which is less than σ .

Theorem 2. *When H is DES, the KDFs are not CPM secure.*

Proof: Firstly, we show that these KDFs are not CPM-secure. Recall from Definition 2, a KDF is CPM-secure if A cannot distinguish whether the challenge output is the derived cryptographic key from the KDF or just a truly random string with the same length after making $q < |SSPACE| \times |CSPACE|$ queries in the learning stage and in the adaptive stage. As discussed in Sect. 2.4, the KDF is not considered CPM-secure if $q \geq \sigma$, where $sl > \sigma$. For DES, if using collisions based on birthday paradox, the adversary can make queries at least σ queries, where σ for DES is 2^{64} .

The A in CPM is an active A who can make queries at most $q < \sigma$. For each query, the A queries different s_i and with a single c , such that $KDF(p, s, c, n) = (H\{p, s_i\}, c)$, where $i = 1, 2, \dots, \sigma$, H is the KDF function which is DES and the A may request the length of cryptographic keys, n is 2^{16} . As DES is broken, with around $q \approx 2^{48}$ queries to DES based KDF and n is 2^{16} [2], A will be able to find collision to the intermediate value PRK . For example, $H(H\{p, s_i\}, c) = H(H\{p, s_{48}\}, c)$, where $i = \tau$ and, $\tau < 2^{48}$. The adversary still can make more queries $(2^{64} - 2^{48})$. Next, A queries pair (s_τ, c') . The challenger computes $K_{2^{48}+1} = H(H\{p, s_\tau\}, c')$ and sends $K_{2^{48}+1}$ to A . During the challenge stage, A queries (s_{48}, c') and A received the challenge output K' . The A wins the game as A will be able to distinguish the challenge output by simply verifying $K' \stackrel{?}{=} K_{2^{48}+1}$. Hence, DES based KDFs are not CPM secure.

Theorem 3. *When H is LILI-128 or SOBER-t32 or RC4, the KDFs are not CPM secure.*

Proof: Here we give that the KDFs are not CPM-secure. Recall from Definition 2, a KDF is CPM-secure if after making $q < |SSPACE| \times |CSPACE|$ queries at the learning stage and at the adaptive stage, the probability for A to distinguish whether the challenge output is the derived cryptographic key or a random string is not greater than $\frac{1}{2} + \epsilon$.

Assume the KDFs are stream cipher based KDFs which follows the two-phase model and which are based on the stream cipher LILI-128, SOBER-t32 and RC4. The security for this KDF is based on the underlying security of these

ciphers namely LILI-128, SOBER-t32 and RC4 which are used to construct the KDF. For instance, LILI-128, SOBER-t32 and RC4 are broken; we can find the collision using about 2^{25} LILI-128 operations [9] with $1M$ of data, 2^{30} SOBER-t32 operations [9] with $100K$ of data and 2^{26} RC4 operations [9] with 2^{26} of data respectively. We denote the collision operations as ϕ . As discussed in Sect. 2.4, the KDF is not considered CPM-secure if $q \geq \sigma$, where $sl > \sigma$. For LILI-128, if using collisions based on birthday paradox, the adversary can make queries at least σ queries, where σ for LILI-128 is 2^{64} . For SOBER-t32, if using birthday paradox the adversary can make $\sigma \approx 2^{128}$ queries. Whereas RC4 the adversary can make $\sigma \approx 2^{1024}$ queries due to birthday paradox attack.

Firstly, we show the RC4 based KDFs are not CPM secure. The A in CPM is an active adversary who can make queries at most $q < \sigma$. For each query, the A queries different s_i and with a single c , such that $KDF(p, s, c, n) = (H\{p, s_i\}, c)$, where $i = 1, 2, \dots, \sigma$ and H is the KDF function which is RC4. A also can request n is 2^{26} . As RC4 is broken, with around $q \approx \phi$ queries to RC4 based KDF, A will be able to find collision to the intermediate value PRK subjected with n is 2^{26} . For example, $H(H\{p, s_i\}, c) = H(H\{p, s_\phi\}, c)$, where $i = \tau$ and, $\tau < \phi$. The A still can make more queries $(\sigma - \phi)$. Next, A queries pair (s_τ, c') . The challenger computes $K_{\phi+1} = H(H\{p, s_\tau\}, c')$ and sends $K_{\phi+1}$ to A . During the challenge stage, A queries (s_ϕ, c') and A received the challenge output K' . The A wins the game as A will be able to distinguish the challenge output by simply verifying $K' \stackrel{?}{=} K_{\phi+1}$. Hence, RC4 based KDFs are not CPM secure.

By the same token, we may show that LILI-128 and SOBER-t32 are used to build the KDFs also are not CPM secure. However, the data requires that is the length of derived cryptographic keys are different for the experiment. For LILI-128, n is $1M$ [9]. For SOBER-t32, n is $100K$ [9]. The A may find the collision of the intermediate value PRK with approximately $q \approx \phi$ which is less than σ .

4 Conclusion

We have described a cryptographic primitive that are used to build the KDFs is broken, it allows the adversarys to distinguish the cryptographic key from the random binary string of the same length. We are confirmed it by results. The consequences of the derived cryptographic keys are distinguishable, the confidentiality of electronic data during transmission over insecure channel will be compromised. Thus, more attention should be paid when choosing or designing cryptographic primitives for building the KDF.

In this research, we analysed the KDF security with a general assumption for private string. For example, one of the KDF application in the standard HIPv2 takes the private input DH-shared secret. DH-shared string is non-uniform distribution. Another interesting area to be investigated is to analyses this KDF with a practical type of private string. An assume distribution is made for the private string (uniformly random or non-uniformly random). We may include the entropy measurement in the security analysis. Shannon entropy and min-entropy are two basic notions of entropy. The values for Shannon entropy and

min-entropy output are the same if the probability of a random variable is uniformly distributed. However, if a random variable has a non-uniform distribution then the min-entropy value is a more conservative estimate of the entropy of the random variable than Shannon entropy. The conservative estimation is of particular importance in KDFs that are safety critical. Min-entropy will be considered for the research rather than the Shannon entropy, as many of the KDF private strings are randomly generated and have a non-uniform distribution. As a result, we may observe another trend of security for a KDF with different types of private strings by considering the entropy of KDFs' private strings.

Acknowledgments. This research was supported by Fundamental Research Grant Scheme (FRGS) 1558, ORICC UTHM.

References

1. Bellare, M., Rogaway, P.: Random oracles are practical: A paradigm for designing efficient protocols. In: Proceedings of the 1st ACM Conference on Computer and Communications Security, pp. 62–73. ACM (1993)
2. Cao, Z.: How to Launch A Birthday Attack Against DES. IACR Cryptology ePrint Archive 2008, vol. 288 (2008)
3. Chen, L.: NIST SP 800-56C: recommendation for key derivation through extraction-then-expansion. Technical report, NIST (2011)
4. Wen, C.C., Dawson, E., González Nieto, J.M., Simpson, L.: A framework for security analysis of key derivation functions. In: Ryan, M.D., Smyth, B., Wang, G. (eds.) ISPEC 2012. LNCS, vol. 7232, pp. 199–216. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-29101-2_14](https://doi.org/10.1007/978-3-642-29101-2_14)
5. Chuah, C.W., Dawson, E., Simpson, L.: Key derivation function: The SCKDF scheme. In: Janczewski, L.J., Wolfe, H.B., Sheno, S. (eds.) SEC 2013. IAICT, vol. 405, pp. 125–138. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39218-4_10](https://doi.org/10.1007/978-3-642-39218-4_10)
6. Dodis, Y., Gennaro, R., Håstad, J., Krawczyk, H., Rabin, T.: Randomness extraction and key derivation using the CBC, cascade and HMAC modes. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 494–510. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-28628-8_30](https://doi.org/10.1007/978-3-540-28628-8_30)
7. Harkins, D., Carrel, D.: RFC 2409: The Internet Key Exchange (IKE). Technical report, Internet Engineering Task Force (1998)
8. Heer, T., Jokela, P., Henderson, T.: Host identity protocol version 2 (HIPv2). Technical report, Internet Engineering Task Force (2015)
9. Hoch, J.J., Shamir, A.: Fault analysis of stream ciphers. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 240–253. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-28632-5_18](https://doi.org/10.1007/978-3-540-28632-5_18)
10. Krawczyk, H.: Cryptographic extraction and key derivation: The HKDF scheme. In: Rabin, T. (ed.) CRYPTO 2010. LNCS, vol. 6223, pp. 631–648. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-14623-7_34](https://doi.org/10.1007/978-3-642-14623-7_34)
11. Krawczyk, H., Eronen, P.: HMAC-based Extract-and-Expand Key Derivation Function (HKDF). Technical report, RFC 5869 (2010)
12. Ma, Z.Q., Gu, D.W.: Improved differential fault analysis of SOSEMANUK. In: 2012 Eighth International Conference on Computational Intelligence and Security (CIS), pp. 487–491. IEEE (2012)

13. Menezes, A.J., Van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1997)
14. Wang, X., Yu, H., Yin, Y.L.: Efficient collision search attacks on SHA-0. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 1–16. Springer, Heidelberg (2005). doi:[10.1007/11535218_1](https://doi.org/10.1007/11535218_1)
15. Wang, X., Yin, Y.L., Yu, H.: Finding collisions in the full SHA-1. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 17–36. Springer, Heidelberg (2005). doi:[10.1007/11535218_2](https://doi.org/10.1007/11535218_2)
16. Xie, T., Liu, F.B., Feng, D.G.: Fast Collision Attack on MD5. IACR Cryptology ePrint Archive 2013:D170 (2013)
17. Zhu, L., Wasserman, M., Mills, W.: PKINIT Algorithm Agility. Technical report, Internet Engineering Task Force (2015)

On the Security of a Privacy Authentication Scheme Based on Cloud for Medical Environment

Chun-Ta Li¹(✉), Dong-Her Shih², and Chun-Cheng Wang²

¹ Department of Information Management, Tainan University of Technology,
529 Zhongzheng Road, Tainan City 71002, Taiwan, ROC
th0040@mail.tut.edu.tw

² Department of Information Management, National Yunlin University of Science
and Technology, 123 University Road, Yunlin 64002, Taiwan, ROC
shihdh@yuntech.edu.tw, jim821112@gmail.com

Abstract. Recently, Chiou et al. proposed a secure authentication scheme which not only ensures message confidentiality and patient anonymity but also provides real telemedicine system implementation. However, in this paper, we found that Chiou et al.'s telemedicine scheme has some security weaknesses such as (1) it fails to protect the confidentiality of patient's inspection report and doctor's treatment record, (2) it fails to provide the property of unlinkability. The above-mentioned design flaws in Chiou et al.'s scheme may lead to privacy exposure and malicious outsider can link and discover the sensitive relationship between the patient and the doctor.

Keywords: Authentication · Cryptanalysis · Cloud · Medical system · Patient privacy · Telemedicine service

1 Introduction

Due to aging society and development of telemedicine, telecare medicine information systems (TMIS) have been proposed for solving distance problem between patient and hospitals [5, 10–12]. In order to diagnose patients' health conditions in telemedicine model, a new kind of cloud-based medical treatment system is introduced to patients via Internet. However, there have some security threats and privacy issues when patients and doctors exchange health data via public channel. Therefore, in various kinds of telecare medicine information systems, ensure security to a health connected care system is the most important feature [6–8].

In the year of 2012, Padhy et al. [13] design a cloud-based rural healthcare information system for facilitating the quality of patient care. In the year of 2013, Banerjee et al. [1] proposed a centralized cloud-based emergency healthcare system and the patient's historical medical records can be retrieved from the cloud database before starting any crucial operation. In the year of 2014, Chen et al. [2] introduced a secure cloud-based medical data exchange system with

mutual authentication and patient privacy. However, their proposed scheme fails to achieve non-repudiation evidence in doctor diagnosis and is unable to provide telemedicine service [9]. In the same year, Chen et al. [3] further proposed a cloud-based privacy authentication scheme for patients and doctors to access medical resources and find medical advice in more convenient way. However, in the year of 2016, Chiou et al. [4] pointed out that Chen et al.'s privacy authentication scheme fails to provide patient anonymity and message authentication and is unable to provide real telemedicine. In order to repair the design flaws of the scheme [3], Chiou et al. further suggest a complete telemedicine system with user authentication, patient unlinkability, and message confidentiality. Unfortunately, in this paper, we found that Chiou et al.'s improved scheme is still vulnerable to inspection report disclosure and treatment record disclosure attacks and fails to achieve the property of unlinkability between the patient and the doctor.

The rest of this paper is organized as follows. In Sect. 2, we first describe the system architecture of cloud-based telemedicine service, which will be helpful for better understanding. In Sect. 3, we provide overview of Chiou et al.'s telemedicine scheme in brief. In Sect. 4, we show three design flaws of Chiou et al.'s telemedicine scheme. Finally, we conclude this paper in Sect. 5.

2 The System Architecture of Cloud-Based Telemedicine Service

In cloud-based telemedicine service system, four participants involved in this system: the patient (P), the healthcare center (H) the doctor (D), and the medical cloud (C). Before accessing the system, every participant must register with the key generation center (KGC) and KGC will issue one pair of public key and private key for every participant. The patient P can authorize and upload his/her personal health records to the cloud C . In addition, P can collect health personal items from body sensors and upload them to the cloud C . On the other hand, the doctor D can download P 's personal health inspection reports and collected personal health items of the sensors from cloud C via P 's authorization. After diagnosing P 's symptom, D can upload P 's treatment records to the cloud C . Finally, in order to realize real telemedicine service, P can download D 's diagnostic records from C without visiting H in person. Figure 1 shows the system architecture of cloud-based telemedicine service.

Step 1. The patient P goes to the healthcare center H and makes a health inspection in person.

Step 2. When P 's inspection report is released, the healthcare center H uploads P 's personal health inspection report to the cloud C via public channels.

Step 3. The body sensors collect P 's personal health items and send them to P 's personal mobile device via secure channels.

Step 4. The patient P uses his/her personal mobile device to upload P 's personal health items to the cloud C via public channels.

Step 5. In the treatment time, the doctor D can download P 's health inspection reports and health information from the cloud C via P 's authorization. Afterward, D can diagnose P 's symptoms and upload P 's treatment record to the cloud C .

Step 6. In order to achieve real telemedicine service, the patient P can use his/her personal mobile device to download D 's diagnose report from the cloud C .

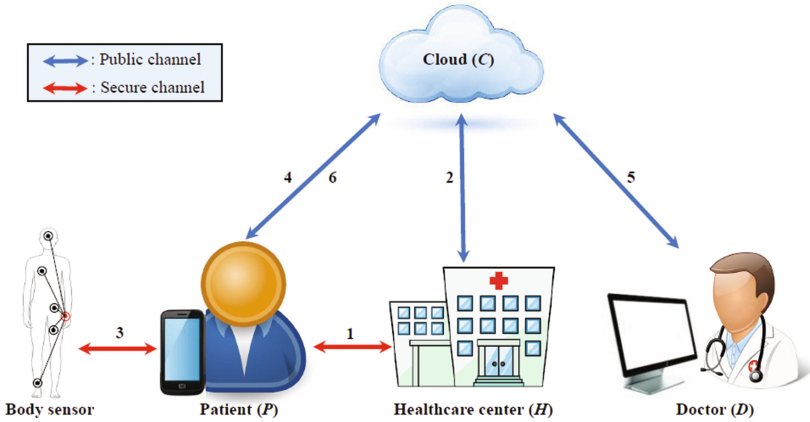


Fig. 1. The system architecture of cloud-based telemedicine service

3 Review of Chiou et al.'s Telemedicine Scheme

In this section, we review Chiou et al.'s scheme [4] based on cloud for telemedicine services. There are four phases involve in their improved scheme: healthcare center uploading phase, patient uploading phase, treatment phase and checking report phase. The notations used throughout this paper are summarized as follows:

- ID_x : The identity of participant x .
- NID_x : The pseudo-random identity of participant x .
- $Data_x$: The health data generated by participant x .
- m_x : The health report generated by participant x .
- PK_x/SK_x : The public and private key pair of participant x .
- key_{xy} : The session key shared between participant x and participant y .
- key_P : The pre-generated key of patient P .
- G_{key} : The group key.
- $e()$: The pairing function.
- $Request_i$: The i th requested message.

- T_x^i : The i th timestamp generated by participant x .
- ΔT : The valid transmission time interval.
- $h(\cdot)$: The one-way hash function.
- $E_k(M)/D_k(M)$: Use the key k to encrypt/decrypt the message M .
- $S_k(M)/V_k(M)$: Use the key k to sign/verify the message M .

3.1 Healthcare Center Uploading Phase

As shown in steps 1 and 2 of Fig. 1, the patient P goes to the healthcare center H to take a health inspection and P uses his/her mobile device to receive a pseudo-random identity NID_P which is allocated by H . When P 's inspection report $m_H = (ID_P, Data_H, T_H^1)$ is released, H carried out the process of mutual authentication with cloud C and uploaded P 's inspection report to C . The detailed steps are described as follows.

Step 1. H uses its private key SK_H to sign P 's m_H and computes $Sig_H = S_{SK_H}(m_H)$, $key_1 = h(e(PK_P, SK_H), NID_P)$ and $C_1 = E_{key_1}(m_H)$, where NID_P is P 's pseudo-random identity. Then H generates a random key $key_{HC} \in_R G_{key}$ as the session key shared between H and C and computes $s_1 = h(e(PK_C, SK_H), T_H^1) \oplus key_{HC}$, $S_2 = h(key_{HC})$ and $C_2 = E_{key_{HC}}(ID_P, NID_P, C_1, Sig_H)$. Finally, H uploads $(ID_H, s_1, s_2, C_2, T_H^1)$ to C .

Step 2. After receiving the messages from H , C checks if the timestamp $T_C^1 - T_H^1 \leq \Delta T$, where T_C^1 is C 's current timestamp. If it holds, C computes $key'_{HC} = h(e(PK_H, SK_C), T_H^1)$ and verifies if $h(key'_{HC}) = s_2$. If it holds, C uses key'_{HC} to reveal $(ID_P, NID_P, C_1, Sig_H)$ by computing $D_{key'_{HC}}(C_2)$. Then C checks whether P is a new user or not by verifying ID_P . If P is a new user, C stores $(ID_P, NID_P, C_1, Sig_H)$ in a new storage space. Otherwise, C updates NID_P and stores C_1 and Sig_H in ID_P 's storage space. Finally, C computes $s_3 = h(key'_{HC} + 1)$ and sends it to H .

Step 3. After receiving s_3 from C , H checks whether $h(key_{HC} + 1) = s_3$. If it holds, H convinces that P 's m_H is successfully uploaded to C . Otherwise, H returns to **Step 1**.

3.2 Patient Uploading Phase

As shown in steps 3 and 4 of Fig. 1, the body sensors are embedded into the patient P 's body and P uses his/her mobile device to collect the measured health items m_B . Then P can make an appointment with D and get an appointment sequence number sn . Moreover, P can download his/her inspection report m_H from C and upload (m_B, m_H) to C . The detailed steps are described as follows.

Step 1. P 's body sensor collects the measured health information m_B and sends m_B to P 's mobile device via a secure channel, where $m_B = (ID_P, Data_B, T_P^1)$. Then P generates a random key $key_{PC} \in_R G_{key}$ as the session key shared between P and C and computes $s_4 = h(e(PK_C, SK_P), T_P^1) \oplus key_{PC}$ and $s_5 = h(key_{PC})$. Finally, P uploads (NID_P, s_4, s_5, T_P^1) to C .

- Step 2.** After receiving the messages from P , C checks if the timestamp $T_C^2 - T_P^1 \leq \Delta T$, where T_C^2 is C 's current timestamp. If it holds, C computes $key'_{PC} = h(e(PK_P, SK_C), T_P^1) \oplus s_4$ and verifies if $h(key'_{PC}) = s_5$. If it holds, C finds P 's (C_1, Sig_H) according to NID_P and sends (s_6, C_1, Sig_H) to P , where $s_6 = h(key'_{PC}, C_1, Sig_H)$.
- Step 3.** After receiving messages from C , P checks whether $h(key_{PC}, C_1, Sig_H) = s_6$. If it holds, P computes $key_1 = h(e(PK_H, SK_P), NID_P)$ and uses it to reveal m_H by computing $D_{key_1}(C_1)$. Then P verifies the validity of m_H by checking whether $m_H = V_{PK_H}(Sig_H)$ holds. If it holds, P computes $key_2 = h(e(PK_D, SK_P), sn)$, $C_3 = E_{key_2}(m_H, m_B)$ and $s_7 = h(key_{PC}, ID_D, sn, C_3)$ and sends (ID_D, sn, s_7, C_3) to C . Finally, P computes $NID_P^{new} = h(NID_P || key_{PC})$ and replaces NID_P with NID_P^{new} .
- Step 4.** After receiving the messages from P , C verifies if $h(key'_{PC}, ID_D, sn, C_3) = s_7$. If it holds, C stores (NID_P^{new}, ID_D) and replaces C_1 with C_3 . Otherwise, C resumes this phase.

3.3 Treatment Phase

As shown in step 5 of Fig. 1, P goes to the hospital and the doctor D obtains P 's ID_P and an appointment sequence number sn . Then D can download P 's medical treatment data from C . After diagnosing P 's symptom, D uploads P 's treatment records to C . The detailed steps are described as follows.

- Step 1.** D generates a random key $key_{DC} \in_R G_{key}$ as the session key shared between D and C and computes $s_8 = h(e(PK_C, SK_D), T_D^1) \oplus key_{DC}$ and $s_9 = h(key_{DC})$. Finally, D uploads (ID_D, s_8, s_9, T_D^1) to C .
- Step 2.** After receiving the messages from D , C checks if the timestamp $T_C^3 - T_D^1 \leq \Delta T$, where T_C^3 is C 's current timestamp. If it holds, C computes $key'_{DC} = h(e(PK_D, SK_C), T_D^1) \oplus s_8$ and verifies if $h(key'_{DC}) = s_9$. If it holds, C confirms that P has an appointment with D and computes $s_{10} = h(key'_{DC}, sn, C_3, Sig_H)$ according to ID_D . Finally, C sends (s_{10}, sn, C_3, Sig_H) to D .
- Step 3.** After receiving messages from C , D verifies if $h(key_{DC}, sn, C_3, Sig_H) = s_{10}$. If it holds, D computes $key'_2 = h(e(PK_P, SK_D), sn)$ and uses it to reveal (m_H, m_B) by computing $D_{key'_2}(C_3)$. Then D verifies the validity of m_H by checking whether $m_H = V_{PK_H}(Sig_H)$ holds. If it holds, D makes a treatment record m_D of P based on (m_H, m_B) and uses his/her private key SK_D to generate a signature $Sig_D = S_{SK_D}(m_D)$, where $m_D = (ID_P, Data_D, T_D^2)$. Finally, D computes $C_4 = E_{key'_2}(m_H, m_B, m_D)$ and $s_{11} = h(key_{DC}, C_4, Sig_D)$ and uploads (s_{11}, sn, C_4, Sig_D) to C .
- Step 4.** After receiving the messages from D , C verifies if $h(key'_{DC}, C_4, Sig_D) = s_{11}$. If it holds, C stores (C_4, Sig_D) . Otherwise, C terminates this phase.

3.4 Checking Report Phase

As shown in step 6 of Fig. 1, in order to provide telemedicine service, P can use his/her mobile device to download m_D from C , where m_D is diagnosed by D . The detailed steps are described as follows.

Step 1. P generates a random key $key_{PC} \in_R G_{key}$ as the session key shared between P and C and computes $s_{12} = h(e(PK_C, SK_P), T_P^2) \oplus key_{PC}$ and $s_{13} = h(key_{PC})$. Then P sends $(NID_P^{new}, s_{12}, s_{13}, T_P^2)$ to C .

Step 2. After receiving the messages from P , C checks if the timestamp $T_C^4 - T_P^2 \leq \Delta T$, where T_C^4 is C 's current timestamp. If it holds, C computes $key'_{PC} = h(e(PK_C, SK_P), T_P^2) \oplus s_{12}$ and verifies if $h(key'_{PC}) = s_{13}$. If it holds, C finds P 's (C_4, Sig_D) according to NID_P^{new} . Then C sends (s_{14}, C_4, Sig_D) to P , where $s_{14} = h(key'_{PC}, C_4, Sig_D)$.

Step 3. After receiving messages from C , P verifies if $h(key_{PC}, C_4, Sig_D) = s_{14}$. If it holds, P uses key_2 to reveal (m_H, m_B, m_D) by computing $D_{key_2}(C_4)$. Then P verifies the validity of m_D by checking whether $m_D = V_{PK_D}(Sig_D)$ holds. If it holds, P gets medical measures according to D 's treatment report and uses the pre-generated key key_P to compute $C_5 = E_{key_P}(m_H, m_B, m_D)$. Finally, P computes $s_{15} = h(key_{PC}, C_5)$ and uploads (s_{15}, C_5) to C .

Step 4. After receiving the messages from P , C verifies if $h(key'_{PC}, C_4) = s_{15}$. If it holds, C replaces C_4 with C_5 . Otherwise, C terminates this phase and P returns to **Step 1**.

4 Cryptanalysis of Chiou et al.'s Telemedicine Scheme

In this section, we demonstrate that Chiou et al.'s telemedicine scheme exposes the patient and the doctor to the flaw of relationship linkability problem and is failing to protect the confidentiality of patient's inspection report and doctor's treatment record. The detailed descriptions of three security weaknesses are as follows.

4.1 Inspection Report Disclosure Attack

In step 2 of patient uploading phase of Chiou et al.'s scheme, the cloud C sends the messages (s_6, C_1, Sig_H) to the patient P through a public channel, where $Sig_H = S_{SK_H}(m_H)$. Due to the public PK_H of the healthcare center H , once a malicious adversary U_A eavesdrops Sig_H from the public channel, U_A can reveal P 's inspection report m_H by performing $m_H = V_{PK_H}(Sig_H) = V_{PK_H}(S_{SK_H}(m_H))$. Similarly, in step 2 of treatment phase of Chiou et al.'s scheme, the cloud C sends the messages (s_{10}, sn, C_3, Sig_H) to the doctor D through a public channel. Therefore, U_A still can reveal P 's m_H by computing $m_H = V_{PK_H}(Sig_H)$. Finally, Chiou et al.'s scheme fails to achieve the confidentiality of P 's inspection report m_H .

4.2 Treatment Record Disclosure Attack

In step 3 of treatment phase of Chiou et al.'s scheme, the doctor D uploads the messages (s_{11}, sn, C_4, Sig_D) to the cloud C through a public channel, where $Sig_D = S_{SK_D}(m_D)$. Due to the public PK_D of the doctor D , once a malicious adversary U_A eavesdrops Sig_D from the public channel, U_A can reveal D 's treatment record m_D by performing $m_D = V_{PK_D}(Sig_D) = V_{PK_D}(S_{SK_D}(m_D))$. As a result, Chiou et al.'s scheme fails to achieve the confidentiality of D 's treatment record m_D .

4.3 Relationship Linkability Problem Between Patient and Doctor

In patient uploading phase of Chiou et al.'s scheme, they claimed that the unlinkability of every session from P to C is guaranteed by using a dynamic pseudo-random identity NID_P which is allotted by H . However, we found that the unlinkability of Chiou et al.'s scheme cannot be protected from message eavesdropping attack during the patient uploading and treatment phases. By launching this attack, a malicious adversary U_A is able to un-intrusively monitor on the public channels between P , C and D and discover some relationships between P and D . We assume that U_A eavesdrops all the communication messages transmitted between P , C and D in cloud-based telemedicine system. In step 3 of patient uploading phase, the messages (ID_D, sn, s_7, C_3) transmitted from P to C are eavesdropped by U_A . In addition, in step 3 of treatment phase, the messages (s_{11}, sn, C_4, Sig_D) transmitted from D to C are eavesdropped by U_A . Note that an appointment sequence number sn is unchanging in patient uploading and treatment phases. In this way, if there is a parameter transmitted between participants containing sn , U_A can easily link and discover the sensitive relationship from P to D by comparing sn with all the eavesdropped messages. Therefore, Chiou et al.'s scheme is vulnerable to relationship linkability problem between P and D .

5 Conclusions

Remote user authentication and patient privacy protection are major concerns over cloud-based telemedicine systems. Recently, Chiou et al. proposed a secure privacy authentication scheme with unlinkability and confidentiality based on bilinear pairing for medical environments. However, we found that Chiou et al.'s telemedicine scheme is vulnerable to inspection report disclosure and treatment record disclosure attacks. Moreover, by eavesdropping the communication messages from public channels, we found their scheme may fail to achieve anonymity and unlinkability between the patient and the doctor. In the future, we plan to propose an enhanced version of their scheme and these security threats should be considered for cloud-based telemedicine services.

Acknowledgements. The authors would like to thank the anonymous reviewers for their valuable suggestions and comments. In addition, this research was partially supported by the National Science Council, Taiwan, R.O.C., under contract no.: MOST 105-2221-E-165-005 and MOST 105-3114-C-165-001-ES.

References

1. Banerjee, A., Agrawal, P., Rajkumar, R.: Design of a cloud based emergency healthcare service model. *Int. J. Appl. Eng. Res.* **8**(19), 2261–2264 (2013)
2. Chen, C.L., Yang, T.T., Shih, T.F.: A secure medical data exchange protocol based on cloud environments. *J. Med. Syst.* **38**(9), 1–12 (2014). article no. 112
3. Chen, C.L., Yang, T.T., Chiang, M.L., Shih, T.F.: A privacy authentication scheme based on cloud for medical environments. *J. Med. Syst.* **38**(11), 1–16 (2014). article no. 143
4. Chiou, S.Y., Ying, Z., Liu, J.: Improvement of a privacy authentication scheme based on cloud for medical environment. *J. Med. Syst.* **40**(4), 1–15 (2016). article no. 101
5. He, D., Zeadally, S.: Authentication protocol for ambient assisted living system. *IEEE Commun. Mag.* **35**(1), 71–77 (2015)
6. He, D., Kumar, N., Chen, J.: Robust anonymous authentication protocol for healthcare applications using wireless medical sensor networks. *Multimedia Syst.* **21**(1), 49–60 (2015)
7. Jiang, Q., Ma, J., Lu, X., Tian, Y.: An efficient two-factor user authentication scheme with unlinkability for wireless sensor networks. *Peer-to-Peer Networking Appl.* **8**(6), 1070–1081 (2015)
8. Jiang, Q., Wei, F., Fu, S., Ma, J., Li, G., Alelaiwi, A.: Robust extended chaotic maps-based three-factor authentication scheme preserving biometric template privacy. *Nonlinear Dyn.* **83**(4), 2085–2101 (2016)
9. Li, C.T., Lee, C.C., Wang, C.C., Yang, T.H., Chen, S.J.: Design flaws in a secure medical data exchange protocol based on cloud environments. In: Wang, G., Zomaya, A., Perez, G.M., Li, K. (eds.) *ICA3PP 2015*. LNCS, vol. 9532, pp. 435–444. Springer, Cham (2015). doi:[10.1007/978-3-319-27161-3_39](https://doi.org/10.1007/978-3-319-27161-3_39)
10. Li, C.T., Weng, C.Y., Lee, C.C.: A secure RFID tag authentication protocol with privacy preserving in telecare medicine information systems. *J. Med. Syst.* **39**(8), 1–8 (2015). article no. 77
11. Li, C.T., Weng, C.Y., Lee, C.C., Wang, C.C.: A hash based remote user authentication and authenticated key agreement scheme for the integrated EPR information system. *J. Med. Syst.* **39**(11), 1–11 (2015). article no. 144
12. Li, C.T., Lee, C.C., Weng, C.Y.: A secure cloud-assisted wireless body area network in mobile emergency medical care system. *J. Med. Syst.* **40**(5), 1–15 (2016). article no. 117
13. Padhy, R.P., Patra, M.R., Satapathy, S.C.: Design and implementation of a cloud based rural healthcare information system model. *Univ. J. Comput. Sci. Eng. Technol.* **2**(1), 149–157 (2012)

Physical Layer Security with Energy Harvesting in Single Hop Wireless Relaying System

Poonam Jindal^(✉) and Rupali Sinha

Department of Electronics and Communication Engineering, National Institute of Technology,
Kurukshetra 136119, Haryana, India
poonamjindal81@nitkkr.nic.in, rupalisinha468@gmail.com

Abstract. In this paper, a single hop wireless relaying system, employing energy harvesting (EH) is proposed, where an information is transmitted from source to destination with the help of a relay in the presence of an eavesdropper. In this proposed system, source and relay utilizes EH technique for obtaining energy from a power beacon. The EH technique employed in this system is time switching EH technique. The secrecy performance of the proposed system is investigated for two cooperative schemes: decode-and-forward (DF), and amplify-and-forward (AF). The proposed system with EH technique provides improvement in secrecy rate, energy efficiency and power consumption as compared to that of the conventional scheme, as in the proposed system, nodes are powered up with EH technique, instead of using individual batteries. The secrecy rate of the proposed system with EH is higher than that of the conventional system by 8.89% for AF relay and by 9.83% for DF relay at a distance of 70 m between the relay and the eavesdropper. Also, it is shown by the resulting analysis that the AF relays have better secrecy rate than that of the DF relays in both proposed system and conventional system.

Keywords: Energy harvesting · Full duplex relay · Jamming · Physical layer security · Secrecy rate

1 Introduction

Due to the broadcast nature of wireless medium, the transmitted information is vulnerable to a number of security attacks. Physical layer security fulfils the purpose of secure transmission of data from the sender to the receiver by exploiting the physical characteristics of the channel [1]. Physical layer security has some advantages over the cryptography technique, as the former does not involve complex distribution and management of keys and has fewer overheads. There are various relaying schemes to increase the coverage of wireless systems and for enhancing the physical layer security, and among these two are widely used: DF and AF [2]. Jamming signals are used to increase the physical layer security, only when jamming power at the eavesdropper is higher than the jamming power at the destination. In conventional relaying systems, the nodes are powered up by individual battery sources to perform their operation. In some networks, sometimes recharging and replacing batteries is either undesirable or not suitable, hence

enhancing lifetime of these networks faces many difficulties [3]. EH technique [4–6] helps in solving this issue. It also provides benefits to low maintenance monitoring, reliability and usability. There are no batteries for replacing and no cost related to replacing those batteries, when the devices are powered up by EH. With the explosive growth of communicating devices, it is the need of the hour to reduce the use of high spectral efficiency and capacity and move towards energy efficient systems. As a result, EH technique is gaining a lot of importance nowadays. There are various EH protocols e.g., power splitting based relaying (PSR) protocol and time switching based relaying protocol (TSR) as shown in [7].

In this paper, a single hop wireless relaying system with energy harvesting is proposed, in which source and relay obtain power from a power beacon. It is also assumed that self interference is perfectly canceled at each node. The secrecy rate, which can be defined as the amount of information that can be transmitted securely from source to destination, of this proposed system is investigated for both AF and DF cooperative schemes.

2 System Model

The system model is a relay network with an eavesdropper, employing EH as shown in Fig. 1. It involves a power beacon B , a source node S , a destination node D , an eavesdropper node E and a relay node R . Let h_{SR}^* , h_{RD}^* , h_{RE}^* and h_{SE}^* represent complex channel gains S to R , from R to D , from R to E and from S to E , respectively. Further the complex channel gains from B to S and from B to R are denoted by h_{BS}^* and h_{BR}^* . Moreover, it is assumed that noise is complex additive white Gaussian noise at each node with variance σ^2 and mean zero and there is no self interference. Further, the relay is uses both full and half duplex operation [8] in this proposed system.

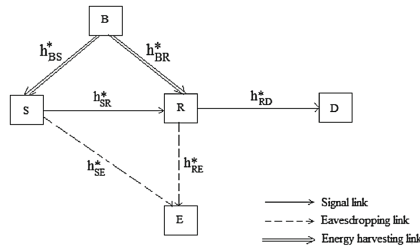


Fig. 1. Proposed secrecy relay network with one eavesdropper employing EH.

2.1 Energy Harvesting Scheme

In this proposed system, the nodes, Source (S) and Relay (R) are harvesting energy from beacon (B) and later this energy is utilized for the transmission of signals from S-R and R-D. For the high throughput, the time switching based EH technique is used. Figure 2 shows this protocol.

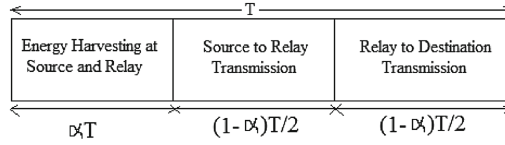


Fig. 2. Time switching based protocol [7].

Energy harvested at S and R , respectively is [9]

$$E_S = \eta P_B \alpha T |h_{BS}^*|^2 \quad (1)$$

$$E_R = \eta P_B \alpha T |h_{BR}^*|^2 \quad (2)$$

Where, the efficiency coefficient of this process is given by $0 < \eta < 1$, the power transmitted by B is represented by P_B and $0 < \alpha < 1$. T denotes the time required for sending a particular block from S to D . As shown in Fig. 2, both S and R harvest energy for a duration of αT . The power transmitted by S and R used in this proposed system are given by [9]

$$P_S = \frac{2\eta P_B |h_{BS}^*|^2 \alpha}{1 - \alpha} \quad (3)$$

$$P_R = \frac{2\eta P_B |h_{BR}^*|^2 \alpha}{1 - \alpha} \quad (4)$$

2.1.1 Decode-and-Forward

It involves two steps. In the first one, as shown in Fig. 3, the source transmits the $x(n)$ signal to the relay, and at the same time, the relay transmits the jamming signal $q(2n)$ to the eavesdropper. In the time slot $2n$, the received signals at R and E are represented as [10]

$$y_R(2n) = \sqrt{P_S} h_{SR}^* x(n) + n_R(2n) \quad (5)$$

$$y_E(2n) = \sqrt{P_S} h_{SE}^* x(n) + \sqrt{P_{RJ}} h_{RE}^* q(2n) + n_E(2n) \quad (6)$$

where, the power of jamming signal of R is denoted by P_{RJ} and the additive white Gaussian noises at R and E are represented by $n_R(2n)$ and $n_E(2n)$, respectively.

As shown in the Fig. 4, in the next time slot, the relay only transmits the previously decoded signal to the legitimate destination and stops receiving any signal. At this instant, the source sends the jamming signal to the eavesdropper E . The received signals at E and D in the time slot $(2n + 1)$ are represented as [10]

$$y_E(2n + 1) = \sqrt{P_R}h_{RE}^*x(n) + \sqrt{P_{Sj}}h_{SE}^*q(2n + 1) + n_E(2n + 1) \tag{7}$$

$$y_D(2n + 1) = \sqrt{P_R}h_{RD}^*x(n) + n_D(2n + 1) \tag{8}$$

where, the power of the jamming signal of S is represented by P_{Sj} and the AWGN at D is given by $n_D(2n + 1)$.

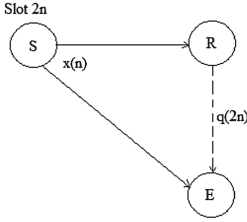


Fig. 3. Illustration of the signals transmitted in the $2n^{\text{th}}$ time slot.

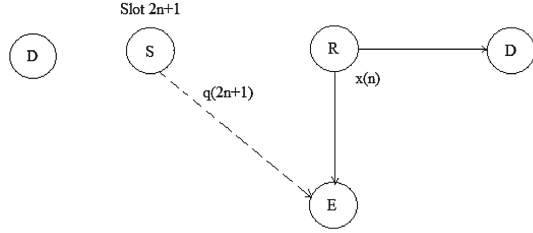


Fig. 4. Illustration of the signals transmitted in the $(2n + 1)^{\text{th}}$ time slot.

2.1.2 Amplify-and-Forward

It also involves two steps, similar to DF technique. The first step is same as that of the DF scheme as shown in Fig. 3. In the $2n^{\text{th}}$ time slot, the signals received at R and E are given by (5) and (6).

In the next time slot, the relay amplifies the signal received by the source and forwards its amplified version to the destination. As shown in Fig. 4, at this instant, the jamming signal is sent by the source to the eavesdropper. Hence, in the $(2n + 1)^{\text{th}}$ time slot, the signals received at D and E can be given as [11]

$$y_D(2n + 1) = G\sqrt{P_S}h_{RD}^*y_R(2n) + n_D(2n + 1) \tag{9}$$

$$y_E(2n + 1) = G\sqrt{P_S}h_{RE}^*y_R(2n) + \sqrt{P_{Sj}}h_{SE}^*q(2n + 1) + n_E(2n + 1) \tag{10}$$

where, $G = \frac{1}{\sqrt{P_S|h_{SR}|^2 + N_o}}$ is the scaling factor [11] and the variance of noise is given by N_o .

3 Achievable Secrecy Rate

3.1 DF Scheme

The rates at D and E using (5) and (6) can be represented as [10]

$$R_d = \frac{1}{2} \log_2(1 + P_R\alpha_{RD}) \tag{11}$$

$$R_e = \frac{1}{2} \log_2 \left(1 + \frac{P_S \alpha_{SE}}{1 + P_{RJ} \alpha_{RE}} + \frac{P_R \alpha_{RE}}{1 + P_{SJ} \alpha_{SE}} \right) \quad (12)$$

where, $\alpha_{RD} = \frac{|h_{RD}|^2}{\sigma^2}$, $\alpha_{SE} = \frac{|h_{SE}|^2}{\sigma^2}$ and $\alpha_{RE} = \frac{|h_{RE}|^2}{\sigma^2}$. It is possible to calculate the achievable secrecy rate as $R_s = \max\{R_d - R_e, 0\}$ using (11) and (12), where

$$R_d - R_e = \frac{1}{2} \log_2 \left(\frac{1 + P_R \alpha_{RD}}{1 + \frac{P_S \alpha_{SE}}{1 + P_{RJ} \alpha_{RE}} + \frac{P_R \alpha_{RE}}{1 + P_{SJ} \alpha_{SE}}} \right). \quad (13)$$

3.2 AF Scheme

The rates at D and E using (5) and (7) are given as [10]

$$R_d = \frac{1}{2} \log_2(1 + G^2 P_S \alpha_{RD}) \quad (14)$$

$$R_e = \frac{1}{2} \log_2 \left(1 + \frac{P_S \alpha_{SE}}{1 + P_{RJ} \alpha_{RE}} + \frac{G^2 P_S \alpha_{RE}}{1 + P_{SJ} \alpha_{SE}} \right). \quad (15)$$

The secrecy rate that can be achieved is represented as $R_s = \max\{R_d - R_e, 0\}$, where

$$R_d - R_e = \frac{1}{2} \log_2 \left(\frac{1 + G^2 P_S \alpha_{RD}}{1 + \frac{P_S \alpha_{SE}}{1 + P_{RJ} \alpha_{RE}} + \frac{G^2 P_S \alpha_{RE}}{1 + P_{SJ} \alpha_{SE}}} \right). \quad (16)$$

4 Numerical Results

This section shows numerical results for investigating secrecy performance of the proposed system employing EH for both DF and AF cooperative schemes. The source S , relay R and destination D are assumed to be present in a line [2] as shown in Fig. 5. Moreover, d_{SR} , d_{RD} , d_{RE} , d_{BS} and d_{BR} represent the distance between nodes S and R , between R and D , between R and E , between B and S and between B and R . The distance between S and E can be given as $d_{SE} = \sqrt{d_{SR}^2 + d_{RE}^2}$, respectively. The channel between any two nodes is the line-of-sight (LOS) channel model $d^{-\frac{c}{2}} e^{j\theta}$, where the distance between the nodes is given by d , θ denotes the random phase which is uniformly distributed within $[0, 2\pi)$, and $c = 3.5$ represents the path loss exponent [2].

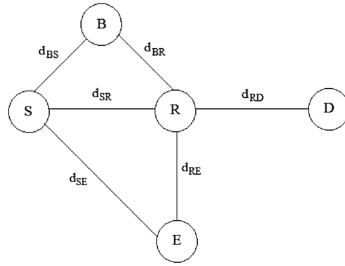


Fig. 5. Illustration of the proposed EH simulation model.

It is assumed that the transmit power of the beacon is $P_B = 37.78$ dBm and the noise power = -70 dBm. Further, it is assumed that $d_{BS} = d_{BR} = 14$ m. Moreover, $\alpha = 0.999$ and $\eta = 1$.

Figures 6 and 7 shows the secrecy rates of DF and AF relay as a function of relay-eavesdropper distance, d_{RE} when, $d_{SR} = 25$ m, $d_{RD} = 30$ m in the proposed system employing EH and in the conventional system without EH. As d_{RE} increases, the E gets shifted away from the line joining S, R and D and the secrecy rate increases for both AF and DF relay, indicating that the transmission of data to destination becomes more secure as the eavesdropper moves away from the line.

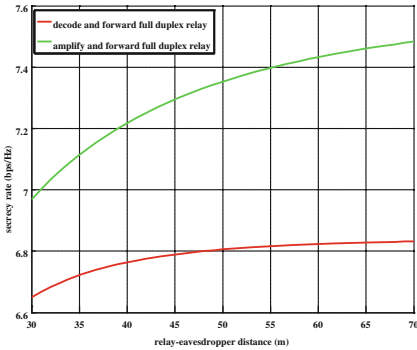


Fig. 6. Secrecy rate versus d_{RE} when $d_{SR} = 25$ m, $d_{RD} = 30$ m, $d_{BS} = 14$ m and $d_{BR} = 14$ m in proposed system employing EH.

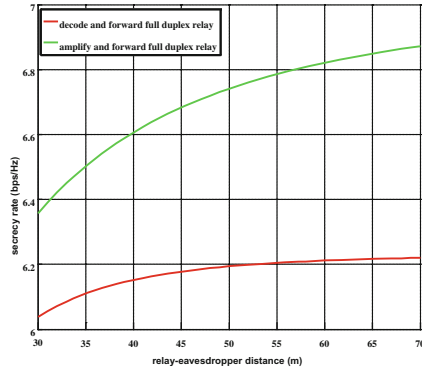


Fig. 7. Secrecy rate versus d_{RE} when $d_{SR} = 25$ m, $d_{RD} = 30$ m in the conventional system without EH.

Figures 8 and 9 depicts the secrecy rate as a function of relay-destination distance d_{RD} when $d_{SR} = 25$ m, $d_{RE} = 30$ m in the proposed system employing EH and in the conventional system without EH for both AF and DF scheme. As D moves away from R , secrecy rate decreases in both the schemes, indicating that the transmission of data to destination becomes less secure, as the distance between the destination and the relay increases.

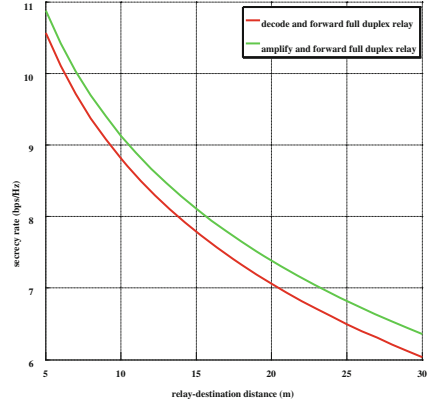
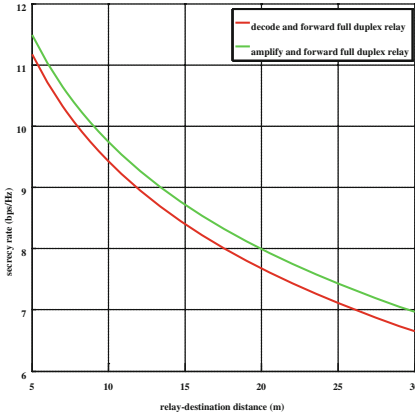


Fig. 8. Secrecy rate versus d_{RD} when $d_{SR} = 25$ m, $d_{RE} = 30$ m, $d_{BS} = 14$ m and $d_{BR} = 14$ m in proposed system employing EH.

Fig. 9. Secrecy rate versus d_{RD} when $d_{SR} = 25$ m, $d_{RE} = 30$ m in the conventional system without EH.

Figure 10 shows the secrecy rate versus path loss for both AF and DF scheme when $d_{SR} = 25$ m, $d_{RE} = 30$ m, $d_{RD} = 30$ m and $d_{BS} = d_{BR} = 14$ m in the proposed system. As the path loss exponent increases, the channel becomes worse and secrecy rate decreases for both cooperative schemes, indicating that the transmission of information becomes less secure, as the channel degrades.

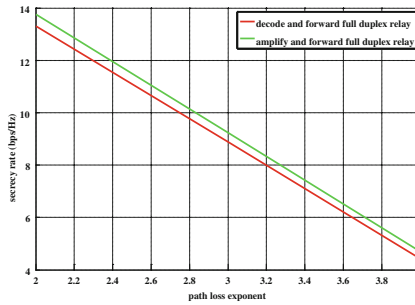


Fig. 10. Secrecy rate versus path loss exponent when $d_{SR} = 25$ m, $d_{RE} = 30$ m, $d_{RD} = 30$ m, $d_{BS} = 14$ m and $d_{BR} = 14$ m in proposed system employing EH.

5 Conclusion

In this paper, a single hop wireless relaying system, employing EH is proposed where both source and relay obtain energy from a power beacon using time switching energy harvesting technique. The secrecy performance of the proposed system is investigated for both AF and DF cooperative schemes. The proposed system with EH technique provides improvement in secrecy rate, energy efficiency and power consumption, as in the proposed system, nodes are powered up with EH technique, instead of using

individual batteries. The secrecy rate of the proposed system with EH is higher than that of the conventional system by 8.89% for AF relay and by 9.83% for DF relay at a distance of 70 m between the relay and the eavesdropper. Also, it is shown by the resulting analysis that the AF scheme outperforms DF scheme in terms of secrecy rate in both proposed system with EH and in the conventional system without EH.

References

1. Wyner, A.D.: The wire-tap channel. *Bell Syst. Tech. J.* **54**(8), 1355–1387 (1955)
2. Dong, L., Han, Z., Petropulu, A.P., Poor, H.V.: Improving wireless physical layer security via cooperating relays. *IEEE Trans. Sig. Process.* **58**(3), 1875–1888 (2010)
3. Zhang, R., Ho, C.K.: MIMO broadcasting for simultaneous wireless information and power transfer. *IEEE Trans. Wirel. Commun.* **12**(5), 1989–2001 (2013)
4. Yuen, C., Elkashlan, M., Qian, Y., Duong, T.Q., Shu, L., Schmidt, F.: Energy harvesting communications: Part 1 [Guest Editorial]. *IEEE Commun. Mag.* **53**(4), 68–69 (2015)
5. Yuen, C., Elkashlan, M., Qian, Y., Duong, T.Q., Shu, L., Schmidt, F.: Energy harvesting communications: Part 2 [Guest Editorial]. *IEEE Commun. Mag.* **53**(6), 54–55 (2015)
6. Yuen, C., Elkashlan, M., Qian, Y., Duong, T.Q., Shu, L., Schmidt, F.: Energy harvesting communications: Part 3 [Guest Editorial]. *IEEE Commun. Mag.* **53**(8), 90–91 (2015)
7. Nasir, A.A., Zhou, X., Durrani, S., Kennedy, R.A.: Relaying protocols for wireless energy harvesting and information processing. *IEEE Trans. Wirel. Commun.* **12**(7), 3622–3636 (2013)
8. Chen, G., Gong, Y., Xiao, P., Chambers, J.A.: Physical layer network security in the full-duplex relay system. *IEEE Trans. Inf. Forensics Secur.* **10**(3), 574–583 (2015)
9. Ngyuyen, N.P., Duong, T.Q., Ngo, H.Q., Hadzi-Velkov, Z., Shu, L.: Secure 5G wireless communications: a joint relay selection and wireless power transfer approach. *IEEE Access* **4**, 3349–3359 (2016)
10. Lee, J.-H.: Full-duplex relay for enhancing physical layer security in multi-hop relaying systems. *IEEE Commun. Lett.* **19**(4), 525–528 (2015)
11. Kumar, N., Bhatia, V.: Performance analysis of amplify-and-forward cooperative networks with best-relay selection over Weibull fading channels. *Wirel. Pers. Commun.* **85**, 641–653 (2015). Springer

A System Design for the Measurement and Evaluation of the Communications Security Domain in ISO 27001:2013 Using an Ontology

Pongsak Sirisom¹, Janjira Payakpate^{1(✉)}, and Winai Wongthai^{1,2}

¹ Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand

pongsaks58@email.nu.ac.th, {janjirap,winaiw}@nu.ac.th

² Research Center for Academic Excellence in Nonlinear Analysis and Optimization, Faculty of Science, Naresuan University, Phitsanulok, Thailand

Abstract. This paper presents a system design using the design and linking semantic technology of ontologies by mapping the structure base and finding identical meanings of each text. The Wu and Palmer method and WordNet database were used for this purpose. The accuracy of the results of the concept are measured by using Recall, Precision, and F-Measure. Then, the proposed designed can be used to developed tools to qualify the security system for communications security domain under the standards of information security management for ISO 27001:2013. However, the cost of certification to organisations to meet international standards is considerable. Our intention was to demonstrate the ontology-based concept for organisations to be able to reduce their certification costs by waiving the requirement for an external consultant to evaluate their standards and policies.

Keywords: ISO 27001:2013 · Ontology mapping · Similarity comparison on ontology

1 Introduction

Information Security Management System: ISMS (ISO 27001) [1] is an international standard for information security management which emphasizes various requirements for avoiding any harmful threats to the data in an organisation's management system. The ISO 27001 standard recommends the creation of an organisational emergency response plan or security plan to reduce losses and maintaining business continuity in the face of such threats to the data and operations of the organisation. This standard has been adopted in many countries including Thailand [2, 3] where both corporations and government units have deployed this standard. A process to obtain certification of standards is often required, which then implies or necessitates the hiring of consultants for evaluating the company's policy. This means a financial burden particularly for small to medium sized enterprises. A significant problem in this context is to ensure that all recommendations of the standard have been addressed in the emergency response plan, and this can be easily and cheaply verified. To achieve this, we suggest the development

of an appropriate taxonomy which maps the plan and enables the comparison against the standard to ensure completeness.

New technology is being continually developed to enable the development of taxonomies for various subject matter areas. This has allowed many researchers to apply the concept of an ontology [5] to the problem of text matching. The ontology technique has been developed for semantic web searching and can be used map the semantics of documents, thus enabling a comparison between documents. It has been used as a tool to measure and evaluate the security operations for data communications in companies, and we have identified the concept as being applicable for the assessment of compliance with the standards published, as text, in the ISO 27001 Standard.

Research Gaps: In recent works [8–11], approaches and methods for measuring and evaluating data communications security, with analysis of ISO 27001, has been discussed. These approaches included an ontology approach and a neural network approach, limited however to mapping the ISO 27001 standards. However, only [12] has addressed the issue of comparing the ISO 27001 standard against another set of standards to identify and evaluate the similarity or congruence of the two standards. Our analysis of these studies suggests that the approaches developed have not used semantic analysis and structural analysis and are therefore, in our view, not sufficiently flexible to be able to fully and concisely measure the congruence of two sets of standards, in particular, a proprietary security plan assessed against the ISO 27001 standard. This is the particular ‘research gap’ that our study has addressed.

Summary of Contribution: Organisations must always be concerned about information security management standards [2, 3]. However, the cost of compliance with ISO international standards are significant and place a financial burden of smaller organisations [9–11] especially when independent consultants are contracted. In our study, we have developed the concept and use of an ontology to enable organisations to assess the completeness of their data and communications security plans, in accordance with the ISO 27001 standards, and, importantly to enable automatic assessment of the level of compliance with that standard, for the purpose of becoming certified. This approach will clearly provide a considerable cost benefit for the organisation by removing the need for hiring external consultants to evaluate the proprietary standards for its content and level of compliance. Our ontology based approach enables a semantic and structural analysis and mapping of a constructed ontology of a proprietary plan and a similarly constructed ontology of the ISO 27001 standard.

2 Related Work

The theoretical and practical basis of our study acknowledges prior work, but also identifies what we consider to be a more versatile approach. Below we identify and discuss these prior publications and the principle frameworks and architectures that have been discussed previously.

2.1 Related Theories

Five important concepts applicable to our work include:

2.1.1 ISO27001

[1] is the international standard for information security management, published by the International Organisation for Standardization (ISO) and the International Electro Technical Commission (IEC). For our purposes, we will refer only to the ISO 27001 nomenclature. The model includes fourteen major rules for security management, twenty-five control objectives, and one hundred and fourteen measurement controls. This research accords to latest release of ISO 27001:2013.

2.1.2 Ontology

[4] is an approach to describing knowledge, to define the boundaries of the knowledge domain (often referred to in systems development as The Universe of Discourse), the knowledge structures of interest, and to enable and identify associations within text, and which allows interpretation of class correlations, class hierarchies, and class properties. The ontology concept will be applied in our work to build relationships between classes in different domains, which, for our purposes means the ISO 27001:2013 and a proprietary set of standards.

2.1.3 Web Ontology Language (OWL)

[5] is a language used to describe an ontology and to define the relationships between data, rules and assertions in the ontology. OWL allows the representation of classes and their details and attributes, and the connections between classes.

2.1.4 WordNet

[6] is an English dictionary database containing verbs, adjectives and adverbs. Wherever appropriate, words are linked as either synonyms or antonyms. Car, for example, has several words that have the same meaning including automobile, vehicle, wheels, and so on. The similar meaning for each word are stored in the dataset, which is called synset. The synset is used to find word pairs of synonyms, antonyms and similar usages.

2.1.5 Simple Protocol and RDF Query Language (SPARQL)

[7] is the standard language for gathering information and showing the result that is stored in the form of RDF (Resource Description Framework) or OWL. SPARQL is defined by the RDF Data Access Working Group (DAWG) by the World Wide Web Consortium (W3C). SPARQL has a language form similar to other query language, such as SQL, or XQuery.

2.2 Related Research

Many researchers have developed various methods for assessing their information security in accordance with the standards of the ISO. Recent work related to our study is summarized as follows:

[9] designed an artificial neural network to implement the ISO 27001 standard for risk management. By training the neural network, a comprehensive set of questions could be answered and deficiencies or unreferenced parts of the standard identified by internal actors, enabling a small company could perform risk management within an appropriate budget. However, this method requires a lot of information from experts to teach the neural network, thus potentially defeating the purpose of using internal resources and reducing the budget requirements. [10] demonstrated an IT security framework using ontologies for small and medium enterprises (SMEs). However, this work was not specifically analyzing compliance with the ISO 27001 standard. [11] introduced an ontology framework to improve and evaluate any proprietary safety standard for compliance with ISO 27001.

This framework was applied to the safety standards in ISO 27001, and was used as a tool to prepare, evaluate, and establish safety policy guidelines in accordance with the security standard in ISO 27001. The mapping techniques in [11] using tags which can reduce the flexibility of attribute matching for example when users input words which may be misspelled. In [12] an ontology was constructed which encapsulated the policies from ISO 27001 and a German IT-security plan of an organisation. The construed ontology is used to build a tool. The tool can give security levels from 0–3 indicating lowest to highest security levels. However, these levels are not suitable in some situations. For example, level of 2.994 should be almost the highest security level, but somehow it is not, as stated by the author of [12].

Notwithstanding the volume of prior research on the subject of ISO 27001, and proprietary standards, and methods for evaluating the compliance of the latter to the former, each prior study has limitations, as indicated in each discussion above. It is these various limitations that we have attempted to overcome by using an ontology based approach for comparison between standards statements, and the level of compliance, one to the other. In general, we are addressing the efficiency and effectiveness of the ontology based approach for this purpose.

3 The Proposed Design

This section represents the workflow of the proposed concept, architecture design, and context similarity comparison on ontology.

3.1 A Workflow of the Proposed Concept

A workflow, using an ontology based approach, for measuring and evaluating a proprietary set of standards for communication security for the level of compliance and conformity to the international standard of Information Security Management ISO

27001:2013. Semantic network concepts and ontology mapping will be applied to achieve the system purpose. The proposed system is illustrated in Fig. 1.

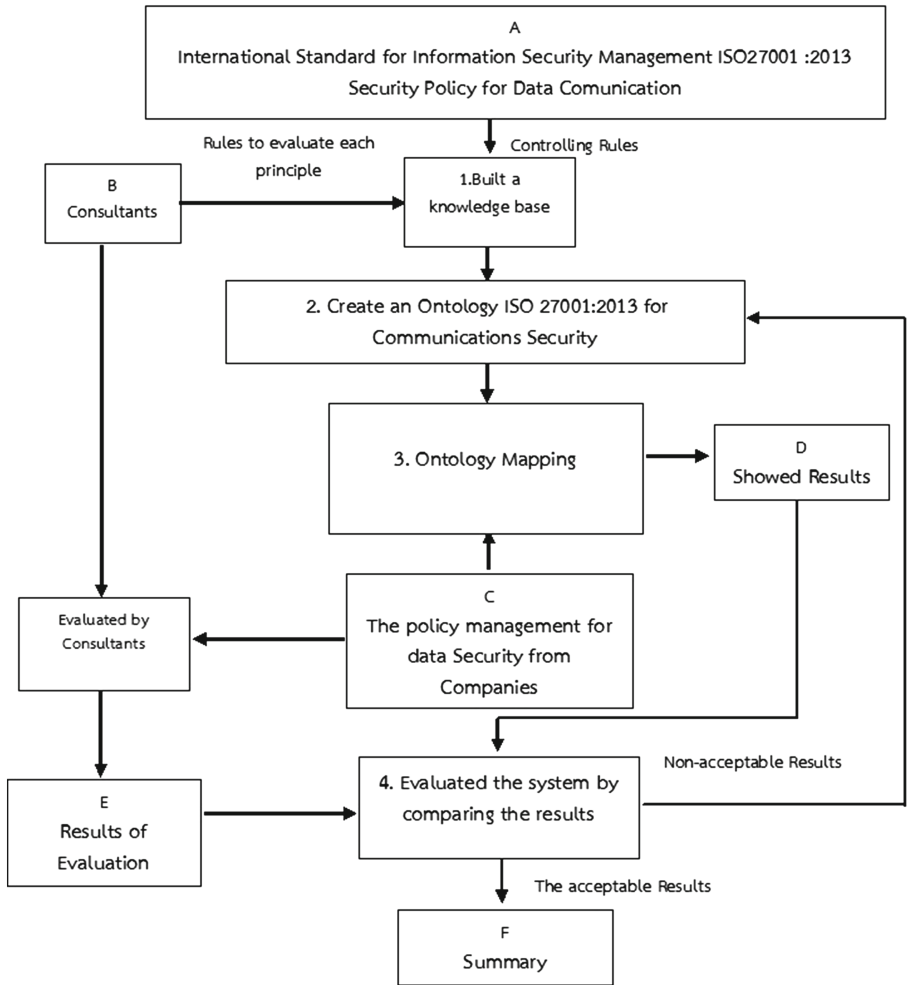


Fig. 1. The workflow of the proposed concept

The development process comprised four steps:

Step 1: Building the knowledge base of the stated objectives of the ISO 27001:2013, and all of the details of that standard of communications security domain (show in Fig. 1A). A conceptual model of an ontology of this knowledge base, including the classes, objects, attributes, constraints and rules to be included in the ontology, was developed. This was guided by expert opinion and perspective (Fig. 1B). Step 2: Creation of the Ontology, implementing the knowledge based decided upon in Step 1. Step 3: Ontology mapping, a model was developed to assess

the level of compliance between ontologies of a proprietary set of standards (show in Fig. 1C) and of the ISO 27001:2013 standard by assessing content similarity in the two ontologies. This will produce results (show in Fig. 1D). Step 4: System evaluating, the acceptance of results will be yielded by comparing between the results from a new system (show in Fig. 1D) and the evaluation results (show in Fig. 1E) from ISO specialists (show in Fig. 1B) for ISO 27001:2013. If the results of the new system are acceptable, the system findings can be summarized effectively (show in Fig. 1F). On the other hand, if the results are not acceptable, the ontology of ISO 27001:2013 should be improved and tested again. Therefore, the process has to be returned to Step 2.

3.2 Architecture Design

To create an architecture design of ontology mapping in Step 3 in Fig. 1, there are four steps: data importing, data converting, associating and presentation of results. The conceptual design for each step of the ontology mapping architecture is shown in Fig. 2. First, the text relating to topics in the standards (Communication Security standards from ISO 27001:2013 and the organisation’s security policy document) is imported as data (circle number 1 of Fig. 2).

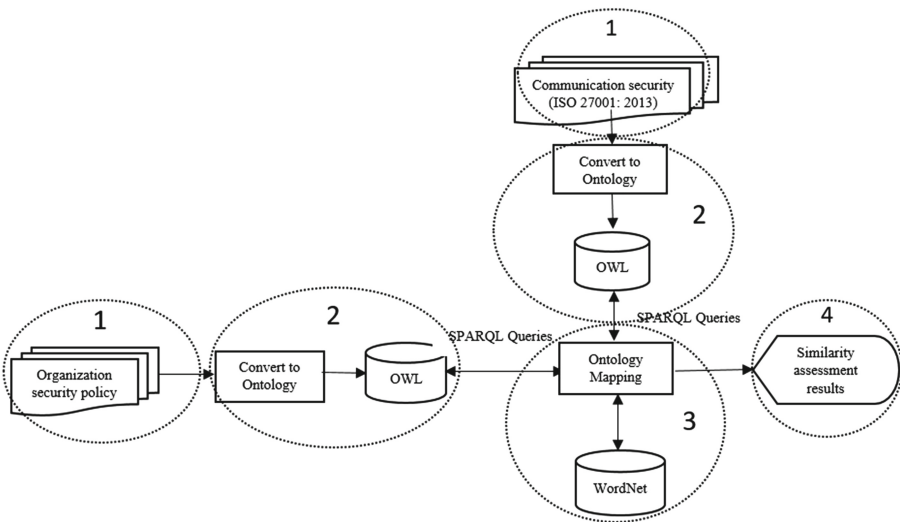


Fig. 2. Architecture design

Second, the imported text is converted into ontology structures (circle number 2 of Fig. 2). This conversion results are OWL files [5]. Third, mapping of associations is done according to the rules stated in the ontology (circle number 3 of Fig. 2). The Wu and Palmer method (see more detail in Sect. 3.3) is used for seeking identical meanings of words. Lastly, the results are presented (circle number 4 of Fig. 2) after the mapping between the two OWL files with WordNet database.

The most important aspect of the architecture design is the ontology mapping in step 3 (circle number 3 of Fig. 2). This step will show whether or not the mapping system in this step works properly. This is presented in detail in Sect. 3.3.

3.3 Context Similarity Comparison on Ontology

There are two steps in this comparison: finding two ontology structures similarity and finding words similarity in these two structures.

First, ontology mapping is a methodology to find similarities between classes in the separate ontologies and class attributes. The mapping can be conducted in several ways [13] such as structure base, syntax base, instance base, and constraint base. In this work, the structure base will be applied since the base can consider as the hierarchical structure within the ontology including classes, subclasses, attributes, and relationships. The results of a comparison between two hierarchical structures of the ontologies will be used as a conclusion whether two structures are identical or not. Second, only when two structures are identical, way to find words similarity in sentences of policies between the two ontologies is by selecting a word from both ontologies and comparing them. Here, the Wu and Palmer method [14, 15] will be used to find the similarity from the WordNet [8] database by considering the depth of the synset for the two words within the structure of a taxonomy. Depth is defined as the number of layers from the concept (in this case a word) of the root in the hierarchical structure of the WordNet. LCS stands for Least Common Subsumer which refers to the distance between two concepts (in this case the two words) in the synset. The similarity value (Sim_{wup}) can be calculated as

$$Sim_{wup} = \frac{2 \times depth(LCS)}{depth(concept1) + depth(concept2)} \quad (1)$$

The value of Sim_{wup} must be between zero and one where zero means no similarity, and if the value of Sim_{wup} is 1 there is complete equivalence; the two concepts, here words, mean exactly the same thing; they are exact synonyms. This is not a binary measure, and measures the degree of similarity.

After the workflow (in Fig. 1) and process mapping ontology (circle number 3 of Fig. 2) have been designed, the validation of the system must be evaluated. The method of evaluation is shown in Sect. 4.

4 Model Validation

In the third step in Sect. 3.2, mapping of associations is done according to the rules stated in the ontology (circle number 3 of Fig. 2). We need to validate the ontology mapping process (see the box inside the circle) using F-Measure approach. [16] also applies the approach with the Wu and Palmer method for validation of similarity of texts. Thus, this paper uses this approach.

A is 1 when the ontology mapping process yields that two texts are a synonym and expert mapping process also yields that the same result as the mapping process. B is 1 when the ontology mapping process yields that two texts are not a synonym but expert mapping

process yields that two texts are a synonym. C is 1 when the ontology mapping process yields that two texts are a synonym but expert mapping process yields that two texts are not a synonym.

To measure the accuracy, values of Recall, Precisions and F-measure [16] will be calculated by using Eqs. (2), (3), and (4) respectively.

$$Recall = \frac{A}{A + B} \times 100\% \quad (2)$$

$$Precisions = \frac{A}{A + C} \times 100\% \quad (3)$$

$$F\text{-measure} = 2 \left(\frac{Precision \times Recall}{Precision + Recall} \right) \quad (4)$$

The expectation is that the accuracy of the final result from Eq. 4 is at least 70% and from Eqs. 3 and 4 are at least 80%. This expectation is based on the approach as discussed in [16].

5 Conclusion

Our proposed concept to comparing the content of two separate texts, in this case the ISO 27001 version 2013 standard and a proprietary standard published in-house, was to construct and populate an ontology for each.

Using used semantic analysis and structural analysis to map the two ontologies to identify similarities in the texts, done according to the recommended approach of Wu and Palmer method and with access to the WordNet database, we proposed concept that should be able to measure the degree of consistency and agreement between the texts. The proposed concept will validate the ontology mapping process using F-Measure approach. The approach also needs to produce at least 70% of the accuracy of the process.

Using our concept, small to medium organisations seeking certification of their adherence to the international standard can demonstrate the level of that adherence without the need for substantial expenditures, which has been a significant problem standing in the way of many organisations seeking certification of compliance with the international standards.

Acknowledgement. Many thanks to Mr. Roy Morien and Mr. Kevin Roebel of the Naresuan University Language Center for his editing assistance and advice on English expression in this document.

References

1. An Introduction to ISO 27001 (ISO27001). <http://www.27000.org/iso-27001.htm>. Accessed October 2016
2. Kanno, Y.: Information Security Measures Benchmark (ISM-Benchmark). IT Security Center, Information-Technology Promotion Agency, Japan (2009)

3. Sharma, N.K., Dash, P.K.: Effectiveness of ISO 27001, as an information security management system: an analytical study of financial aspects. *Far East J. Psychol. Bus.* **9**(3), 42–55 (2012)
4. Uschold, M., Gruninger, M.: *Ontologies principles methods and applications*. *Knowl. Eng. Rev.* **11**(2), 93–155 (1996)
5. World Wide Web Consortium: *OWL Web Ontology Language* (2004). <http://www.w3.org/TR/owl-features/>
6. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: an on-line lexical database. *Int. J. Lexicogr.* **3**, 235–244 (1990)
7. Pérez, J., Arenas, M., Gutierrez, C.: *Semantics and complexity of SPARQL*. Universidad de Talca, Chile (2009)
8. Lovrić, Z.: Model of simplified implementation of PCI DSS by using ISO 27001 standard. In: *Central European Conference on Information and Intelligent Systems*, 19–21 September 2012
9. Shrivastava, A.K.: ISO27001 compliance via artificial neural network. In: *13th IEEE International Symposium on Pacific Rim Dependable Computing* (2007)
10. Fenz, S., Weippl, E.: *Ontology based IT-security planning*. In: *Secure Business*, Austria (2006)
11. Fenz, S., Goluch, G., Ekelhart, A., Riedl, B., Weippl, E.: Information security fortification by ontological mapping of the ISO/IEC 27001 standard. In: *13th IEEE International Symposium on Pacific Rim Dependable Computing* (2007)
12. Fenz, S.: *Ontology-based generation of IT-security metrics*. In: *SAC 2010*, Sierre, Switzerland, 22–26 March 2010
13. Liu, X., Cao, L., Dai, W.: Overview of ontology mapping and approach. In: *2011 4th IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT)*, pp. 592–595, 28–30 October 2011
14. Wu, Z., Palmer, M.: Verb semantic and lexical selection. In: *Proceeding of 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Las Cruces, US, pp. 133–138 (1994)
15. Corley, C., Mihalcea, R.: *Measuring the semantic similarity of texts*. Department of Computer Science, University of North Texas (2005)
16. Fernando, S., Stevenson, M.: *A semantic similarity approach to paraphrase detection*. Department of Computer Science, University of Sheffield, Sheffield, UK (2008)

Timing Side Channel Attack on Key Derivation Functions

Chai Wen Chuah^(✉) and Wen Wen Koh

University Tun Hussein On Malaysia, Parit Raja, Malaysia
cwchuah@uthm.edu.my, ai140157@siswa.uthm.edu.my

Abstract. A key derivation function is a function that generate one or more arbitrary length of cryptographic keys from a private string together with some public information. The generated cryptographic key(s) from this key derivation function proposals are generally indistinguishable from random binary strings of the same length based on formal mathematical proof. To date, there are designed of key derivation function proposals using cryptographic primitives such as hash functions, block ciphers and stream ciphers. However, there are limited security analysis of side channel attacks for the key derivation function proposals. This paper is to investigate the timing side channel attacks towards these three types of cryptographic primitives based key derivation function. Key derivation functions based on stream ciphers and block ciphers are input-dependent execution, the experiment results have shown that both key derivation functions proposals are vulnerable against timing side channel.

Keywords: Key derivation function · Timing side channel attack · Hash function · Block cipher · Stream cipher

1 Introduction

A key derivation function (KDF) is a fundamental component of many cryptographic systems. The KDFs are used to generate one or more arbitrary length of cryptographic keys. These cryptographic keys are used with a cryptographic algorithm for protecting electronic data during transmission over insecure channels. The KDFs take the private string together with some optional public strings as inputs in order to generate these cryptographic keys. The private string can be a password, Diffie-Hellman (DH) shared secret or non-uniformly random source material [8, 12, 14]. The public strings can be random salt value and/or context information. Note that the private strings cannot be used directly as encryption keys, as these private strings are not properly distributed. The derived cryptographic keys are required indistinguishable from random binary strings of the same length [5, 11].

Single phase and two-phase are the design's for KDFs. The inputs to single phase KDF are private string and/or public strings. The latest design is two-phases key derivation functions consist of extractor and expander. The inputs

for extractor are the private string combine with a random salt value. The inputs for expander are the output from the extractor with context information. The two-phase KDF proposals consisting of an extractor and an expander are the current preferable design trends as it can be designed and analysed separately.

To date, there are KDF proposals are composed by using hash functions [4,12], block ciphers [4] and stream ciphers [5]. Hash functions and block cipher based MACs transform a variable-size input into a fixed-length output, modification for both ciphers are required to follow properties of KDF which is the generation of arbitrary length of cryptographic keys. Generally, the approach is to produce multiple output blocks until the required length has been obtained and to discard any bits in excess of the required length. The keystream generator for a stream cipher is designed to take two inputs: a short secret key and some public information, and produce a output sequence of arbitrary length. A modification for keystream generator for a stream cipher is carried out to transform a variable-size input as for KDF to generate arbitrary length of cryptographic keys.

Side channel attacks are attack that target on the observation of the cryptographic's implementation for obtaining useful information in order to discover the secret of cryptographic algorithms. The physical observables that come from the investigation of an adversary may result in the information useful during cryptanalysis. This useful information is known as side channel information. Power consumption, timing analysis, and electromagnetic radiation are examples of side channel information. This research project is focusing on timing side channel attack on KDFs. The goal of this research project is to investigate by applying the timing side channel attack on KDFs, which types of information that the adversary can be extracted from the experiments.

2 Related Works

2.1 KDFs Based on Steam Ciphers

The stream cipher based KDF (SCKDF) is a two-phase model where both the extractor and the expander are based on ideal pseudorandom keystream generators [5]. The pseudorandom keystream generator takes two inputs: a key and an initial vector (IV). In SCKDF, the pair of inputs to the pseudorandom keystream generator (key, IV) are replaced with the input pair (p , s). If s is not null, the private string p is divided into blocks which same length with key, else p is divided into blocks with the total length of key and the length of IV. The pseudorandom keystream generator executes entire blocks of p and generates an intermediate string, namely PRK . The length of PRK is same with the length of the key. The PRK is the input to the extractor phase together with c . If the length of c is greater than the length of pseudorandom keystream generator IV, then c is divided into the blocks same length of IV. The pseudorandom keystream generator executes entire blocks of c then generates the n bits length of cryptographic key. Noted that the pseudorandom keystream generator for extractor and expander

can be same type pseudorandom keystream generators as well as combination two different pseudorandom keystream generators.

2.2 KDFs Based on Hash Functions

Krawczyk proposed a KDF using HMAC-SHA families (HKDF) [11]. The proposed KDF is two-phase KDF which consists of a computational extractor and a pseudorandom expander. The extractor function is $Ext_p(s) : F((s \oplus opad) \| F((s \oplus ipad) \| p))$, where F denotes a hash function. The output for this phase (PRK) is based on the length of hash digest. The s is proposed has the same length as the hash digest of F , we denoted it as fl . If $sl < fl$ or $sl > fl$, s is hashed to have $sl = fl$. The expander phase of the HKDF functions is $Exp_{PRK}(c, n) : K(1) \leftarrow F(PRK \oplus opad) \| F((PRK \oplus ipad) \| c \| 0)$ and F is the hash function. If $kl < fl$ or $kl > fl$, s is hashed to have $kl = fl$. This scenario can be happened when two different types of hash function are used to construct the extractor and the expander. For example, SHA256 is used to construct the extractor and SHA512 is used to build the expander. The output length for the expander phase is fl . If $n > fl$, second or more iterations are necessary until the required length has been obtained, for example $fl \geq n$. The extractor function is as below: $K(i+1) \leftarrow F((PRK \oplus opad) \| F((PRK \oplus ipad) \| K(i) \| c \| i))$, $1 \leq i < t$, where $t = \lceil \frac{n}{fl} \rceil$. The cryptographic key is the concatenation string such that $K(1) \| K(2) \| \dots \| K(t-1)$. The first n bits are used as the cryptographic key and the remaining bits are discarded.

2.3 KDFs Based on Block Ciphers

AES-CMAC based KDF is a two phase KDF [15]. CMAC is a keyed hash function that is based on a symmetric key block cipher, such as AES [17]. The AES block cipher supports key sizes of 128, 192 and 256 bit. The output size for AES is 128 bits. The AES-CMAC based extractor can be either AES-128, 192 or 256, but the expansion is fixed to use AES-128 as recommended by the authors.

The extractor function for AES-CMAC is $PRK_i = F_s(PRK_{i-1} \oplus p_i)$, where F is AES (128 or 192 or 256), $1 \leq i \leq t$, $t = \lceil \frac{pl}{128} \rceil$ and $PRK_0 = 0^{128}$. In the extractor phase, p is divided into 128 bits per block, together with key (s) as the input to the AES. Output from the block processed is XORed with the next input block together with the key and processed by using AES. The process is continuing until the last block of input. There is slightly different operation for processing the last block of the p . If the last block is a padding block, then the subkey is K_2 else the subkey is K_1 , such that $p_t = p_t \oplus K_b$, $b \in \{1, 2\}$. The output from this extraction phase is 128 bit and we denoted it as PRK .

The inputs to the expander is PRK and c . The block ciphers that are used to build this expander are AES-128. PRK is used as the key to AES as well as subkey generation, K_1 and K_2 . The extractor function is as below: $K(i) \leftarrow F_{PRK}(K_{i-1} \oplus c_i)$ where F is AES-128, $1 \leq i \leq t$, $t = \lceil \frac{cl}{128} \rceil$ and $K(0) = 0^{128}$. The c is divided into 128 bits equally size of block. Each block is processed

sequentially by using AES and PRK as the key. Again, if the last block of c is padding block, it will XOR with subkey K_2 else the last block is XORed with subkey K_1 , such that $c_t = c_t \oplus K_b$, $b \in \{1, 2\}$. Once $i = t$, the function output 128 bit of string. If the system requires 150 bits on cryptography key. This mean the output n is greater than 128 bit, another iteration with the same PRK is used to produce the next 128 bit. As a result, the expander produces 256 bit but the system only takes 150 bits and the remaining bits are discarded.

2.4 Side Channel Attack

A side channel attack is an attack that use side channel information in order to retrieve secret data from a cryptosystem [7]. Side channel attacks aim to completely bypass the mathematical security of a cryptographic system, but focusing on observing the side-effects of the cryptosystem's implementation. The side-effect that disclose from the implementation of cryptosystem may expose some correlation with internal secret parameters. There are three major types of side channel attacks namely power analysis attack [9], timing attack [2, 10] and electromagnetic attack [13]. For power analysis attack, one analyses the relationship between the power consumption and the instructions that executed by a processor. Besides that, the relationship between the data in which the processor operate and the power consumption is exploits. For timing attack, timing attack focus on the correlation between the data that the algorithms operate and the time taken to perform an operation in order to recover the secret parameters. For the electromagnetic attack, electromagnetic attack can be performed by measuring the emitted electromagnetic radiation from device use and then analyse the signal produce. The method and techniques use by electromagnetic attack in order to exploit some information are quite similar to power attack. In next section, all three types of side channel attacks are explained in details.

Timing Attack. Timing attack enable the attacker to reveal the vulnerabilities of the cryptosystem by analyse the amount of time required to execute cryptographic algorithms. This is because, the amount of time taken is depend on the type of input. Cryptosystem needs slightly different amounts of time to process different inputs. Performance optimizations, branching and conditional statements, RAM cache hits and instructions of processor are the factors that influence the amount of information leakage. Kocher and Paul had showed that an attacker may be able to find fixed Diffie-Hellman exponents and factor RSA keys in order to break other cryptosystems by carefully measuring the time taken required to perform private key operations [10]. OpenSSL is another example which had been successfully attack by the timing attacks which demonstrated by Brumley and Boneh [2]. These two example had shown that he main concept of timing attack is analysing the relationship between the data involved in carry out the operations and the variations in time execution.

3 Timing Side Channel Attacks

The timing side channel attacks against the KDFs based on stream ciphers, hash functions and block ciphers are conducted by measuring the execution time taken to generate cryptographic keys of length n from p , s and c . The length of p is between 10 bytes until 128 bytes. For KDFs based on stream ciphers, the length of s and c are based on the length of IV. Another experiment for KDFs based on stream ciphers is the s is null and the c is remains same length with IV for that particular stream ciphers. For KDFs based on hash functions, the length of s is based on the length of the key (HMAC). For KDFs based on block ciphers, the length of s is based on the block ciphers key’s length. When the s is null for KDFs based on hash functions and block ciphers, both will set the s and zero with the length as the key (see Sect. 2). Therefore, we are not simulate a nul s for both cases. This experiment is to capture any information that we can get through the time side channel attacks, so we just discarded the time to generate the n length of cryptographic key.

We constructed the KDFs based on stream ciphers, hash functions and block ciphers. The stream ciphers are Trivium [3] and Rabbit [1]. The hash functions are SHA256 and SHA512. and block cipher used is AES128. All these KDFs designs described in Sect. 2. The code of the stream ciphers, hash functions and block ciphers are retrieved from [6, 16, 18] respectively. For each experiment, measurement were taken 100 times. The execution time was captured using CLOCK_MONOTONIC. The average time (mean) is recorded. All the simulations were performed at a machine with the following specifications: Intel (R) core (TM) i7-4790 CPU @ 3.60 GHz 3.60 GHz, 8 GB RAM and in 64 bit OS.

4 Result and Discussion

Figures 1 and 2 show the experiment results for KDFs based on stream ciphers. The stream ciphers are Trivium and Rabbit. Figures 3 and 4 show the experiment results for KDFs based on hash functions. The hash functions are SHA256 and SHA512. Figure 5 presents the experiment result for KDFs based on block ciphers. The block cipher is the AES.

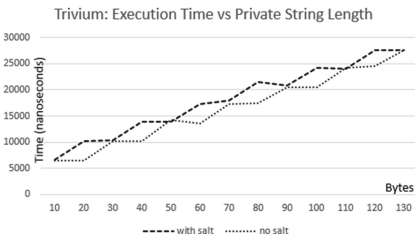


Fig. 1. Trivium based KDFs.



Fig. 2. Rabbit based KDFs.

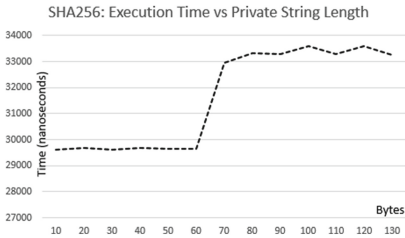


Fig. 3. SHA256 based KDFs.

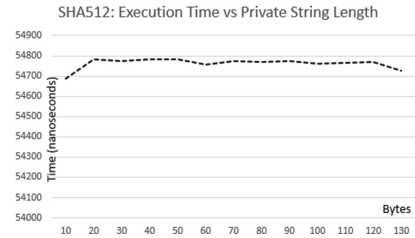


Fig. 4. SHA512 based KDFs.

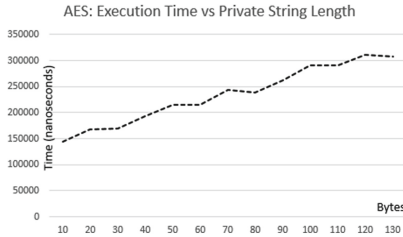


Fig. 5. AES based KDFs.

Looking at the graphs for KDFs based on stream ciphers (Figs. 1 and 2) for both with salt and without salt execution time, with s execution time is slower than without s . The difference in the execution time are distinct for the adversary to observe whether the inputs is either include s or vise versa. However, for certain cases the adversary may not be able to distinguish whether the key generation include s or not. Take an example the Rabbit based KDF when the p size is 30 bytes. The execution time for s and without s are similar. The reason is the key size (v) and IV size (w) for Rabbit are 16 bytes and 8 bytes respectively, this mean $v + w = 24$ bytes. The number of looping for both cases (s and without s) are the same, which is two times loops. For instance, p is 30 bytes, s is 8 bytes. First round v takes 16 bytes of p , w takes 8 bytes. The second round v takes the remaining 14 bytes which comes up total 30 bytes of p . The remaining length of v and w are padded with '0'. On the other case, where s is null and the p is 30 bytes. First round $v + w$ takes 24 bytes of p . The remaining 6 bytes has to be executed for the second rounds. So, total both cases have to perform two times loops to completely execute all bytes of p . Once we may observe for these two figures are the increment for Trivium based KDF is gradually. While Rabbit based KDF is same when p is between 30 bytes until 60 bytes. One of the reason is because the key size and IV size for Trivium are same which is 10 bytes each. The Rabbit key size is 16 bytes and IV is 8 bytes.

Next, the total execution time of KDFs based on hash functions, the block size for SHA256 and SHA512 is 64 bytes and 128 bytes respectively. Hash functions execute the input in fixed block size. In Fig. 3, there turns out two major execution time for p , that is from 10 bytes until 60 bytes and from 70 bytes until

128 bytes. The execution time for p between 70 bytes to 128 bytes are double as SHA256 needs to loop twice to completely execute the p . In Fig. 4, all execution time for all ranges of p are similar. This is because SHA512 execute fixed block size which is 128 bytes. More specifically, we can say that KDFs based on SHA256 and KDFs based on SHA512, the adversary is hard to distinguish the size of p if the length of p is not multiply of 64 bytes and 128 bytes respectively.

Finally, the graph of AES based KDF (see Fig. 5). One may observe from this experiment is that the execution time increase when the p length increase gradually. Since the block size for AES is 16 bytes, the p is divided in 16 bytes per block, if the size of p increase, then the number blocks are increased. Hence, the number of iterations increase gradually parallel with number blocks produced by dividing p into blocks.

In conclusion, the KDFs based on stream ciphers and the KDFs based on block ciphers which designs are input length dependent execution allow the adversary to gain side information about KDFs which is the length of p . This timing side channel attacks can be exploited by the adversary to recover the length of p thereby providing a significant advantage over a traditional brute-force attack towards p . However, the attacks does not leak any information about the p as the KDFs are design not to contents dependency. Hence, to retrieve the information of p , the adversary requires to attack the cryptographic primitives that are used to build the KDFs.

5 Conclusion

We have simulated timing side channel attacks towards KDFs based on stream ciphers, block ciphers and hash functions. The results have shown that KDFs based on stream ciphers and KDFs based on block ciphers are vulnerable to the timing side channel attacks. For both KDFs based on stream ciphers and block ciphers, the side information that the adversary may know the length of p is used to generate the cryptographic key. The additional information that the adversary can gain from KDFs based on stream ciphers is that whether the inputs are including the salt or without salt. However, this side channel attacks are failed to expose the information about p as the KDFs are designed not contents based dependency.

Nowadays, KDFs have been used in many application. The most common use of KDFs are to take the passwords together with a salt in order to derive cryptographic keys. This application is known as the password based KDFs (PBKDF). The PBKDF mainly use salt and repeated hash computation to carry out the operation [8]. If the private string is password, once the adversary can gain the information about the length of p using the timing side channel attacks, an interesting point is to investigate how fast the adversary is able to recover the p compare with the traditional brute force attack.

Acknowledgments. This research was supported by Fundamental Research Grant Scheme (FRGS) 1558, ORICC UTHM and eGates UTHM.

References

1. Boesgaard, M., Vesterager, M., Zenner, E.: The Rabbit stream cipher. In: Robshaw, M., Billet, O. (eds.) *New Stream Cipher Designs*. LNCS, vol. 4986, pp. 69–83. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-68351-3_7](https://doi.org/10.1007/978-3-540-68351-3_7)
2. Brumley, D., Boneh, D.: Remote timing attacks are practical. *Comput. Netw.* **48**(5), 701–716 (2005)
3. Cannière, C., Preneel, B.: TRIVIUM. In: Robshaw, M., Billet, O. (eds.) *New Stream Cipher Designs*. LNCS, vol. 4986, pp. 244–266. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-68351-3_18](https://doi.org/10.1007/978-3-540-68351-3_18)
4. Chen, L.: NIST SP 800-56C: Recommendation for Key Derivation through Extraction-then-Expansion. Technical report, NIST (2011)
5. Chuah, C.W., Dawson, E., Simpson, L.: Key derivation function: the SCKDF scheme. In: Janczewski, L.J., Wolfe, H.B., Sheno, S. (eds.) *SEC 2013*. IAICT, vol. 405, pp. 125–138. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-39218-4_10](https://doi.org/10.1007/978-3-642-39218-4_10)
6. Eastlake, D., Hansen, T.: RFC 6234: US Secure Hash Algorithms. Technical report, Internet Engineering Task Force (2011)
7. Joye, M., Olivier, F.: Side-channel analysis. In: van Tilborg, H.C.A., Jajodia, S. (eds.) *Encyclopedia of Cryptography and Security*, pp. 1198–1204. Springer, Heidelberg (2011)
8. Kaliski, B.: RFC 2898: PKCS# 5, Password-based Cryptography Specification version 2.0. Technical report, Internet Engineering Task Force (2000)
9. Kocher, P., Jaffe, J., Jun, B., Rohatgi, P.: Introduction to differential power analysis. *J. Cryptographic Eng.* **1**(1), 5–27 (2011)
10. Kocher, P.C.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Kobitz, N. (ed.) *CRYPTO 1996*. LNCS, vol. 1109, pp. 104–113. Springer, Heidelberg (1996). doi:[10.1007/3-540-68697-5_9](https://doi.org/10.1007/3-540-68697-5_9)
11. Krawczyk, H.: Cryptographic extraction and key derivation: the HKDF scheme. In: Rabin, T. (ed.) *CRYPTO 2010*. LNCS, vol. 6223, pp. 631–648. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-14623-7_34](https://doi.org/10.1007/978-3-642-14623-7_34)
12. Krawczyk, H., Eronen, P.: HMAC-based Extract-and-Expand Key Derivation Function (HKDF). Technical report, RFC 5869 (2010)
13. Longo, J., Mulder, E., Page, D., Tunstall, M.: SoC It to EM: electromagnetic side-channel attacks on a complex system-on-chip. In: Güneysu, T., Handschuh, H. (eds.) *CHES 2015*. LNCS, vol. 9293, pp. 620–640. Springer, Heidelberg (2015). doi:[10.1007/978-3-662-48324-4_31](https://doi.org/10.1007/978-3-662-48324-4_31)
14. McGrew, D., Weis, B.: *Key Derivation Functions and Their Uses* (2010)
15. SP NIST. 800-108: Recommendation for Key Derivation Using Pseudorandom Functions (2009)
16. Robshaw, M.: The eSTREAM project. In: Robshaw, M., Billet, O. (eds.) *New Stream Cipher Designs*. LNCS, vol. 4986, pp. 1–6. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-68351-3_1](https://doi.org/10.1007/978-3-540-68351-3_1)
17. Song, J., Poovendran, R., Lee, J., Iwata, T.: The AES-CMAC algorithm. Technical report (2006)
18. Song, J.H., Poovendran, R., Lee, J., Iwata, T.: RFC 4493: The AES-CMAC Algorithm. Technical report, Internet Engineering Task Force (2005)

A Security Aware Fuzzy Embedded ACO Based Routing Protocol (SAFACO) in VANETs

Hang Zhang^(✉), Xi Wang, and Dieter Hogrefe

Institut fuer Informatik, Georg-August-Universitaet Goettingen,
Goldschmidtstrasse 7, 37077 Goettingen, Germany
{hang.zhang, hogrefe}@cs.uni-goettingen.de,
xi.wang@stud.uni-goettingen.de

Abstract. Intelligent vehicle technologies have been developed rapidly in recent year. Smart vehicles can exchange the essential information, such as the road and weather conditions, to guarantee the road safety. However, vehicles communicate with each other or the Road Side Unit (RSU) via wireless channel, which poses many challenges to fulfill the security requirements. In the data routing process, attackers can analyze, modify or drop the data packets to reveal the vehicles' privacy or disturb the normal communication. In this paper, we devise a Security Aware Fuzzy embedded ACO routing algorithm (SAFACO) in VANETs, which combines digital signatures with fuzzy logic embedded ACO based routing protocol. After introducing the main structure of SAFACO, we also provide a detailed discussion of its security mechanism. The proposed secure communication scheme is able to provide the authentication, ensure the data consistency, preserve the privacy of vehicles, also to detect and isolate malicious vehicles.

Keywords: Security · Fuzzy logic · ACO · VANETs

1 Introduction

1.1 VANETs

Vehicular Ad hoc Networks (VANETs) are special MANETs, which usually consist of fast moving vehicles and some well-equipped Road Side Units (RSUs). Due to the high mobility of vehicles, the topology of a VANET is changing frequently. Therefore, the transmission time of data packets in this kind of networks is strongly limited. However, vehicles normally move along the roads. Based on observed mobility patterns, their position could be reasonably predicted.

There are three different types of communications in VANETs: The Vehicle-to-Vehicle communication (V2V), the Vehicle-to-Infrastructure communication (V2I) and the communication between the roadside devices. In order to communicate with the RSUs, On-Board Units (OBUs) are installed in the vehicles. It is a core element of next generation intelligent transportation systems (ITS), which aiming to manage vehicle traffic and assist drivers with safety information etc. VANETs can be distinguished from other types of ad hoc networks by the following features [1, 2]:

1. *Dynamic and rapidly-changing network topology:*

Due to the high speed of vehicles and limited transmission range of vehicles and RSUs, the topology of the vehicular networks is highly dynamic changing. According to the 801.11p specification, 1000 m is the maximum wireless transmission range of each vehicle [1]. Suppose that if there are two vehicles moving in opposite directions, the connectivity of the network may change rapidly

2. *Predictability of vehicular pathway:*

Nowadays, most vehicles are equipped with a Global Positioning System (GPS), which enables the utilization of current position information for communication. Vehicular nodes are also usually constrained by pre-built highways, roads and streets in the map, so the vehicle pathway can be predicted according to the mobility model. For example, Kinetic Graphs framework [3] can be used to predict and manage the mobility by modeling a vehicle's trajectory to make forwarding decision, which enhances the network performance.

3. *Sufficient energy and resources and hard delay constraints:*

The vehicle engine offers continuous power for communication devices, which means rich resources are available in VANETs. GPS and digital map can be used by the vehicle to obtain location, velocity and direction information. Moreover, due to the fast moving and limited efficient transmission range, vehicular networks require a reliable multi-hop scheme to minimize communication delay for active safety applications.

In recent years, intelligent vehicle technologies have been developed rapidly and hence modern vehicles are getting much smarter. Vehicles can communicate with not only infrastructures, such as RSUs, but also with other vehicles. Along one journey, a vehicle usually needs to communicate with other vehicles or RSUs to get a variety of essential information, such as traffic and environment monitoring information. There exist various applications to benefit the drivers in VANETs. Da Cunha et al. have classified the common applications in VANETs into following five categories [4]: safety applications, efficiency applications, comfort applications, interactive entertainments and urban sensing. One example from the safety applications category is the vehicle onboard collision avoidance system [5]. In this approach vehicles, which move in the vicinity of an accident scenario are informed by the vehicle onboard collision avoidance system. There are also many other customer oriented comfort applications in VANETs, which can give the drivers information like the nearest restaurant's location providing the user's favorite, coupons from a nearby market, road charging, city leisure information, tourist information and so on [4]. However, all this data is exchanged via wireless channel, which poses challenges to fulfill the security requirements. In order to support these applications, a proper routing protocol is needed for guarantee the secure data transmission in VANETs. Due to the inherent characteristics of VANETs, the conventional routing protocols for MANETs are not suitable for this kind of networks. A reliable and secure routing protocol plays a critical role in ensuring the security of data transmission in VANETs. For this purpose, routing protocol needs to be designed with particular consideration to adapt the dynamic changing in VANETs.

1.2 Ant Colony Optimization

The Ant Colony Optimization (ACO) meta-heuristic is part of the swarm intelligence field and it is inspired by the foraging behavior of ants in the nature. Biological ants lay down pheromone on the traveled path to transport information. Once an ant deposits pheromone along the path, the trail value is reinforced and this might attract more ants to follow. Thus, the pheromone realizes the indirect information exchange between the individual ants, such as the shortest path to the food source, as shown in Fig. 1.

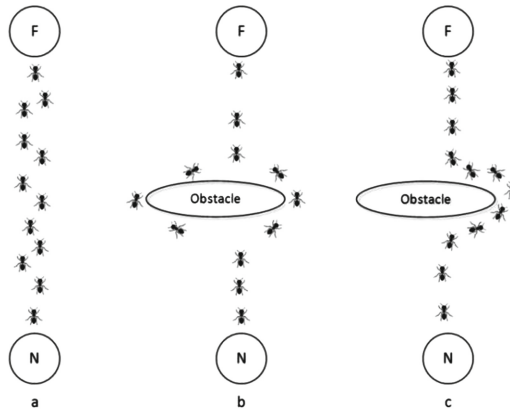


Fig. 1. Ant follows the pheromone to find the shorter path.

In ACO, the artificial ants communicate with each other in a way similar to the biological ants. While exploring the network, the artificial ants mark the nodes they have passed with an artificial pheromone. For choosing the next hop, they are usually attracted by the node with the highest pheromone value. In general, there are two basic artificial ants: Forward ANTs (FANTs) and Backward ANTs (BANTs). FANTs discover a path to a randomly chosen destination node. Once FANTs reach the destination, BANTs are sent back to the source node following the reverse path. BANTs update the local models of the network status at each intermediate node. It has been shown that ACO based routing can be successfully applied to both wired and wireless networks, also to VANETs [6–9]. ACO routing procedures can take advantages of the vehicle’s position information to adapt the rapidly dynamic networks.

1.3 Fuzzy Logic

Fuzzy logic has been widely utilized in many areas of our daily life, such as automatic control, automobile production, academic education, industrial manufacturing and so on. It represents and manipulates the linguistic information in a natural way via membership functions and fuzzy rules. In general, fuzzy logic system consists of three sub-components: fuzzifier, Fuzzy Interface System (FIS) and defuzzifier, as shown in Fig. 2. Fuzzifier calculates all input values into fuzzy membership functions. FIS executes all applicable rules in the rule-base on the output of fuzzifier to compute the fuzzy output functions. And then

defuzzification refers to the way a crisp value is extracted from a fuzzy set as a representation value. Fuzzy theory applies well in control decision procedures of VANETs, due to the uncertainty of nodal mobility, unstable links, and limited resources. It can either improve the performance or to handle the problem that conventional theory cannot approach successfully because latter relies on a valid and accurate model.

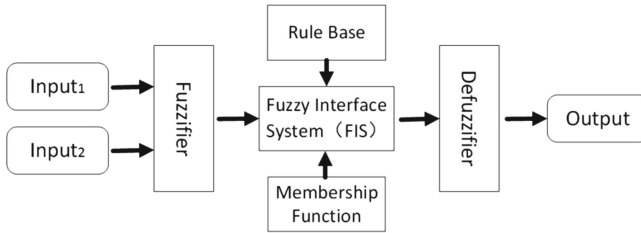


Fig. 2. Fuzzy logic system.

2 Related Work

MAR-DYMO [6] proposed by Correia et al. is the first ant based algorithm that adapted the Dynamic MANETs On-demand (DYMO) routing protocol to VANETs. The idea of this paper is to use vehicles' position and speed to update the pheromone and help making routing decision that apply well in VANETs. The authors modify the reactive DYMO protocol by adding the pheromone level, evaporation rate and the predicted lifetime for each route to the routing table. It also works in multi-path model, the actual route is chosen based on their pheromone levels. Moreover, it guarantees both link quality and link stability. MAZACORNET [7] proposed by Rana et al. is the first ACO based routing algorithm that uses a concept of zones. It subdivides the networks into zones to achieve scalability and uses proactive approach to find routes within the zones and a reactive approach between zones. Balaji et al. introduce a hierarchical approach which combines a clustering architecture with ACO routing procedures in VANETs [8]. ACO based routing protocols in VANETs are still a hot issue in recent years. They are not only adapted in V2V communications, but also associated with devices like Road Side Units (RSUs). S-AMCQ [9] proposed by Eiza et al. applies ACO algorithm to calculate the feasible routes which satisfy multiple QoS constraints determined by data traffic types in VANETs' communications. It is designed for V2I communications, and the authentication process applied in the protocol relies on a Certification Authority (CA), which the local transportation authority or vehicle manufacturer can act as.

3 Attack Model

As aforementioned section, many ACO based routing protocols have been proposed in VANETs and shown good results. However, few of them have considered the security in the routing process. Now we want to introduce a scenario in which the attacker can disrupt or even destroy the normal communications in VANETs.

We suppose that a vehicle V_1 is now in a suburb district S_1 in which there are few RSUs. The driver of vehicle V_1 wants to find out where is the nearest restaurant which provides his favorite food. In order to get this information, V_1 has to query either a RSU or another vehicle which has cached this information in his nearby area. In this area S_1 , two roadside units R_1 and R_2 provide directly the desirous information and vehicle V_2 has cached this information from RSU R_2 before. As shown in Fig. 3, all of the three objects are out of V_1 's transmission range. The two roadside units R_1 and R_2 are both far away from V_1 . Vehicle V_2 is closer to V_1 . In this case, V_1 will query V_2 for the information. In a normal case, V_1 sends out the query packet via vehicle V_3 or V_5 to V_2 based on a particular routing algorithm which does not consider any security issues. It can possibly receive multiple paths, such as V_1 - V_3 - V_4 - V_2 , V_1 - V_5 - V_6 - V_2 , V_1 - V_7 - V_8 - V_9 - V_2 and so on. However, if vehicle V_4 is an attacker or selfish one, it can launch many different kinds of attacks. For example, it can impersonate V_2 to send back some bogus information, e.g. guide V_1 to visit another restaurant which is far way (attack type 1), or directly modify the information packet received from V_2 to disrupt the ongoing communication (attack type 2), or just evaporate the data packet to collect privacy related sensitive data from both V_1 and V_2 (attack type 3), or simply drop the query packet to save energy (attack type 4).

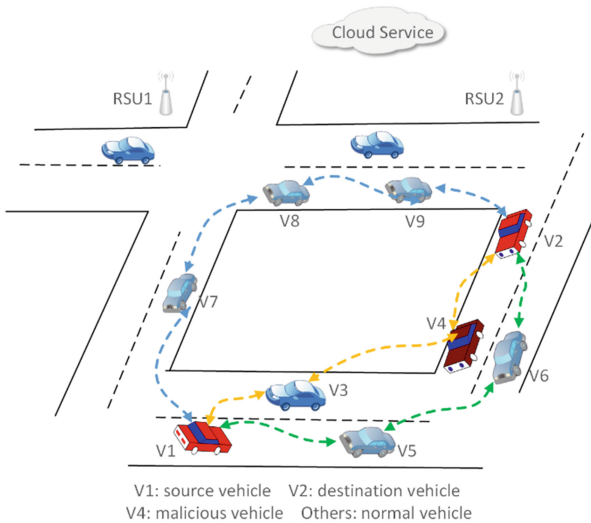


Fig. 3. An attack model in VANETs.

As introduced above, we consider four different kinds of basic attack models in VANETs. The first one is known as masquerading attack in which the adversary actively pretends to be a candidate vehicle to either reply or relay the messages. The second type of attack harms the data consistency. The third one can disclosure the privacy of vehicles participant in the communication and the selfish vehicles can strongly affect the functionality of the whole network.

4 Proposed Idea

Considering a secure and reliable transmission of queried data packets, we propose a framework to design a security aware ACO based routing protocol for VANETs, which aims not only to find out the optimal reliable route to transmit the queried data, but also protect the privacy of the vehicles.

The proposed secure communication system in VANETs focus on protecting the network against the above mentioned four basic attacks. In order to prevent attack type 1 and 2, digital signatures based authentication mechanism is designed in our proposal. Anonymous key pairs can preserve the privacy of vehicles. Finally, fuzzy logic algorithm is applied to detect and isolate the selfish or malicious vehicles in the routing process. We describe our solution in detail in the following subsections.

4.1 Authentication

As discussed in [10], digital signatures is recommended as a convenient and reliable solution for message authentication, hence we choose the Elliptic Curve Cryptography (ECC) for applying the digital signatures in our proposal. Since VANETs are infrastructural networks, the public key certificates are issued by a trusted authority, e.g. a RSU. Before sender vehicle sends out the query message, it signs the message by using its private key and it also adds his public key certificate which is issued by a Certificate Authority (CA) into the message. Receiver vehicle can use the CA's public key to extract the sender vehicle's public key from the certificate and then further verify whether the extracted key matches with the signature in this message. In this way, the authentication of message is complete. The overhead caused by the verification can be reduced by selecting the relevant messages and verifying only the signatures with known public keys.

4.2 Data Consistency

Meanwhile, data consistency can be also guaranteed by the digital signatures. Attackers can modify the message contents, but can't create a corresponding signature for the modified message, because he doesn't know the private key of message sender. Once a vehicle receives this message, it can easily verify the signature and recognize the modification.

4.3 Privacy

Anonymous public/private key pairs are applied for protect the privacy of vehicles. These key pairs allow vehicles to digitally sign messages and hence authenticate themselves to others. Anonymous key pairs are public/private key pairs which are authenticated by the CA in the network, but they don't contain information about the actual vehicle's identity (i.e. its electronic license plate). Vehicles can possess a set of anonymous public keys to prevent tracking.

4.4 Detection of Malicious Vehicles

Since vehicles have sufficient power and hardware supports (i.e. On Board Unit (OBU) or processors equipped in the vehicle for operating the vehicle's functionalities) for the computation, fuzzy logic algorithm adapts well in vehicular networks. Sethi et al. have proposed a fuzzy embedded ACO based routing protocol in MANETs [11]. However, the connection between the output value from fuzzy logic system and the pheromone in ACO routing procedure is not described clearly. We inspired from this idea and propose a Security Aware Fuzzy embedded ACO based routing protocol (SAFACO) for VANETs.

In SAFACO, once a vehicle V_1 wants to send out query message to V_2 , V_1 first sends out forward ants to discover the reliable routes between the sender and receiver. As shown in Fig. 3, V_1 sends out ants to its neighboring vehicles V_3 , V_5 and V_7 . Ants follow the highest pheromone value on each neighbor vehicle for the next hop as introduced in Sect. 1.2. After forward ants have reached V_2 , backward ants are generated and sent back to V_1 following exactly the path made by forward ants, but in the opposite direction. In a scenario without attackers, V_1 receives three reliable routes as demonstrated in Sect. 3: V_1 - V_3 - V_4 - V_2 , V_1 - V_5 - V_6 - V_2 and V_1 - V_7 - V_8 - V_9 - V_2 .

As assumed before, if vehicle 4 is the attacker, then it drops all the received forward ants sent by V_1 . However, the selfish behavior can be monitored and affects the pheromone values which are assigned by other vehicles to V_4 . During the route discovery phase, ants gathering the information of each intermediate node. We consider all the observations into the fuzzy system to output a trust value, as shown in Fig. 2. This trust value represents to what extent is the vehicle trustable. In order to efficiently use pheromone in the ACO routing structure, we observe not only the packet drop rate, but also many other metrics, such as, the authentication fail rate. Once a message fails to be authenticated, then the pheromone value of the message sender should be reduced. Other Quality of Service (QoS) metrics like the link stability, the packet transmission delay, etc., can be also applied. We consider all the mentioned metrics as the input values for our fuzzy system. Finally, the output trust value will be integrated into the pheromone value by applying the following formula:

$$Ph_{(new)} = [a*Ph_{(old)} + b*T_{(Fuzzy)}]/2 \quad (1)$$

Where $Ph_{(new)}$ is the new pheromone value; $Ph_{(old)}$ is the old pheromone value and $0 \leq Ph_{(new)}, Ph_{(old)} \leq 1$; $T_{(Fuzzy)}$ is the trust value output from the fuzzy system and $0 \leq T_{(Fuzzy)} \leq 1$; a, b are weights and $a + b = 1$.

Once all the pheromone values assigned to V_4 is reduced to under the predefined threshold value, then V_4 will be isolate from the network. In this way, after the detection of above attacks, normal vehicles will isolate the malicious nodes voluntarily, and select another route with relative high pheromone value.

5 Conclusion and Future Work

In this paper we have introduced a novel design for secure communication system architecture in VANETs. We have proposed a model which combines the digital signature authentication mechanism with a fuzzy logic embedded ACO based routing protocol in VANETs to detect and isolate malicious or anomalous vehicles. We have explained how the proposed scheme can protect the network from four types of attacks we introduced. The structure of the proposed fuzzy logic embedded ACO based routing protocol is also introduced. As future work, we are going to implement the proposed system in a conventional network simulator, such as NS-3 or QualNet, and test its performance in means of detection rate, overhead and data transmission delay.

References

1. Liu, J., Wan, J., Wang, Q., Deng, P., Zhou, K., Qiao, Y.: A survey on position-based routing for vehicular ad hoc networks. *Telecommun. Syst.* **62**, 15–30 (2016)
2. Li, F., Wang, Y.: Routing in vehicular ad hoc networks: a survey. *IEEE Veh. Technol. Mag.* **2**(2), 12–22 (2007)
3. Härrri, J., Bonnet, C., Filali, F.: Kinetic mobility management applied to vehicular ad hoc network protocols. *Comput. Commun.* **31**(12), 2907–2924 (2008)
4. Cunha, F., Villas, L., Boukerche, A., Maia, G., Viana, A., Mini, R.A.F., Loureiro, A.A.F.: Data communication in VANETs. *Ad Hoc Netw.* **44**(C), 90–103 (2016). <http://dx.doi.org/10.1016/j.adhoc.2016.02.017>
5. Hansen, J.H., Boyraz, P., Takeda, K., Abut, H. (eds.): *Digital Signal Processing for In-Vehicle Systems and Safety*. Springer, Heidelberg (2012)
6. Correia, S.L.O.B., Celestino, J., Cherkaoui, O.: Mobility-aware ant colony optimization routing for vehicular ad hoc networks, pp. 1125–1130 (2011)
7. Rana, H., Thulasiraman, P., Thulasiram, R.K.: MAZACORNET: mobility aware zone based ant colony optimization routing for VANET. In: *Evolutionary Computation*, pp. 2948–2955 (2013)
8. Balaji, S., Sureshkumar, S., Saravanan, G.: Cluster based ant colony optimization routing for vehicular ad hoc networks. *Int. J. Sci. Eng. Res.* **4**(6), 26–30 (2013)
9. Eiza, M.H., Owens, T., Ni, Q.: Secure and robust multi-constrained QoS aware routing algorithm for VANETs. *IEEE Trans. Dependable Secure Comput.* **13**(1), 32–45 (2016)
10. Raya, M., Hubaux, J.-P.: Securing vehicular ad hoc networks. *J. Comput. Secur.* **15**(1), 39–68 (2007)
11. Sethi, S., Udgata, S.K.: Fuzzy-based trusted ant routing (FTAR) protocol in mobile ad hoc networks. In: Sombatheera, C., Agarwal, A., Udgata, S.K., Lavangnananda, K. (eds.) *MIWAI 2011*. LNCS, vol. 7080. Springer, Heidelberg (2011)

Cryptanalysis of “An Efficient Searchable Encryption Against Keyword Guessing Attacks for Shareable Electronic Medical Records in Cloud-Based System”

Chun-Ta Li¹, Cheng-Chi Lee^{2,3}(✉), Chi-Yao Weng⁴, Tsu-Yang Wu⁵,
and Chien-Ming Chen⁶(✉)

¹ Department of Information Management, Tainan University of Technology,
No. 529, Zhongzheng Road, Tainan City 71002, Taiwan, R.O.C.
th0040@mail.tut.edu.tw

² Department of Library and Information Science, Fu Jen Catholic University,
No. 510, Jhongjheng Road, New Taipei City 24205, Taiwan, R.O.C.
cclee@mail.fju.edu.tw

³ Department of Photonics and Communication Engineering, Asia University,
No. 500, Lioufeng Road, Taichung City 41354, Taiwan, R.O.C.

⁴ Department of Computer Science, National Pingtung University, No. 4-18,
Min-Sheng Road, Pingtung City 90003, Taiwan, R.O.C.
cyweng@mail.nptu.edu.tw

⁵ Fujian Provincial Key Laboratory of Big Data Mining and Applications,
Fujian University of Technology, No. 3, Xueyuan Road, Fuzhou City 350118, China
wutsuyang@gmail.com

⁶ Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen University Town, Nanshan District, Shenzhen 518055, China
chienming.taiwan@gmail.com

Abstract. Recently, Wu et al. proposed a secure channel free searchable encryption (SCF-PEKS) scheme which not only can guard against keyword guessing and record disclosure attacks but also can provide much better performance than other related scheme for shareable EMRs. However, in this paper, we demonstrated that Wu et al.’s SCF-PEKS scheme has some design flaws and security weaknesses such as (1) it fails to ensure the properties of message authentication and untraceability, (2) it fails to prevent the malicious outsider from forging a fake EMR as the sender, (3) it fails to prevent the privileged cloud insider from revealing sender’s secret keyword and sensitive record. The aforementioned security flaws in Wu et al.’s scheme may lead to privacy exposure and the receiver misled the contents of this fake record.

Keywords: Cryptanalysis · Cloud storage service · Electronic medical record · Keyword guessing attack · Keyword search

1 Introduction

In recent years, cloud-based E-health care system has gained increasing attention because it allows patient to access variety online medical services, such as telemedicine and utilization of electronic medical records (EMRs). The architecture of cloud-based E-health care system is shown in Fig. 1. In order to preserve the privacy of EMR, the sender encrypts the EMR and generates a secure index involving some keywords before uploading it to the cloud server. Then the authorized receivers can search on the sender’s encrypted EMR by sending a trapdoor associated with a certain query keyword. If the query keyword is involved in sender’s index, the receivers are permitted to download the sender’s encrypted EMR from the cloud server. Recently, many researchers proposed the public key encryption keyword search method without secure channels [1–6], referred to as SCF-PEKS. In 2016, Wu et al. [6] proposed an efficient and secure SCF-PEKS scheme based on pairings and they claimed their proposed scheme is secure against keyword guessing attacks (IND-KGA). Unfortunately, in this paper, our analysis show that the scheme in [6] is susceptible to outsider attacks, privileged-insider attacks, and not achieves IND-KGA.

The remainder of the paper is encryption based searchable encryption organized as follows. In Sect. 2, we briefly review Wu et al.’s SCF-PEKS scheme. We further demonstrate two kinds of attacks on Wu et al.’s scheme in Sects. 3 and 4, respectively. Finally, we conclude this paper in Sect. 5.

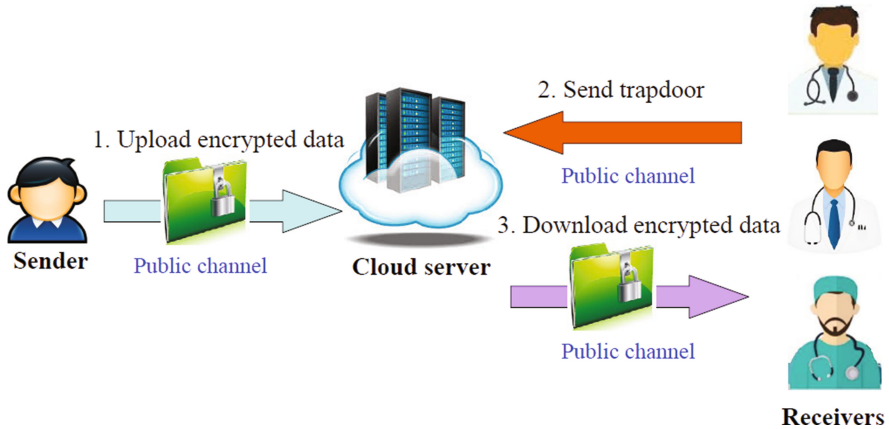


Fig. 1. The architecture of cloud-based EMRs system

2 Review of Wu et al.’s SCF-PEKS Scheme

In this section, we review Wu et al.’s secure channel free public key encryption with keyword search (SCF-PEKS) scheme [6]. Three entities involved in this scheme: the sender, the receivers and the cloud server. In addition, four

stages involve in their scheme: *Initialization*, *Data Processing*, *Search* and *Record Retrieval*. Wu et al. define eight algorithms used in their scheme and the detail of each algorithm is described as follows:

- **GlobalSetup(λ):**

The algorithm takes the security parameter λ as input and outputs the global parameter gp .

- **KeyGen(gp):**

The algorithm takes the global parameter gp as input and outputs a public/secret key pair (pk, sk) .

- **Enc(gp, M, sk_S):**

The algorithm takes the global parameter gp , an electronic medical record M and a sender's secret key sk_S as input and outputs the encrypted ciphertext \mathcal{C} .

- **IndexGen(gp, sk_S, \mathcal{W}):**

The algorithm takes the global parameter gp , a sender's secret key sk_S and a keyword set \mathcal{W} as input and outputs the secure index \mathcal{I} .

- **ReKeyGen(gp, sk_S, pk_R):**

The algorithm takes the global parameter gp , a sender's secret key sk_S and a receiver's public key pk_R as input and outputs a re-encryption key rk .

- **Trapdoor(gp, sk_R, w'):**

The algorithm takes the global parameter gp , a receiver's secret key sk_R and a query keyword w' as input and outputs the trapdoor $T_{w'}$.

- **Search($gp, \mathcal{I}, T_{w'}, rk$):**

The algorithm takes the global parameter gp , the index \mathcal{I} , the trapdoor $T_{w'}$ and a re-encryption key rk as input and outputs 1 if $w = w'$, otherwise outputs 0. Note that keyword w is involved in \mathcal{I} and query keyword w' is involved in $T_{w'}$.

- **Dec($gp, \mathcal{C}, sk_R, rk$):**

The algorithm takes the global parameter gp , a ciphertext \mathcal{C} , a receiver's secret key sk_R and a re-encryption key rk as the input and outputs the electronic medical record M if each input parameter is correct.

2.1 Initialization

In this stage, the cloud server takes the security parameter λ and runs $\text{GlobalSetup}(\lambda)$ to generate two cyclic groups \mathbb{G}_1 and \mathbb{G}_2 with the same prime order p , having g as a generator of \mathbb{G}_1 . In addition, the cloud server initializes a bilinear map $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$, chooses a hash function $H : \{0, 1\}^* \rightarrow \mathbb{Z}_p$, selects a random value sk_C from \mathbb{Z}_p as the secret key and computes the corresponding public key $pk_C = g^{sk_C}$. Finally, the global parameter can be denoted as $gp = \{\mathbb{G}_1, \mathbb{G}_2, g, p, \hat{e}, H, pk_C\}$.

After that, each sender selects a secret value sk_S from \mathbb{Z}_p as the secret key and runs $\text{KeyGen}(gp)$ to generate a public/secret key pair (pk_S, sk_S) , where $pk_S = g^{1/sk_S}$. Similarly, each receiver $R_i \in \mathcal{R}$ generates his/her secret key $sk_{R_i} \in \mathbb{Z}_p$ and the corresponding public key $pk_{R_i} = g^{1/sk_{R_i}}$, where \mathcal{R} is the receiver set.

2.2 Data Processing

In this stage, the sender first selects a keyword set \mathcal{W} from the electronic medical record M and runs $\text{IndexGen}(gp, sk_S, \mathcal{W})$ to generate a secure index \mathcal{I} . More precisely, for each keyword $w \in \mathcal{W}$, the sender computes $\tau_w = pk_C^{sk_S \cdot H(w)}$. After that, the index can be denoted as $\mathcal{I} = \{\tau_w\}_{w \in \mathcal{W}}$. Besides that, in order to protect the electronic medical record $M \in \mathbb{G}_2$, the sender runs $\text{Enc}(gp, M, sk_S)$ to generate the encrypted record \mathcal{C} . Thus, the sender chooses a random value k from \mathbb{Z}_p and computes $C_1 = M \oplus \hat{e}(g^{sk_S}, g^k), C_2 = g^k$. After that, the encrypted record can be denoted as $\mathcal{C} = \{C_1, C_2\}$. Moreover, the sender generates a re-encryption key $rk_{S \rightarrow R_i}$ for each receiver $R_i \in \mathcal{R}$ by running the algorithm $\text{ReKeyGen}(gp, sk_S, pk_{R_i})$ and computes $rk_{S \rightarrow R_i} = pk_{R_i}^{sk_S}$. Thus, the re-encryption key set can be denoted as $\mathcal{RK} = \{rk_{S \rightarrow R_i}\}_{R_i \in \mathcal{R}}$. Finally, the sender uploads the secure index \mathcal{I} , the encrypted record \mathcal{C} and the re-encryption key set \mathcal{RK} to the cloud server.

2.3 Search

In this stage, the receiver R_i needs to generate the trapdoor for a query keyword w' by running $\text{Trapdoor}(gp, sk_{R_i}, w')$. More precisely, the receiver chooses a random value $r \in \mathbb{Z}_p$ and computes $T_1 = pk_C^r, T_2 = pk_C^{H(w') \cdot r \cdot sk_{R_i}}$. After that, the trapdoor can be denoted as $T_{w'} = \{T_1, T_2\}$ and the receiver sends the trapdoor $T_{w'}$ to the cloud server.

After receiving $T_{w'}$ from R_i , the cloud server runs $\text{Search}(gp, \mathcal{I}, T_{w'}, rk_{S \rightarrow R_i})$ to check whether the encrypted record \mathcal{C} involves the keyword w' . More precisely, for each τ_w in \mathcal{I} , the cloud server checks if

$$\begin{aligned}
 \hat{e}(\tau_w, T_1) &= \hat{e}(pk_C^{sk_S \cdot H(w)}, pk_C^r) \\
 &= \hat{e}(g^{H(w) \cdot r \cdot sk_C}, g^{sk_S \cdot sk_C}) \\
 &= \hat{e}(g^{H(w) \cdot r \cdot sk_C}, g^{\frac{sk_S}{sk_{R_i}} \cdot sk_{R_i} \cdot sk_C}) \\
 &= \hat{e}(g^{H(w) \cdot r \cdot sk_C \cdot sk_{R_i}}, g^{\frac{sk_S}{sk_{R_i}} \cdot sk_C}) \\
 &= \hat{e}(pk_C^{H(w) \cdot r \cdot sk_{R_i}}, pk_{R_i}^{sk_S \cdot sk_C}) \\
 &= \hat{e}(T_2, rk_{S \rightarrow R_i}^{sk_C}).
 \end{aligned} \tag{1}$$

The Search algorithm outputs 1 only if Eq. (1) holds, which implies $w = w'$ and the cloud server sends $\{\mathcal{C}, rk_{S \rightarrow R_i}\}$ to the receiver R_i . Otherwise, the cloud server outputs 0 and sends \perp to R_i .

2.4 Record Retrieval

After receiving $\{\mathcal{C}, rk_{S \rightarrow R_i}\}$ from the cloud server, the receiver runs $\text{Dec}(gp, \mathcal{C}, sk_R, rk)$ to retrieve the record M and the record can be retrieved as follows:

$$\begin{aligned}
M &= C_1 \oplus \hat{e}(rk_{S \rightarrow R_i}, C_2)^{sk_{R_i}} \\
&= M \oplus \hat{e}(g^{sk_S}, g^k) \oplus \hat{e}(g^{sk_S/sk_{R_i}}, g^k)^{sk_{R_i}} \\
&= M \oplus \hat{e}(g^{sk_S}, g^k) \oplus \hat{e}(g^{sk_S}, g^k) \\
&= M
\end{aligned} \tag{2}$$

3 Outsider Attacks on Wu et al.'s SCF-PEKS Scheme

In this section, we illustrate that Wu et al.'s scheme fails to provide the property of untraceability. Moreover, their scheme is vulnerable to outsider attack and record forgery attack. Firstly, we demonstrate that the malicious outsider \mathcal{A} is able to control the public communication channels to utilize the message eavesdropped and message exchanged between all entities in the EMR system. More precisely, \mathcal{A} may replace or insert a self created message between two communicating entities. Moreover, the public keys of both the sender and the receiver are known to \mathcal{A} .

3.1 Fails to Provide Message Authentication

In the data processing stage, after receiving $\{\mathcal{I}, \mathcal{C}, \mathcal{RK}\}$ from the sender, the cloud server does not check the validation of received messages before storing them. Assuming an outsider \mathcal{A} who alters $\{\mathcal{I}, \mathcal{C}, \mathcal{RK}\}$ to $\{\mathcal{I}^{\mathcal{A}}, \mathcal{C}^{\mathcal{A}}, \mathcal{RK}^{\mathcal{A}}\}$ and sends $\{\mathcal{I}^{\mathcal{A}}, \mathcal{C}^{\mathcal{A}}, \mathcal{RK}^{\mathcal{A}}\}$ to the cloud server, the cloud server is unable to recognize the alteration. The same weakness also exists in the end of search stage. Therefore, the message authentication is not provided in Wu et al.'s scheme.

3.2 Fails to Provide Untraceability

In the data processing stage of Wu et al.'s scheme, the sender's re-encryption key $rk_{S \rightarrow R_i}$ is uploaded to the cloud server via a public channel. Similarly, in the record retrieval stage of Wu et al.'s scheme, the re-encryption key $rk_{S \rightarrow R_i}$ is downloaded to the receiver via a public channel. Thus, the untraceability will not be ensured if there is a parameter transmitted between the sender and the receiver containing $rk_{S \rightarrow R_i}$. Thus, Wu et al.'s scheme is failed to ensure untraceability and the outsider \mathcal{A} can discover the relation of a connection between the sender and the receiver as long as the upload and download requests transmitted over the public channels contains $rk_{S \rightarrow R_i}$.

3.3 Record Forgery Attacks

Continued to the Sects. 3.1 and 3.2, once the outsider \mathcal{A} has learned the relation between the sender and the receiver, \mathcal{A} may intentionally forge a fake record M' and replace sender's real record M with outsider's M' during the record retrieval

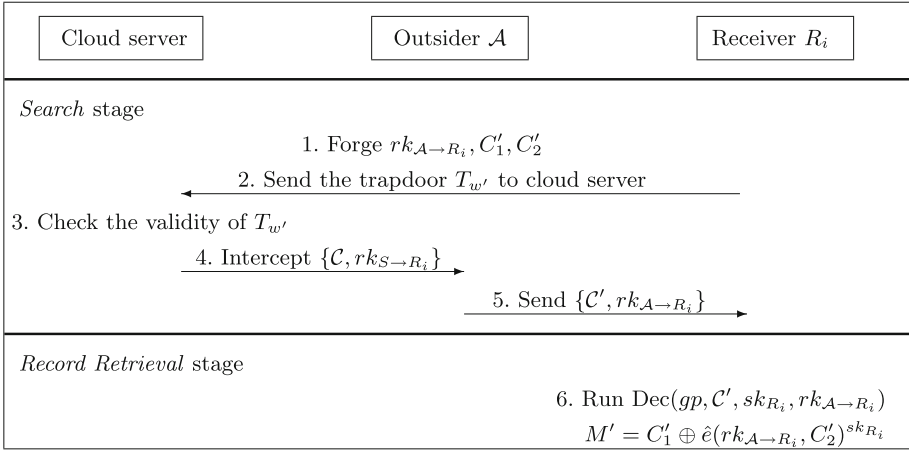


Fig. 2. Record forgery attack on Wu et al.’s scheme

stage. For clarity, the details of this weakness are depicted in Fig. 2. The detailed steps of this attack are described as follows:

1. First, the outsider \mathcal{A} forges a fake electronic medical record M' . Then \mathcal{A} chooses two random values r' and k' from \mathbb{Z}_p and computes $rk_{\mathcal{A} \rightarrow R_i} = pk_{R_i}^{r'} = g^{r'/sk_{R_i}}$ and $C'_1 = M' \oplus \hat{e}(g^{r'}, g^{k'}), C'_2 = g^{k'}$. Afterwards, the encrypted M' can be denoted as $\mathcal{C}' = \{C'_1, C'_2\}$.
2. In the search stage, we assume the receiver R_i sends the trapdoor $T_{w'}$ to the cloud server and the outsider \mathcal{A} notices there is a download request between the cloud server and the receiver R_i . If R_i 's $T_{w'}$ is valid, the cloud server will send $\{\mathcal{C}, rk_{S \rightarrow R_i}\}$ to the receiver R_i via the public channel. In this moment, \mathcal{A} intercepts $\{\mathcal{C}, rk_{S \rightarrow R_i}\}$ and sends $\{\mathcal{C}', rk_{\mathcal{A} \rightarrow R_i}\}$ back to R_i . Note that parameters \mathcal{C} and $rk_{S \rightarrow R_i}$ are replaced with \mathcal{C}' and $rk_{\mathcal{A} \rightarrow R_i}$, respectively.
3. When receiving the message $\{\mathcal{C}', rk_{\mathcal{A} \rightarrow R_i}\}$, in the record retrieval stage, R_i will decrypt the ciphertext \mathcal{C}' to retrieve the record M' by invoking $\text{Dec}(gp, \mathcal{C}', sk_{R_i}, rk_{\mathcal{A} \rightarrow R_i})$. Finally, the fake record M' will be retrieved as follows:

$$\begin{aligned}
 M' &= C'_1 \oplus \hat{e}(rk_{\mathcal{A} \rightarrow R_i}, C'_2)^{sk_{R_i}} \\
 &= M' \oplus \hat{e}(g^{r'}, g^{k'}) \oplus \hat{e}(g^{r'/sk_{R_i}}, g^{k'})^{sk_{R_i}} \\
 &= M \oplus \hat{e}(g^{r'}, g^{k'}) \oplus \hat{e}(g^{r'}, g^{k'}) \\
 &= M'
 \end{aligned} \tag{3}$$

Therefore, the record forgery attack is successful launched on Wu et al.’s SCF-PEKS scheme.

4 Privileged-Insider Attacks on Wu et al.’s SCF-PEKS Scheme

In Wu et al.’s SCF-PEKS scheme, they claimed that their scheme can guard against keyword guessing attack and only legitimate receiver can correctly decrypt the record M . However, we found that Wu et al.’s scheme cannot prevent the inside keyword guessing attack from a malicious privileged-insider due to the cloud server has plentiful information to run $\text{Enc}(gp, M, sk_S)$ algorithm. Moreover, the privilege-insider can perform guessing attack to disclosure M without knowing sk_{R_i} . In the following, we demonstrate the details of inside keyword guessing and record disclosure attacks on Wu et al.’s scheme and the detailed steps of our proposed attack are depicted in Fig. 3.

1. When receiving the uploaded messages $\{\mathcal{I}, \mathcal{C}, \mathcal{RK}\}$ from the sender, the privileged cloud server first selects a guessed keyword $w^* \in \mathcal{W}$. Then the cloud server computes $H(w^*)$ and its multiplicative inverse element $H(w^*)^{-1}$ such that $H(w^*) \cdot H(w^*)^{-1} \equiv 1 \pmod{\phi(p)}$, where $\phi()$ is Euler phi function.
2. In this step, the privileged cloud server computes $g^{sk_S^*} = \tau_W^{sk_C^{-1} \cdot H(w^*)^{-1}}$ and $\hat{e}(g^{sk_S^*}, g^k)$, where τ_W and g^k are collected from sender’s uploaded messages.
3. Afterwards, the privileged cloud server computes $M^* = C_1 \oplus \hat{e}(g^{sk_S^*}, g^k)$. If M^* is a plaintext, it indicates the correct guessing of sender’s keyword and the privileged-insider succeeds to guess the keyword w and disclose the sender’s record M . Otherwise, it indicates the incorrect guessing of sender’s keyword and the privileged-insider repeats the above steps until the sender’s keyword is successfully guessed.

Finally, the privileged cloud insider succeeds to derive the keyword $w = w^*$ and disclosures the record $M = M^*$ and Wu et al.’s scheme is vulnerable to inside keyword guessing and record disclosure attacks.

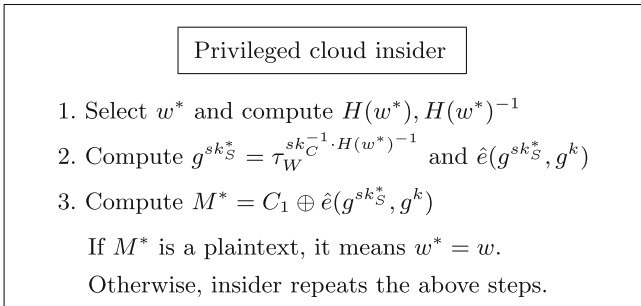


Fig. 3. The keyword guessing and record disclosure attacks are performed by the privileged cloud insider

5 Conclusions

In this paper, we have first reviewed Wu et al.'s SCF-PEKS scheme. Then some weaknesses on design aspects and security of Wu et al.'s scheme for shareable EMRs in cloud-based E-health care system have been pointed out. Wu et al.'s scheme is failure to provide message authentication and untraceability between the sender and the receiver. Moreover, their scheme is vulnerable to record forgery, inside keyword guessing and record disclosure attacks and is not easily repairable. In our future works, we would try to propose an improved version of their SCF-PEKS scheme and the aforementioned security aspects should be carefully considered for cloud-based E-health care system.

References

1. Baek, J., Safavi-Naini, R., Susilo, W.: Public key encryption with keyword search revisited. In: Gervasi, O., Murgante, B., Laganà, A., Taniar, D., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2008. LNCS, vol. 5072, pp. 1249–1259. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-69839-5_96](https://doi.org/10.1007/978-3-540-69839-5_96)
2. Fang, L.M., Susilo, W., Ge, C.P., Wang, J.D.: Public key encryption with keyword search secure against keyword guessing attacks without random oracle. *J. Syst. Softw.* **238**, 221–241 (2013)
3. Guo, L., Yau, W.C.: Efficient secure-channel free public key encryption with keyword search for EMRs in cloud storage. *J. Med. Syst.* **39**(2), 1–11 (2015). doi:[10.1007/s10916-014-0178-y](https://doi.org/10.1007/s10916-014-0178-y). Article no. 11
4. Li, C.T., Lee, C.W., Shen, J.J.: An extended chaotic maps based keyword search scheme over encrypted data resist outside and inside keyword guessing attacks in cloud storage services. *Nonlinear Dyn.* **80**(3), 1601–1611 (2015)
5. Rhee, H.S., Park, J.H., Susilo, W., Dong, H.L.: Trapdoor security in a searchable public-key encryption scheme with a designated tester. *J. Syst. Softw.* **83**(5), 763–771 (2010)
6. Wu, Y., Lu, X., Su, J., Chen, P.: An efficient searchable encryption against keyword guessing attacks for searchable electronic medical records in cloud-based system. *J. Med. Syst.* **40**(12), 1–9 (2016). doi:[10.1007/s10916-016-0609-z](https://doi.org/10.1007/s10916-016-0609-z). Article no. 258

eDSDroid: A Hybrid Approach for Information Leak Detection in Android

Hoang Tuan Ly¹, Tan Cam Nguyen^{2(✉)}, and Van-Hau Pham¹

¹ Information Security Lab, University of Information Technology,
Vietnam National University, Ho Chi Minh City, Vietnam
{tuanly, haupv}@uit.edu.vn

² Faculty of Science and Technology, Hoa Sen University, Ho Chi Minh City, Vietnam
camnt@grad.uit.edu.vn

Abstract. Leaking personal information on mobile devices is a serious problem. Work on information leak detection for mobile devices, until now, mostly focus on action within a single application, while the coordinated action of several applications for the malicious purpose is becoming popular. This study proposes a hybrid approach that combines static and dynamic analysis to detect information leak as a result of the coordinated action of multiple applications. In this text, we call it inter-application malware. The analysis takes place in two stages. In the first stage, we use static analysis to determine the chains of sensitive actions on multiple applications. The chain of sensitive actions is the sequential user's actions that may lead to information leakage. In the second stage, we validate whether the chain of sensitive actions indeed leaks user's data by using the dynamic analysis. In fact, the applications in question are forced to execute after the chains of sensitive actions detected in the first stage. We monitor the sensitive actions to determine which actions make information leak. In order to do so, we modify the Android Emulator to trigger and monitor any action of any applications running on it. We have evaluated our tool, namely eDSDroid, on the famous Toyapps test case. The test result shows the correctness and effectiveness of our tool.

1 Introduction

Nowadays, Android is the most popular mobile operating system. It is integrated on almost kinds of mobile devices, e.g. mobile phone, tablet, smartwatch, etc. Unfortunately, the popularity of Android has made them become the target of malware. According to F-Secure statistic, the number of new malware on Android accounted for 99% of the mobile operating system [11]. That makes the Android security protection has become an urgent problem. Techniques for malware analysis can be classified into the following categories:

- Static analysis [2, 7, 8, 10, 15] is an analysis approach that focuses to examine the application without execution.
- Dynamic analysis [9, 12, 17, 18] is an analysis approach in which analyzed application is executed in a sandbox. The captured action is then analyzed to detect malicious behavior.

- Hybrid of static and dynamic analysis [4, 13, 16] is an analysis approach that combines the static and dynamic analysis. Static analysis is often used to direct dynamic analysis. It helps to limit the scope of dynamic analysis.

Each technique has its pros and cons. Static analysis helps to know the detail about the behaviors of application, but it may not be able to analyze correctly if the application is obfuscated. Dynamic analysis can identify the abnormal behaviors exactly but it can not capture all the potential behaviors. Taking advantages from both sides makes hybrid approach promising. Work on information leak detection for mobile devices, until now, mostly focus on action within a single application, while the coordinated action of several applications for the malicious purpose is becoming popular. In fact, there are two types of the coordinated action of malware:

- Confused deputy attack: the normal application is misused by malware.
- Collusion attack: the malicious applications coordinate their actions with each other. For example, they can combine their permissions to perform the actions which cannot be made independently.

To tackle the inter-application information leak, this study proposes a new hybrid of static and dynamic analysis for information leak detection of inter-application malware on Android platform. Our approach is deeply inspired from SmartDroid [4] which uses static analysis to direct the dynamic one. The difference is that our approach aims to detect information leak by inter-application malware whereas SmartDroid works on a single application. This study makes the following contributions:

- We propose a new hybrid approach which uses static analysis to direct dynamic analysis to analyze inter-application malware. Static analysis determines the potential information leakage on multiple applications. Dynamic analysis is used as a double check to eliminate the false positive of the static stage. Static analysis also helps to limit the scope of dynamic analysis.
- We have implemented and evaluated our tool – namely eDSDroid on the famous ToyApps dataset. The test result shows the correctness and effectiveness our tool in detecting inter-application malware.

2 Related Work

Epicc [5] uses static analysis on inter-component communication (ICC). It analyzes the properties of Android Intents to monitor ICC. However, Epicc does not handle URI which used for Android component communication.

FlowDroid [6, 14] uses static taint flow analysis to monitor sensitive data-flow in each Android component. The sensitive data-flow is the flow of data that can make leakage. The sensitive data-flow passes from Source where sensitive data is read (for example Device ID, contacts, GPS location, etc.) and ends in Sink where sensitive data is moved out of component (for example: Internet, text messages, files system, etc.). The data is considered as leaked if it really passes from source to sink. The major weakness of FlowDroid is that it only analyzes on a single application.

As an extended version of FlowDroid, DidFail [3] combines FlowDroid and Epicc analysis to determine the sensitive data-flow on multiple applications. The biggest advantage of DidFail is that it can detect inter-application malware. However, it also has some limitations such as DidFail only supports for Activity and Service; it is difficult to control all source code because Epicc is not an open source.

SmartDroid adopts the hybrid approach that combines the static and dynamic analysis to detect the sensitive behavior of Android malware. The analysis focuses on the user interface (UI) interactions of Android application. The authors use static analysis to extract the chains of sensitive UI interactions that may lead to information leakage, then they use dynamic analysis to explore and perform these chains automatically. The sensitive UI interactions are monitored to get runtime information that is used to analyze and determine which UI interaction is malware behavior. However, SmartDroid only supports to analyze on a single application.

3 Proposed System

Get the inspiration from SmartDroid, we propose a system, named eDSDroid, that uses a new hybrid approach of the static and dynamic analysis to detect information leakage on inter-applications. In particular, we use static analysis to determine the chains of sensitive actions on multiple applications and then we use dynamic analysis to explore and perform these chains automatically. Besides that, we monitor and collect the runtime information from the sensitive actions. The information tracked from these actions helps us to determine which actions make information leak.

eDSDroid system has four modules: ESA Builder, ESA Receiver, Application Interactor, and Log Analyzer (see Fig. 1).

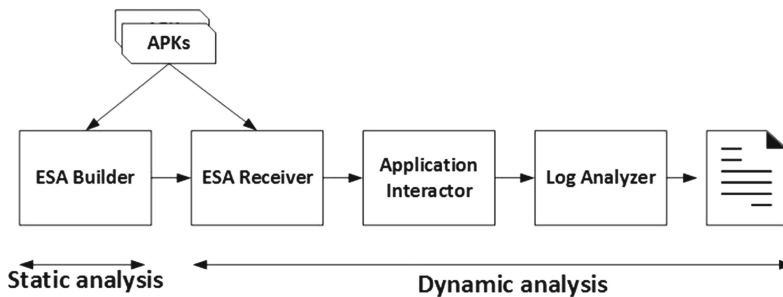


Fig. 1. eDSDroid model

3.1 ESA Builder

The ESA Builder (Expected Sensitive Action Builder) used to extract the chains of user actions that may lead to information leak on multiple applications. In this study, we call these chains are the Expected Sensitive Actions – ESAs.

The process of the ESA Builder is showed in 4 steps:

- Step 1: generates sensitive data-flows from apk files by using DidFail. The result of this step is the list of possible source-sink pairs (see Fig. 2).

```

Data Flow 1:
Source: WriteFile - Location.getLastKnownLocation()
Sink: WriteFile - Log.i()

Data Flow 2:
Source: WriteFile - String.getString()
Sink: WriteFile - Log.i()

Data Flow 3:
Source: SendSMS - TelephonyManager.getDeviceId()
Sink: SendSMS - Log.i()

Data Flow 4:
Source: WriteFile - String.getString()
Sink: WriteFile - FileOutputStream.write()
    
```

Fig. 2. Sensitive data-flows

- Step 2: uses apktool [1] to decompile apk files to Smali Byte Code files.
- Step 3: creates AFCG tree for each API function in the list of source-sink pairs (in step 1) based on Smali Byte Code (in step 2). AFCG - Activity Function Call Graph is a tree graph of the relationship between activities and functions in the application. The root node is the main activity. The leaf nodes are sensitive APIs. Figure 3 is an example for AFCG.

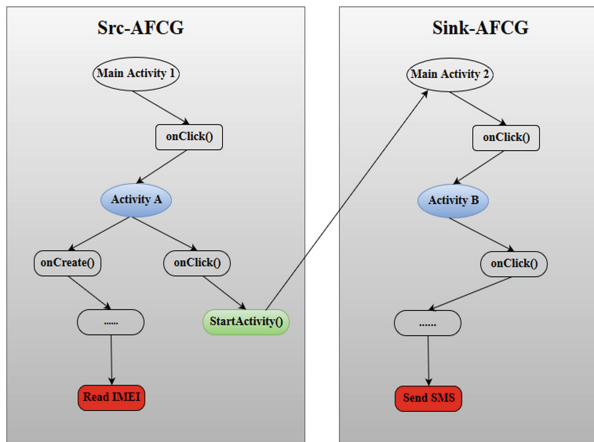


Fig. 3. Example for AFCG tree

The detail of the step 3 is below:

- The first, we create AFCG trees for Source (the left tree in Fig. 3) and Sink (the right tree in Fig. 3). To create AFCG tree, we locate the API function in Smali Byte Code

files and mark it as the leaf node of the tree. Next, we find its parent function and add this function as a parent node. The finding will be finished if the parent function is a member of the Main Activity class.

- The second, we combine these trees to create the full AFCG tree for source-sink (see Fig. 3).

The result of this step is the list of AFCG trees for the corresponding source-sink pairs.

- Step 4: extracts ESAs. We perform depth-first searching algorithm in AFCG tree to get ESAs on multiple applications. The ESAs will be used to direct the dynamic analysis. For example, from the tree in Fig. 4, we can extract the ESAs as below:
 - ESA for source: Main Activity 1 → onClick() → Activity A → onCreate() → Read IMEI.
 - ESA for sink: Main Activity 1 → onClick() → Activity A → onClick() → StartActivity → MainActivity 2 → onClick → Activity B → onClick() → Send SMS.

```
Data Flow 9:
  Source: echoer - String.getString()
  Sink: echoer - Log.i()

Application Interactor log:
#####eDSDroid Log: Start Flow 9

Data Flow 10:
  Source: SendSMS - LocationManager.getLastKnownLocation()
  Sink: echoer - Log.i()

Application Interactor log:
#####eDSDroid Log: Start Flow 10
Src: org.cert.WriteFile.SubActivity->LocationManager getLastKnownLocation() called
Sink: org.cert.echoer.MainActivity->Log i(String, String) called with params pl =
```

Fig. 4. Tracer's log

3.2 ESA Receiver

ESA Receiver (Expected Sensitive Action Receiver) is an Android application which sends ESAs to Application Interactor. This application is also used to control the running of applications.

3.3 Application Interactor

Application Interactor module used to interact automatically with the applications according to the behaviour in ESAs. Beside that, we also monitor the sensitive API functions during interaction process. In this study, we implemented this module by modifying Android Framework. This module has three parts: Activity Controller, UI Controller, and Sensitive API Tracer.

- Activity Controller: We modified Activity class of Android Framework to restrict the activities which not in ESAs.

- **UI Controller:** We modified View class of Android Framework to control all Android UI controls and make them perform after ESAs.
- **Sensitive API Tracer:** We modified the sensitive API functions of Android Framework to add the tracers which are the print commands uses to print out the tracked information to the console. The tracked information is caught by Android Logcat tool.

3.4 Log Analyzer

Log Analyzer used to collect and analyze the tracer's information from Application Interactor module to determine which sensitive data-flow indeed leaks. We use Android Logcat tool (a command-line tool which used for android debugging) to collect the tracer's information and export them into the log file (see Fig. 4).

As mentioned in the previous sections, eDSDroid determines a list of the sensitive data-flow and their corresponding ESAs during the static analysis process. In dynamic analysis stage, it will interact automatically with the applications according to the behavior in these ESAs. During the interaction, the information related to sensitive API functions will be logged if these functions are called. If sensitive API functions appeared in the logs, it means their corresponding sensitive data flow come be true. An example in Fig. 4 shows that the data-flow 10 exists the log of both source and sink. So, we can confirm this data-flow makes the leakage.

4 Evaluation

4.1 Dataset

We use Toyapps [3] to validate eDSDroid. Toyapps is a famous dataset that used to evaluate sensitive data leakage on Android. Toyapps is the set of three applications. They work together to leak the user data as device ID, user location, etc.

- *SendSMS*: gets and sends device ID to Echoer app, then gets it again from Echoer app and leaks it via SMS.
- *WriteFile*: gets and sends user location to Echoer app, then gets it again from Echoer app and writes it to file.
- *Echoer*: receives devive ID and user location from SendSMS/WriteFile apps, then forwards back them to these apps.

4.2 Result

We have analyzed the Toyapps dataset by using eDSDroid and DidFail.

According to the analysis results in Table 1, DidFail determined 11 data-flows, while eDSDroid determined 10 data-flows. In particular:

- For data-flows on a single application: DidFail determined 9 data-flows. There are only 8 data-flows really leaks detected by eDSDroid. The dynamic analysis state of eDSDroid helped to remove the data-flow which not occur at run-time.

- For data-flows on multiple applications: both tools have the same results with two data-flows detected.

Table 1. The analysis result of DidFail and eDSDroid.

Data-flow	DidFail	eDSDroid
Data-flows on single app	9	8
Data-flows on multiple apps	2	2
Total data-flows	11	10

From these results, we have a comparison between these tools, show in Table 2.

Table 2. The comparison between the tools.

Comparison	DidFail	eDSDroid
Technical analysis	Static	Hybrid of static and dynamic
Inter-application analysis	Yes	Yes

5 Conclusions and Future Work

In this paper, we propose a new hybrid approach which uses static analysis to direct dynamic analysis to analyze inter-application malware. We have also implemented a new analysis tool – namely eDSDroid which can analysis and detects information leakage on multiple applications. According to the testing result on Toyapps, eDSDroid shows the correctness and effectiveness in detecting inter-application malware.

Some limitations of eDSDroid should be resolved in the future, such as it only supports for Activity and Service. And it does not process for data dependency on UI elements. For example, our tool can not detect in case: the malware application sends SMS only if a checked variable is equal to true and this variable is only set to true when the user clicks on the button. In future works, we will continue to perform more testing on the real world application to complete the eDSDroid tool.

Acknowledgments. This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2016-26-01.

References

1. ApkTool. <https://ibotpeaches.github.io/Apktool/>
2. Sanz, B., Santos, I., Ugarte-Pedrero, X., Laorden, C., Nieves, J., Bringas, P.G.: Instance-based anomaly method for android malware detection. In: SECURE, pp. 387–394. SciTePress (2013)
3. CERT Division of the Software Engineering Institute (SEI), DidFail: Android Taint Flow Analysis. <https://www.cert.org/secure-coding/tools/didfail.cfm>

4. Zheng, C., Zhu, S., Dai, S., Gu, G., Gong, X., Han, X., Zou, W.: SmartDroid: an automatic system for revealing UI-based trigger conditions in android applications. In: Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, (SPSM 2012), pp. 93–94 (2012)
5. Oceau, D., et al.: Effective inter-component communication mapping in android with Epicc: an essential step towards holistic security analysis. USENIX Security (2013)
6. Bodden, E.: FlowDroid Taint Analysis, Secure Software Engineering. European Center for Security and Privacy by Design. sseblog.ec-spride.de/tools/FlowDroid/
7. Chin, E., Felt, A.P., Greenwood, K., Wagner, D.: Analyzing inter-application communication in android. In: Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services, MobiSys 2011, pp. 239–252 (2011)
8. Felt, A.P., Chin, E., Hanna, S., Song, D., Wagner, D.: Android permissions demystified. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, pp. 627–638. ACM (2011)
9. Tchakounte, F., Dayang, P.: System call analysis of malwares on android. Int. J. Sci. Technol. 2(9), 669–674 (2013)
10. Fuchs, A.P., Chaudhuri, A., Foster, J.S.: ScanDroid: automated security certification of Android applications. Technical report, University of Maryland (2009)
11. Paul, I.: F-secure says 99 percent of new mobile malware targets android, but don't worry too much. <http://www.greenbot.com/article/2148521/99-percent-of-new-mobile-malware-is-on-android-but-good-luck-catching-it.html>. Accessed 2 Sept 2014
12. Hoffmann, J., Neumann, S., Holz, T.: Mobile malware detection based on energy fingerprints — a dead end? In: Stolfo, S.J., Stavrou, A., Wright, C.V. (eds.) RAID 2013. LNCS, vol. 8145, pp. 348–368. Springer, Heidelberg (2013). doi: [10.1007/978-3-642-41284-4_18](https://doi.org/10.1007/978-3-642-41284-4_18)
13. Graa, M., Cuppens-Boulahia, N., Cuppens, F., Cavalli, A.: Detecting control flow in smartphones: combining static and dynamic analyses. In: Xiang, Y., Lopez, J., Kuo, C.-C., Jay, Zhou, W. (eds.) CSS 2012. LNCS, vol. 7672, pp. 33–47. Springer, Heidelberg (2012). doi: [10.1007/978-3-642-35362-8_4](https://doi.org/10.1007/978-3-642-35362-8_4)
14. Arzt, S., et al.: FlowDroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. In: PLDI (2014)
15. Schmidt, A.-D., Bye, R., Schmidt, H.-G., Clausen, J., Kiraz, O., Yuksel, K.A., Camtepe, S.A., Albayrak, S.: Static analysis of executables for collaborative malware detection on android. In: IEEE International Conference on Communications, ICC 2009, pp. 1, 5, 14–18, June 2009
16. Nair, S.K., Simpson, P.N.D., Crispo, B., Tanenbaum, A.S.: A virtual machine based information flow control system for policy enforcement. Electron. Notes Theor. Comput. Sci. 197, 3–16 (2008)
17. van der Veen, V., Rossow, C., Bos, H.: TraceDroid: a fast and complete android method tracer. Hack In The Box, HITB, Malaysia, October 2013
18. Enck, W., Gilbert, P., Chun, B.-G., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.N.: TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In: Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation, OSDI 2010, pp. 1–6 (2010)

Detect Sensitive Data Leakage via Inter-application on Android by Using Static Analysis and Dynamic Analysis

Nguyen Tan Cam^{1(✉)}, Van-Hau Pham², and Tuan Nguyen²

¹ Faculty of Science and Technology, Hoa Sen University, Ho Chi Minh City, Vietnam
cam.nguyentan@hoasen.edu.vn

² Faculty of Computer Network and Communications, University of Information Technology,
Vietnam National University, Ho Chi Minh City, Vietnam
{haupv, tuanna}@uit.edu.vn

Abstract. Mobile malwares (especially spyware) target heavily Android operating system. Data is leaked if it exists a sensitive data flow (Data propagation from sensitive source to critical sink). Usually, a sensitive data flow is executed by a chain of actions. In most cases, sensitive data flows are begun and finished in the same application. However, there exist cases where these flows can pass to multi-applications by using inter-application communication. Standalone application analysis can not detect such data flows. Static analysis faces limitations when malware code is obfuscated. Besides, certain actions only take place when receiving input from user. It means that the information related to sensitive data flows is depended on the input data. Which is not available at analysis time when using static analysis technique. In this study, we propose uitHyDroid system that allows to detect sensitive data leakage via multi-applications by using hybrid analysis. uitHyDroid uses static analysis to collect sensitive data flows in each application. Meanwhile, dynamic analysis is used to capture inter-application communications. In this study, to evaluate our approach, we use the extended of DroidBench dataset and applications downloaded from GooglePlay. The experimental results show that almost of sensitive data leakages in the first dataset are correctly detected. Beside that, the proposed system detects several malwares in real-world applications.

Keywords: Android security · Hybrid analysis · Inter-application communication · Sensitive data leakage detection

1 Introduction

According to Symantec [1], there are 3944 new malwares on Android in 2015, surging 77% year-on-year. In a report of IDC, the Android operating system (OS) accounted for 82.8% of the market shared [2]. From the other analysis of Symantec [3], the act of stealing data and tracking the users are accounted for the highest proportion with 36%.

Android applications can transfer data to each other by using inter-application communication (IAC) such as Intent, shared file. Detecting sensitive data leakage by analysis IAC is a technique adopted in several studies such as IccTA [4], DIDFAIL [5], ApkCombiner [6], and IACDroid [7]. However, these studies only use static analysis.

One of the weakness of static analysis is that the input data is not available at analysis time. It means that static analysis can not analyze IACs that are generated at run time. Moreover, it is difficult to analyze applications use obfuscation technique. To resolve the limitations of static analysis, some other studies use dynamic analysis to detect sensitive behaviors of Android applications [8–10]. Dynamic analysis can be supported by several frameworks that allow to interact with application [11–14].

Monkey [12] is the popular interaction tool. It uses the random interaction strategy. While still adopting random interaction strategy, Dynodroid [13] is more advanced thank to it’s ability to generate system events. However, it needs to modify Android framework to gather information of the registered applications.

Some other studies combine dynamic analysis and static analysis to analyze Android applications. Intent Fuzzer [11] uses FlowDroid [15] to determinate system calls that related to Intent. On device, the system generates data related to the intent. This data is sent to the target application using Android’s SDK. This method requires source code of the Android applications to analyze. It is the main limitation of this study.

A3E [14] allows to explore applications while running on actual device. To explore applications, this tool uses the relationships of screens of these applications and the list of all UI elements in each screens. The relationships of screens are determined by ScanDroid [16]. This approach uses fuzzing interaction with all UI elements. However, several UI elements relate to sensitive behaviors. To determine UI elements related to chain of sensitive behaviors, SmartDroid [10] uses static analysis to determine chain of sensitive behaviors. After that, it uses dynamic analysis to verify these behaviors. In dynamic analysis phase, the authors modify emulator to perform the expected behaviors. The experimental result of this study shows that it is more efficient than the other studies that use fuzzing interaction. However, this work only focus on standalone application. It is not able to analyze application that use IAC to leak sensitive data.

To summarize, Table 1 shows the characters of related works. Only SmartDroid focuses on related UI elements in dynamic analysis (on target interaction). However it analyzes standalone application instead of inter-application.

Table 1. Main characters of related works.

Related works	Single application	Inter-application	Static analysis	Dynamic analysis	Hybrid analysis	Interaction approach
IccTA [4]		✓	✓			
DIDFAIL [5]		✓	✓			
ApkCombiner [6]		✓	✓			
IACDroid [7]		✓	✓			
Dynodroid [13]	✓			✓		Fuzzing
A3E [14]	✓			✓		Fuzzing
SmartDroid [10]	✓				✓	On target
Intent Fuzzer [11]		✓			✓	Fuzzing

In this study, we propose a system, named uitHyDroid, which is used to detect sensitive data leakage via multi applications without modifying emulator. This system uses static analysis to determine the sets of events that are performed to generate possible sensitive data

flows. These flows can be linked with the other flows in the other application to create an inter-application flow by using IAC. To bypass the encrypted code technique and code obfuscation technique, we use dynamic analysis to capture the IAC's data. This data is used to complete inter-application sensitive data flows. By interacting with related UI elements, that are collected in static analysis phase, the proposed system is faster in dynamic analysis.

To evaluate the accuracy of uitHyDroid, we use DroidBench dataset [17], our samples and 100 selected real applications from Android market. The results show that uitHyDroid has the high accuracy.

This study has the following main contributions:

- We propose an approach to detect inter-application sensitive data leakage by combining static analysis and dynamic analysis. Static analysis phase determines all possible chain of behaviors in each application. In dynamic phase, the proposed system monitors data of IAC to determine possible chain of behaviors via multi-applications.
- We enrich DroidBench Dataset [17] by adding 20 samples that demonstrate sensitive data leakage through IAC by using code obfuscation technique and dynamic data generation.
- We evaluate the effectiveness of uitHyDroid on the extended dataset and real-world applications in the wild. The experimental results show that the proposed system is promising.

The rest of this paper is organized as follows: The motivation examples are introduced in Sect. 2. The architecture uitHyDroid is presented in Sect. 3. Implementation and evaluation are mentioned in Sect. 4. Section 5 concludes the paper.

2 Motivation Examples

There are four different types of application components, i.e. Activities, Services, Content Providers, and Broadcast receivers [18]. Each type of application components serves a distinct purpose. An activity in Android application represents a single screen with a user interface. Each activity can contain one or more UI elements. End user can interact with these elements to create a chain of behaviors.

The example in Fig. 1 shows a case study of a spyware. This application provide a normal function like any word counter utility. Beside that, it performs sensitive data leakage based on interaction of end users. In this example, we have four action chains that are listed in Table 2. Each action chain includes a list of UI elements in ordered that end user need to interact with the application.

Table 2. The action chains of the example application.

No	Action chain	Is sensitive action chain	Number of interaction
1	“Word Count”	×	1
2	“Character Count”	×	1
3	“About” → “History”	×	2
4	“About” → “Update”	✓	2

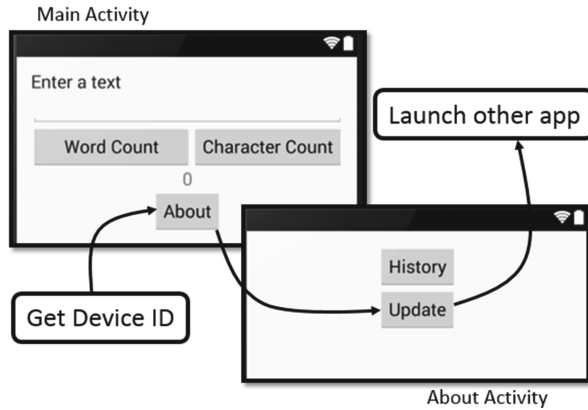


Fig. 1. The motivation example.

In this example, when we click on the button *About* in *Main Activity*, the application gets the device ID then sends this data to *About Activity*. This data is leaked to other application when we click on button *Update* in *About Activity*. It means that, only action chain number 4 can leak sensitive data with two interaction times. We call this action chains is sensitive action chain. If we use the fuzzing interaction approach, we need to interact with the application six times (the total value of the last column in Table 2). From this analysis, the first research question we want to answer is: “How is the dynamic analysis affected if the chain of sensitive actions are determined before?”

Beside that, static analysis can not know data that generated during application execution. Figure 2 shows an example that uses real-time input data to perform several corresponding behaviors. In this example, the malicious behavior is generated by clicking on button *OK* that is depended on content of the edit text.

From this analysis, the second research question is: “Can dynamic analysis be used to know dynamic generated data on runtime to solve the limitations of static analysis?”.

To monitor dynamic generated data at runtime, we can modify Android framework or application source code [8, 10]. The third research question is: “Can we capture dynamic generated data at runtime without framework modification or repackaging?”

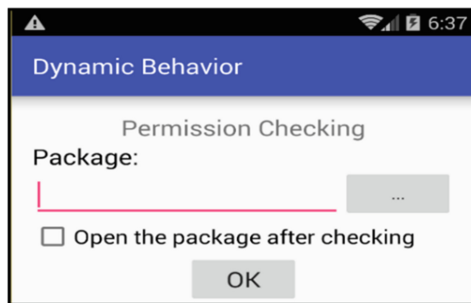


Fig. 2. The main activity of *Dynamic Behavior* package.

3 uitHyDroid System

As depicted in Fig. 3, our uitHyDroid system consists three modules: Data Path Collector, Automation Operator and IAC Collector.

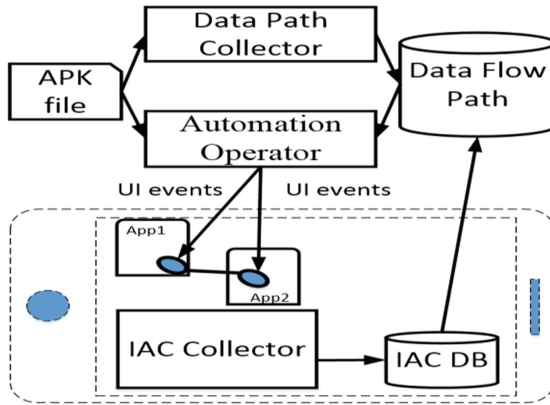


Fig. 3. The uitHyDroid architecture.

Data Path Collector module analyzes apk files to determine the chains of sensitive actions that may lead to sensitive data leakage. The information of these chains (SA) consists of the following elements: chain of ordered functions (F), list of related UI elements (U), and list of activities (A). It means that $SA = \{F, U, A\}$. F contains functions in source-sink list [19]. The information of SA represent the chains of events that we need to interact with application to leak sensitive data. In this context, SA is called Data flow path. Data Path Collector is depicted in Fig. 4.



Fig. 4. Data Path Collector Module.

In this module, we use ApkTool [20] to convert apk file to smali code and Android-Manifest.xml. Sensitive data leakage is performed by invoking system calls used to get sensitive source and system calls used to send the data out of device. When an application performs sensitive data leakage, it means that these system calls exist in its smali code files. Firstly, this module determines location of these system calls in smali code files. From these locations, it traces back to collect chain of ordered functions (F). Then it determines UI elements (U) that related to F . Lastly, this module determines list of activities (A) that contain elements in U .

If F contains a sink function for inter-application communication purpose such as *startActivity*, *startActivityForResult*, the module uses string analysis to collect possible

related data. For example, the module determines target component and target applications in Intent (one of parameters of *startActivity* function).

Automation Operator module interacts with Android applications by using information of SA from Data Path Collector. To interact with related UI elements in each activity in *A*, we use Selendroid [21], a framework that allows to interact with an application on Android device or Emulator from outside. This module determines the first activity in each *A*. Then it interacts with related UI elements in this activity. If the interaction launches any next activity and the activity is a member of *A*, the interaction process is continued.

To complete data flows via multi-applications, we need to determine inter-application communications (IACs). Each type of inter-application communication has the corresponding information. For example, when using function *startActivity* to perform IACs, we need information of attributes of Intent object.

IAC collector module collects information of IACs during application execution by using our own Android service. This service uses Xposed [22] to hook the system calls related to IACs. Xposed is a framework that allows to intercept a method, change parameters for the method call, modify the return value or skip the call to the method completely. In this study, we intercept methods that related to inter-application communications to collect related data. This monitoring can capture dynamic generated data on runtime and bypass code obfuscation technique that static analysis can not do.

The related data from IAC Collector is saved into a database, named IAC DB. This database is send back to Automation Operator module to perform interacting with other applications related to the IACs.

4 Evaluation

In this study, to evaluate the uitHyDroid system, we use samples in Inter-application communication category of DroidBench [17] and our samples. Our sample is proposed to demonstrate case studies of dynamic data generation and code obfuscation. Table 3 depicts our 20 samples. The first column of Table 3 represents the application groups that cooperate to leak sensitive data. For simplicity, we use Intent to leak IMEI number in these samples.

The experimental results show that uitHyDroid detects all sensitive data leakage case studies in Inter-application Communication category of DroidBench and our dataset. The main contribution of this work is the proposed system is able to detect sensitive IACs in our samples that use dynamic content generation technique and code obfuscation technique. To evaluate the capacity of the proposed system on real-world applications, we use 100 applications in the wild, that include several applications detected the existing of sensitive data leakage in previous work [7]. The analysis results show that uitHyDroid can detect sensitive data flows in real-world applications.

Table 3. Our own samples in dynamic content category and code obfuscation category.

Sample name	Source	Sink	Category
ntcDynamicContent1.apk → ntcDynamicContent2.apk	IMEI	SMS	Dynamic Content
ntcDynamicContent3.apk → ntcDynamicContent4.apk	IMEI	SMS	Dynamic Content
ntcDynamicContent5.apk	–	–	Dynamic Content
ntcDynamicContent6.apk → ntcDynamicContent7.apk	IMEI	SMS	Dynamic Content
ntcDynamicContent8.apk → ntcDynamicContent9.apk	IMEI	SMS	Dynamic Content
ntcDynamicContent10.apk	–	–	Dynamic Content
ntcCodeObfuscation1.apk → ntcCodeObfuscation2.apk	IMEI	SMS	Obfuscation
ntcCodeObfuscation3.apk → ntcCodeObfuscation4.apk	IMEI	SMS	Obfuscation
ntcCodeObfuscation5.apk	–	–	Obfuscation
ntcCodeObfuscation6.apk → ntcCodeObfuscation7.apk	IMEI	SMS	Obfuscation
ntcCodeObfuscation8.apk → ntcCodeObfuscation9.apk	IMEI	SMS	Obfuscation
ntcCodeObfuscation10.apk	–	–	Obfuscation

5 Conclusion

In this study, we proposed a hybrid analysis system, named uitHyDroid, to detect inter-application sensitive data leakage in Android applications. The system can know information of inter-application communication to solve the limitations of static analysis when facing code obfuscation technique and dynamic generated data. The test results show that the uitHyDroid system can successfully detect sensitive data leakages in Inter-application communication category of DroidBench and our dataset. Moreover, this system has the capacity to detect sensitive data leakage in real-world applications in the wild. However, within the limited scope of the study, we only test the proposed system on these simple datasets. This system must be evaluated on the larger dataset in the future.

Acknowledgments. This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2016-26-01.

References

1. Symantec: Internet Security Threat Report, vol. 21, April 2016. <https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf>
2. Corporation, I.D. (2015). <http://www.idc.com/prodserv/smartphone-os-market-share.jsp>

3. Symantec: 2015 Internet Security Threat Report, vol. 20 (2015). http://www.symantec.com/security_response/publications/threatreport.jsp. Accessed May 2015
4. Li, L., et al.: IccTA: detecting inter-component privacy leaks in android apps. In: The 37th International Conference on Software Engineering (ICSE), Firenze, Italy (2015)
5. Klieber, W., et al.: Android taint flow analysis for app sets. In: Proceedings of the 3rd ACM SIGPLAN International Workshop on the State of the Art in Java Program Analysis, pp. 1–6. ACM, Edinburgh, United Kingdom (2014)
6. Li, L., Bartel, A., Bissyandé, Tegawendé, F., Klein, J., Traon, Y.L.: ApkCombiner: combining multiple android apps to support inter-app analysis. In: Federrath, H., Gollmann, D. (eds.) SEC 2015. IAICT, vol. 455, pp. 513–527. Springer, Heidelberg (2015). doi: [10.1007/978-3-319-18467-8_34](https://doi.org/10.1007/978-3-319-18467-8_34)
7. Cam, N.T., Hau, P., Nguyen, T.: Android security analysis based on inter-application relationships. In: Kim, K.J., Joukov, N. (eds.) Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 689–700. Springer, Heidelberg (2016). doi: [10.1007/978-981-10-0557-2_68](https://doi.org/10.1007/978-981-10-0557-2_68)
8. Enck, W., et al.: TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In: Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation, pp. 1–6. USENIX Association, Vancouver, BC, Canada (2010)
9. Shabtai, A., et al.: “Andromaly”: a behavioral malware detection framework for android devices. *J. Intell. Inf. Syst.* **38**(1), 161–190 (2012)
10. Zheng, C., et al.: SmartDroid: an automatic system for revealing UI-based trigger conditions in android applications. In: Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, pp. 93–104. ACM, Raleigh, North Carolina, USA (2012)
11. Sasnauskas, R., Regehr, J.: Intent fuzzer: crafting intents of death. In: Proceedings of the 2014 Joint International Workshop on Dynamic Analysis (WODA) and Software and System Performance Testing, Debugging, and Analytics (PERTEA), pp. 1–5. ACM, San Jose, CA, USA (2014)
12. UI/Application Exerciser Monkey (2016). <http://developer.android.com/tools/help/monkey.html>
13. Machiry, A., Tahiliani, R., Naik, M.: Dynodroid: an input generation system for Android apps. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, pp. 224–234. ACM, Saint Petersburg, Russia (2013)
14. Azim, T., Neamtiu, I.: Targeted and depth-first exploration for systematic testing of android apps. *SIGPLAN Not.* **48**(10), 641–660 (2013)
15. Arzt, S., et al.: FlowDroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for Android apps. In: Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 259–269. ACM, Edinburgh, United Kingdom (2014)
16. Fuchs, A.P., Chaudhuri, A., Foster, J.S.: Scandroid: automated security certification of android (2009)
17. Spride, E.: DroidBench – Benchmarks. (2016). <http://sseblog.ec-spride.de/tools/droidbench/>. Accessed 10 March 2016
18. Android.com. <https://developer.android.com/guide/components/fundamentals.html>. Accessed 5 May 2016
19. Rasthofer, S., Arzt, S., Bodden, E.: A machine-learning approach for classifying and categorizing android sources and sinks (2014)
20. ApkTool (2015). <https://github.com/iBotPeaches/Apktool>
21. Selendroid: Selenium for android (2016). <http://selendroid.io>. Accessed 2016
22. Xposed framework (2016). <http://repo.xposed.info/module/de.robv.android.xposed.installer>. Accessed 5 July 2016

Known Bid Attack on an Electronic Sealed-Bid Auction Scheme

Kin-Woon Yeow^(✉), Swee-Huay Heng, and Syh-Yuan Tan

Faculty of Information Science and Technology,
Multimedia University, Cyberjaya, Melaka, Malaysia
yeowkinwoon@gmail.com, {shheng, sytan}@mmu.edu.my

Abstract. In this paper, we cryptanalyze a receipt-free electronic sealed-bid auction scheme and show that it is forgeable under the known bid attack. Specifically, we show that a malicious sealer can forge the sealed-bid with non-negligible probability. Besides, we also propose a possible fix for the attack.

Keywords: Sealed-bid · Sealer · Electronic auction · Known bid attack

1 Introduction

Auction is a process of trading goods or services with competition among bidder to achieve the highest bid in a fair manner between the auctioneer and bidders. The most prevailing auctions [10] are English auction, Dutch auction, First-price sealed-bid auction and Second-price sealed-bid auction [14]. First-price sealed-bid auction is a type of auction process in which all potential buyers concurrently submit sealed-bids in a single round, every submission of bid is confidential and the highest bidder is rewarded with the item for the price that the bidder bid [10].

Franklin and Reiter [3] presented the pioneer (first-price) sealed-bid auction in 1996. Their proposed scheme possessed the property of anonymity but cannot overcome the bid-rigging problem. Generally, bid-rigging takes place when the coercers order other bidders to bid at low prices in order for him to win the auction easily [4]. To overcome the attack, Benaloh and Tuinstra [1] introduced the security properties of receipt-freeness by using homomorphic encryption. To achieve stronger anonymity, Sakurai and Miyazaki [12] proposed an electronic sealed-bid auction scheme with bulletin board that can provide anonymity, confidentiality, and non-repudiation. In 2000, Viswanathan et al. [15] proposed a three phased scheme for electronic sealed-bid auction that based on rules of interaction to reveal the winning bid and achieve higher efficiency compared to [12, 13].

In 2008, Wu et al. [16] cryptanalyzed Liaw et al. [9] electronic auction scheme besides proposing a electronic sealed-bid auction scheme that enhances the efficiency and functionality of previous sealed-bid auction schemes [2, 7, 9]. In 2009, Howlader et al. [4] introduced a new entity, namely, sealers in sealed-bid auction to prevent bid-rigging attack. Later, Lee et al. proposed an efficient scheme under GDH assumption that is proven secure against impersonation attack [8].

In 2014, Montenegro and Lopez [11] used secure multiparty computation to propose a electronic sealed-bid auction scheme which enables bidder to hide the bidding information from the auctioneer. Recently, Howlader and Mal [5] published a receipt-free electronic sealed-bid auction mechanism that could withstand exposed public communication channel.

In the context of electronic sealed-bid auction, Howlader and Mal [5] adopted the similar bid structure as in Howlader, Ghosh, and Pal [4] and Howlader, Roy, and Mal [6] which make use of the entity sealers. In [6], Howlader, Roy, and Mal made comparison on the security assumption of [4–6] and showed that electronic sealed-bid auction schemes do not need to assume an honest auctioneer.

1.1 Our Contribution

We present the known bid attack that can be mounted on Howlader et al.’s electronic sealed-bid auction (ESBA) scheme [4]. The scheme is based on the assumption that some sealers are honest, i.e., some sealers are malicious. Besides that, Howlader et al. claimed that the auctioneer cannot open the bid before the scheduled time and any sealed-bid provided by dishonest sealers to coercers in any intermediate stage of the auction will not reveal any information.

In this paper, we present a known bid attack mounted by the malicious first sealer that seals the bid for the winning bidder in Howlader et al.’s ESBA scheme [4]. In particular, we show that after the first round of the auction, the malicious first sealer can successfully forge a valid bid in the subsequent auction under the known bid attack by manipulating the encrypted bid-vector $\langle {}_i T_j, {}_i C_j \rangle$.

Besides, we propose a fix for Howlader et al.’s ESBA scheme [4] which strengthens the bid’s commitment value to prevent the removal of hash value that contains bidder’s identification.

1.2 Organization

We organize the paper as follows. The basic definition of sealed-bid auction is introduced in Sect. 2. In Sect. 3, we recall Howlader et al.’s ESBA scheme. In Sect. 4, we present and discuss the known bid attack on the ESBA scheme. The fix is proposed and discussed in Sect. 5, followed by the conclusion in Sect. 6.

2 Sealed-Bid Auction Scheme

In this section, we briefly describe the definitions of the electronic sealed-bid auction [4].

Definition 1. *An electronic sealed-bid auction (ESBA) scheme is a tuple of algorithms (Setup, Bid, Open), specify by:*

1. **Setup:** System setting and parameters distribution that includes private key of each entity (bidder, sealer, auctioneer) and publication of their public key.

2. **Bid:** Bidder selects a bidding price from the price list given by the auctioneer. Bidder generates a bid based on the selected price and sends to the sealer for sealing operation through anonymous channel.
 - **Seal:** Sealer receives the bid from the bidder and starts sealing through sequential layers of sealers. The last sealer publishes the sealed-bid to the bulletin board.
 - **Verify:** Bidder verifies the sealed-bid from the bulletin board.
3. **Open:** Auctioneer opens the winning sealed-bid and confirms with the winning bidder.

3 Howlader et al.’s ESBA Scheme [4]

We first describe the Howlader et al.’s ESBA scheme as follows before showing the known bid attack in the next section. The ESBA scheme consists of three common algorithms in sealed-bid auction, namely, (Setup, Bid, Open):

- **Setup:** Auctioneer A selects subgroup \mathbb{G}_q of order q from \mathbb{Z}_p^* , where p and q are large primes and $p|q - 1$. Let $g \in \mathbb{G}_q$ be a generator of group \mathbb{G}_q . G_y and G_n are two independent generators of group \mathbb{G}_q , which indicate bidding and not bidding at price p_j respectively. Bidder B_i ’s private key is x_{B_i} and public key is $h_{B_i} = g^{x_{B_i}}$. A chooses his private key x_A and publishes his public key $h_A = g^{x_A}$. Sealer S_i has a private key x_{S_i} and publishes his public key $h_{S_i} = g^{x_{S_i}}$. $h_S = \prod_{i=1}^t h_{S_i}$ is the shared public key of t sealers. $h_{S/S_1, S_2, \dots, S_t} = h_S/h_{S_1}h_{S_2} \dots h_{S_t}$ is the shared public key of sealers excluding S_1, S_2, S_t .
- **Bid:** The bidder B_i selects his bidding price from the list P and generates an encrypted bid-vector. The bidder sends the encrypted bid-vector $\langle {}_i\Gamma_j, {}_iC_j \rangle$ to the sealers through an anonymous channel. The sealers S_1, S_2, \dots, S_t perform the sealing operation and write the receipt-free sealed bid-vector on a public board. The sealing operation is performed by all the t sealers and the bidder can verify his sealed bid-vector after that. The detail description of the bidding process is as follows.
 Bidder B_i decides his bidding price $p_j \in P$ from the price list published by the auctioneer and encrypts the price as follows:

$${}_i\Gamma_j = ({}_iX_j, {}_iY_j) = (g^{i r_j}, h_A^{i r_j} h_S^{i r_j} G) \text{ for } 1 \leq j \leq n$$

where $i r_j \in_R \mathbb{Z}_q$ is randomly selected by bidder B_i . The notation ${}_i\Gamma_j$ denotes the j^{th} encrypted price of the bidder B_i . Furthermore, the notation G indicates G_y for bidding at price, p_j or G_n for not bidding at price, p_j .

$${}_iC_j = h_A^{i r_j H({}_i\Gamma_j) x_{B_i}} \text{ for } 1 \leq j \leq n$$

The hash function, H in above equation is a one-way hash function. The bidder B_i constructs the encrypted bid-vector ${}_i\Gamma_j$ and sends to the first sealer, S_k through an anonymous channel. When S_k receives the encrypted bid-vector,

he engraves his random seed and computes the partial bid as $\langle {}_i\Gamma_{j,k}, {}_iC_j, {}_i\mathbb{R}_j \rangle$ where ${}_i\Gamma_{j,k} = ({}_iX_{j,k}, {}_iY_{j,k})$:

$$\begin{aligned} {}_iX_{j,k} &= {}_iX_j \cdot g^{i r_{j,s_k}}, & {}_iY_{j,k} &= {}_i r_{j,s_k} \cdot {}_iY_j \cdot h_A^{i r_{j,s_k}} \cdot h_{S/S_k}^{i r_{j,s_k}} \cdot {}_iX_j^{-x_{S_k}} \\ &= g^{i r_j} \cdot g^{i r_{j,s_k}}, & &= {}_i r_{j,s_k} \cdot h_A^{i r_j} h_S^{i r_j} \cdot h_A^{i r_{j,s_k}} \cdot h_{S/S_k}^{i r_{j,s_k}} \cdot h_{S_k}^{-i r_j} \cdot G \\ &= g^{i r_j + i r_{j,s_k}}, & &= {}_i r_{j,s_k} \cdot h_A^{i r_j + i r_{j,s_k}} \cdot h_{S/S_k}^{i r_j + i r_{j,s_k}} \cdot G \end{aligned}$$

where ${}_i r_{j,s_k} \in_R Z_q$ is randomly selected by S_k . Next, the response ${}_i\mathbb{R}_j = \langle {}_iR_j, {}_i\alpha_j, {}_i\beta_j \rangle$ is computed by S_k by selecting random numbers ${}_i w_{j,k}$ for $1 \leq j \leq n$ and computing:

$$\begin{aligned} {}_iR_j &= {}_i w_{j,k} + {}_iC_j \cdot {}_i r_{j,s_k} \\ {}_i\alpha_j &= g^{i w_{j,k}} \\ {}_i\beta_j &= ({}_i r_{j,s_k})^{-i C_j} \cdot h_A^{i w_{j,k}} \end{aligned}$$

Then, S_k sends $\langle {}_i\Gamma_{j,k}, {}_iC_j, {}_i\mathbb{R}_j \rangle$ to the next sealer that has not done the sealing process yet. The next sealer repeats the exact process until the last sealer, S_t who seals the bid and publishes $\langle {}_i\Gamma_{j,t}, {}_iC_j, {}_i\mathbb{R}_j \rangle$ on the public bulletin board. The bidder B_i checks the validity of the sealed bid as follows:

$$\begin{aligned} g^{i C_j i r_j + i R_j} &\stackrel{?}{=} {}_i\alpha_j ({}_iX_{j,t})^{i C_j} \\ h_A^{i C_j i r_j + i R_j} G^{i C_j} &\stackrel{?}{=} {}_i\beta_j ({}_iY_{j,t})^{i C_j} \end{aligned}$$

If the verification fails, the bidder can raise a complaint.

- **Open:** When the bidding phase ends, the bids are opened as per the scheduled time. All sealers compute $V = \prod_{i=1}^m {}_i r_{j,s_k, \dots, t}$ for $1 \leq j \leq n$, which is the product of all random seeds used to seal the value of bidding price and send privately to the auctioneer. After receiving V from t sealers, auctioneer opens the bid-vectors and declares the winning price:

$$\begin{aligned} \mathbb{Y}_j &= \frac{\prod_{i=1}^m {}_iY_{j,t}}{\prod_{i=1}^m {}_iX_{j,t}^{x_A}}, & \mathbb{P}_j &= \frac{\mathbb{Y}_j}{\prod_{k=1}^t V} \\ &= \left(\prod_{i,v=1}^{m,t} {}_i r_{j,s_v} \right) \cdot G, & &= G_n^{m-l} G_y^l \end{aligned}$$

$\mathbb{P}_j = G_n^{m-l} G_y^l$ for $l \geq 0$. Auctioneer declares the winning price as p_j , for the j where $l \geq 1$ appears first.

Auctioneer computes ${}_iC_j = {}_iC_j^{-H(i\Gamma_j)x_A} = {}_iX_j^{x_{B_i}}$. The verification process includes the execution of interactive zero-knowledge protocol to verify that ${}_iC_j$ and h_{B_i} have common exponent as x_{B_i} .

- B_i selects δ_j for $1 \leq j \leq n$ randomly and computes $a_j = g^{\delta_j}$, $b_j = {}_iX_j^{\delta_j}$ and sends (a_j, b_j) to the auctioneer.

- Auctioneer selects random challenges c_j for $1 \leq j \leq n$ and sends to the bidder B_i .
- B_i computes $\gamma_j = \delta_j + c_j x_{B_i}$ and replies to the auctioneer.
- Auctioneer verifies:

$$g^{\gamma_j} \stackrel{?}{=} a_j \cdot h_{B_i}^{c_j}, \quad \text{for } 1 \leq j \leq n$$

$${}_i X_j^{\gamma_j} \stackrel{?}{=} b_j \cdot {}_i C_j^{c_j}$$

4 The Known Bid Attack

In this section, we show how to mount a known bid attack on Howlader et al.'s scheme [4] by forging a valid sealed-bid. In particular, we consider a malicious first sealer who can forge a valid sealed-bid without knowing the private key of the winning bidder. Using the knowledge of system public parameters $({}_i G_j, h_{B_i}, h_A, h_{S_i})$, the malicious sealer F mounts the attack as follows:

1. The bidder B places the bid by passing the value of $\langle {}_i G_{j,t}, {}_i C_j \rangle$ to F at the beginning of bid sealing process, where G indicates G_y for bidding at price, p_j or G_n for not bidding at price, p_j .

$${}_i G_j = ({}_i X_j, {}_i Y_j) = (g^{i r_j}, h_A^{i r_j} h_S^{i r_j} G)$$

$${}_i C_j = h_A^{i r_j H({}_i G_j) x_{B_i}}$$

2. F receives the value $\langle {}_i G_j, {}_i C_j \rangle$ and starts the normal sealing operation. The value of $\langle {}_i G_j, {}_i C_j \rangle$ will be kept for imminent use.
3. Eventually, the sealing operation ends up publishing the sealed-bid on bulletin board.
4. After the **Open** algorithm, F knows the price bid by the winning bidder. If the winning bidder is B , the value of G in ${}_i G_j = (g^{i r_j}, h_A^{i r_j} h_S^{i r_j} G)$ is known and F can change the bidding price, G_y to designated position of his forged price. For illustration of identifying the bidding price, assuming that the bidding price is at p_{10} where price $P = \{p_1, p_2, p_3, \dots, p_n\}$, value for list of bidding price ${}_i G_j$ looks like below:

$${}_i G_1 = (g^{i r_1}, h_A^{i r_1} h_S^{i r_1} G_n)$$

$${}_i G_2 = (g^{i r_2}, h_A^{i r_2} h_S^{i r_2} G_n)$$

$$\vdots$$

$${}_i G_{10} = (g^{i r_{10}}, h_A^{i r_{10}} h_S^{i r_{10}} G_y)$$

$$\vdots$$

$${}_i G_n = (g^{i r_n}, h_A^{i r_n} h_S^{i r_n} G_n)$$

5. Since G (G_y and G_n) is known, F can forge a valid bid by replacing the list of bidding price ${}_i G'_j = (g^{i r_j}, (h_A h_S)^{i r_j} G')$, whereby G' is the replaced bidding

price. F continues to calculate the new commitment value ${}_iC'_j$ by removing the hash value of the previous encrypted price such that:

$$\begin{aligned} {}_iC'_j &= {}_iC_j^{\frac{H({}_i\Gamma'_j)}{H({}_i\Gamma_j)}} \\ &= h_A^{{}_i r_j H({}_i\Gamma_j) x_{B_i} \frac{H({}_i\Gamma'_j)}{H({}_i\Gamma_j)}} \\ &= h_A^{{}_i r_j H({}_i\Gamma'_j) x_{B_i}} \end{aligned}$$

6. We can see that the forged values of encrypted bid-vector $\langle {}_i\Gamma'_j, {}_iC'_j \rangle$ are of a valid bid. Thus, F can impersonate the winning bidder to place a bid of any price in the subsequent auction.

Note that in the subsequent auction, the actual bidder is not able to verify F 's bid as the previous exponent ${}_i r_j$ can be rerandomized such that ${}_i X'_j = (g^{i r_j})^{r'}$, ${}_i Y'_j = (h_A h_S)^{i r_j r'} G$ where $r' \in \mathbb{Z}_q^*$ is randomly chosen. Although the bidder cannot verify the forged bid himself, the validity of the bid can be verified by the auctioneer through the zero-knowledge protocol which shows the value of the bidder's private key is engraved in the bid.

4.1 Discussion

As stated by the authors in [4], the proposed scheme do not assumes the auctioneer is honest and they allow some sealers to be malicious, i.e., some sealers are corruptible. Thus, there exists the possibility that the first sealer is dishonest.

We term the above attack a known bid attack as upon knowing the bidding price of a sealed-bid, the malicious first sealer can always mount the known bid attack with 100% success probability to forge a valid bid of the winning bidder in the subsequent auction. In the attack demonstrated, an attack conducted by the malicious first sealer, F does not requires the help of auctioneer to perform the attack and the attack has the same success rate in the case of either the auctioneer is honest or dishonest. This shows that our known bid attack does not deviate from the security assumption in [4] and it is indeed a flaw of the ESBA scheme. The main problem is that the bidding information (bidding price and identity of bidder) is not tightly bound in the bid. In particular, the ${}_i\Gamma_j$ and ${}_iC_j$ of [4] is malleable.

We note that although the similar bid structure is applied on Howlader and Mal [5] and Howlader et al. [6], the known bid attack is not applicable in these schemes as the bid indicator, G is used differently and the commitment ${}_iC_j$ is absent from the bid. A potential fix for the attack is proposed in the next section.

5 The Fix

In this section, we present a fix towards the known bid attack. A minor change is made towards the original commitment value ${}_iC_j = h_A^{i r_j H({}_i\Gamma_j) x_{B_i}}$. The new commitment value, ${}_iC_j^*$ is as follows:

$${}_iC_j^* = h_A^{(i r_j H(i \Gamma_j) + H(I)) x_{B_i}} \text{ for } 1 \leq j \leq n$$

We denote I in the above equation as the auction item’s reference number. In practice, I can include extra information such as auction round information, bid information and timestamp. In completing this fix, a minor modification in the **Open** algorithm is made, where the auctioneer computes ${}_iC_j^*$ for the verification process of interactive zero-knowledge protocol. The detail of the change is as follows:

$$\begin{aligned} {}_iC_j^* &= (({}_iC_j^*)^{x_A})^{\frac{1}{x_A}} (h_{B_i})^{-H(I)} \frac{1}{H(i \Gamma_j)} \\ &= {}_iX_j^{x_{B_i}} \end{aligned}$$

The replacement of ${}_iC_j$ with ${}_iC_j^*$ will not affect the rest of the algorithms.

5.1 Discussion

Our fix is able to prevent the known bid attack from the malicious first sealer because the new commitment value, ${}_iC_j^*$ is able to prevent the removal of $H(i \Gamma_j)$. Our fix is efficient as only a minor tweak was made to the original scheme. Therefore, the security properties claimed in the original scheme are not affected.

However, there is a limitation on the proposed fix if we consider a stronger attack. There are two approaches in mounting the known bid attack, (i) the attack performed by the first sealer alone, and (ii) the attack performed through the collusion between the auctioneer and the first sealer. More precisely, the former known bid attack is mounted by a malicious first sealer without any help from other entity; the latter attack is a collaborated known bid attack resulted from the collusion of the first sealer and the auctioneer.

Recall that in Howlader et al.’s ESBA [4] scheme, they made two assumptions, namely, not all sealers have to be honest and the auctioneer can be malicious. However, they did not point out if collusion is allowed between the malicious sealers and the auctioneer. In the proposed fix, we are able to resolve the first sealer known bid attack but not the collaborated version.

When they collude, the malicious first sealer runs the step 1 to step 6 in the attack as in Sect. 4. In the middle of step 5, the malicious first sealer provides the hash value below to the auctioneer:

$$\begin{aligned} H(i \Gamma_j)' &= H({}_iX_j, {}_iY_j)' \\ &= H(g^{i r_j}, h_A^{i r_j} h_S^{i r_j} G')' \end{aligned}$$

The auctioneer then completes step 5 to forge a valid commitment value, ${}_iC_j^{* \prime}$ as follows:

$$\begin{aligned} {}_iC_j^{* \prime} &= {}_iC_j^{x_A H(i \Gamma_j)'} h_{B_i}^{x_A H(I)} \\ &= {}_iX_j^{x_{B_i} x_A H(i \Gamma_j)'} g^{x_{B_i} x_A H(I)} \\ &= g^{x_A x_{B_i} i r_j H(i \Gamma_j)' + x_A x_{B_i} H(I)} \\ &= g^{x_A (i r_j H(i \Gamma_j)' + H(I)) x_{B_i}} \end{aligned}$$

We reserve the full fix to overcome the collusion between the first sealer and the auctioneer as our immediate future work.

6 Conclusion

We presented a known bid attack on Howlader et al.'s scheme. We showed that a malicious first sealer of the winning bidder can always succeed in mounting the known bid attack to forge a valid bid for the subsequent auction. We proposed a partial fix for the scheme where the fixed scheme is secure as long as the auctioneer remains honest.

Acknowledgment. The authors would like to convey gratitude towards the Malaysia government's Fundamental Research Grant Scheme (FRGS/2/2014/ICT04/MMU/03/1) for supporting this work.

References

1. Benaloh, J., Tuinstra, D.: Receipt-free secret-ballot elections (extended abstract). In: Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing, pp. 544–553 (1994)
2. Chang, C.C., Chang, Y.F.: Efficient anonymous auction protocols with freewheeling bids. *Comput. Secur.* **22**(8), 728–734 (2003)
3. Franklin, M.K., Reiter, M.K.: The design and implementation of a secure auction service. *IEEE Trans. Softw. Eng.* **22**(5), 302–312 (1996)
4. Howlader, J., Ghosh, A., Pal, T.D.R.: Secure receipt-free sealed-bid electronic auction. In: Ranka, S., Aluru, S., Buyya, R., Chung, Y.-C., Dua, S., Grama, A., Gupta, S.K.S., Kumar, R., Phoha, V.V. (eds.) IC3 2009. CCIS, vol. 40, pp. 228–239. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-03547-0_22](https://doi.org/10.1007/978-3-642-03547-0_22)
5. Howlader, J., Mal, A.K.: Sealed-bid auction a cryptographic solution to bid-rigging attack in the collusive environment. *Secur. Commun. Netw.* **8**(18), 3415–3440 (2015)
6. Howlader, J., Roy, S.K., Mal, A.K.: Practical receipt-free sealed-bid auction in the coercive environment. In: Lee, H.-S., Han, D.-G. (eds.) ICISC 2013. LNCS, vol. 8565, pp. 418–434. Springer, Cham (2014). doi:[10.1007/978-3-319-12160-4_25](https://doi.org/10.1007/978-3-319-12160-4_25)
7. Hwang, M.S., Lu, E.J.L., Lin, I.C.: Adding timestamps to the secure electronic auction protocol. *Data Knowl. Eng.* **40**(2), 155–162 (2002)
8. Lee, C.C., Lin, T.H., Tsai, C.S.: Cryptanalysis of a secure and efficient authentication protocol for anonymous channel in wireless communications. *Secur. Commun. Netw.* **5**(12), 1375–1378 (2012)
9. Liaw, H.T., Juang, W.S., Lin, C.K.: An electronic online bidding auction protocol with both security and efficiency. *Appl. Math. Comput.* **174**(2), 1487–1497 (2006)
10. McAfee, R.P., McMillan, J.: Auctions and bidding. *J. Econ. Lit.* **25**(2), 699–738 (1987)
11. Montenegro, J.A., Lopez, J.: A practical solution for sealed bid and multi-currency auctions. *Comput. Secur.* **45**, 186–198 (2014)
12. Sakurai, K., Miyazaki, S.: An anonymous electronic bidding protocol based on a new convertible group signature scheme. In: Dawson, E.P., Clark, A., Boyd, C. (eds.) ACISP 2000. LNCS, vol. 1841, pp. 385–399. Springer, Heidelberg (2000). doi:[10.1007/10718964_32](https://doi.org/10.1007/10718964_32)

13. Schoenmakers, B.: A simple publicly verifiable secret sharing scheme and its application to electronic voting. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 148–164. Springer, Heidelberg (1999). doi:[10.1007/3-540-48405-1_10](https://doi.org/10.1007/3-540-48405-1_10)
14. Vickrey, W.: Counterspeculation, auctions, and competitive sealed tenders. *J. Finance* **16**(1), 8–37 (1961)
15. Viswanathan, K., Boyd, C., Dawson, E.: A three phased schema for sealed bid auction system design. In: Dawson, E.P., Clark, A., Boyd, C. (eds.) ACISP 2000. LNCS, vol. 1841, pp. 412–426. Springer, Heidelberg (2000). doi:[10.1007/10718964_34](https://doi.org/10.1007/10718964_34)
16. Wu, C.C., Chang, C.C., Lin, I.C.: New sealed-bid electronic auction with fairness, security and efficiency. *J. Comput. Sci. Technol.* **23**(2), 253–264 (2008)

Perceptual 3D Watermarking Using Mesh Saliency

Jeongho Son, Dongkyu Kim, Hak-Yeol Choi, Han-Ul Jang,
and Sunghee Choi^(✉)

School of Computing, KAIST, 291, Daehak-ro,
Yuseong-Gu, Daejeon, Republic of Korea
sunghee@kaist.edu

Abstract. In this paper, we introduce a novel blind 3D mesh watermarking method which focuses on preserving the appearance of the watermarked model. Despite the high transparency achieved by existing 3D watermarking schemes, we observe that only a small amount of geometric error can bring a significant impact to appearance of 3D models, especially in visually important regions. We integrate this human perceptual importance, called saliency, to control the distortions on surfaces. Our method enhances the imperceptibility while maintaining the efficiency of processing spatial information by conjugating spatial and spectral regions. We use the vertex norm distribution and solve the quadratic error minimization problem to insert watermark bits. Experimental results demonstrate that our method performs well for perceived visual quality and also robustness against various geometric attacks.

Keywords: 3D watermarking · Mesh saliency · Perceptual watermark

1 Introduction

One of the main concerns of 3D blind polygonal mesh watermarking is that, despite the usefulness, it would necessarily make geometric distortions on the surface of the mesh and cause damage to the model appearance, which is a very challenging problem. This issue can be seriously treated in some applications such as additive manufacturing and medical imaging, where a slight change can lead to a significant difference for a human perception system. However, only few previous works consider the visual differences between the original mesh and the watermarked mesh by estimating the perceptually-correlated distortions [9, 15].

One of the standard techniques to hide inevitable artifact is to use unnoticed areas to preserve the important regions. The importance of a point or local region is different throughout the 3D surfaces for the human eye. To ensure the imperceptibility of 3D mesh watermarking, we should preserve geometries of perceptually important region during the embedding process. Therefore, mesh saliency can provide valuable information and bring affirmative effects to the perceived quality of 3D polygonal mesh watermarking.

In this paper, we propose a novel blind 3D polygonal mesh watermarking method, which improves the visual quality of the watermarked model in consequences of integrating mesh saliency. The proposed method employ the distribution of the vertex norm histogram as a watermark embedding domain, which is firstly introduced by Cho et al. [2]. We formulate our notion of perceptual 3D mesh watermarking as an optimizing problem to minimize the saliency-weighted sum of squared distance error, and provide a quadratic programming (QP) formulation to solve the problem. Compared to other 3D watermarking techniques, our method gives good performances on preserving perceptually important areas, and also provides reasonable robustness against various attacks.

This paper is organized as follows: we present related works in Sect. 2. We formalize the embedding and extracting process in Sect. 2, including an objective to be minimized and a QP framework. Section 3 presents extensive experiments regarding perceptually based quality, and also robustness of our saliency-oriented watermarking scheme.

2 Related Work

Since Ohbuchi et al. [8] first introduced the concept of 3D polygonal mesh watermarking, there have been various approaches to achieve satisfactory results for practical applications. According to the recent benchmark [16], Cho et al. [2] and Wang et al. [15] show reasonable performances in terms of robustness. While Wang et al. [15] uses the mesh local volume moment as a watermarking domain and modifies the low-frequency components of the surface, Cho et al. uses the spatial domain that uses the distribution of the vertex norm. There are similar approaches to alter the distribution of the radial distances from a certain center point to the surface [1, 5, 9]. Although the vertex norm distribution offers a lot of computational cost saving due to its simplicity, this kind of methods would not guarantee the visual quality of the watermarked object [16].

In the past few decades, there have been made noticeable progress to measure mesh saliency by merging criteria inspired by low-level human visual cues [7, 11, 13]. The perceived quality enhancement method using mesh saliency has been successfully applied to mesh processing applications such as simplification [7, 13] and segmentation [11]. With the increasing concerns to human visual system, recent trend of the 3D mesh quality assessment pays attention not only to the amount of geometric error but also to the perceptually-based distance [4, 6, 17]. We use these concepts of human visual attention to hide distortion errors caused by watermark bits.

3 Perceptual Watermarking Scheme

In this section, we explain how the concept of saliency can be used in 3D mesh watermarking, and suggest an objective minimization problem for the robust watermark embedding.

3.1 Mesh Saliency

In Song et al. [13], the mesh saliency on a 3D mesh surface is defined as the difference from expected behavior of the log-Laplacian spectrum of the mesh in frequency domain. A frequency of a mesh can be defined as eigenvalues of the Laplacian matrix, L , as introduced by Taubin [14]. The Laplacian spectrum is defined as

$$H(f) = \{\lambda_f, 1 \leq f \leq n\}, \quad (1)$$

where λ_f is the eigenvalue of the Laplacian matrix L and n is the number of vertices of the mesh. Following from Lee et al. [7], we use the log-Laplacian spectrum

$$L(f) = \log(|H(f)|), \quad (2)$$

and the final saliency map S can be approximated from the difference of the spectrum and its local averaging norm:

$$S = |L(f) - \frac{1}{n}[1 \ 1 \ \dots \ 1]L(f)|. \quad (3)$$

Note that the saliency map S is in the spatial domain, not in the frequency domain. The resulting saliency value for a vertex v_i is the sum of elements in the i -th row of S .

3.2 Watermark Embedding Using Saliency

We adopt the standard technique that embeds the watermark bits to the distribution of vertex norm, to take advantage of the mesh saliency. The problem is defined as follows. A 3D polygonal mesh is defined as a graph (V, F) , where V consists of n vertices $\{v_1, \dots, v_n\}$, and F is the set of m triangle facets. With respect to the center of mass C , the spherical coordinates of vertices v_i are denoted by $(\rho_i, \theta_i, \phi_i)$, where ρ_i is the Euclidean distance from C to v_i .

Assuming that the watermark bits can be denoted by a sequence $W = (w_1, \dots, w_L)$, where L refers to the watermark payload and $w_i \in \{-1, +1\}$ for $1 \leq i \leq L$. As mentioned above, the distribution of $\{\rho_i\}$ is used to modify the spatial information of the vertices. The histogram h divides the vertex norm distribution into $L + 2$ bins of exactly same size Δk :

$$\Delta k = \frac{\max\{\rho_i\} - \min\{\rho_i\}}{L + 2}. \quad (4)$$

Within all bins, the vertex norms are normalized to have values among $[0, 1)$. The reason why two additional bins are exist is to discard the left and right end of the histogram, which often contain noisy inputs or outliers.

Several blind watermarking schemes found that the invariability of the center of mass must be achieved due to the causality problem. In this work, the center of mass is defined as the volume moment ratios

$$C = \left(\frac{m_{100}}{m_{000}}, \frac{m_{200}}{m_{000}}, \frac{m_{300}}{m_{000}} \right), \quad (5)$$

where m_{100} is the sum of volume weights $w(f) = \frac{1}{6} \det(v_1^{(f)}, v_2^{(f)}, v_3^{(f)})$ over all faces f in the mesh. The detail expressions can be found from Sheynin and Tuzikov [10].

The embedding process is done simultaneously by solving an optimization problem. The objective of the optimization is a weighted sum of squared distance error of each vertex norm, brought from the hidden information in histogram bins:

$$E_{wss} = s_1 \Delta \rho_1^2 + \dots + s_n \Delta \rho_n^2. \quad (6)$$

Here, the weights s_i come from the mesh saliency map S . Careful selection for these weight can preserve salient regions and directly affect visual quality. We adopt an amplification operator A with control parameter λ to set saliency weights

$$s_i = A(S(v_i), \lambda) = \begin{cases} \lambda S(v_i), & \text{if } S(v_i) \geq s_{threshold} \\ S(v_i), & \text{otherwise} \end{cases}, \quad (7)$$

where $s_{threshold}$ is a constant in the open interval $(0, 1)$. Based on our observation, $\lambda = 100$, $s_{threshold} = 0.45$ have shown reasonable results for general purposes.

Let H_{wss} be a $n \times n$ matrix whose diagonal entries are s_1, \dots, s_n and zero for the others. A QP framework is introduced to get optimal vertex norm displacements

$$\arg \min_{\Delta \rho} = \frac{1}{2} \Delta \rho^T H_{wss} \Delta \rho, \quad (8)$$

together with the constraints such as the invariability of the center of mass, the intensity of mean value modifications exceeding the watermark strength α , and the insurance that the histogram bin entities are same as before.

3.3 Watermark Extraction

The watermark extraction process is not different from existing method. Given a 3D stego mesh, the histogram of the vertex norm distribution is revisited. Then, excluding the two bins at each end, a histogram bin B_i provides the i -th bit of the embedded watermark:

$$w_i = \begin{cases} 1, & \text{if } \frac{\sum_{v \in B_i} \rho(v)}{\sum_{v \in B_i} 1} > 0.5 \\ -1, & \text{if } \frac{\sum_{v \in B_i} \rho(v)}{\sum_{v \in B_i} 1} < 0.5 \end{cases}, \quad (9)$$

Therefore, the receiver can confirm whether the given 3D mesh is watermarked or not by using the intrinsic characteristics of the mesh geometry without any additional information, making our method blind.

4 Experimental Results

We have tested our perceptual 3D watermarking using various 3D polygonal mesh models. For the parameters, we choose $L = 64$ as a capacity of the watermark, and set the watermark strength $\alpha = 0.05\%$. As commented before, $\lambda = 60$,

$s_{threshold} = 0.45$ are used, but these control parameters can be adjusted freely to suit the user’s preferences. All experiments have been performed on a PC equipped with Intel Core i7 3.40 GHz CPU with 8 GB RAM, and conducted in MATLAB. The models are courtesy of the AIM@SHAPE Shape Repository.

4.1 Human Perceptually-Based Quality

We studied performance on general 3D meshes with randomly generated watermark bits. Figures 1 and 2 depict the effects of adopting mesh saliency. Using mesh saliency, our method does not make any significant displacement in the salient regions in contrast to other methods. Moreover, the error distributions generated by our method are rather smooth throughout the surface. This effect comes from the frequency-based nature of the mesh saliency map we applied;

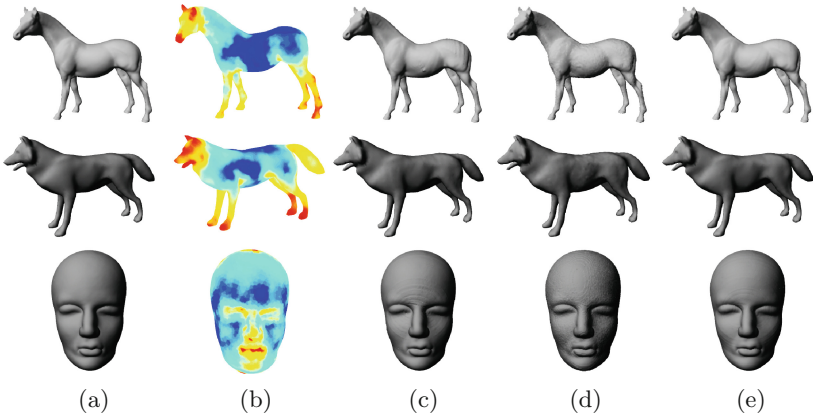


Fig. 1. Experimental results. (a) Original models; (b) saliency; (c) our results; (d) Cho et al.; (e) Rolland-Neviere et al.

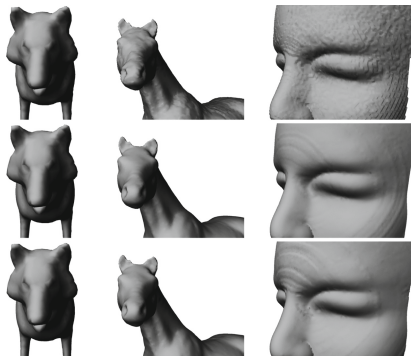


Fig. 2. Stego mesh models. Top: results from [2]. Middle: results from [9]. Bottom: results from ours.

Table 1. Evaluation of stego meshes.

Method	Model	MRMS	MSDM	FMPD
[2]	Horse	3.177×10^{-3}	0.6369	0.8820
	Mask	2.795×10^{-3}	0.8342	1.0000
	Dog	2.626×10^{-3}	0.4406	0.3195
	Kitten	3.093×10^{-3}	0.5883	0.5798
[9]	Horse	1.480×10^{-3}	0.2992	0.1911
	Mask	1.038×10^{-3}	0.5136	0.2641
	Dog	0.837×10^{-3}	0.1974	0.0900
	Kitten	1.537×10^{-3}	0.1964	0.1334
Ours	Horse	2.903×10^{-3}	0.3197	0.1982
	Mask	4.123×10^{-3}	0.5086	0.2334
	Dog	1.707×10^{-3}	0.2308	0.1062
	Kitten	3.048×10^{-3}	0.1963	0.1667

the proposed method complements disadvantages of spatial approaches with the spectral weight map.

Table 1 shows the estimations with three representative distortion metrics, compared to other methods. The Hausdorff distance is a popular metric to capture the objective geometric embedding distortion [3]. Even though the geometric error is not a main concern, the proposed method shows a reasonable result. The Mesh Structural Distortion Measure (MSDM) [6] and the Fast Mesh Perceptual Distance (FMPD) [17] are used to measure the perceptual distance. However, these two assessments cannot exactly reflect saliency; the proposed method shows reasonable quality values while providing superior saliency-preserving performance.

4.2 Robustness Against Attacks

We evaluate the robustness against additive noise, Laplacian smoothing, quantization and mesh simplification. These are common attacks for 3D meshes and often used to evaluate watermarking methods as benchmark statistics [16].

Figure 3 demonstrates the experimental results of the attack scenarios. For each test result, we use the average of bit error ratios (BER) among 100 trials of the watermark embedding and extracting processes. We observe that our proposed method shows good performance compared to the existing methods which are commonly used.

For geometric attacks, the proposed method achieves high performance results. Since the displacements caused by the embedding are relatively smooth, the proposed method obtains more robustness, especially for the smoothing attacks. However, there is a trade-off between the robustness and the amount of geometric error; users can choose the amplification parameters to control which of these two objectives is to be emphasized.

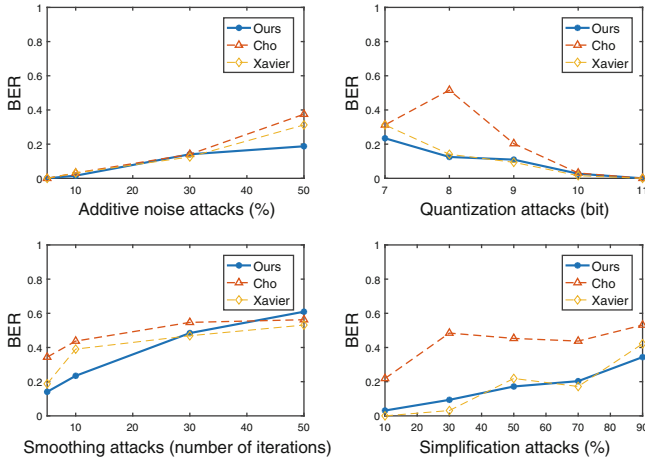


Fig. 3. Robustness against several attacks, in terms of BER.

5 Conclusion

This paper presents a novel approach for 3D polygonal mesh watermarking using mesh saliency. Unlike previous methods which focus on geometric distortions regardless of the differences of visual importance among the surface, we capture the visual appearance by preserving regions having high saliency. Combining the spatial domain-based watermarking and the frequency-based weight map, our method induce some positive effects, which result in a smooth surface and high robustness.

In future work, applying mesh saliency to frequency domain-based watermarking methods can bring interesting results. For the assessment, there is no specific metric for a perceptually-based mesh quality that reflects mesh saliency. Moreover, our method can be extended to take into account the preferable viewpoint, which produces huge impact in certain practical situations. Similar to the strategy in [12, 18], hiding artifacts in unobtrusive area can efficiently reduce the attentions to the noise.

Acknowledgments. This work was supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. R0126-16-1024, Managerial Technology Development and Digital Contents Security of 3D Printing based on Micro Licensing Technology).

References

1. Bors, A.G., Luo, M.: Optimized 3D watermarking for minimal surface distortion. *IEEE Trans. Image Process.* **22**(5), 1822–1835 (2013)
2. Cho, J.-W., Prost, R., Jung, H.-Y.: An oblivious watermarking for 3-d polygonal meshes using distribution of vertex norms. *IEEE Trans. Signal Process.* **55**(1), 142–155 (2007)

3. Cignoni, P., Rocchini, C., Scopigno, R.: Metro: measuring error on simplified surfaces. In: *Computer Graphics Forum*, vol. 17, pp. 167–174. Wiley Online Library (1998)
4. Guo, J., Vidal, V., Baskurt, A., Lavoué, G.: Evaluating the local visibility of geometric artifacts. In: *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pp. 91–98. ACM (2015)
5. Roland, H., Rondao-Alface, P., Macq, B.: Constrained optimisation of 3D polygonal mesh watermarking by quadratic programming. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1501–1504. IEEE (2009)
6. Lavoué, G., Gelasca, E.D., Dupont, F., Baskurt, A., Ebrahimi, T.: Perceptually driven 3D distance metrics with application to watermarking. In: *International Society for Optics and Photonics SPIE Optics+ Photonics*, p. 63120L (2006)
7. Lee, C.H., Varshney, A., Jacobs, D.W.: Mesh saliency. *ACM Trans. Graph. (TOG)* **24**, 659–666 (2005). ACM
8. Ohbuchi, R., Masuda, H., Aono, M.: Watermarking three-dimensional polygonal models through geometric and topological modifications. *IEEE J. Sel. Areas Commun.* **16**(4), 551–560 (1998)
9. Rolland-Neviere, X., Doërr, G., Alliez, P.: Triangle surface mesh watermarking based on a constrained optimization framework. *IEEE Trans. Inf. Forensics Secur.* **9**(9), 1491–1501 (2014)
10. Stanislav, A.: Sheynin and Alexander V Tuzikov.: explicit formulae for polyhedra moments. *Pattern Recogn. Lett.* **22**(10), 1103–1109 (2001)
11. Shilane, P., Funkhouser, T.: Distinctive regions of 3D surfaces. *ACM Trans. Graph. (TOG)* **26**(2), 7 (2007)
12. Son, J., Choi, S.: Orientation selection for printing 3D models. In: *2015 International Conference on 3D Imaging (IC3D)*, pp. 1–6. IEEE (2015)
13. Song, R., Liu, Y., Martin, R.R., Rosin, P.L.: Mesh saliency via spectral processing. *ACM Trans. Graph. (TOG)* **33**(1), 6 (2014)
14. Taubin, G.: A signal processing approach to fair surface design. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 351–358. ACM (1995)
15. Wang, K., Lavoué, G., Denis, F., Baskurt, A.: Robust and blind mesh watermarking based on volume moments. *Comput. Graph.* **35**(1), 1–19 (2011)
16. Wang, K., Lavoué, G., Denis, F., Baskurt, A., He, X.: A benchmark for 3D mesh watermarking. In: *Shape Modeling International Conference (SMI)*, pp. 231–235. IEEE (2010)
17. Wang, K., Torkhani, F., Montanvert, A.: A fast roughness-based approach to the assessment of 3D mesh visual quality. *Comput. Graph.* **36**(7), 808–818 (2012)
18. Zhang, X., Le, X., Panotopoulou, A., Whiting, E., Wang, C.C.L.: Perceptual models of preference in 3D printing direction. *ACM Trans. Graph. (TOG)* **34**(6), 215 (2015)

Perceptual Watermarking for Stereoscopic 3D Image Based on Visual Discomfort

Sang-Keun Ji, Ji-Hyeon Kang, and Heung-Kyu Lee^(✉)

School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{skji,jhkang,hklee}@mmc.kaist.ac.kr

Abstract. As 3D content including images and videos has been common and popular, the demand for copyright protection has been increased. To protect the copyright of 3D content, 3D image watermarking schemes have been proposed. Given that visual discomfort can occur in 3D content during the watermark embedding process due to binocular mismatch, unlike in 2D content, 3D watermarking schemes should consider visual discomfort because it can decrease the performance of the human vision system (HVS). In this paper, a perceptual watermarking scheme for stereoscopic 3D images considering the issue of visual discomfort is introduced. The proposed scheme analyses the factors that cause visual discomfort during the watermark embedding process. In order to minimize visual discomfort and prevent quality degradation, perceptual masking using an occluded map and a defocused map is applied. Experimental results show that the proposed scheme offers low visual discomfort while preserving the robustness against attacks.

Keywords: Stereoscopic 3D image · Digital watermarking · Perceptual embedding · Binocular mismatch · Color mismatch · Sharpness mismatch

1 Introduction

Given the advances in 3D technologies and displays, 3D content including images and videos has become common and popular. Unlike 2D content, 3D content provides depth perception to instill a feeling of reality by displaying two images with different perspectives [1]. There are two formats for 3D content distribution: DIBR (Depth-Image-Based Rendering) and S3D (Stereoscopic 3D) [2]. DIBR 3D images, consisting of a depth map and center image, generate left and right images using DIBR algorithms. Unlike DIBR, S3D images simply consists of left and right images. While DIBR was preferred over S3D due to data storage limitations and difficulties in 3D shooting techniques in the past, S3D content is more widely used and commercialized due to advances in computing technologies. S3D content has an advantage in that it can present high-definition 3D content without the quality degradation issue arising.

Due to the demand for copyright protection of 3D content, much research on stereoscopic watermarking schemes has been conducted [3]. However, not only

are most outcomes DIBR-based watermarking schemes, but also relatively few S3D-based watermarking schemes which consider the HVS and the discomfort. In one study [4], a watermarking scheme based on a visual sensitivity model for HD stereo images in the DCT domain was proposed. This scheme used a visual sensitivity model based on what is known as the just noticeable distortion (JND), which presents the maximum distortion thresholds in pixels using HVS characteristics. However, the authors of that study did not consider binocular characteristics because the JND model only considers a single image. In another work [5], a stereo image watermarking method based on the binocular just noticeable model (BJND) which consider the binocular visibility of stereo images was proposed. Because the BJND model can describe the sensitivity of the HVS to luminance changes in stereo images, unlike the JND model, this scheme embeds a watermark while considering changes between the original and watermarked images to be lower than the corresponding BJND values. However, the BJND model cannot be regarded as a completely objective measure because it was developed based on psychophysical experiments and modelling.

In this paper, a perceptual watermarking method with low visual discomfort for stereoscopic 3D images is introduced. The proposed method analyses the characteristics of S3D images and the factors that cause visual discomfort during the watermark embedding process to reduce visual discomfort and prevent quality degradation. The rest of paper is organized as follows. In Sect. 2, the background knowledges about visual discomfort and characteristics of S3D contents is handled. In Sect. 3, the proposed watermarking scheme is described. In Sect. 4, the experimental setup and results are shown and Sect. 5 concludes.

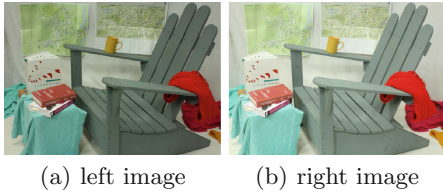
2 Background

To design a S3D watermarking scheme, we consider two issues related to visual discomfort for perceptual watermarking: visual discomfort assessment and the Depth-of-Field (DoF).

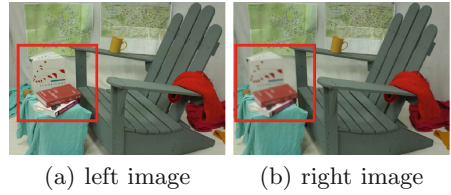
2.1 Visual Discomfort Assessment

When a viewer views 3D content through a stereoscopic display, visual discomfort that presents a perceived degree of annoyance can occur [6]. Distortion can arise during the watermarking process, as it is a type of noise addition. This distortion can be considered as binocular mismatches among visual discomfort factors due to increases in the photometry differences between the left and right images. In binocular mismatch cases, there are mainly brightness, gamma, contrast, color, and sharpness mismatches [7].

Voronov et al. proposed metrics for evaluating color mismatches and sharpness mismatches when analyzing visual discomfort of S3D contents [8]. The color mismatch metric can evaluate noticeable color differences between left and right images caused by inconsistencies in the camera settings and shooting environment. An example of the color mismatch is shown as in Fig. 1. The larger the



(a) left image (b) right image



(a) left image (b) right image

Fig. 1. Examples of color mismatch

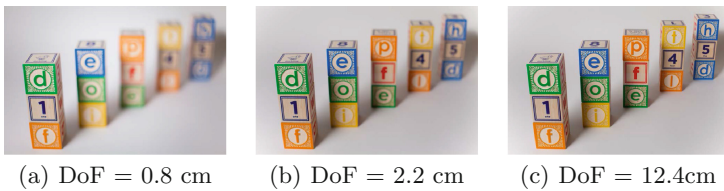
Fig. 2. Examples of sharpness mismatch

color mismatch metric is, the higher the visual discomfort becomes. The sharpness mismatch metric can evaluate differences at high frequencies caused by focus mismatches and inaccurate post-processing steps, as shown in Fig. 2. While the process is similar to that of the color mismatch metric, high-frequency information is used instead of the RGB color space. Likewise, the larger the sharpness mismatch metric is, the higher the visual discomfort becomes.

Because 3D watermarking schemes should reduce binocular mismatches caused by the watermark embedding process, the proposed method focuses on two types of mismatch, the color and sharpness mismatches, to mitigate visual discomfort.

2.2 The Depth-of-Field

The Depth-of-Field (DoF), which is called the focus range, is the distance of the focused region. The higher the DoF is, the larger the area that can be seen clearly becomes, as shown in Fig. 3. When the DoF in 3D contents increases, however, visual discomfort can occur due to the accommodation-vergence conflict [9]. In [10], the synthetic blur that decrease the DoF was activated to a defocused region for visual comfort in VR applications. Therefore, the proposed method detects defocused regions, which don't contain visually important information, and embeds a watermark into them to improve the invisibility.



(a) DoF = 0.8 cm (b) DoF = 2.2 cm (c) DoF = 12.4cm

Fig. 3. Examples with different Depth-of-Fields

3 Proposed Method

In the proposed method, there are three main steps: perceptual masking construction, watermark embedding, and watermark extraction. The process of perceptual mask construction and watermark embedding is shown in Fig. 4.

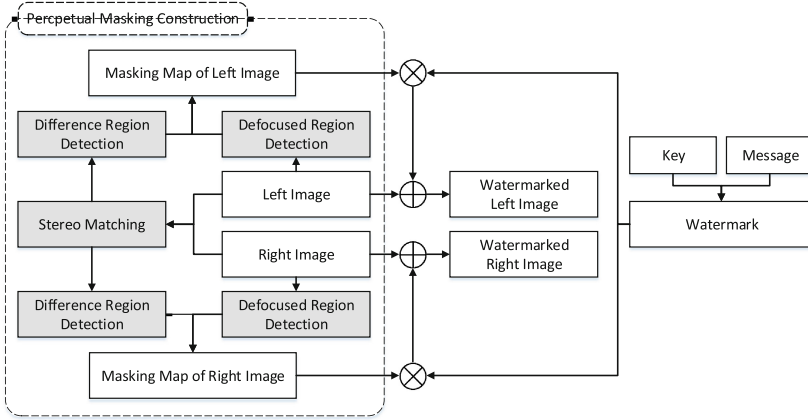


Fig. 4. The process of perceptual mask construction and watermark embedding

3.1 Perceptual Masking Construction

To minimize the increase in visual discomfort occurred by binocular mismatches during the watermark embedding process, a perceptual mask consisting of an occluded map and a defocused map is constructed. This process includes occluded region detection and defocused region detection.

Occluded Region Detection. This process finds regions that can minimize the increase in the color and sharpness mismatches. These mismatches tend to increase when distortion occurs in regions where the luminance difference is greater rather than smaller between the left and right images. This tendency can be explained by hidden pixels occurring in DIBR images [11]. In other words, a region where the luminance difference is greater can be considered as an occluded region, which is not common in left and right images but is visible to only one view. Therefore, the watermark is embedded with a high weight in regions with greater luminance differences, whereas it is embedded with low weight in regions with smaller luminance differences.

The process of occluded region detection is as follows. First, stereo matching is performed between the left and right images. For each view, the reconstructed image I' of a certain view I is reconstructed from the other view using matching information. Subsequently, the occluded map M_o of each view is computed by the following equation:

$$M_o^i(I) = \begin{cases} 1, & \text{if } D_y^i(I) > t_y \\ 0, & \text{otherwise} \end{cases}, \text{ where } D_y^i(I) = \frac{\sum^{n \times n} (Y(I_i) - Y(I'_i))^2}{n^2} \quad (1)$$

Here, I' is the reconstructed image of image I , Y is the luminance of the image, and n is the block size. D_y^i is the luminance difference of the i -th block between

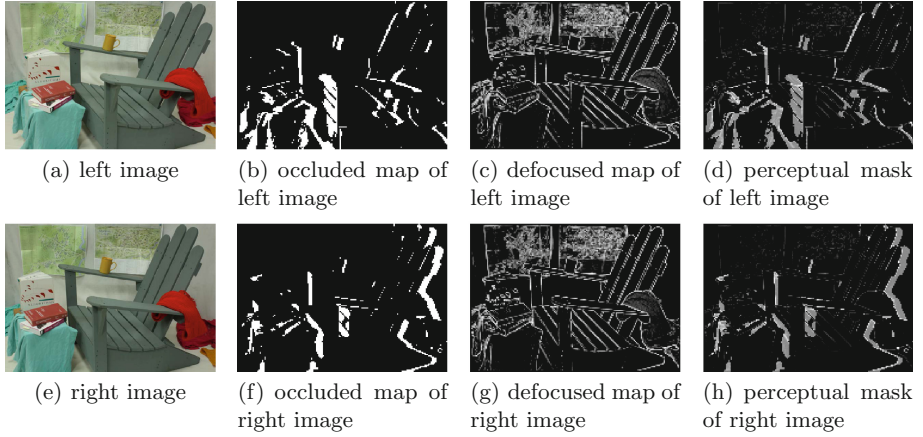


Fig. 5. Examples of perceptual masking

I and I' , and t_y is a predefined threshold. Examples of an occluded map are shown in Figs. 5(b) and (f).

Defocused Region Detection. Because regions apart from the region-of-interest (ROI) are blurred for visual comfort, as noted in Sect. 2, the watermark is embedded with a high weight for imperceptibility. To find defocused region, the differences D_b between an image and a blurred image using DoF blur is calculated by:

$$D_b^i = \frac{\sum \{Y(I_i) - b(Y(I_i))\}^2}{n^2}, \quad (2)$$

where b denotes the blur kernel and D_b^i is the luminance difference in the i -th block between the two images. Regions with low differences are blurred regions; however, they can be regions at low frequencies, such as a flat region. Thus, it is necessary to consider the color distribution d_{rgb} to remove flat regions.

$$d_{rgb}^i = \sqrt{\sum (I_r^i - \mu_r^i)^2} + \sqrt{\sum (I_g^i - \mu_g^i)^2} + \sqrt{\sum (I_b^i - \mu_b^i)^2} \quad (3)$$

Here, μ_r^i , μ_g^i , and μ_b^i are the average values of the i th block, respectively, and d_{rgb}^i is the color distribution of the i th block.

Finally, the defocused map M_d that presents the defocused region is detected by the following equation:

$$M_d = s_{min} + \frac{D_b \cdot d_{rgb} \times (s_{max} - s_{min})}{max(D_b \cdot d_{rgb}) - min(D_b \cdot d_{rgb})} \quad (4)$$

Here, s_{min} and s_{max} represent the lower and upper bounds of the defocused map. Examples of an defocused map are shown in Figs. 5(c) and (g).

3.2 Watermark Embedding and Extraction

The proposed method embeds a watermark into the DCT domain using a spread-spectrum method. The watermark embedding process is performed as follows. First, the watermark pattern W , which has a normal distribution with a zero mean and unit variance, is generated by a pseudo-random number generator using a key. After a host image is transformed to the DCT domain, mid-frequency coefficients V are selected to be watermarked. The watermark W is then embedded into V using the following equation:

$$v'_i = v_i + \alpha |v_i| w_i, \quad (0 \leq i \leq N) \quad (5)$$

In this equation, $V = \{v_1, v_2, \dots, v_n\}$ is the vector of the DCT coefficient and V' is the watermarked vector. $W = \{w_1, w_2, \dots, w_n\}$ is the vector of the watermark and α denotes the watermark strength. After embedding the watermark, the watermarked image I' is obtained using the inverse DCT. Finally, the perceptually watermarked image I_w is obtained by applying perceptual masking to I' using the following equation.

$$I_w = I \times \{1 - (\alpha_o M_o + \alpha_d M_d)\} + I' \times (\alpha_o M_o + \alpha_d M_d), \text{ where } \alpha_o + \alpha_d = 1 \quad (6)$$

Here, α_o and α_d are the weights of the occluded map and defocused map, respectively.

To extract the embedded watermark, the DCT coefficient V^* at a fixed position using the watermark embedding process is extracted from a suspected image. A normal correlation is then calculated between V^* and W . If the correlation exceeds a predefined threshold, which is experimentally determined to ensure a low false positive rate, it determined that a watermark is detected.

4 Experimental Results

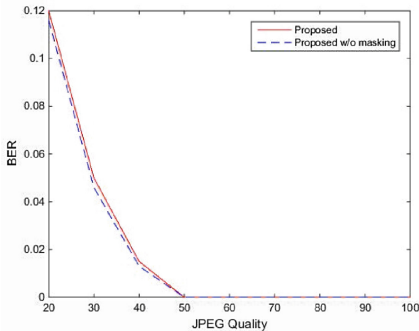
For evaluation, we used S3D images, which have the resolution of 2872×1984 , from Middlebury Stereo Datasets [13] that have been widely used for stereo evaluation. The parameter values used in the experiment are: $n = 64$, $t_y = 192$, $[s_{min}, s_{max}] = [0, 1]$. α_o and α_d are 0.5, respectively. To verify the performance, the proposed method is compared with the BJND method [5].

4.1 Visual Quality

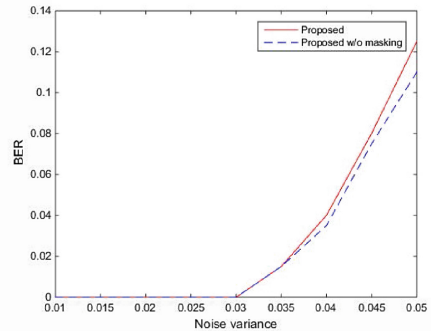
To verify visual quality, the structural similarity (SSIM) [14], color mismatch, and sharpness mismatch are evaluated by setting the PSNR to 48dB. As shown in Table 1, the proposed method show that the SSIM is higher than other methods. Compared with the BJND method, the increases in color and sharpness mismatches are reduced by 61% and 3%, respectively. To measure the effect of perceptual masking, the proposed method without masking was compared. As a result, we find that perceptual masking is effective in mitigating visual discomfort during the watermark embedding process, as the increases in color and sharpness mismatches are reduced by 66% and 65%, respectively.

Table 1. SSIM, color and sharpness mismatches with the PSNR 48 dB

		Original image	BJND method	w/o masking	Proposed method
SSIM		-	0.9861	0.9858	0.9904
Color Mismatch	value	27.6083	28.0084	28.0688	27.7631
	increase	-	(+0.4002)	(+0.4606)	(+0.1548)
Sharpness Mismatch	value	0.6255	0.6486	0.6898	0.6479
	increase	-	(+0.0232)	(+0.0644)	(+0.0224)



(a) JPEG compression



(b) Gaussian noise addition

Fig. 6. Examples with different Depth-of-Fields

4.2 Robustness

To measure the robustness, two experiments were carried out on JPEG compression and Gaussian noise addition, respectively. Experimental results show that the proposed method with perceptual masking preserves the robustness against attacks similar to that without perceptual masking. Therefore, there was hardly any decrease in the robustness due to masking.

5 Conclusion

In this paper, we proposed a perceptual watermarking for S3D image based on visual discomfort assessment. To minimize visual discomfort caused by binocular mismatch during the watermark embedding process, the proposed method applied perceptual masking, which consists of occluded map and defocused map, that exploits the characteristics of S3D content. To measure visual discomfort, we adopted the color and sharpness mismatch that evaluate binocular mismatch between left and right images. Experimental results show that the proposed method has low visual discomfort while preserving the robustness against attacks. For the future work to improve the proposed scheme, various visual discomfort assessments should be considered for evaluating quality degradation.

Acknowledgments. This research was supported by a grant from the Advanced Technology Center R&D Program funded by the Ministry of Trade, Industry & Energy of Korea (10042252) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B2009595)

References

1. Reichelt, S., Haussler, R., Futterer, G., Leister, N.: Depth cues in human visual perception and their realization in 3D displays. In: SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, p. 76900B (2010)
2. Vetro, A., Tourapis, A.M., Muller, K., Chen, T.: 3D-TV content storage and transmission. *IEEE Trans. Broadcast.* **57**(2), 384–394 (2011)
3. Chammem, A., Mitrea, M., Preteux, F.: Stereoscopic video watermarking: a comparative study. *Ann. Telecommun. Annales des Telecommunications* **68**(11–12), 673–690 (2013)
4. Niu, Y., Soudidene, W., Beghdadi, A.: A visual sensitivity model based stereo image watermarking scheme. In: European Workshop on Visual Information Processing (EUVIP), pp. 211–215 (2011)
5. Bitaghsir, S.A., Karimi, N., Azizi, S., Samavi, S.: Stereo image watermarking method based on binocular just noticeable difference. In: International ISC Conference on Information Security and Cryptology, pp. 33–38 (2014)
6. Lambooi, M., Fortuin, M., Heynderickx, I., IJsselsteijn, W.: Visual discomfort and visual fatigue of stereoscopic displays: a review. *J. Imaging Sci. Technol.* **53**(3), 30201–1 (2009)
7. Li, J., Barkowsky, M., Le Callet, P.: Visual Discomfort in 3DTV: definitions, causes, measurement, and modeling. In: Kondoz, A., Dagiuklas, T. (eds.) *Novel 3D Media Technologies*, pp. 185–209. Springer, New York (2015)
8. Voronov, A., Vatolin, D., Sumin, D., Napadovsky, V., Borisov, A.: Methodology for stereoscopic motion-picture quality assessment. In: IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, p. 864810 (2013)
9. Tam, W.J., Speranza, F., Yano, S., Shimono, K., Ono, H.: Stereoscopic 3D-TV: visual comfort. *IEEE Trans. Broadcast.* **57**(2), 335–346 (2011)
10. Hillaire, S., Lcuyer, A., Cozot, R., Casiez, G.: Depth-of-field blur effects for first-person navigation in virtual environments. In: Proceedings of the 2007 ACM Symposium on Virtual Reality Software and Technology, pp. 203–206 (2007)
11. Lee, M.J., Lee, J.W., Lee, H.K.: Perceptual watermarking for 3D stereoscopic video using depth information. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp. 81–84 (2011)
12. Barni, M., Bartolini, F., Cappellini, V., Piva, A.: A DCT-domain system for robust image watermarking. *Sig. Process.* **66**(3), 357–372 (1998)
13. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Jiang, X., Hornegger, J., Koch, R. (eds.) *GCPR 2014. LNCS*, vol. 8753, pp. 31–42. Springer, Cham (2014). doi:[10.1007/978-3-319-11752-2_3](https://doi.org/10.1007/978-3-319-11752-2_3)
14. 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)

Fingerprint Spoof Detection Using Contrast Enhancement and Convolutional Neural Networks

Han-Ul Jang, Hak-Yeol Choi, Dongkyu Kim, Jeongho Son,
and Heung-Kyu Lee^(✉)

School of Computing, KAIST, 291, Daehak-ro,
Yuseong-gu, Daejeon 34141, Republic of Korea
{hanulj,sonjh}@kaist.ac.kr, {hychoi,dkim,hklee}@mmc.kaist.ac.kr

Abstract. Recently, as biometric technology grows rapidly, the importance of fingerprint spoof detection technique is emerging. In this paper, we propose a technique to detect forged fingerprints using contrast enhancement and Convolutional Neural Networks (CNNs). The proposed method detects the fingerprint spoof by performing contrast enhancement to improve the recognition rate of the fingerprint image, judging whether the sub-block of fingerprint image is falsified through CNNs composed of 6 weight layers and totalizing the result. Our fingerprint spoof detector has a high accuracy of 99.8% on average and has high accuracy even after experimenting with one detector in all datasets.

Keywords: Biometrics · Fingerprint spoof detection · Convolutional neural networks · Multimedia security

1 Introduction

Fingerprint recognition is an automation technique that proves whether two human fingerprints match. Fingerprint recognition scans human fingerprints that have different shapes for each person in a short period of time, and then releases security or other functions if it is determined by the fingerprint of the same user [1].

However, fingerprint recognition technology has a problem of leakage of biometric information. Biometric information including fingerprints are unique information that can not be changed. If leaked once, malicious users may impersonate and threaten security [2]. The problem of leakage of biometric information is a serious security threat, and the importance of technology for verifying actual biometric information is rapidly increasing. Hence, in the fingerprint recognition system, it is essential to distinguish whether the fingerprint to be authenticated is an alive part of a person or a forged fingerprint.

The fingerprint spoof detection technique is divided into hardware and software techniques depending on whether additional sensors are used or not [3]. Among them, the software technique has a merit that it can be used in a general

fingerprint recognition device because it judges whether or not the fingerprint is falsified by using the fingerprint image. The software techniques can be classified as feature based and deep learning based. The feature based detection method is mainly a research to discriminate fingerprint liveness with a single feature point in early stage [1, 3], and it has not been shown good performance for various fake materials [4, 5].

Feature based detection with multiple features was proposed to detect various fingerprints liveness, and Dubey et al. used the Speeded-Up Robust Features (SURF), Pyramid Histogram of Oriented Gradients (PHOG), and texture features to detect fingerprint liveness [4]. Rattani et al. proposed an automatic adaptation of a liveness detector to new spoof materials [6]. They used Gray Level Co-occurrence Matrix (GLCM), Histogram of Oriented Gradients (HOG), Binary Statistical Image Features (BSIF), Local Phase Quantization (LPQ), Binary Gabor Patterns (BGP) and AdaBoost using Local Binary Patterns (LBP) for fingerprint liveness detection.

Fingerprint detection using deep learning has recently been studied. Marasco et al. researched fingerprint liveness detection using CaffeNet, GoogLeNet, and Siamese networks, and showed high robustness against various fake fingerprint materials [7]. Nogueira et al. conducted a research on forgery fingerprint detection using CNN [8]. They used CNN-Alexnet and CNN-VGG learned by using natural images and fine-tuned them to identify fingerprint forgery. Both techniques are used to determine whether a fingerprint is falsified by learning entire-size fingerprint images. Also, the preprocessing including contrast enhancement or segmentation that improves the detection performance was not applied to those techniques.

In order to improve the detection performance, the proposed method performs contrast enhancement and divides the fingerprint image into several blocks to judge whether or not each block is falsified. The result of the forgery of the blocks included in the fingerprint image is integrated into the majority voting system to determine whether the fingerprint image is falsified.

The proposed technique improves contrast by applying histogram equalization to fingerprint images. Then, each fingerprint image is divided into several non-overlapped blocks, and each falsification of the block is judged. The detection results of blocks are combined to reduce the detection errors of the fingerprint images.

Section 2 describes the proposed technique, and Sect. 3 shows the experimental results. In Sect. 4, we conclude the paper and explain future works.

2 Proposed Method

The proposed detection method is represented in Fig. 1. This technique largely proceeds to the preprocessing process and the fake fingerprint detection process. The preprocessing process involves increasing the contrast of the fingerprint image and dividing the image into multiple blocks. The fingerprint spoof detection process generates falsification detection result of fingerprint image by

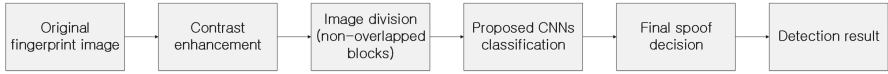


Fig. 1. Block diagram of proposed detection method

performing spoof detection in blocks of the image and totalizing the results of block spoof detection.

In this method, histogram equalization, which is a contrast enhancement technique, is performed in the preprocessing process to improve the recognition rate of the fingerprint image. Then, the fingerprint image is divided into several non-overlapped blocks, and each falsification of a block is judged. The falsification detection results of blocks are combined to reduce falsification detection errors of the fingerprint images.

2.1 Preprocessing

In the preprocessing process of fingerprint image, contrast enhancement and image division proceed. As a contrast enhancement method, histogram equalization is applied. Histogram equalization is used as a preprocessing step in many fingerprint recognition studies as a way to increase the contrast of images [9]. Histogram equalization is a process that makes the probability of the histogram of an image uniform. The probability density function of a pixel intensity level is calculated by the following equation:

$$p_r(r_k) = \frac{n_k}{n}, \quad (0 \leq p_r(r_k) \leq 1, \quad 0 \leq k \leq 255) \tag{1}$$

where r_k, k, n_k are a pixel intensity level, gray-level, and the number of pixels at r_k , respectively. The equation that generates the normalized histogram is derived by using $p_r(r_k)$ is as follows:

$$s_k = \sum_{j=0}^k \frac{n_j}{n} = \sum_{j=0}^k p_r(r_j) \tag{2}$$

where s_k is the new intensity value of level k .

In the case of a fingerprint image, the fingerprint can be divided into a ridge area that touches the scanning device and a background area where the fingerprint does not touch. The histogram equalization increases the difference between the ridge area and the background area so that the fingerprint shape can be better recognized.

The fingerprint image with enhanced contrast is divided into small non-overlapped blocks to determine whether or not each sub-block of the fingerprint image is falsified. Then, the total result of falsification detection of blocks reduces the detection errors that can occur when detecting the entire-size image at one time.

2.2 Convolutional Neural Networks

Convolutional Neural Networks are similar to general Neural Networks, except that they generate meaningful characteristics from data through convolution. The neurons of CNNs have learnable weights and biases. In addition, each neuron calculates the output by performing a dot product between inputs and weights, adding biases, and applying non-linearity. The process of calculating the output of each neuron is as follows:

$$z = g(W \cdot u + b) \tag{3}$$

where u and z are input and output, respectively. W and b are weight and bias, and are learnable parameters. $g(\cdot)$ is an activation function (or non-linearity), such as sigmoid, ReLU, or Leaky ReLU. While general Deep Neural Networks provide fully-connected layers and require a large number of weights and biases, CNNs have the advantage of using a patch-wise convolution to reduce many weights.

The proposed CNN architecture is illustrated in Fig. 2. The input of the CNNs is a block with a size of 16×16 of a gray-scale fingerprint image, and the output consists of two classes to distinguish the real fingerprint from the fake fingerprint. The proposed CNNs have 6 weight layers, consisting of 4 convolution layers and 2 fully-connected layers. Since the size of the block image used as input is 16×16 , it is designed as a 6-layer model and inspired by the architecture of Visual Geometry Group Networks (VGG-Net) [10].

The proposed technique uses Batch Normalization (BN) to improve the performance of CNNs. BN is a technique that stabilizes the training process and accelerates learning speed by reducing the internal covariate shift [11]. BN has its own regularization effect.

BN normalizes $x = W \cdot u + b$. BN calculates the mean and standard deviation for each dimension of x , normalizes it, and generates a normalized value using scale and shift factors. When normalizing x , bias b is ignored because the effect

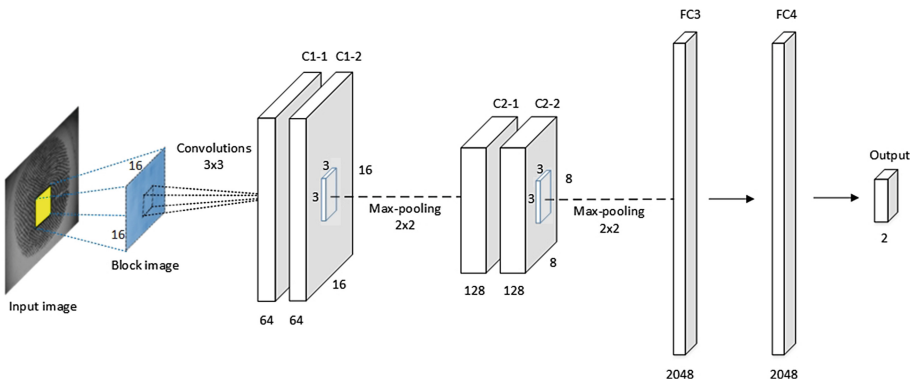


Fig. 2. Architecture of proposed CNNs

of b will be canceled by the subsequent mean subtraction. The BN transform can be expressed as:

$$z = g(\text{BN}(W \cdot u)) \quad (4)$$

where BN is BN transform. Through the CNNs with improved performance, our model is trained to detect the falsification of each sub-block and the test is conducted using the trained model.

2.3 Final Spoof Decision

The trained model is used to detect falsification for each sub-block of fingerprint images and the falsification detection results are totalized. In the proposed technique, Majority Voting System (MVS) is used to totalize the results of falsification detection of blocks. We classify the results of detection of all sub-blocks included in the fingerprint image into two classes, real and fake, and totalize the votes. The class corresponding to the majority is finally determined as the class of the fingerprint image. If the votes of the two classes are the same, then determine the fingerprint image as a fake fingerprint.

3 Experimental Results

3.1 Datasets

We used the ATVS database [2] to test the proposed fingerprint spoof detection technique. The database is contained in captured original and fake fingerprints images. In the case of fake fingerprints, the ‘with cooperative set’ consists of fingerprint images generated by using human original fingers and moldable materials. In the ‘without cooperative set’, fake fingerprints created by scanning the remaining fingerprint traces on the CD. In this experiment, we employed fingerprint images of a Biometrika FX2000 instrument, which is one of the most popular flat optical fingerprint scanners, and the examples of fingerprint images are illustrated in Fig. 3. For experimental images, 256 real fingerprints and 256 fake fingerprints that taken with Biometrika FX2000 are used.

For evaluation in fingerprint spoof detection competitions, the real/fake fingerprint ratio is set to 1:1 and equally distributed to training and testing sets. The size of the fingerprint image of the ‘with cooperative set’ is 400×560 pixels and the size of the ‘without cooperative set’ is 296×560 pixels. Since the sub-block size is set to 16×16 , each fingerprint image of the ‘with cooperative set’ has 875 sub-blocks, and the ‘without cooperative set’ has 630 sub-blocks.

3.2 Performance Metrics

Fingerprint spoof detection results were rated as the Average Classification Error (ACE). The ACE is the standard metric used for evaluation in biometrics liveness detection competitions. It is defined as

$$ACE = \frac{SFPR + SFNR}{2} \quad (5)$$

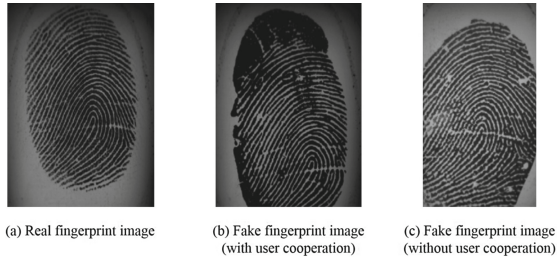


Fig. 3. Typical examples of real and fake fingerprint images that can be obtained from the public ATVS database used in the experiments. Figure extracted from [3].

Spoof False Positive Rate (SFPR) is the percentage at which the fake fingerprint is misclassified as a real fingerprint, and the Spoof False Negative Rate (SFNR) is the percentage at which the real fingerprint is misclassified as a fake fingerprint.

3.3 Implementation Details

During preprocessing, the fingerprint image is divided into non-overlapped blocks with a size of 16×16 pixels. ReLU was used for the activation function for training the model. The proposed CNNs were trained using the Caffe [12] framework, which provides very fast CPU and GPU implementations. In this paper, GPU instances were used for training and the maximum number of iteration was set to 100,000.

3.4 Results

The fingerprint spoof detection results measured for each dataset are described in Table 1. The ACE of the proposed scheme is 0.20% on average, which is much better than the existing scheme [3]. In addition, the error rate of the experiments with ‘without cooperative dataset’, which is a scanned fingerprint image, is lower than the error rate of them with ‘with cooperative dataset’.

Table 2 shows the experimental result with or without contrast enhancement. The average error rate of the detection without contrast enhancement is 1.17%, while that with contrast enhancement is 0.20%. The contrast enhancement process has relatively high performance improvement.

Table 1. Average classification error on testing datasets (%)

Dataset	Galbally’s method [3]	Proposed method
Cooperative	7.0	0.39
Non-cooperative	4.6	0
Average	5.8	0.20

Table 2. Average classification error on contrast enhancement (%)

Dataset	Without contrast enhancement	With contrast enhancement
Cooperative	1.95	0.39
Non-cooperative	0.39	0
Average	1.17	0.20

Table 3. Average classification error on one classifier using all datasets (%)

Dataset - Train	Dataset - Test	Proposed method
Both	Cooperative	0.39
Both	Non-cooperative	0
Both	Both	0.20

Table 3 shows the experimental result when one detector is generated by learning all datasets together. ‘Both’ dataset contains ‘with cooperative dataset’ and ‘without cooperative dataset’. For various test datasets, error rates are between 0%–0.39% and the experimental result shows high accuracy.

When learning the proposed CNNs, error rates of blocks with increasing number of iterations are dramatically reduced. For various test datasets, the error rates of blocks converged to around 10% after 60,000 times of iteration. The proposed method obtains an average accuracy of 99.8% by judging whether or not it is real fingerprint in block image and accumulating the results.

4 Conclusion

In this paper, we propose a technique to detect fingerprint spoof using contrast enhancement and CNNs. The proposed method uses histogram equalization as a contrast enhancement technique to improve the recognition rate of fingerprint images and detects fake fingerprints by judging whether or not the sub-block of fingerprint image is forged through CNNs. The proposed CNNs is composed of 6 weight layers and totalizing the results.

The experimental results show that the average accuracy is 99.8%. Also, the detection method with contrast enhancement has relatively high performance improvement from 98.83% to 99.8%.

In the future, we plan to use fingerprint images taken from various devices, and we plan to research highly scalable techniques with high accuracy for untrained fingerprint devices.

Acknowledgments. This work was supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No. R0126-16-1024, Managerial Technology Development and Digital Contents Security of 3D Printing based on Micro Licensing Technology), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B2009595).

References

1. Jain, A., Chen, Y., Demirkus, M.: Pores and ridges: high-resolution fingerprint matching using level 3 features. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(1), 15–27 (2007)
2. Galbally, J., Fierrez, J., Alonso-Fernandez, F., Martinez-Diaz, M.: Evaluation of direct attacks to fingerprint verification systems. *Telecommun. Syst.* **47**(3–4), 243–254 (2011)
3. Galbally, J., Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J.: A high performance fingerprint liveness detection method based on quality related features. *Future Gener. Comput. Syst.* **28**(1), 311–321 (2012)
4. Dubey, R., Goh, J., Thing, V.: Fingerprint liveness detection from single image using low level features and shape analysis. *IEEE Trans. Inf. Forensics Secur.* **6013**(c), 1 (2016)
5. Huang, Q., Chang, S., Liu, C., Niu, B., Tang, M., Zhou, Z.: An evaluation of fake fingerprint databases utilizing SVM classification. *Pattern Recogn. Lett.* **60**, 1–7 (2015)
6. Rattani, A., Ross, A.: Automatic adaptation of fingerprint liveness detector to new spoof materials. In: *IEEE International Joint Conference on Biometrics*, pp. 1–8. IEEE (2014)
7. Marasco, E., Wild, P., Cukic, B.: Robust and interoperable fingerprint spoof detection via convolutional neural networks. In: *IEEE Symposium on Technologies for Homeland Security (HST)*, 1–6. IEEE (2016)
8. Nogueira, R.F., de Alencar Lotufo, R., Machado, R.C.: Fingerprint liveness detection using convolutional networks. *IEEE Trans. Inf. Forensics Secur.* **11**(6), 1206–1213 (2016)
9. Greenberg, S., Aladjem, M., Kogan, D.: Fingerprint image enhancement using filtering techniques. *Real-Time Imaging* **8**(3), 227–236 (2002)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*, pp. 1–14(2015)
11. Ioffe, S., Szegedy, C.: Batch normalization accelerating deep network training by reducing internal covariate shift, pp. 1–11, [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
12. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe. In: *Proceedings of the ACM International Conference on Multimedia - MM 2014*, pp. 675–678. ACM Press, New York (2014)

Content Recapture Detection Based on Convolutional Neural Networks

Hak-Yeol Choi, Han-Ul Jang, Jeongho Son, Dongkyu Kim,
and Heung-Kyu Lee^(✉)

School of Computing, Korea Advanced Institute of Science and Technology,
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
{hychoi,hanulj,dkim,hklee}@mmc.kaist.ac.kr, sonjh@kaist.ac.kr

Abstract. Detecting recaptured images has been considered as an important issue. The previous techniques tried to make hand-crafted features represent the statistical characteristics of the recaptured images. Different to the existing methods, the proposed method solves the recapturing detection problem based on a deep learning technique which shows high performance for various applications in recent image processing. Specifically, we propose a recaptured image classification scheme based on a convolutional neural networks (CNNs). To our best knowledge, this is the first work of applying CNNs into the recaptured image detection. For reliable performance evaluation, we used high-quality database for training and testing. The experimental results show high performance compared to the state-of-the-art methods.

Keywords: Convolutional neural networks · Multimedia forensic · Image recapture detection · Deep learning

1 Introduction

With the rapid development of digital cameras, user can easily generate high quality images at low cost. In the same time, the resolution of the LCD screen is rapidly increasing. With the rapid development of these two technologies, it is possible to recapture a considerably high-quality image by simply taking an image from the LCD monitor with the digital camera such as a DSLR. Since the quality of recaptured image is very high, it is difficult to separate completely those two types of images with the naked eye [1].

The image recapturing could lead to two problems. First, illegal acquisition and distribution of images which protected by digital right management (DRM) technology is possible. As of now, there is no technology that can perfectly prevents capturing images from the LCD screen. Second, the attacker can disable the existing forensic technique by changing the characteristics of the images by image recapturing.

Various image forensic studies have been conducted during overdecade to detect such image acquisition. The recapturing process generates various types

of traces such as aliasing, blurriness, noise and color/luminance change. Those traces are difficult to identify with the human eye, but they leave different statistical properties in the images. Conventional forensic techniques classified both original captured images and recaptured images using the distinguishable statistical properties.

For a long time, deep learning technology has shown a great effect in the field of image recognition [2]. The deep learning framework can automatically distinguish images with different characteristics by learning the features of the image without detailed human instructions.

Researchers have created various architectures such as deep Boltzmann machines [3], deep auto-encoders [4], and convolutional neural networks [2] to train multi-layer networks. Among the various deep learning methods, CNNs have shown high performance in diverse pattern recognition problems. CNNs are expected to be more general and robust than methods based on hand-crafted feature because it learns multiple features integrally without being limited to specific features.

In this paper, we propose image recapturing detection technology based on CNNs. Since the proposed method is based on CNNs, it can distinguish the recaptured images from original captured images by collectively judging the various traces occurring in the recapturing process.

The structure of this paper is as follows. First, the various existing recapturing detection technologies are introduced in Sect. 2. In Sect. 3, we propose a novel image recapturing detection system based on CNNs. Following Sect. 4, we show the experimental setting, performing and results. In the final Sect. 5, we conclude the paper and offer future works.

2 Related Work

Here, we discuss various previous image forensic techniques that have been attempted to distinguish between original captured images and recaptured images. Farid and Lyu proposed 216 statistical features extracted from a multi-scale wavelet decomposition [5]. Those 216 features are designed to be applicable to a various forensic applications including recapturing detection. Bai et al. have modelled the specular component in the printed and LCD displayed image [6]. They classify the images using specular component through a support vector machine (SVM) classifier. Gao et al. also proposed a method of distinguishing recaptured images taken by low-resolution mobile devices [7]. In this method, they proposed a technique distinguishes the images based on the physical features including specularity, gradient, color and contrast of the recaptured images. Cao and Kot modelled the statistical features of common anomalies generated during the camera recapturing process and vectorized them to distinguish through the SVM [1].

Mahdian et al. proposed a method for detecting periodic patterns generated by a regular sampling grid on the LCD screen [8]. They employed the cyclostationarity theory to detect the periodic pattern. Thongkamwitoon et al. proposed

a detection technique using learned edge blurriness [9]. In this study, two sets of dictionaries are trained using the K-singular value decomposition from the line spread profiles. Then, the SVM is built using dictionary approximation errors and the mean edge spread width from the training images.

The common limitation of existing techniques is that existing schemes are difficult to adapt to environmental changes. For example, the emergence of new technologies in capturing or displaying devices may change the property of the recaptured images. In such a new environment, the methods must be redesigned to accommodate the changes. On the other hand, the proposed method can overcome the limitation through new learning in response to the new environment.

3 Proposed Method for Recaptured Image Detection

3.1 The Architecture of CNN

As we mentioned above, CNNs automatically learn the features of the image and use them to classify the images. CNN has a deep architecture that includes multi-level non-linear operations. It contains convolution layers, pooling layers and fully connected layers. Each convolution layers generate the output feature map that incorporates convolution result from multiple inputs and deliver them to the pooling layer. The convolution results from multiple input is transformed into element-wise non-linearity [10]. The output feature map of the convolution layer can be viewed as a specific representation of the input image. The pooling layer reduces the spatial resolution of each feature map and makes the information more global [11]. Among several pooling methods, max pooling method is known to enable fast convergence and generalization. The output feature vector passed through the convolution and pooling layer is delivered to the fully connected layer. Finally, the fully connected layer outputs the probability of the sample classified into each class by softmax connection.

3.2 Proposed CNNs Architecture

The input of the proposed CNN architecture is RGB 3 channels 64×64 image sub-block. After extracting the sub-blocks from each image, the sub-blocks are shuffled randomly to be the input with a label. Figure 1 represents the proposed CNN architecture. The proposed architecture are designed with inspiration from VGGNet [12]. The depth of the network layer is designed in three layer since it operates with small size sub-blocks. In the proposed architecture, two convolution layers and one pooling layer are repeated three times. In addition, there are three fully-connected layers at the end.

Convolution Layer : The convolution layer includes two operations, convolution and non-linearity. The convolution operation can be expressed as follows.

$$x_j^l = \sum_{i=1}^n x_i^{l-1} * \omega_{ij}^{l-1} + b_j^l \quad (1)$$

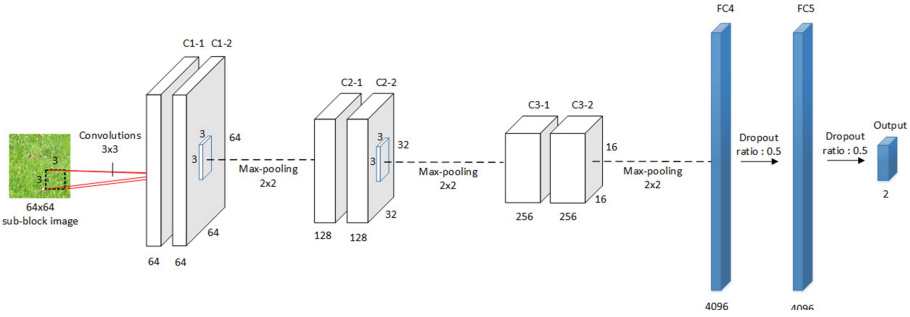


Fig. 1. The proposed CNN architecture

where $*$ is convolution operator and x_j^l is j -th output map in layer l . ω_{ij}^{l-1} is weight connecting the i -th output map in layer $l - 1$ and the j -th output map in layer l . Also, b_j^l is the bias parameter for the j -th output map in layer l . The convolution operation helps reduce the number of free variables, which increases the generalization performance of the network. In the proposed scheme, two convolution layers and one pooling layer are repeated three times. In the first set convolution layer, 64 kernels of size 3×3 are included. Thus, the output size is $64 \times 64 \times 64$ which means the number of feature map is 64 and the resolution of feature map is 64×64 . The convolution layer of the second set contains 128 kernels of size 3×3 and the last set contains 256 kernels of size 3×3 . In the convolution layer, a non-linearity operation is obtained through an element-wise non-linear activation function. In the proposed method, the Rectified Linear Units (ReLU) is used as the activation function. ReLUs are known to help grow the speed of convergence in training with large models. The ReLUs function can be expressed as follows.

$$\begin{cases} f_{m,n} = 0 & \text{for } x < x_{m,n}^l \\ f_{m,n} = x_{m,n}^l & \text{for } x \geq x_{m,n}^l \end{cases} \quad (2)$$

where (m, n) is the pixel index of feature map and $x_{m,n}^l$ is the input patch's location (m, n) of layer l . ReLUs are applied to the outputs of all convolution layers.

Pooling Layer: After obtaining the feature map from the convolution layer, the resolution of the feature map is reduced at the pooling layer. Using all the feature maps obtained from the convolution layer can cause excessive computational complexity and overfitting problem. Therefore, the pooling layer with window size 2×2 and stride 2 are included after the convolution layer of all sets.

The proposed architecture uses max pooling. It passes only max values to the next step from the local region of feature map using 2×2 window. This reduction in spatial resolution results in CNNs being able to represent a higher-level feature [11].

Fully-Connected Layer: In this layer, the learned features pass through the two fully connected layers. The features will be fed to the top layer of the CNNs

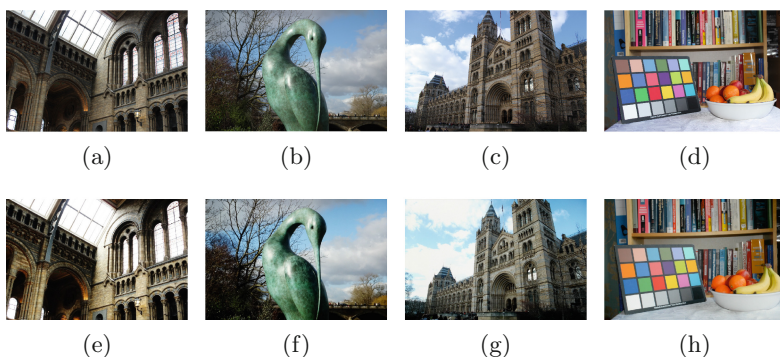


Fig. 2. The sample images used in the experiments (a)–(d): The original captured images, (e)–(h): The corresponding recaptured images

where a softmax activation function is used to the classification. After that, the weight and bias of the convolution layer and the fully connected layer are updated through a back propagation algorithm. Through this whole process, CNNs are learned and the accuracy of classification is increased. In the proposed architecture, the first two fully-connected layer contain 4096 neurons and the last fully-connected layer has 2 neurons. The final output is fed into a two-way softmax. A dropout technique is used after the first two fully-connected layer [13]. The dropout solves the overfitting problem by preventing excessive co-adaptation of hidden units by ignoring the hidden units probabilistically at the training stage. It allows the improvement of the performance since the dropout helps get more robust features.

4 Experimental Results

For objective performance evaluation, we used a public high-quality database [9]. This database contains original images taken from indoor and outdoor using 9 different cameras. In addition, the eight difference cameras are used for recapturing. In the recapturing process, the composition, light, shaking, and other variables were strictly controlled. Also, the capture distance, lens aperture and focal length are adjusted to minimize perceived aliasing. Figure 2 shows some examples used in the experiments.

We used 1,000 images for experiments. For training, we extracted 140,000 sub-blocks if size 64×64 from 500 set of images. 70,000 sub-blocks from 250 original captured images and 70,000 sub-blocks from recaptured images were used for training. Each sub-block is randomly extracted to minimize the dependency on a specific image. 500 images were used for the accuracy test. Among them, 250 of images are original captured images and 250 of images are recaptured images. All images were resized to width 2048 pixels. The images for testing are broken into 64×64 sub-block since the proposed CNN architecture determines a 64×64 sub-block is from recaptured image or not.

Table 1. The detection results of proposed method of sub-block from original captured images and recaptured images

	Accuracy
Original captured image-block	87.61%
Recaptured image-block	83.84%

Table 2. The comparison results of the detection accuracy of original captured images and recaptured images

	Original captured images	Recaptured images
Cao and Kot [1]	83.67%	92.02%
Farid and Lyu [5]	87.56%	90.04%
Thongkamwitoon et al. [9]	94.89%	99.03%
Proposed	98.00%	95.20%

We compared Cao and Kot [1], Farid and Lyu [5] and Thongkamwitoon et al. [9] for objective evaluation of performance. Those schemes are often used for the performance benchmark of the recaptured image detection. All the experiments were executed through the GPU, and the proposed CNN architecture was implemented and tested using the Caffe framework [14].

As mentioned above, the proposed method determines block-by-block an image is whether recaptured one or not. Therefore, the majority voting system (MVS) is used for the final decision through the block unit results.

Table 1 shows the block unit detection result of the proposed method. The proposed method shows higher detection rate for the original captured image sub-block.

Table 2 shows the comparison results between the proposed and existing schemes. CNN-based method mostly has higher performance than existing state-of-the-art method. For the Thongkamwitoon's method, the result for the original captured image is high and the result for the recaptured image is low. Nevertheless, as mentioned before, the proposed scheme has a relative advantage that it is easily adaptable for new environments through retraining.

Figure 3 shows false detection samples of the original captured images. The green blocks mean correctly detected regions and the remaining blocks are the false detected regions. In Fig. 3(a) and (b), it can be seen that false detection occurs in the region of blurred area due to the out of focusing and the sky area that is flat.

Figure 4 shows false detection samples of the recaptured images. In the same way, the green blocks are correctly detected region and the remaining blocks are falsely detected region. Similar to the results of original captured image, the false detection occurs in the flat sky region.

Those results are very similar with that the false alarms are occurred in the flat areas in the blur region detection studies. One of the most powerful traces

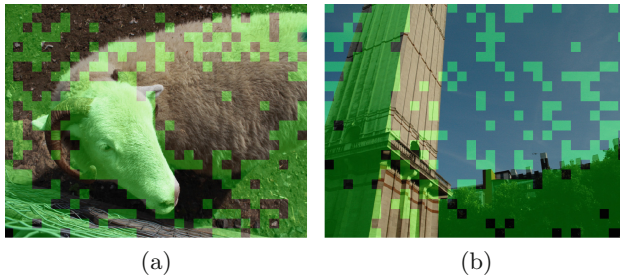


Fig. 3. The samples of false detection of the original captured images of the proposed method. (The green colored blocks are correctly judged one)



Fig. 4. The samples of false detection of the recaptured images of the proposed method. (The green colored blocks are correctly judged one)

in the recapturing process is blurring [9]. The blurred effect is clearly revealed in the edge region, but is unclear in the flat region. Therefore, it can be understood that the false detection frequently occurs in the flat regions.

5 Conclusion

In this paper, we proposed a recapture image detection based on deep learning technology. Unlike the existing recapture image detection method, the proposed method automatically learns features of recaptured images distinguished from the original captured image through well-designed CNN architecture. Because the proposed technique learns multiple features integrally, it is robust and general compared to hand-crafted feature-based method. For the experiments, we used a high-quality database to confirm reliable results. Although the proposed CNN architecture is not complicated, the performance of the proposed method is generally high. As a future work, a study to improve the false detection error in the flat area could be conducted.

Acknowledgments. This work was supported by the Institute for Information & communications Technology Promotion(IITP) grant funded by the Korean government(MSIP) (No. R0126-16-1024, Managerial Technology Development and Digital

Contents Security of 3D Printing based on Micro Licensing Technology) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2016R1A2B2009595)

References

1. Cao, H., Kot, A.C.: Identification of recaptured photographs on LCD screens. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1790–1793 (2010)
2. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2323 (1998)
3. Salakhutdinov, R., Hinton, G.: Deep boltzmann machines. In: International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 1, pp. 448–455 (2009)
4. Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P.: Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.* **1**, 1–40 (2009)
5. Farid, H., Lyu, S.: Higher-order wavelet statistics and their application to digital forensics. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 8, pp. 1–8 (2003)
6. Bai, J., Ng, T.T., Gao, X., Shi, Y.Q.: Is physics-based liveness detection truly possible with a single image? In: 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems (ISCAS), pp. 3425–3428 (2010)
7. Gao, X.T., Ng, T.T., Qiu, B., Chang, S.F.: Single-view recaptured image detection based on physics-based features. In: 2010 IEEE International Conference on Multimedia and Expo (ICME), pp. 1469–1474 (2010)
8. Mahdian, B., Novozamsky, A., Saic, S.: Identification of aliasing-based patterns in re-captured LCD screens. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 616–620 (2015)
9. Thongkamwitoon, T., Muammar, H., Dragotti, P.L.: An image recapture detection algorithm based on learning dictionaries of edge profiles. *IEEE Trans. Inf. Forensics Secur.* **10**, 953–968 (2015)
10. Liu, Y., Yao, X.: Evolutionary design of artificial neural networks with different nodes. In: Proceedings of IEEE International Conference on Evolutionary Computation, pp. 913–917 (1996)
11. Scherer, D., Müller, A., Behnke, S.: Evaluation of pooling operations in convolutional architectures for object recognition. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010. LNCS, vol. 6354, pp. 92–101. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15825-4_10](https://doi.org/10.1007/978-3-642-15825-4_10)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, pp. 1–14 (2015)
13. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors, pp. 1–18 (2012). [arXiv:1207.0580](https://arxiv.org/abs/1207.0580)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia (ACMMM), pp. 675–678 (2014)

Secret Sharing Deniable Encryption Technique

Mohsen Mohamad Hata, Fakariah Hani Mohd Ali^(✉), and Syed Ahmad Aljunid

Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia
mosen_cray04@yahoo.com, {fakariah, aljunid}@tmsk.uitm.edu.my

Abstract. Deniable Encryption was introduced to ensure that the sender and/or receiver in the communication able to create and encrypt fake messages into different ciphertexts to protect the real messages from a coercing adversary. Numerous past works have been proposed to cater the issues of coercible communication. To date, only Bi-Deniable Encryption has catered the issue of sender and receiver coercion. However there can be more than one receiver in a communication at a time. Multi-Party Computation addresses the problem of dishonest users that can be corrupted by the adversary. In some cases, Deniable Encryption may fails due to the coercer already know that it is applied in the communication protocol. A new deniability technique is needed to solve these problems. This research proposed Secret Sharing Deniable Encryption Technique. Secret Sharing technique is used to hide the secret key by creating shares and distributed among users.

Keywords: Deniable encryption · Secret sharing · Cryptography · Multi-Party Computation · Public-Key Infrastructure

1 Introduction

There are many tools, protocols and algorithms have been introduced in secure communication. It can be categorized by the type of attacks the communication system may experience or just by the flaws or errors the system may has. Encryption is one of the security elements being used in communication. It provides protection towards privacy and integrity of information.

Multi-Party Computation is an issue that can be addressed in a Public-Key Infrastructure. Although in a Public-Key Infrastructure users are registered, they may not share a common interest. There exist users with minimum trust against each other and yet wish to compute information together while still keeping their inputs private.

Relatively in a Public-Key Infrastructure network, the communication channel is secure. But in the case of coercible communication, an adversary has the power or authority to approach the users in PKI after the message has been sent. So in this case, Secure Multi-Party Computation methods that considered the communication channel is insecure are also desired in designing an incoercible communication.

Another issue in Multi-Party Computation is where; there exists an adaptive adversary that can corrupt the dishonest users. Majority of honest or dishonest users can

determine successfulness of the computation. This is important to keep the result or output of the computation remains secret among the users throughout chains of secret messages.

Deniable Encryption was introduced by Canetti et al. (1997). It is an encryption scheme that achieved deniability if the sender or receiver can create a fake message and encrypt it into a different ciphertext, thus keeping the real message private from the coercer.

It was first designed to be implemented in electronic secret voting schemes. However, its implementation also suits other applications that involve data encryption in the presence of coercing adversary. Deniable Encryption has been designed to be implemented for authentication, communication and file system.

There have been many works on Deniable Encryption in the past. The works are mainly to introduce a new deniability technique to improve past works. The next section will explain the characteristics used in designing a Deniable Encryption technique.

2 Forms of Deniability Characteristics

Deniability can be defined as the ability to deny information to someone who especially has no prior knowledge of it. Deniability characteristics are implemented in encryption algorithms so that it has the properties of deceiving and indistinguishable. Current forms of deniability characteristics are as follows:

1. Plausible. A steganography technique is a plausible form of deniability to disguise the encrypted data so that it will look like a plain data with different meanings (Amin et al. 2008). Figure 1 illustrates the example.



Fig. 1. Example of plausible deniability

2. Hidden. The data is stored in a hidden encrypted storage or volume that is not visible (Karstens 2006). Figure 2 illustrates the example.



Fig. 2. Example of hidden deniability

3. A Non-committing Encryption. Encrypting a message would derive two or more ciphertexts (Durmuth and Freeman 2011). Figure 3 illustrates the example.

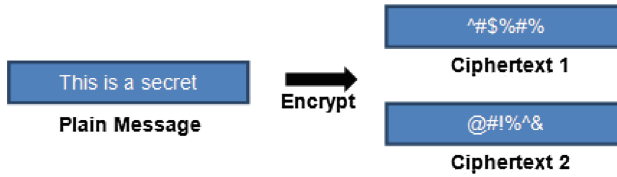


Fig. 3. Example of non-committing encryption deniability

To date, these forms have been implemented in Digital Communication, Authentication, Digital File System and many others. The implementations depend on the area and objectives of the Deniable Encryption technique. The next section will explain some of related works on Deniable Encryption techniques that have been proposed in the past.

3 Related Works

Deniable Encryption has been quite popular for the past five years. Numerous deniability techniques have been proposed in the effort of correcting and improving Canetti et al. (1997) works on Deniable Encryption. Their techniques have challenged the communication implementation in Public Key Infrastructure. They defined the basic protocol of Deniable Encryption that later being the guidelines of recent works. They also claim that their work is strong against adaptive adversary. Following works on Deniable Encryption proposed new or enhanced techniques.

A multi-layer deniability technique has been introduced by Klonowski et al. (2008). The work proposes a technique if there exist a coercing adversary that already knows Deniable Encryption scheme is used in the communication. So the coercer in this situation can demand the real message instead of the fake one. The idea of the solution is, to construct layers of fake messages. The first layer is revealed and with the demand of the coercer, reveals the second layer. Coercer will obtained fake messages, where the real message is still hidden in a deeper layer. The idea of constructing deniability in communication's bandwidth has broadened the application of Deniable Encryption in network layers. But it also brings up new issue of bandwidth management.

Ibrahim (2009a) got the motivation from past works on Deniable Encryption schemes that are inefficient in providing strong deniability notion. "A Method for Obtaining Deniable Public-Key Encryption" research proposes three versions of Deniable Encryption scheme. The schemes are constructed to achieve efficient deniability notions to single bit and multi-bit message encryption. All of the schemes rely on quadratic residuosity of a two-prime modulus. The objectives are to achieve a high level of deniability that equivalent to the factorization of a large two-prime modulus and reduce bandwidth consumption. Later works has also managed to improve his works due to new approach of bandwidth management.

Sender-side Deniable Encryption schemes have been proposed by many from past works. Where sender-side coercion is a popular issue, receiver-side coercion is also important. Ibrahim (2009b) in "Receiver-Deniable Public-Key Encryption" highlighted the importance of receiver-side coercion. A sender does not have the information to

decrypt an encrypted message. This is because, the decryption key is held by the receiver. In this research, proposed Deniable Encryption technique uses mediated-RSA for Public-Key Infrastructure. The mediated-RSA has the property of fast key revocation. So by this, the receiver will not have the full encryption information. Thus, the coercer will not gain the information to decrypt the encrypted messages. The Receiver-Side Deniable Encryption approach can be debatable as in the claim of the effect of Sender-Side coercion is more critical.

The research by Howlader and Basu (2009) has been concentrated on the practical implementation in communication. An important property of deniability is that it must be deceiving to the coercer. In a communication between the sender and receiver, the messages being exchanged must have a particular context or subject. For this reason, this research proposed a strong notion of deniability where the fake messages have the same context with real messages. By this, the fake messages will be indistinguishable and deceivable to the coercer. This research has introduced new deniability characteristic in the effort to strengthened Deniable Encryption technique.

Past works of Deniable Encryption schemes have been consider inefficient and impractical. By this motivation, Meng and Wang (2009) proposed a new Deniable Encryption scheme to cater the problem for practicality in large scales networks. The idea expands from the solution of Klonowski et al. (2008) and includes BCP (Bresson, Catalano and Pointcheval) commitment scheme. A BCP cryptosystem has been used in public-key encryption. It consists of two different decryption procedures based on two different trapdoors. This works however only focuses on the network scales view of implementation rather than introducing new technique of deniability.

Bi-Deniable is a term used by O'Neill and Peikert (2010) that referred to both sender and receiver are coercible simultaneously. In other words, both sender and receiver participate in constructing the fake messages. This Deniable Encryption scheme involves a plan-ahead technique where the numbers of alternative fake messages are decided prior to the communication. It is a true public-key scheme that is non-interactive and has no involvement of third parties. A lattice-based Trapdoor Functions and Identity-Based encryption are used to construct the deniability scheme. The technique proposed on Trapdoor Function has always been a crucial criteria in implementing Deniable Encryption.

O'Neill et al. (2011) highlighted the issue of past works on Deniable Encryption scheme that only achieved limited forms of deniability. Through this research, a Bi-Deniable Encryption scheme is proposed, where both sender and receiver are coercible. The scheme is based on simulatable encryption and lattice-based. It is applicable for a multi-distributional model and is immediately a non-committing encryption.

Durmuth and Freeman (2011) proposed a new scheme that takes up the issue of negligible detection probability in Deniable Encryption. In this scheme, samplable (sample-able) public key bit encryption is introduced where a non-committing encryption is generated while still maintaining the semantic security's objective.

In this research, Multi-Party Computation solution becomes an objective for introducing a new Deniable Encryption technique. The next section will shows how the original work of Secret Sharing technique is used in Deniable Encryption.

4 Secret Sharing Scheme

Secret Sharing Scheme has been introduced by Adi Shamir in (1979). It was proposed to cater the issue in a Multi-Party Computation for Symmetric Key encryption. Adi Shamir has designed a new key management scheme for a group of users to share secrets. The proposed key management scheme is a shared system.

The scheme uses a threshold (k, n) . Where k is the number of desired shares (threshold) to recover the secret and n is the total of users/shares. For example total of 5 users are sharing a key to an encrypted message. The threshold limit that has been agreed is 3 out of 5 shares needed to recover the key.

For this situation the threshold will be $k = 3$ and $n = 5$. By using an interpolation formula, 5 data points will be generated based on the key. These data points are the shares that will be distributed among 5 users. In retrieving the key, at least 3 data points are needed. To date, numerous of recent works have been done enhancing the technique of Secret Sharing in modern Network Communication.

5 Secret Sharing Deniable Encryption

In this research, Secret Sharing will be implemented as a form of deniability. The main objective of this proposed Deniable Encryption technique is to enhance the method of securing the real key by creating shares of fake keys. Even if some users are compromised the real key will remain hidden.

A threshold of (k, n) is defined prior the communication takes place. Where n is the total of users and k is the agreed number (threshold) of users. So number of users that is equal or more than k would reveal the secret key. The coercer may approach more than two users (a sender and multiple receivers) and force them into revealing the encryption and decryption keys. Corrupted users (less than k) will not reveal the real keys as long as it does not satisfy the threshold k .

Secret Sharing technique had only been applied in a symmetric key system. But in current environment, most transactions or communications are implemented in a Public-Key Infrastructure. To apply Secret Sharing technique in the Public-Key Infrastructure, a new method is proposed.

Secret Sharing properties will be used as a Deniable Encryption technique. Unlike the previous implementation of Secret Sharing scheme, shares of the Secret (decryption) key will be generated as pairs of Public and Private Keys. An interpolation process is needed in to generate the threshold shares.

In this work, two main processes have been identified to create the Secret Sharing Deniable Encryption technique. The first process is Fake Keys Generation. LaGrange Polynomial is used for interpolation process. Below is the process to generate the fake keys (Fig. 4).

1. LaGrange Polynomial Interpolation formula:

$$f(x) = a_0x^0 + a_1x^1 + a_2x^2 + \dots + a_{k-1}x^{k-1}$$

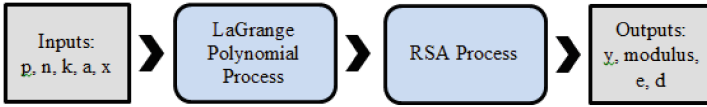


Fig. 4. Fake keys generation process

2. The Private (decryption) Key of the sender is used in the interpolation formula.
3. Based on the total users (n), threshold limit (k) will be $0 < k < n$.
4. A large random prime number is used for every n (total users/shares) to generate the shares of data points (x, y).
5. To generate fake keys from data points, RSA key generation formula will be used:

$$n = p * q$$

$$\Phi(modulo) = (p - 1) * (q - 1)$$

Find e (Public Key), such that:

$$1 < e < \Phi(modulo)$$

Find d (Private Key), such that:

$$(d * e) \% \Phi(modulo) = 1$$

6. Using the data points, the x value will replace the value of p and q will have the next random prime number of x .
7. This RSA key generation process is repeated to create all fake keys for each user.

The second process is Real Key Recovery. The process involves reverse RSA and LaGrange Basis Polynomials. Following is the process of recovering the real key.

1. Gathering up k shares (fake Private Keys (d), fake Public Keys (e), fake Modulus N ($modulo$) and y values).
2. Find p and q values, from reverse RSA formulas:

$$(d * e) \% \Phi(modulo) = 1$$

3. $(d * e)$ is congruent to $1 \% \Phi(modulo)$. So $(d * e)$ is near to the value of multiple of Φ ; and Φ has the value near to value of $modulo$. From here we can summarize $(e.d)/modulo$ is 1 is less than a factor r . $r = 1 + (e.d)/modulo$ and $\Phi = (e * d - 1)/r$.
4. From $modulo = p * q$

$$\begin{aligned} \Phi(modulo) &= (p - 1) * (q - 1) \\ &= (p * q) + 1 - (p + q) \end{aligned}$$

$p * q = \text{modulo}$ So,

$$\Phi(\text{modulo}) = (\text{modulo} + 1) - (p + q)$$

Substitute $p + q$ with sum
 where $sum = \text{modulo} + 1 - \Phi$

- At this point, use quadratic formula to solve p and q
 $delta = sum^2 - 4 * \text{modulo}$ So,

$$p = (sum + delta) / 2$$

$$q = (sum - delta) / 2$$

- With the value $p = x$ recovered and we already know the y , we can compute the points (x, y) for all the k -numbers of shares (ℓ) with the LaGrange Basis Polynomials formula:

$$f(x) = \sum_{j=0}^2 y_j \cdot \ell_j(x)$$

- The result will be the real decryption (Private, Pr) key.

The threshold limit (k) of a Secret Sharing must be well chosen. If too small value of k threshold, the shares would be easily compromise. And if too big value of k threshold, it will be hard to gather the required shares. The criteria in choosing the value of k threshold are highly depending on number of users in the group.

In coercion, the adversary will force the sender and receiver to give up the public (encryption) key and private (decryption) key. For a normal case, the sender and receiver will give the adversary fake keys. But if the adversary already knows that Deniable Encryption is applied, sender and/or receiver can be eventually corrupted. In this case, the information of the Deniable Encryption technique has been revealed.

But since a threshold has been chosen wisely, it is hard for the adversary to gather all the required k number of shares from the other users in the group. This is because the location of the other users of the group is not accessible by the adversary and among the users as well.

6 Implementation

There are twenty users in a group of Public Key Infrastructure network. Five top level users wishes to communicate in secret that can only be known only to them. Acknowledging that even in a Public Key system the message could be accessed by someone with higher authority they will communicate using Deniable Encryption technique. The detail process is as follow:

- The sender (User1) encrypts a secret message using his Public Key (PbU1) and Modulus N (NU1) to be sent to four users (User2, User3, User4 and User5).

PbU1: 5

NU1:

780802694301684407224933535432360665080820616014861144529180248765
 003674980349760241687981112323625747209375567279180847921533445493
 4752733394992500492631.

2. The Certificate Authority will act as a dealer to generate fake keys. The sender (User1) will give his Private Key (PrU1) to the Certificate Authority and create threshold (3, 5) of fake keys for five users.

PrU1:

123210777206737628899734141729442660323282464059444578116720995060
 014699921327948265066053338498015523793744825313369591885317070231
 451868849348311747725.

3. The Certificate Authority will use two formulas to create the Fake Keys:
 - a. LaGrange Polynomial Interpolation
 From the LaGrange formula:

$$f(x) = a_0x^0 + a_1x^1 + a_2x^2 + \dots + a_{k-1}x^{k-1}$$

The formula will be run 5 (n) times to generate data points (x and y values) for each user. The value of k will have the value of 3. PrU1 will be inserted at a_0 while the other a_n will be random numbers. The x values will be a large prime number randomly generated and unique for each user. All the shares will be mod by a LaGrange Prime value ($Lp \in \mathbb{P}: Lp > S, p > n; Lp =$ LaGrange Prime, $S =$ Secret, $n =$ Total Users). This is done so that the shares will be in Finite Field Arithmetic values.

- b. RSA Key Generation

To generate the fake Public and Private Keys for users, RSA key generation algorithm is used. From the base formula $n = p * q$, two large prime numbers are needed to get the value of Modulus $N(n)$. From the previous value of LaGrange formula, x will be used as p and q will be the next prime of x . Next step is to solve:

$$\Phi(modulo) = (p - 1) * (q - 1)$$

Compute Public Key (e):

$$1 < e < \Phi(modulo);$$

$$gcd(e, \Phi(modulo)) = 1$$

Compute Private Key (d):

$$d = e^{-1} mod(\Phi(modulo));$$

$$(d * e) \% \Phi(\text{modulo}) = 1$$

4. LaGrange Prime, Y values, Fake Modulus N, Fake Public Key and Fake Private Key (FNU1, FPbU1, FPrU1; FNU2, FPbU2, FPrU2; FNU3, FPbU3, FPrU3; FNU4, FPbU4, FPrU4 and FNU5, FPbU5, FPrU5) will be distributed to each user including the sender by the Certificate Authority.
5. The sender (User1) will use Fake Modulus N and Fake Public Keys of the receivers (FNU2, FPbU2; FNU3, FPbU3; FNU4, FPbU4 and FNU5, FPbU5) to encrypt a fake message. Both encrypted secret and fake messages will be sent to the receivers (User2, User3, User4 and User5).
6. In case of coercion, the users will hand in the fake keys and fake message. Even if either of them gets compromised, the coercer will have difficulties to find out the threshold of shared keys and gathering them.
7. The receivers will have to gather at least 3 (k) shares (Fake Modulus N, Public Keys and Private Keys) from other user to recover the real key (PRU1) and decrypt the secret message.

7 Research Analysis and Findings

Secret Sharing scheme has been successfully implemented in the past. In this research, a few findings have been identified from the implementation of the scheme for Deniable Encryption. Determining the values of threshold (k , n) are crucial in designing the technique in assuring the aspects of Correctness and Privacy of Secret Sharing scheme.

For generating RSA 512 bit fake keys, the p and q are dependable with the value of x ($p = x$) from the LaGrange formula. Higher total of users (n) will create a large value since the x value is already 256 bit. Therefore, in this research the number of users are limited up to 15. This has been done in optimizing the time of shares generation. If more than 15 users, the shares generation process might be exposed to a side-channel attack.

The threshold value of k has been widely discussed in the past researches of Secret Sharing scheme. If the value of k is too large, the process of gathering of shares will be tedious since the shares might not be in near distance from each other. If the value of k is too small, then the shares are easily compromised by the attacker or unintentionally among the users themselves.

From the polynomial process, the value of k has an exponential effect. Since $k < n$, the value of k is dependable to n . So, from this research a recommendation has been made in choosing the value of k . The threshold value of k is chosen to be 60% from the total users (n). This means that 60% of n will be the value of k (i.e. 60% of 15 users = 9). This recommendation would satisfy the problem of the value k too large or too small.

8 Conclusions

Deniable Encryption has been quite popular for the past five years. Canetti's (1997) contribution on proposed solution of Deniable Encryption gets a lot of critics. Recent

works on Deniable Encryption have introduced many forms of deniability characteristics. These forms are used to create strong Deniable Encryption techniques. Depending on the objective of the technique, many past works achieved incoercible communication. However, Multi-Party Computation issue can still occurred in either of the techniques proposed.

Secret Sharing Deniable Encryption Technique proposed in this works will provide a new stronger solution in achieving incoercible communication. The Secret Sharing technique with the threshold property will ensure that the adversary cannot retrieve the real key and secret message. Corrupted sender and/or receiver will not be able to satisfy the threshold and thus cannot provide the adversary enough information to gather the shares.

Acknowledgment. We would like to thank UiTM Shah Alam, Malaysia for their contributions in this research. This research was supported by the Research Management Institute, Universiti Teknologi MARA and registered under the LESTARI #600-IRMI/DANA 5/3/LESTARI(0107/2016).

References

- Shamir, A.: How to share a secret. *Mag. Commun. ACM* **22**(11), 612–613 (1979)
- Canetti, R., Friege, U., Goldreich, O., Naor, M.: Adaptively secure multi-party computation. Technical Report (1996)
- Canetti, R., Dwork, C., Naor, M., Ostrovsky, R.: Deniable Encryption. In: Kaliski, B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 90–104. Springer, Heidelberg (1997). doi:[10.1007/BFb0052229](https://doi.org/10.1007/BFb0052229)
- Kerner, M.: Uncoercible communication and deniable encryption, or how to lie with impunity. CSE P 590TU Final Projects (2006)
- Ao, J., Liao, G., Ma, C.: A novel non-interactive verifiable secret sharing scheme. In: International Conference on Communication Technology ICCT 2006 (2006)
- Klonowski, M., Kubiak, P., Kutylowski, M.: Practical deniable encryption. In: Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M. (eds.) SOFSEM 2008. LNCS, vol. 4910, pp. 599–609. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-77566-9_52](https://doi.org/10.1007/978-3-540-77566-9_52)
- Howlader, J., Basu, S.: Sender-side public key deniable encryption scheme. In: ARTCOM 2009 Proceedings of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing (2009)
- Ibrahim, M.H.: A method for obtaining deniable public-key encryption. *Int. J. Netw. Secur.* **8**(1), 1–9 (2009a)
- Ibrahim, M.H.: Receiver-deniable public-key encryption. *Int. J. Netw. Secur.* **8**(2), 159–165 (2009b)
- Meng, B., Wang, J.Q.: A receiver deniable encryption scheme. In: International Symposium on Information Processing (ISIP 2009) (2009)
- Choi, S.G., Dachman-Soled, D., Malkin, T., Wee, H.: Improved non-committing encryption with applications to adaptively secure protocols. In: Matsui, M. (ed.) ASIACRYPT 2009. LNCS, vol. 5912, pp. 287–302. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-10366-7_17](https://doi.org/10.1007/978-3-642-10366-7_17)
- Meng, B., Wang, J.Q.: An efficient receiver deniable encryption scheme and its applications. *J. Netw.* **5**(6), 683 (2010)
- O'Neill, A., Peikert, C.: Bi-deniable encryption. Eurocrypt 2010 Rump Session (2010)

- O'Neill, A., Peikert, C., Waters, B.: Bi-deniable public-key encryption. In: Advances in Cryptology - CRYPTO 2011 – 31st Annual Cryptology Conference (2011)
- Ayyasamy, R., Subramani, P.: An enhanced distributed certificate authority scheme for authentication in mobile ad-hoc networks. *Int. Arab J. Inf. Technol.* **9**(3) (2012)
- Gao, C.-Z., Xie, D., Li, J.: Deniably information-hiding encryptions secure against adaptive chosen ciphertext attack. In: 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems (2012)

Improved 3D Mesh Steganalysis Using Homogeneous Kernel Map

Dongkyu Kim, Han-Ul Jang, Hak-Yeol Choi, Jeongho Son, In-Jae Yu,
and Heung-Kyu Lee^(✉)

School of Computing, KAIST,
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
{dkim,hanulj,hychoi,ijyu}@mmc.kaist.ac.kr,
{longman,heunglee}@kaist.ac.kr

Abstract. Steganalysis targets to detect the existence of hidden information in a given content. In this paper we propose to use a local feature set which is designed to enhance discrimination of features obtained from a cover and a stego mesh. The proposed feature captures the fine deformation of the 3D mesh surface induced by a steganography or watermarking method. In our 3D steganalysis approach, in addition, we apply the homogeneous kernel map to the local feature set, which make it possible to bring much more discrimination via non-linear mapping. The proposed feature set and its combination with the homogeneous feature map have shown good performance on two different steganography and watermarking algorithm with a well known and widely used 3D mesh database through repeated experiments.

Keywords: Steganalysis · 3D mesh · Homogeneous kernel map

1 Introduction

Steganalysis is a technique used to identify whether a given content holds a hidden message such as a copyright, secret information, etc. So far, many steganalysis techniques have been studied along with steganography and watermarking schemes for different types of content.

Steganalysis has importance in two aspects. First, we can employ it to improve a steganography or watermarking method i.e., the algorithm can be refined to make it more difficult to identify whether a certain content has hidden messages or not. Next, it is important in terms of security as well. Some government agencies such as the National Security Agency can monitor and censor a communication which is suspected of being steganographically processed by people, who have malicious intention, such as terrorist organization [1].

As the 3D printer market has grown recently, the amount of 3D content has increased rapidly on the Internet, and the importance of information hiding techniques on 3D model has been growing as well. Ohbuchi et al. [2] proposed a watermarking scheme that employed a spread-spectrum approach to modify the

sequence of values obtained using mesh spectral analysis. Cho et al. [3] suggested a watermarking algorithm based on a histogram of vertex norm. They modified the distribution of vertex norms to embed a watermark in host 3D meshes. As a steganography scheme, meanwhile, Chao et al. [4] proposed a high capacity embedding technique based on projection of vertex onto the principal axis and using multi-layer concept.

To the best of our knowledge, Yang and Ivrišimtzis [5] proposed the first 3D mesh steganalysis using a feature set which consists of various statistics for position of vertices, dihedral angle between adjacent faces, face normal, etc. In [6, 7], a steganalytic algorithm was designed for a specific watermarking technique [3], which can estimate the number of bins and perceive the presence of secret messages. Recently, Li and Bors [8] suggested using additional local features e.g., vertex normal, Gaussian curvature and curvature ratio combining with [5]. For discriminative 3D mesh steganalysis, in this paper, we propose to use some further features such as edge normal, mean and total curvature which make up for other features.

On the other hand, in image steganalysis, Boroumand and Fridrich [9] investigated how to boost the performance of the existing steganalytic methods by transforming their features with explicit feature maps. However, the feature mapping called Nyström’s approximation [10] used by them is data-dependent and requires an extra training process to find the feature map. Those properties make the approach not available in the extant 3D mesh steganalytic algorithms because of their difficulties in securing data. To overcome them, we propose to use the homogeneous kernel map [11] which is a data-independent approximation, while not requiring an additional learning.

The rest of the paper is organized as follows. In the next section, we explain the feature extraction process in 3D mesh steganalysis. Section 3 describes the homogeneous kernel map and how to combine it with the feature set. In Sect. 4, the experimental results show the effectiveness of our proposed method. Finally, the conclusion is drawn in Sect. 5

2 3D Mesh Statistical Feature Extraction

We describe 3D mesh statistical features used to identify the changes induced by information hiding and how to extract them in brief. All the features introduced here follows mesh normalization and calibration process.

The mesh normalization step makes a 3D mesh fit into unit cube centered at $(0.5, 0.5, 0.5)$. Next, in the calibration process, we generate the reference mesh \mathcal{M}' of the normalized mesh \mathcal{M} by applying some smoothing technique on \mathcal{M} . All features are extracted from the differences between \mathcal{M} and \mathcal{M}' . We assume that the difference of the stego mesh is unlike that of the cover mesh. We expect that both smoothed versions of the cover and stego meshes should be almost same, and provide the same and fair reference. In addition, the alterations induced by information embedding take the form of noise. The difference between the cover mesh and its smoothed version is smaller than the difference between the stego mesh and its correspondent.

Now, we extract the features with pairs of each mesh and its corresponding reference obtained in the manner mentioned above. We use mean and total curvature, and edge normal as additional features with the 13 features proposed by [5, 8]. The absolute differences between x , y , and z -coordinates of vertices of \mathcal{M} and \mathcal{M}' for the Cartesian coordinate system are calculated to yield feature vectors $\mathbf{f}_1, \mathbf{f}_2$ and \mathbf{f}_3 whose dimension is equal to the number of vertices. In a similar way, feature vectors $\mathbf{f}_4, \mathbf{f}_5$ and \mathbf{f}_6 are generated for the Laplacian coordinate system. In addition, we regard the absolute differences in terms of ℓ^2 -norm of the vertices on the Cartesian and Laplacian coordinate system as \mathbf{f}_7 and \mathbf{f}_8 , respectively. A feature vector \mathbf{f}_9 is generated using dihedral angles between adjacent faces. The angles between the corresponding face and vertex normals are considered as \mathbf{f}_{10} and \mathbf{f}_{11} . We also take into account Gaussian curvature and curvature ratio as \mathbf{f}_{12} and \mathbf{f}_{13} , respectively.

The additional features which we consider are edge normal, mean curvature, and total curvature [12]. An edge normal is the weighted sum of the normals of the faces which have the edge and is defined as follows:

$$\vec{n}_{e(i)} = \sum_{f \in F_{e(i)}} \text{Area}(f) \cdot \vec{n}_f \quad (1)$$

where $F_{e(i)}$ represents the faces containing edge $e(i)$. $\text{Area}(f)$ and \vec{n}_f denote the area and the face normal for the face f , respectively.

In the same way as \mathbf{f}_{12} and \mathbf{f}_{13} , the fourteenth feature is generated by calculating the angle between edge normals of the edge $e(i)$ of \mathcal{M} and its corresponding edge $e'(i)$ of \mathcal{M}' .

$$\mathbf{f}_{14}(i) = \arccos \frac{\langle \vec{n}_{e(i)}, \vec{n}_{e'(i)} \rangle}{\|\vec{n}_{e(i)}\| \cdot \|\vec{n}_{e'(i)}\|} \in [0, \pi], \quad (2)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\|\cdot\|$ is ℓ^2 -norm.

In addition, we use other curvature information, mean curvature and total curvature [12], to complement the aforementioned curvature information and enhance discrimination of features.

$$H(v(i)) = \frac{\kappa_1(v(i)) + \kappa_2(v(i))}{2} \quad (3)$$

$$T(v(i)) = |\kappa_1(v(i))| + |\kappa_2(v(i))| \quad (4)$$

where $\kappa_1(v(i))$ and $\kappa_2(v(i))$ denote principal curvatures called the maximum and minimum curvature, respectively. $v(i)$ represents i -th vertex of a mesh.

$$\mathbf{f}_{15}(i) = |H(v(i)) - H(v'(i))| \quad (5)$$

$$\mathbf{f}_{16}(i) = |T(v(i)) - T(v'(i))| \quad (6)$$

where $v(i)$ and $v'(i)$ are i -th vertex of \mathcal{M} and \mathcal{M}' , respectively.

Since a 3D mesh model does not have a tractable unit corresponding to a pixel in an image, the feature vectors of different dimensions are obtained from

each model and thus they cannot be used directly. Instead, we use statistics such as mean, variance, skewness, and kurtosis for the features. Using these statistics is very natural because it is common and shows good performance in image steganalysis [13]. Before the moments are calculated, like [5, 8], we apply logarithm transformation to all the feature vectors, $\log(\mathbf{f}_i + \epsilon)$ where \mathbf{f}_i consists of positive or zero values and thus a small number ϵ is added to avoid taking the logarithm of zero values. The log function is commonly used to input data which contains values with large and small order-of-magnitude, and it can be replaced with other power functions such as square root. By calculating the four statistical moments for each feature vector \mathbf{f}_i and stacking all of them, we finally have a 64-dimensional local feature set, called LFS64.

3 Feature Mapping Using Homogeneous Kernel Map

The features we created in Sect. 2 may not be linearly separable because they are not directly related to the embedding domain of watermarking or steganography, which can make many classifiers used in steganalysis poorly work. Although the FLD ensemble [14], which has been widely used in recent years, has non-linearity, it is known to have almost the same performance as linear classifiers [15].

Therefore, before training a detector, we apply explicit maps to the features instead of using a kernelized classifier considered as an implicit approach. In our 3D steganalytic algorithm, we perform the feature mapping using the homogeneous kernel map. There is an alternative way called Nyström’s approximation [9, 10] which has shown good performances in image steganalysis recently but we do not employ it due to its following disadvantages: First, it is data dependent. Second, it requires additional learning separately. In addition, [9] pointed out that it may not be effective if the features are low dimensional and dense. In 3D mesh steganalysis, it is appropriate to use the homogeneous kernel map which is data-independent and approximated without any additional learning since the number of data to be acquired is relatively small and the extracted features are low dimensional.

Given a cover and stego feature $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$, we want to find the feature map $\Psi(\cdot)$ called the homogeneous kernel map such that $K(\mathbf{x}, \mathbf{y}) = \langle \Psi(\mathbf{x}), \Psi(\mathbf{y}) \rangle_{\mathcal{H}}$ where \mathcal{H} stands for a Hilbert space. The kernel satisfies additivity and homogeneity, i.e., $K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^D k(x_j, y_j)$ where $k(x_j, y_j) = \sqrt{x_j y_j} k(\sqrt{y_j/x_j}, \sqrt{x_j/y_j})$, and x_j and y_j are j -th components of \mathbf{x} and \mathbf{y} , respectively. Again, we can express $k(x_j, y_j)$ using the so-called kernel signature $\mathcal{K}(\lambda)$ as follows.

$$k(x_j, y_j) = \sqrt{x_j y_j} \mathcal{K}(\log y_j - \log x_j). \quad (7)$$

According to Bochner’s theorem, any positive-definite function $\mathcal{K}(\lambda)$ can be expressed as

$$\mathcal{K}(\lambda) = \int_{\mathbb{R}} e^{-i\omega\lambda} d\mu(\omega), \quad \lambda \in \mathbb{R} \quad (8)$$

where $\mu(\omega)$ denotes a non-negative symmetric measure, and we assume that $\mu(\omega) = \kappa(\omega)d\omega$, i.e., $\kappa(\omega)$ is the spectrum of the kernel signature in the frequency domain. The following expression can be obtained by replacing the kernel signature $\mathcal{K}(\cdot)$ of Eq. (7) with (8).

$$\begin{aligned} k(x_j, y_j) &= \int_{\mathbb{R}} [e^{-i\omega \log x_j} \sqrt{x_j \kappa(\omega)}]^* [e^{-i\omega \log y_j} \sqrt{y_j \kappa(\omega)}] d\omega \\ &= \int_{\mathbb{R}} [\Psi(x_j)]_{\omega}^* [\Psi(y_j)]_{\omega} d\omega, \quad [\Psi(x_j)]_{\omega} = e^{-i\omega \log x_j} \sqrt{x_j \kappa(\omega)} \end{aligned} \quad (9)$$

$\Psi(x_j)$ is a continuous and infinite dimensional vector. Thus, we need to approximate its discrete and finite one $\hat{\Psi}(x_j)$ by sampling and scaling it.

$$[\hat{\Psi}(x_j)]_k = \sqrt{\omega_0} [\Psi(x_j)]_{k\omega_0}, \quad k = -N, \dots, N \quad (10)$$

where ω_0 is the so-called fundamental frequency. Meanwhile, the multi-dimensional feature map $\hat{\Psi}(\mathbf{x})$ is generated by stacking the scalar ones as

$$\hat{\Psi}(\mathbf{x}) = \bigoplus_{j=1}^D \hat{\Psi}(x_j) \quad (11)$$

Note that the final feature map $\hat{\Psi}(\mathbf{x})$ is a $D(2N+1)$ -dimensional vector because the dimensionality of $\hat{\Psi}(x_j)$ for x_j is $(2N+1)$ by Eq. (10). Now, we have non-linear feature sets of $64 \times (2N+1)$ dimensions for the cover and stego meshes.

4 Experimental Results

In this section, we demonstrate and analyze the effectiveness of our approach. The experiments were conducted under the following environments. The Princeton Mesh Segmentation database was used to generate dataset for training and testing, which contains 380 various shaped 3D meshes.

Two different information hiding methods were considered here. The first scheme was a high capacity steganography method [4] for which we set parameters i.e., interval number and the number of layers to 5,000 and 10, respectively. We treated all vertices as carriers except 3 base vertices. Next, using the mean based watermarking method [3], we embedded randomly generated messages of 64-bit length into all the meshes with the incremental step size $\Delta k = 0.001$ and the strength factor $\alpha = 0.03$. Those two methods yielded the embedded meshes with 0.0550 ± 0.0053 and 0.0859 ± 0.0069 in the fast mesh perceptual distance (FMPD) [16], respectively. It is difficult to make a visual distinction between an embedded mesh with FMPD below 0.2 and its cover.

We employed the FLD ensemble classifier [14] which is commonly used in image steganalysis. The classifier builds some base learners on *randomly* sampled subspaces of the feature space, and the database we used is relatively small. Thus, to remove the chance factor and consider the statistical importance of results,

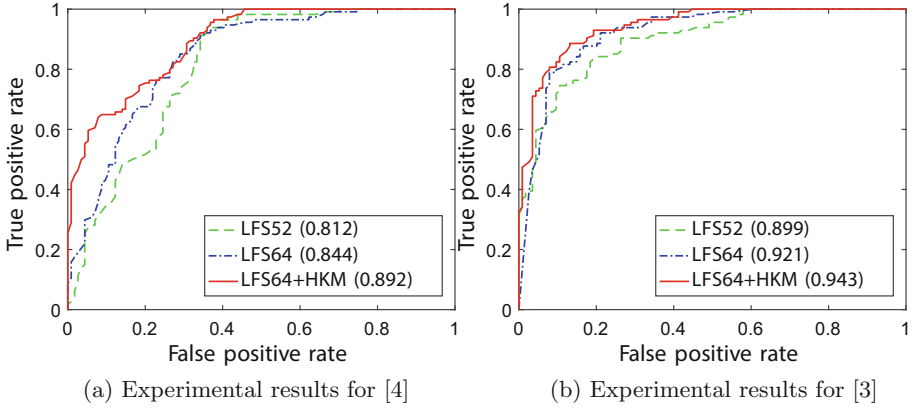


Fig. 1. Receiver operating characteristic curves of the different feature types. The numbers in parentheses represent the areas under the curves (AUC).

the experiments were performed for 100 datasets constructed by randomly and repeatedly splitting the whole data into 70% for training and 30% for test. As measures for performance evaluation, we used receiver operating characteristic (ROC) curves and box-and-whisker plots for detection error which is defined as follows.

$$ERR = \frac{FP + FN}{TP + TN + FP + FN} \quad (12)$$

where the terminology comes from [17]. The ROC curves in Fig. 1 were drawn at the median values of the detection errors obtained from 100 repeated experiments.

We first compared LFS64, our proposed feature set, with LFS52 [8] which shows the state-of-the-art performance. For the method of [4], as seen in Fig. 1, LFS64 showed a larger area under the curve (AUC) and a higher true positive rate for a small false positive rate than LFS52. Also, we observed a slight improvement with the smaller median value as shown in Fig. 2. On the other hand, the performance improvement for [3] was noticeable and the detection error decreased by about 2%. This shows that the proposed feature captures the fine deformation of the surface due to the change of the vertex norm.

Next, we demonstrated the results when applying the homogeneous kernel map (HKM) to LFS64. The homogeneous kernel used here was χ^2 with the parameter $N = 3$ in Eq. (10) accounting for time complexity. For both information hiding methods, the local feature set, or LFS64 combined with HKM yielded higher AUC from the ROC curves, and had less upper, lower quartile, and median in the detection error distributions than the other feature sets. The improvement of the combination is clearer and more conspicuous against [3] than [4]. As shown in Fig. 2b, the upper quartile of LFS64+HKM is equal to the lower one of LFS64, i.e., its interquartile range (IQR) does not fall on LFS64's IQR.

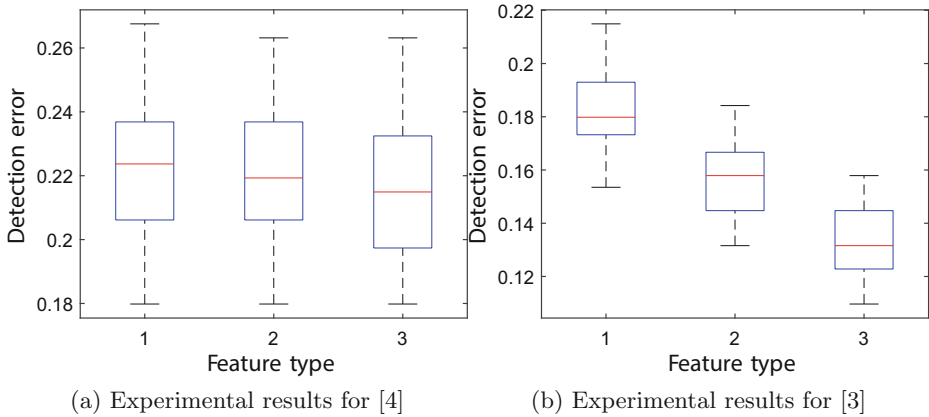


Fig. 2. Box-and-whisker plots for detection errors. Feature types 1 to 3 indicate LFS52, LFS64 and LFS64+HKM, respectively.

5 Conclusion

In this paper, we proposed a local feature set, LFS64 and its combination with the homogeneous kernel map to improve performance for 3D mesh steganalysis. The three local features, i.e., edge normal, mean and total curvatures were adopted to recognize more subtle changes induced by information hiding on the surface of a mesh. In addition, we applied the homogeneous kernel map to the feature set, which helped a classifier find non-linear separation boundaries more efficiently. For the two representative steganography and watermarking techniques, the proposed method showed significantly improved performances in terms of AUC or detection error. In the future, more sophisticated 3D mesh steganalytic algorithm should be designed to detect a stego mesh which is generated by an information hiding method changing its connectivity without distortions.

Acknowledgments. This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT1402-05.

References

1. Bateman, P., Schaathun, H.G.: Image steganography and steganalysis. Master's thesis, University of Surrey, United Kingdom
2. Ohbuchi, R., Takahashi, S., Miyazawa, T., Mukaiyama, A.: Watermarking 3D polygonal meshes in the mesh spectral domain. In: Graphics Interface (GI), vol. 2001, pp. 9–17. Citeseer (2001)
3. Cho, J.W., Prost, R., Jung, H.Y.: An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms. *IEEE Trans. Sig. Process. (TSP)* **55**(1), 142–155 (2007)
4. Chao, M.W., Lin, C.H., Yu, C.W., Lee, T.Y.: A high capacity 3D steganography algorithm. *IEEE Trans. Vis. Comput. Graph. (TVCG)* **15**(2), 274–284 (2009)

5. Yang, Y., Ivrişsimtziş, I.: Mesh discriminative features for 3D steganalysis. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **10**(3), 27 (2014)
6. Yang, Y., Pintus, R., Rushmeier, H., Ivrişsimtziş, I.: A steganalytic algorithm for 3D polygonal meshes. In: *IEEE International Conference on Image Processing (ICIP)*, pp. 4782–4786. IEEE (2014)
7. Yang, Y., Pintus, R., Rushmeier, H., Ivrişsimtziş, I.: A 3D steganalytic algorithm and steganalysis-resistant watermarking. *IEEE Trans. Vis. Comput. Graph. (TVCG)* (in press). doi:[10.1109/TVCG.2016.2525771](https://doi.org/10.1109/TVCG.2016.2525771)
8. Li, Z., Bors, A.G.: 3D mesh steganalysis using local shape features. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2144–2148. IEEE (2016)
9. Boroumand, M., Fridrich, J.: Boosting steganalysis with explicit feature maps. In: *ACM Workshop on Information Hiding and Multimedia Security (IH & MMSec)*, pp. 149–157. ACM (2016)
10. Perronnin, F., S nchez, J., Xerox, Y.L.: Large-scale image categorization with explicit data embedding. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2297–2304. IEEE (2010)
11. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **34**(3), 480–492 (2012)
12. Peyre, G., Cohen, L.: Surface segmentation using geodesic centroidal tessellation. In: *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pp. 995–1002. IEEE (2004)
13. Lyu, S., Farid, H.: Detecting hidden messages using higher-order statistics and support vector machines. In: Petitcolas, F.A.P. (ed.) *IH 2002*. LNCS, vol. 2578, pp. 340–354. Springer, Heidelberg (2003). doi:[10.1007/3-540-36415-3-22](https://doi.org/10.1007/3-540-36415-3-22)
14. Kodovsky, J., Fridrich, J., Holub, V.: Ensemble classifiers for steganalysis of digital media. *IEEE Trans. Inf. Forensics Secur. (TIFS)* **7**(2), 432–444 (2012)
15. Cogramne, R., Fridrich, J.: Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory. *IEEE Trans. Inf. Forensics Secur. (TIFS)* **10**(12), 2627–2642 (2015)
16. Wang, K., Torkhani, F., Montanvert, A.: A fast roughness-based approach to the assessment of 3D mesh visual quality. *Comput. Graph.* **36**(7), 808–818 (2012)
17. Fawcett, T.: An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**(8), 861–874 (2006)

From Sealed-Bid Electronic Auction to Electronic Cheque

Kin-Woon Yeow^(✉), Swee-Huay Heng, and Syh-Yuan Tan

Faculty of Information Science and Technology,
Multimedia University, Melaka, Malaysia
yeowkinwoon@gmail.com, {shheng, sytan}@mmu.edu.my

Abstract. In this paper, we establish a relation between sealed-bid e-auction and e-cheque by proposing a transformation technique which transforms a secure sealed-bid e-auction into a secure e-cheque scheme. Although the application scenario differs, we notice that the scheme structure and fundamental security properties in both schemes are similar, namely, unforgeability, anonymity and indistinguishability. As a proof of concept, we apply the transformation technique on the classic Sakurai and Miyazaki sealed-bid e-auction scheme and obtain a secure e-cheque scheme.

Keywords: Sealed-bid · Auction · Cheque · Privacy · Transformation

1 Introduction

The first e-cheque scheme is published by Chaum et al. in 1988 [3] which is based on digital signature and it achieved the security properties of untraceability, undeniability and unforgeability. In the creation of e-cheque of Chaum et al.'s scheme, the data redundancy can only be identified after a few repetitions of data, hence heavy computation overhead are generated. A year later, Chaum et al. enhanced the verification algorithm of the scheme [2]. Based on RSA signature, Chen further reduced the computational cost of an e-cheque scheme by replacing modular exponentiation with hash function to allow selection of payee's credential during writing of an e-cheque [5].

In 2006, Lu et al. proposed a new e-cheque scheme by using identity-based signature scheme as an alternative to achieve one-timeness and computational efficiency [14]. In 2009, Su proposed an e-cheque scheme that increases the efficiency using short signature size of 160 bits [20]. More e-cheque schemes were proposed as listed in Table 1. Similar with e-cheque, sealed-bid e-auction has develop through time. Franklin and Reiter [7] presented the first sealed-bid e-auction scheme in 1996 with the property of anonymity. Their scheme made use of the verifiable signature sharing, thus the auctioneer knows bids of all the bidders at the end of the auction. In 1999, a sealed-bid e-auction scheme was introduced by Kikuchi et al. which improved on the anonymity of sealed-bid e-auction and proposed a tie-breaking algorithm. When two or more highest bids

Table 1. Recent developments of e-Cheque

Scheme	Technique
Pasupathinathan et al. [17]	Schnorr signature
Chen et al. [4]	Blind signature
Hinarejos et al. [8]	Fair exchange protocol
Sun et al. [21]	Chameleon hashing

are placed, the use of tie-breaking algorithm that consists of multiple auction rounds will decrease the number of highest bids.

A prevailing concern in sealed-bid e-auction is bid-rigging attack which occurs when coercers give command towards other bidders to bid at low prices for the coercers to win the auction. In view of this, Sakurai and Miyazaki [18] proposed a sealed-bid e-auction scheme which solved the bid-rigging attack with undeniable signature and the technique of bulletin board. Although Sakurai and Miyazaki's scheme provides non-repudiation, confidentiality and anonymity to the bid, Viswanathan et al. showed that the anonymity is only enjoyed by losing bidders but not the winning bidder [23]. Viswanathan et al. also inspired by Schoenmakers's scheme [19] which is based on verifiable secret sharing that does not preserve the anonymity of users. In view of that, Viswanathan et al. proposed a publicly verifiable auction scheme for sealed-bid e-auction that preserve the anonymity of both winning and losing bidders. A lot secure sealed-bid e-auction schemes have been proposed since then as described in Table 2.

Table 2. Recent developments of Seal-Bid e-Auction

Scheme	Technique
Abe and Suzuki [1]	Chameleon bit-commitments
Chen et al. [6]	Homomorphic encryption
Wu et al. [24]	Symmetric encryption
Lee et al. [12]	Group signature
Xiong et al. [25]	Ring signature and verifiable encryption
Li et al. [13]	Modified group signature protocol
Montenegro and Lopez [15]	Secure multiparty computation
Nojournian and Stinson [16]	Verifiable secret sharing
Howlader and Mal [10]	MIX and deniable signature

1.1 Our Contributions

Although e-cheque and sealed-bid e-auction are significantly differ in their application scenario, we notice a high similarity between their security properties and this leads us to generalize a generic transformation from sealed-bid e-auction to

e-cheque. This transformation allows one to easily construct a secure e-cheque scheme from a secure sealed-bid e-auction scheme, instead of building from scratch.

Firstly, we formally define the security models of sealed-bid e-auction and e-cheque which are needed by the transformation. The security models defined for sealed-bid e-auction are based on three security properties, namely, unforgeability, anonymity, and confidentiality. Similarly, the same security properties are required to create a secure construction of e-cheque.

Secondly, we present the transformation from a secure sealed-bid e-auction to a secure e-cheque. In precise, we show that when the underlying sealed-bid e-auction scheme fulfills the security properties of a secure sealed-bid e-auction scheme in our transformation framework, we can directly obtain a secure transformed e-cheque scheme.

1.2 Organization

We organize the paper as follows. Firstly, we provide the definition of algorithms and security models for sealed-bid e-auction in Sect. 2. Secondly, we provide the definition of algorithms and security models for e-cheque in Sect. 3. Followed by, generic transformation from sealed-bid e-auction to e-cheque in Sect. 4. In Sect. 5, we show an instance of the transformation. In Sect. 6, conclusion on the transformation is provided.

2 Sealed-Bid e-Auction

As the first step of initializing the transformation framework, we formally define the security definitions of sealed-bid e-auction (SBEA). To the best of our knowledge, this is the first rigorous definitions of SBEA.

2.1 Definition

A SBEA scheme consists of three algorithms, namely, $SBEA = \{\mathbf{Setup}, \mathbf{Bid}, \mathbf{Open}\}$:

Setup (1^k): A Trusted Third Party (TTP) generates public parameters ($param$) that includes public key (pk) and secret key (sk) for each entity (auctioneer (AU) and bidder (BR)).

Bid: Each BR chooses his bidding price (P). Then, BR hides P and creates the bid (σ) from the P . BR places σ by submitting σ to AU. In some schemes, BR is allowed to verify the σ placed.

Open(σ): After the duration of bidding ends, AU declares and opens the winning bid. BR has to prove the possession of the winning bid.

Note: In some cases, auctioneer and bidder generate their own pk and sk [22]. However, our model does not conflict with theirs as we can view the TTP as a Certification Authority (CA) which certifies the public key of every entity.

2.2 Security Model

We define the security properties of SBEA, namely, unforgeability, anonymity, and confidentiality in this section.

Unforgeability. In SBEA scheme is defined as the incapability of impersonating any other BR to forge a valid bid [24]. We rigorously define the security notion of SBEA, namely existential unforgeable under chosen price attack (EUF-CPA). One can view this as an analogy to the EUF-CMA security notion of digital signature. We define the EUF-CPA as a game between challenger (C) and adversary (A_1) as follows:

Registration phase

Setup Query: When the game starts, C gives system parameters to A_1 and registers A_1 .

Training phase

Bid Query: A_1 queries for chosen price, P to obtain bid (σ_i) from C , where i changes accordingly to the number of training iteration.

Open Query: A_1 queries for the verification details of (P, σ_i) from C .

Forging phase

Forge: After certain time of training, A_1 selects P^* and forges σ^* . If σ^* is a valid bid, then A_1 wins the game.

We say that a SBEA scheme is EUF-CPA if there is no polynomial time adversary A_1 has probability in winning the game above.

Anonymity. In SBEA scheme is defined as the incapability to gain access to other BRs' information, excluding those approved parties that store BRs' information [24]. We rigorously define the security notion of SBEA, namely bidder anonymity under chosen price attack (BA-CPA). One can view this analogy to the SA-CMA security notion of undeniable signature. We adopt the same **Registration phase** and **Training phase** as in Sect. 2.2 to define BA-CPA as a game between C and A_1 as follows.

Identifying phase

Identify: After using several identities for training, A_1 gives P^* to C which returns σ^1 and σ^2 where one of them is a random element in bid space. A_1 analyses and determines the correct bid σ^* from the two choices.

We say that a SBEA scheme is BA-CPA against adversary A_1 when A_1 has negligible probability of selecting the valid σ^* in the game above.

Confidentiality. In SBEA scheme is defined as the secrecy of all bidding prices excluding winning price. The losing prices are not reveal to anyone including AU [24]. We rigorously define the security notion of SBEA, namely indistinguishability under chosen bid attack (IND-CBA). One can view this game as a SBEA variant of IND-CCA in encryption scheme. Adopting the same **Registration phase** and **Training phase** as in Sect. 2.2, the IND-CBA game between C and A_1 is as follows.

Identifying phase

Identify: After using several identities for training, A_1 gives P^1 and P^2 to C , which returns σ^* as a valid bid from either P^1 or P^2 . A_1 analyses and determines the correct price P^* .

We say that a SBEA scheme is IND-CBA against adversary A_1 when A_1 has negligible probability of selecting the valid P^* in the game above.

3 e-Cheque

To the best of our knowledge, we provide the first rigorous definition of e-cheque.

3.1 Definition

An e-cheque scheme consists of three algorithms, namely, $\text{e-cheque} = \{\mathbf{Register}, \mathbf{Write}, \mathbf{Transfer}\}$:

Register (1^k): A TTP generates public parameters ($param$) that includes public key (pk) and secret key (sk) for each entity (Bank (B), Payer (PR), and payee).

Write: PR upon receiving the invoice containing payment details, writes the payment details and hides PR's account information (I). PR creates a valid e-cheque (ϑ). Then, ϑ is submitted to B.

Transfer(ϑ): At the end of the day, PR proves to B the ownership of ϑ . After B checks for sufficient amount in PR's account for clearance, B debits from PR's account and credits into payee's account.

Note: In some cases, bank and user generate their own pk and sk [4]. However, our model does not conflict with theirs as we can view the TTP as a CA which certifies the public key of each entity.

3.2 Security Model

We define the security properties of e-cheque, namely, unforgeability, anonymity, and confidentiality in this section.

Unforgeability. In e-cheque scheme is defined as the incapability for user to forge a valid signed e-cheque of another user [17]. We rigorously define the security notion of e-cheque, namely existential unforgeable under chosen account attack (EUF-CAA). One can view this analogy to the EUF-CMA security notion of digital signature. We define EUF-CAA as a game between challenger (C) and adversary (A_2) as follows:

Registration phase

Register Query: When the game starts, C gives system parameters to A_2 and registers A_2 .

Training phase

Write Query: A_2 queries using account information, I to obtain a cheque, ϑ_j from C , where j changes accordingly to the number of training iteration.

Transfer Query: A_2 queries for the verification details of (I, ϑ_j) from C .

Forging phase

Forge: A_2 selects I^* and forge ϑ^* . If ϑ^* is a valid cheque, A_2 wins the game.

The e-cheque scheme is said to be EUF-CAA if any probabilistic polynomial time adversary A_2 has negligible probability in winning the game above.

Anonymity. In e-cheque protocol is defined as the incapability to gain access to PRs identity, except B [17]. We rigorously define the security notion of e-cheque, namely payer anonymity under chosen account attack (PA-CAA). One can view this analogy to the SA-CMA security notion of undeniable signature. Here, we provide a brief version of anonymity game as the same **Registration phase** and **Training phase** as in Sect. 3.2.

Identifying phase

Identify: A_2 gives I^* to C which returns ϑ^1 and ϑ^2 , where one of them is a random element in cheque space. A_2 analyses and determines the correct cheque ϑ^* from the two choices.

We say that an e-cheque scheme is PA-CAA against adversary A_2 when A_2 has negligible probability of selecting the valid ϑ^* in the game above.

Confidentiality. In e-cheque is defined as the secrecy of all unused or invalid cheque excluding valid cheque [17]. The unused and invalid cheques are not revealed to anyone including B. We rigorously define the security notion of e-cheque, namely indistinguishability under chosen cheque attack (IND-CCeA). One can view this game as an e-cheque variant of IND-CCA in encryption scheme. Adopting the same **Registration phase** and **Training phase** as in Sect. 3.2, the IND-CCeA game between C and A_2 is as follows.

Identifying phase

Identify: After using several identities for training, A_2 gives I^1 and I^2 to C , which returns ϑ^* as a valid cheque from either I^1 or I^2 . A_2 analyses and determines the correct price I^* .

We say that an e-cheque scheme is IND-CCeA against adversary A_2 when A_2 has negligible probability of selecting the valid I^* in the game above.

4 Generic Transformation from Seal-Bid e-Auction to e-Cheque

In this section, we show the transformation of a SBEA scheme to an e-cheque scheme. Firstly, we consider a single bidder in SBEA and change it to e-cheque as follows: bid \rightarrow cheque, bidder \rightarrow payer, auctioneer \rightarrow bank, and one round of SBEA \rightarrow a submission of cheque. Despite these similarities, there are two implementation issues remain in the transformation: (1) Bulletin board in SBEA is not required in e-cheque; (2) A passive party (payee) is required in e-cheque.

As a solution, the bulletin board in SBEA is now the online verification in e-cheque. Furthermore, the additional party, namely, payee is treated as a passive party who can observe the transaction between the payer and the bank to confirm the delivery of the e-cheque. It remains to check if the transformed e-cheque scheme is secure or not. We analyze the security of transformed scheme and show in Theorems 1, 2, and 3 as follows:

Theorem 1. *Let $SBEA = \{\text{Setup}, \text{Bid}, \text{Open}\}$ be a SBEA scheme where bids are required to be sealed for bidding. Let $e\text{-cheque} = \{\text{Register}, \text{Write}, \text{Transfer}\}$ be the e-cheque scheme. Then e-cheque is secure against EUF-CAA if the underlying SBEA is secure against EUF-CPA.*

Proof (Sketch). Let A_2 be an adversary who (t', ϵ') -breaks the EUF-CAA of e-cheque scheme, we show that there exists an adversary A_1 who (t, ϵ) -breaks the EUF-CPA of SBEA scheme where t represents the time needed to complete the attack and ϵ is the probability of success in the attack.

During training phase, A_1 received *params* and passes to A_2 which starts the write query. As A_2 issues payment details I into the write query, A_1 uses $P = I$ as its bid query and forwards the answer $\sigma = \vartheta$ from the bid oracle to A_2 . A_2 also issues transfer query on ϑ . Since the result of write query and bid query are inter-related, A_1 could input the ϑ into open query which performs the verification process. Suppose open query outputs (valid/invalid), A_1 returns (valid/invalid) to A_2 .

Finally, when A_2 outputs ϑ^* , A_1 sets the forged bid as $\sigma^* = \vartheta^*$ and returns as result of the EUF-CPA attack. If $\text{Open}(\sigma) = \text{accept}$, this indicates A_1 breaks the EUF-CPA SBEA scheme successfully.

Theorem 2. *Let $SBEA = \{\text{Setup}, \text{Bid}, \text{Open}\}$ be a SBEA scheme where bids are required to be sealed for bidding. Let $e\text{-cheque} = \{\text{Register}, \text{Write}, \text{Transfer}\}$ be the e-cheque scheme. Then e-cheque is secure against PA-CAA if the underlying sealed-bid e-auction is secure against BA-CPA.*

Proof (Sketch). Let A_2 be an adversary who (t', ϵ') -breaks the PA-CAA e-cheque scheme, we show that there exists an adversary A_1 who (t, ϵ) -breaks the BA-CPA SBEA scheme, where t represents the time needed to complete the attack and ϵ is the probability of success in the attack. The same training phase as in Theorem 1 is performed.

A_1 starts his identifying phase when A_1 receives the challenge $P^* = I^*$ from A_2 . A_1 starts his identifying phase and receives (σ^1, σ^2) . A_1 sets $(\vartheta^1 = \sigma^1, \vartheta^2 = \sigma^2)$ and passes $(\vartheta^1, \vartheta^2)$ to A_2 as the challenge. When A_2 outputs its choice of ϑ^* , A_1 runs **Transfer**(ϑ) with A_2 to verify the ϑ^* . A_1 sets $\sigma^* = \vartheta^*$ and runs **Open**(σ) with challenger. If A_2 made a correct guess, A_1 will output the correct σ^* , passes the **Open**(σ) and break the BA-CPA of SBEA scheme successfully.

Theorem 3. *Let $SBEA = \{\text{Setup}, \text{Bid}, \text{Open}\}$ be a SBEA scheme where bids are required to be sealed for bidding. Let e-cheque = $\{\text{Register}, \text{Write}, \text{Transfer}\}$ be the e-cheque scheme. Then e-cheque is secure against IND-CCeA attack if the underlying SBEA is secure against IND-CBA.*

Proof (Sketch). Let A_2 be an adversary who (t', ϵ') -breaks the IND-CCeA e-cheque scheme, we show that there exists an adversary A_1 who (t, ϵ) -breaks the IND-CBA SBEA scheme, where t represents the time needed to complete the attack and ϵ is the probability of success in the attack. The same training phase as in Theorem 1 is performed.

A_1 starts his identifying phase when A_1 receives the challenge $P^1 = I^1$ and $P^2 = I^2$ from A_2 . A_1 starts his identifying phase and receives (σ^*) . A_1 sets $(\vartheta^* = \sigma^*)$ and passes (I^1, I^2, ϑ^*) to A_2 as the challenge. When A_2 outputs its choice of I^* , A_1 runs **Transfer**(ϑ) with A_2 to verify the I^* . A_1 sets $\sigma^* = \vartheta^*$ and runs **Open**(σ) with challenger. If A_2 made a correct guess, A_1 will output the correct P^* , pass the **Open**(σ) and break the IND-CBA of SBEA scheme successfully.

5 An Instance

Using Sakurai and Miyazaki’s SBEA scheme [18] as instance, we perform the transformation as follows:

Register: The Registration Authority issues the secret key S_G to each member of the registered group via a secure channel. Let $P_G = \alpha^{S_G} \pmod p$ be its corresponding public-key, S_A be payer $_A$ ’s (PR_A) individual private key and $P_A = \alpha^{S_A} \pmod p$ be the corresponding PR_A ’s individual public-key, and $Cert_A$ as certificate. Thus, $pk = \{P_G, P_A, Cert_A\}$ and $sk = \{S_G, S_A\}$.

Write: The bank publishes a public key, w_a for the generation of PR_A ’s account information. Then, PR_A sets the transfer amount and generates three random numbers $x, k, \mu \in Z_q^*$ and computes:

$$\begin{aligned}
 h &= \alpha^x \pmod p & r &= \alpha^k \pmod p & \tilde{r} &= (r^r)^x \pmod p & \lambda &= \alpha^\mu \pmod p \\
 \tilde{\lambda} &= \lambda^x \pmod p & c &= H(w_k, \tilde{r}, \lambda, \tilde{\lambda}) & \sigma &= rk - cS_A \pmod q
 \end{aligned}$$

Next, PR_A calculates the group signature, $t = \frac{(\tilde{r}\mu + rk)}{(S_G + H(w_a, \tilde{r}, r, \lambda))} \pmod{q}$ and sends $(P_A, Cert_A, h, \tilde{r}, \lambda, \tilde{\lambda}, \sigma, t)$ to the bank. The bank validates the validity of $Cert_A$ and publishes $(h, \tilde{r}, \sigma, t)$. (Note: since the bank has not known the account information, thus the transfer cannot proceed)

Transfer: For the process of submitting an e-cheque, PR_A need to delivers the e-cheque and proves the validity of w_a to the bank via confirmation protocol, if w_a is invalid the bank can prove it through disavowal protocol by setting $\beta = \alpha^\sigma P_A^{H(w_a, \tilde{r}, \lambda, \tilde{\lambda})}$ as follows:

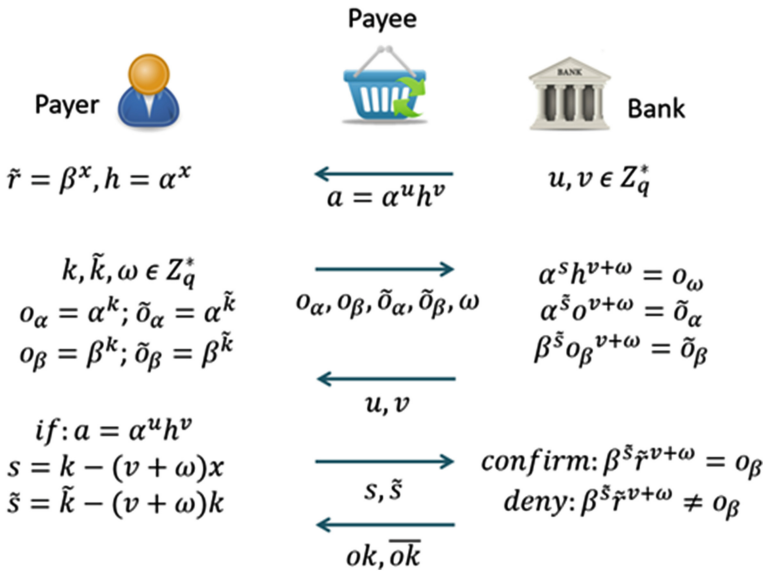


Fig. 1. Confirmation and disavowal protocol

Figure 1 shows the process of proving the validity of w_a , the payee is able to view the transaction between PR_A and bank. PR_A sends w_a and x to the bank. Bank verifies the validity of signature using public-key P_A with Formula 1.

$$\tilde{r} = (\alpha^\sigma P_A^{H(w_a, \tilde{r}, \lambda, \tilde{\lambda})})^x \tag{1}$$

Then, bank computes $r = \tilde{r}^{x^{-1}}$ and verifies the group signature $\tilde{\lambda} \tilde{r}^r = (P_G \alpha^{H(w_a, \tilde{r}, \lambda)})^{tx}$. In the end of w_a validation process, if both individual signature and group signature are correct, bank notifies both PR_A and payee that the e-cheque transaction has completed.

Since the SBEA scheme [18] contains the security properties of unforgeability, anonymity and confidentiality. Following Theorems 1, 2, and 3, the transformed e-cheque scheme is a secure scheme which enjoys the same security properties as the original SBEA scheme.

6 Conclusion

We presented a generic transformation of SBEA to e-cheque. As an instance, we perform transformation on Sakurai and Miyazaki's SBEA scheme [18] and show that we obtain a secure transformed Sakurai and Miyazaki's e-cheque scheme.

Acknowledgment. The authors would like to convey gratitude towards the Malaysia government's Fundamental Research Grant Schemes (FRGS/2/2014/ICT04/MMU/03/1) and (FRGS/1/2015/ICT04/MMU/03/5) for supporting this work.

References

1. Abe, M., Suzuki, K.: Receipt-free sealed-bid auction. In: Chan, A.H., Gligor, V. (eds.) ISC 2002. LNCS, vol. 2433, pp. 191–199. Springer, Heidelberg (2002). doi:[10.1007/3-540-45811-5_14](https://doi.org/10.1007/3-540-45811-5_14)
2. Chaum, D., Boer, B., Heyst, E., Mjølsnes, S., Steenbeek, A.: Efficient offline electronic checks. In: Quisquater, J.-J., Vandewalle, J. (eds.) EUROCRYPT 1989. LNCS, vol. 434, pp. 294–301. Springer, Heidelberg (1990). doi:[10.1007/3-540-46885-4_31](https://doi.org/10.1007/3-540-46885-4_31)
3. Chaum, D., Fiat, A., Naor, M.: Untraceable electronic cash. In: Goldwasser, S. (ed.) CRYPTO 1988. LNCS, vol. 403, pp. 319–327. Springer, New York (1990). doi:[10.1007/0-387-34799-2_25](https://doi.org/10.1007/0-387-34799-2_25)
4. Chen, C.L., Wu, C.H., Lin, W.C.: Improving an on-line electronic check system with mutual authentication. In: International Conference on Advanced Information Technologies (2010)
5. Chen, W.: Efficient on-line electronic checks. *Appl. Math. Comput.* **162**(3), 1259–1263 (2005)
6. Chen, X., Lee, B., Kim, K.: Receipt-free electronic auction schemes using homomorphic encryption. In: Lim, J.-I., Lee, D.-H. (eds.) ICISC 2003. LNCS, vol. 2971, pp. 259–273. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-24691-6_20](https://doi.org/10.1007/978-3-540-24691-6_20)
7. Franklin, M.K., Reiter, M.K.: The design and implementation of a secure auction service. In: Software Engineering, IEEE, pp. 302–312 (1996)
8. Hinarejos, M.F., Ferrer-Gomila, J.L., Draper-Gil, G., Huguet-Rotger, L.: Anonymity and transferability for an electronic bank check scheme, information. In: 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, pp. 427–435 (2012)
9. Howlader, J., Ghosh, A., Pal, T.D.R.: Secure receipt-free sealed-bid electronic auction. In: Ranka, S., Aluru, S., Buyya, R., Chung, Y.-C., Dua, S., Grama, A., Gupta, S.K.S., Kumar, R., Phoha, V.V. (eds.) IC3 2009. CCIS, vol. 40, pp. 228–239. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-03547-0_22](https://doi.org/10.1007/978-3-642-03547-0_22)
10. Howlader, J., Mal, A.K.: Sealed-bid auction: a cryptographic solution to bid-rigging attack in the collusive environment. *Secur. Commun. Netw.* **8**(18), 3415–3440 (2015)
11. Kikuchi, H., Hakavy, M., Tygar, D.: Multi-round anonymous auction protocols (Special issue on internet technology and its applications). In: IEICE Transactions on Information and Systems, pp. 769–777 (1999)
12. Lee, C.C., Ho, P.F., Hwang, M.S.: A secure e-Auction scheme based on group signatures. *Inf. Syst. Front.* **11**(3), 335–343 (2009)

13. Li, M.J., Juan, J.S.T., Tsai, J.H.C.: Practical electronic auction scheme with strong anonymity and bidding privacy. *Inf. Sci.* **181**(12), 2576–2586 (2011)
14. Lu, R., Cao, Z., Dong, X.: Efficient ID-based one-time proxy signature and its application in e-Cheque. In: *Cryptology and Network Security*, pp. 153–167 (2006)
15. Montenegro, J.A., Lopez, J.: A practical solution for sealed bid and multi-currency auctions. *Comput. Secur.* **45**, 186–198 (2014)
16. Nojournian, M., Stinson, D.R.: Efficient sealed-bid auction protocols using verifiable secret sharing. In: Huang, X., Zhou, J. (eds.) *ISPEC 2014*. LNCS, vol. 8434, pp. 302–317. Springer, Cham (2014). doi:[10.1007/978-3-319-06320-1_23](https://doi.org/10.1007/978-3-319-06320-1_23)
17. Pasupathinathan, V., Pieprzyk, J., Wang, H.: Privacy enhanced electronic cheque system. In: *Seventh IEEE International Conference on E-Commerce Technology, CEC 2005*, pp. 431–434 (2005)
18. Sakurai, K., Miyazaki, S.: An anonymous electronic bidding protocol based on a new convertible group signature scheme. In: Dawson, E.P., Clark, A., Boyd, C. (eds.) *ACISP 2000*. LNCS, vol. 1841, pp. 385–399. Springer, Heidelberg (2000). doi:[10.1007/10718964_32](https://doi.org/10.1007/10718964_32)
19. Schoenmakers, B.: A simple publicly verifiable secret sharing scheme and its application to electronic voting. In: Wiener, M. (ed.) *CRYPTO 1999*. LNCS, vol. 1666, pp. 148–164. Springer, Heidelberg (1999). doi:[10.1007/3-540-48405-1_10](https://doi.org/10.1007/3-540-48405-1_10)
20. Su, R.: A new efficient self-certified one-time short signature scheme from bilinear pairings. In: *ChinaCOM Fourth International Conference on Communications and Networking in China*, pp. 1–5 (2009)
21. Sun, Y., Chai, J., Liang, H., Ni, J., Yu, Y.: A secure and efficient e-Cheque protocol from chameleon hash function. In: *5th International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, pp. 470–475 (2013)
22. Suzuki, K., Kobayashi, K., Morita, H.: Efficient sealed-bid auction using hash chain. In: Won, D. (ed.) *ICISC 2000*. LNCS, vol. 2015, pp. 183–191. Springer, Heidelberg (2001). doi:[10.1007/3-540-45247-8_15](https://doi.org/10.1007/3-540-45247-8_15)
23. Viswanathan, K., Boyd, C., Dawson, E.: A three phased schema for sealed bid auction system design. In: Dawson, E.P., Clark, A., Boyd, C. (eds.) *ACISP 2000*. LNCS, vol. 1841, pp. 412–426. Springer, Heidelberg (2000). doi:[10.1007/10718964_34](https://doi.org/10.1007/10718964_34)
24. Wu, C.C., Chang, C.C., Lin, I.C.: New sealed-bid electronic auction with fairness, security and efficiency. *J. Comput. Sci. Technol.* **23**(2), 253–264 (2008)
25. Xiong, H., Qin, Z., Li, F.: An anonymous sealed-bid electronic auction based on ring signature. *Int. J. Netw. Secur.* **8**(3), 235–242 (2009)

Enhanced Database Security Using Homomorphic Encryption

Connor Røset^(✉), Van Warren, and Chia-Chu Chiang

Computer Science, University of Arkansas at Little Rock,
2801 S University Ave, Little Rock, AR 72204, USA
{cbroset, cxchiang}@ualr.edu, van@wdv.com

Abstract. When sensitive data is stored on publicly available areas, privacy of that data becomes a concern. Organizations may wish to move data to public servers so the data is more accessible by their employees or consumers. It is important that this data be encrypted to ensure it remains confidential and secure. When this data is encrypted, it becomes difficult or impossible to perform calculations on a publicly stored database. A solution to this is homomorphic encryption, which allows an unlimited number of computations on encrypted data. This project analyzes an N-tier rotation scheme which allows an unlimited number of addition and subtractions of encrypted data, along with an unlimited number of scalar multiplications and divisions. This scheme is inspired by a combination lock and features multiple levels of security depth. The result of the proposed algorithm is a fast encryption scheme which allows data to be manipulated post encryption.

Keywords: Cryptography · Database security · Encryption · Fully homomorphic · Homomorphism

1 Introduction

In the modern world of advanced computing and networking technology, more and more individuals are moving sensitive data to public servers for easier user access. Applications which allow users to remotely access this sensitive data over different networks should ensure the data maintains three things: its confidentiality, its integrity, and its availability (Pfleeger and Pfleeger 2006).

The first aspect, confidentiality, can be achieved via Encryption. Encryption algorithms encode plaintext, or human-readable information, into undecipherable format called ciphertext. Later, the process of decryption converts that ciphertext back into its original plaintext. While there are many existing encryption algorithms to provide data confidentiality, these methods require that the stored and secure ciphertext be decrypted before processing. This leaves the sensitive data stored on a database exposed, which could provide potential hackers a venue to attain that sensitive information. A current solution to this problem is homomorphic encryption. Homomorphic encryption works by allowing computation on encrypted data. This also removes the need for decryption, leaving no room for a hacker to gain access to any sensitive data.

Homomorphic encryption is broken into two types: fully homomorphic or partially homomorphic. Fully homomorphic encryptions allow for an unlimited number of arithmetic operations on encrypted data while partially homomorphic encryption only allows a limited number of arithmetic operations. Most current homomorphic encryption systems are partially homomorphic and only support a limited number of computations. In 2009, Gentry introduced a fully homomorphic encryption scheme which allows an arbitrary number of arithmetic operations on encrypted data (Gentry 2009b). However, Gentry's encryption scheme was found to have significant performance problems (Gentry et al. 2012; Lauter et al. 2011). According to one source, this scheme took up to 5 s per multiplication. It had a memory space increase from 1 Mb to over 10 Gb (Barthelemy 2016). For reference, the encrypted ciphertext became over 1,250 times larger than its corresponding unencrypted data.

This paper introduces an N-tier rotation based encryption scheme. This scheme allows an arbitrary number of arithmetic operations on encrypted data with multiple levels of security depth. The encrypted results of these operations, after decryption, are the same as the expected results from computation on the original data.

2 Problem

There exists the need for an encryption scheme which allows an unlimited number of computations on encrypted data stored a publicly available system. Such an encryption scheme is further necessary to allow businesses and organizations to transfer their existing sets of data to publicly available systems in order to save costs and make their data more easily accessible to employees or other organization members.

This project further seeks to solve a limitation specified in the report "Securing Databases with a Combination of Encryption and Information Hiding" (Wooldridge et al. 2016). This report explained their approach for a secure database utilizing encryption and information hiding had not been tested with partially or fully homomorphic encryption. The proposed algorithm of this paper is intended to be an implantation of a homomorphic encryption solution in their described database such that standard SQL commands can be applied to said database.

3 Survey of Existing Methods

Encryption algorithms can be divided into two categories: symmetric and asymmetric. An encryption scheme is symmetric when it uses the same key for encryption and decryption. An encryption scheme is asymmetric when it uses two separate keys for encryption and decryption. Asymmetric encryption schemes utilize a publicly available encryption key, which is posted so others can encrypt and send the owner sensitive information, and a privately kept decryption key, which remains private to the owner so this key holder can decrypt any ciphertext which was made with the publicly available encryption key. As discussed earlier, encryption schemes can be partially or fully homomorphic. An encryption scheme is partially homomorphic if it exhibits either additive or multiplicative properties, but not both (Gentry 2009b). There are two popular partially

homomorphic encryption schemes and one popular fully homomorphic encryption schemes currently in use, Paillier, RSA, and Gentry, respectively.

3.1 Paillier

The Paillier algorithm (Paillier 1999) is a partially homomorphic encryption which allows additive arithmetic, but not multiplicative arithmetic.

Algorithm 1. Paillier Key generation

Input Data: Inputs $(p, q, n, \lambda, g, \mu)$ for key generation

Result: Encryption key (n, g) , Decryption key (λ, μ)

1. Randomly generate the encryption and decryption keys
 - 1.1 Randomly generate p and q such that p and q are very large prime numbers and such that $gcd(p * q, (p - 1)(q - 1)) = 1$
 - 1.2 Calculate n such that $n = p * q$
 - 1.3 Calculate $n^2 = n * n$
 - 1.4 Calculate $\lambda = lcm((p - 1), (q - 1))$
 - 1.5 Randomly generate g such that $g \in \mathbb{Z}_{n^2}^*$ and ensure n divides g 's order
 - 1.6 Find μ such that $\mu = (((g^\lambda \bmod n^2) - 1) / n)^{-1} \bmod n$
 2. Return the public encryption key and private decryption key to the user
-

Algorithm 2. Paillier based encryption algorithm

Input Data: Encryption key (n, g) , input number $x \in \mathbb{Z}_n$

Result: An encrypted value $c \in \mathbb{Z}_{n^2}^*$

1. Encrypt the input number x
 - 2.1 Randomly generate r such that $r \in \mathbb{Z}_n^*$
 - 2.2 compute ciphertext c such that $c = g^x r^n \bmod n^2$
 2. Return the encrypted value of x which is c to the user.
-

Algorithm 3. Paillier based decryption algorithm

Input Data: Decryption key (λ, μ) , ciphertext c

Result: Decrypted ciphertext c , which is x

1. Decrypt the ciphertext c
 - 1.1 $x = (((c^\lambda \bmod n^2) * \mu \bmod n) - 1) \bmod n$
 2. Return the decrypted value of c to the user, which is x
-

3.2 RSA

Rivest, Adleman, and Dertouzos introduced the concept of “privacy (partial) homomorphism” in 1978 (Rivest et al. 1978). This concept allowed multiplicative arithmetic, but not additive arithmetic, to be performed on encrypted data.

Algorithm 4. RSA key generation algorithm

Input Data: inputs p , q , n , $\varphi(n)$, e , d

Result: Encrypted key e and decryption key d

1. Randomly generate the encryption and decryption keys
 - 1.1 Randomly generate large prime numbers p and q
 - 1.2 computer $n = p * q$
 - 1.3 find $\varphi(n) = (p - 1) (q - 1)$
 - 1.4 Randomly generate e such that $1 < e < \varphi(n)$ and $\gcd(e, \varphi(n)) = 1$
 - 1.5 Find $d = e^{-1} \bmod \varphi(n)$
 2. Return public key e and private key d
-

Algorithm 5. RSA encryption algorithm

Input Data: Encryption key and an input number x ;

Result: An encrypted value $c \in \mathfrak{R}$;

1. Encrypt the input x
 - 1.1 $c = x^e \bmod n$
 2. Return the encrypted value of x which is c to the user.
-

Algorithm 6. RSA decryption algorithm

Input Data: A decryption key d and an encrypted value c

Result: A decrypted ciphertext c , which is x

1. Decrypt the input c
 - 1.1 $x = c^d \bmod n$
 2. Return the decrypted value of c to the user, which is x
-

3.3 Gentry

In Gentry’s 2009 thesis, a fully homomorphic encryption algorithm is defined which provides an unlimited number of additions and multiplications to be performed on encrypted data (Gentry 2009a). A review of Gentry’s homomorphic algorithm by Ryan Howard in 2013 showed the algorithm to take seconds for addition and subtraction, minutes for multiplication, and hours for division, proving to be cost and time prohibitive (Howard 2013).

3.4 Warren

Van Warren proposed a method of encryption using homomorphic shape encryption. In his method, a private key is used to perform arithmetic including addition, subtraction, multiplication, and division (Warren 2016). This method allows for addition and subtraction of encrypted values along with scalar multiplication and division of encrypted values. Warren's description is the inspiration for the N-tier rotation encryption scheme.

4 Approach

The inspiration for the N-tier rotation scheme comes from a combination lock. Such a lock uses a sequence of numbers in succession to open the lock. The sequence of numbers is entered by rotating the dial on the lock in the correct direction and order (Fig. 1).

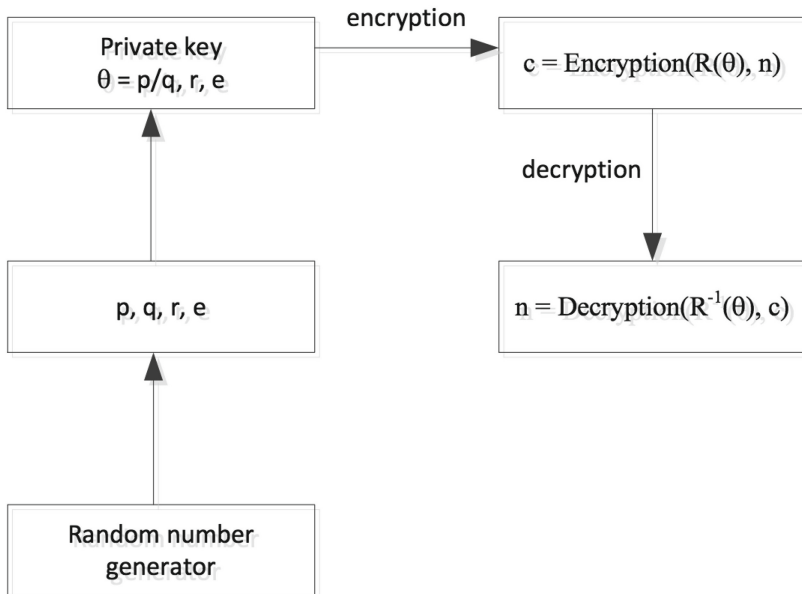


Fig. 1. Encryption algorithm

4.1 Algorithm

The N-tier rotation-based encryption algorithm execute the following steps for encryption and decryption in the following figure:

However, this algorithm only provides one level of depth of security and is not the most secure. Given enough time, one would be able to ascertain what the original numbers were by shifting them back to their original locations. As such, an “N-tier” approach is necessary. The N-tier algorithm uses multiple private keys to encrypt the data, providing multiple levels of security depth. In the above figure, the number e controls the number of tiers for the

encryption, or depth. When multiple rotations are used, different values of θ will be generated for each individual rotation to use. Each θ is found such that $\theta_i = p_i/q_i$ for the encryption step of each c_i where $1 \leq i \leq e$, where each p and q is randomly generated. The encrypted value of $c_i = \text{Encryption}(R(\theta_i), c_i)$. The algorithm for N-tier rotation based encryption is shown below in Algorithm 1. The decryption is shown below in Algorithm 2.

Algorithm 7. The N-tier rotation based encryption algorithm

Input Data: An encryption key (p, q, r, e) and an input number $n \in \mathfrak{R}$;

Result: An encrypted value $c \in \mathfrak{R}$;

1. Randomly generate the key values (p, q, r, e) :

1.1 Randomly generate e for the number of rotations and $e \in \mathfrak{R}$;

1.2 Randomly generate private key $(p_i, q_i) = \begin{bmatrix} p_1 & q_1 \\ \dots & \dots \\ p_e & q_e \end{bmatrix} \Rightarrow [\dots]$ where $\theta_i = p_i/q_i$

and $1 \leq i \leq e$ and p_i and $q_i \in \mathfrak{R}$;

1.3 Randomly generate private key $r_i = \begin{bmatrix} r_1 \\ \dots \\ r_e \end{bmatrix}$ where $1 \leq i \leq e$ and $r_i \in \mathfrak{R}$;

2. Compute the encrypted value of the input number n in $\begin{bmatrix} n \\ r_1 \end{bmatrix}$;

For $k = 1$ to e Do

$$\begin{bmatrix} c_{k+1} \\ m_{k+1} \end{bmatrix} = R_k(\theta_k) \times \begin{bmatrix} c_k \\ m_k \end{bmatrix} \text{ where } c_1 = n \text{ and } m_k = r_k \text{ and } 1 \leq k \leq e;$$

3. Return the encrypted value of n which is c_{e+1} to the user.

Algorithm 8. The N-tier rotation based decryption algorithm

Input Data: A decryption key (p, q, r, e, d) and an encrypted value $c \in \mathfrak{R}$;

1. Retrieve the value of e

2. Retrieve the key (p_i, q_i) where $\begin{bmatrix} p_1 & q_1 \\ \dots & \dots \\ p_e & q_e \end{bmatrix} \Rightarrow [\dots]$ where $\theta_i = p_i/q_i$ and $1 \leq i \leq e$;

3. Retrieve the value of $r_i = \begin{bmatrix} r_1 \\ \dots \\ r_e \end{bmatrix}$ where $1 \leq i \leq e$ and $r_i \in \mathfrak{R}$;

4. Retrieve the value d ;

5. $c_e = c$;

6. For $k = e$ to 1 Do

6.1 Compute $m_k = ((d \times r_k) - [R_k^{-1}(\theta_k)]_{2,1} \times c_k) / [R_k^{-1}(\theta_k)]_{2,2}$;

6.2 Compute $c_{k-1} = ([R_k^{-1}(\theta_k)]_{1,1} \times c_k) + ([R_k^{-1}(\theta_k)]_{1,2} \times m_k)$;

7. Return the decrypted value of c_1 to the user

4.2 Decision for Java

Java was chosen as the implementation programming language for two reasons: The BigDecimal class and the potential to export the implementation to other systems easily. First, the BigDecimal class included in Java is a tool which allows numbers of any size to be used with an arbitrarily decided level of precision. Second, this algorithm is intended to be used by anyone, so it is necessary that it run on many platforms. As such, Java was an obvious choice due to easy cross-platform capabilities.

5 Implementation

Encryption and decryption were done via two classes: A node and a rotator. Each node contained a member “data” and a member “d.” Data was the actual data of the node instance and d was the mentioned d value from the encryption scheme. The d value outlined in the encryption scheme is conveniently kept inside the node. For reference, the d value keeps track of the number of operations done to the node in the following manner: for n many additions, $d = d + n$, for n many subtractions, $d = d - n$, for n many multiplications, $d = d * n$, for n many divisions, $d = d / n$. Both variables were of type BigDecimal. The rotator for this implementation included 5 randomly generated keys, and 5 randomly generated r values, enabling 5 rotations or tiers or security, although the rotator supported other number of tiers. The rotator then could perform encryption and decryption of a particular node.

It should be noted that the rotator in this implementation is dependent on the number of digits in the data of the nodes. For example, if the nodes have data that are five digits long, the rotator should have keys reasonably similar in size. If the node has data that is fifty digits long, the rotator should, again, have keys that are similar in size. Failure to do so would lead to sporadic decryption behavior, which is discussed later (Fig. 2).

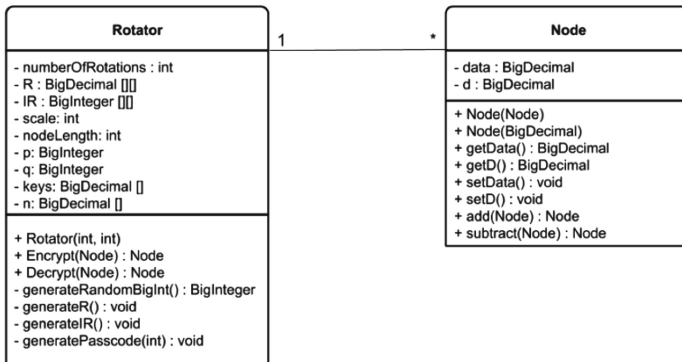
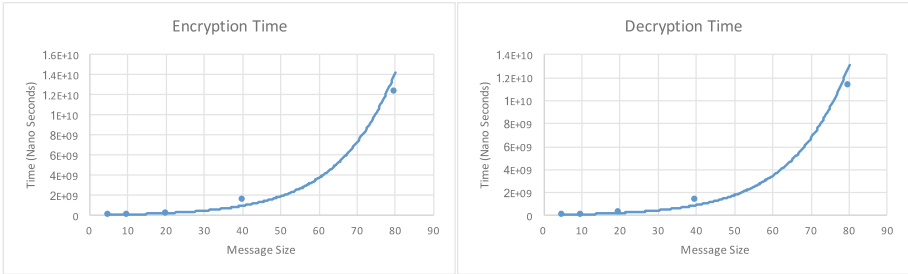


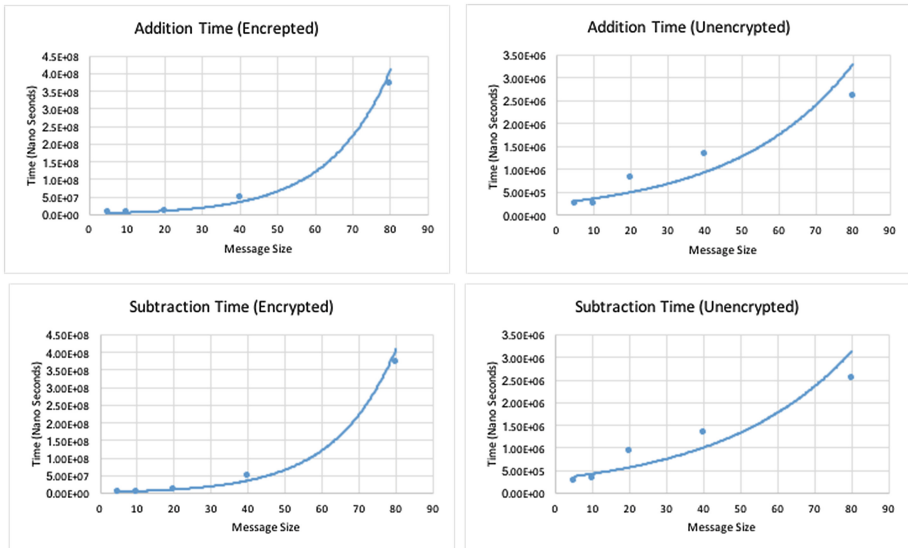
Fig. 2. Class diagram

6 Performance Analysis

Performance is limited by the overall size of the message in the node, along with the number of tiers, or depth, for security. The longer the node and/or the more depth, the slower the program runs. For example, if there is only 5 digits in the node's data, the encryption process is fairly quick, as shown below. Alternatively, if the data had 80 digits, the program took an order of magnitude longer to encrypt and decrypt, also shown below. Encryption and decryption are roughly similar in runtime.



For addition and subtraction, there was not a notable speed difference between addition and subtraction of encrypted messages versus the unencrypted nodes. The encrypted values took longer than unencrypted messages and this is likely due to the messages becoming much longer, or noisy, after the encryption process to ensure they decrypt correctly. In this context, noise in the data is the extra digits added to the head and tail of a node's data when encrypted. This is useful in that it makes the encrypted data difficult to decipher since it is not at all like its original self. However, the consequence of this is that it takes more time to manipulate the encrypted data versus the unencrypted data.



6.1 Limitations

Within the scope of this program, there is an apparent limitation of the `BigDecimal` class at messages of size 100 or more. The encryption and decryption process leads to characteristics of overflow should the program have been implemented using `double` or `float` types with smaller messages. These characteristics manifest as large numbers which will decrypt to incorrect messages which usually vary from the original message by a few numbers.

The rotator implementation requires the keys for encryption and decryption to be sufficiently large, yet not too large, for encryption and decryption to be successful. For an analogy, imagine a player in a soccer field with a ball. In this context, the player is the rotator and the ball is the data. When they are the right size, the player can kick or “encrypt” the ball. Should the ball be very small, say the size of a pea and the player normal size, the player would smash the pea when trying to kick it. If the ball were very large, say its diameter is equal in length to the field’s width, the player would be unable to move the ball. In this implementation, data would be checked to see if its size was sufficiently large, sixteen digits as mentioned earlier, and make keys of that size. If the data size was small, keys of size sixteen would be used. Should the key not be made sufficiently large, or be too large, encryption and decryption behavior would be sporadic and unreliable.

6.2 Drawbacks: Analysis of Overflow and Effectiveness

One of the significant drawbacks of this implementation is the memory space requirement of `BigDecimal`. By design, `BigDecimal` is intended to fill available system memory in order to store very large numbers. As a result, the system can quickly become full and be incapable of storing information correctly. For numbers to be able to encrypt and decrypt correctly, they need to be able to store very large encrypted numbers. So, while `BigDecimal` can accurately store these encrypted numbers by design, the limitations of the hardware to store them quickly come into effect.

The memory space problem of `BigDecimal` leads to the next major drawback of this implementation. With multiple rotations or numbers with very large, such as numbers with 100 or more digits, the runtime and memory space of the algorithm becomes very large. As such, the program will lose precision when encrypting very large numbers and will decrypt them incorrectly, as discussed in the limitations section.

One more drawback of this current implementation, and the algorithm in general, is the significant rise in processing time for very large numbers. It is discussed in later sections about how this algorithm will need to be improved to accommodate large numbers in a timely manner.

Despite these drawbacks, this algorithm in its current state is still applicable for a database situation. Even at 80 digit numbers, this program could still accurately decrypt computed results to their correct value. To put this in perspective, the lowest 80-digit number is 1×10^{80} . The Gross World Product is estimated to be \$73.7 trillion (CIA 2016), which is a mere 7.37×10^{12} . As such, this current implementation could comfortably handle real world situations.

7 Conclusions

As it currently stands, this encryption scheme is useful in the appropriate scenario. It currently needs an application programming interface such that it could support an input of any size. Such an implementation would allow long strings of text to be converted to their corresponding ASCII value and then be stored and encrypted in the N-tier encryption scheme. Furthermore, such an implementation would make for a simple application programming interface so future users could integrate the encryption scheme into their own programs and databases.

The first major goal of this encryption scheme for future research is a programming implementation to support any size input. This would allow for any particularly large number or a large number created by concatenating many strings into a single input via their ASCII codes. This process will require either more advanced Java programming implementations or further study about the nature of the relationship between encryption keys to the inputs to be encrypted, such as the requirements that rotation keys and input be similar in size. Future research on this algorithm should find the optimal key length for the rotator class given the number of digits in a particular input.

Next, it is necessary to provide a mathematical proof for the security of the algorithm to ensure its security before further proceedings. Such a proof would ensure the algorithm is a viable solution for both personal and enterprise levels of security.

Finally, one final item to be addressed is developing a tool that could be used in conjunction with existing SQL submission and retrieval tools, such as MySQL Workbench. This tool is the compatibility layer defined in the Applications section of this report. It would need to be designed in such a way that users would have to make no changes to their current workflow, but their new queries would be over an encrypted database.

Acknowledgments. We would like to make a special thanks to the Arkansas Department of Higher Education for providing funding for this research via the Student Undergraduate Research Fellowship. We would also like to thank the University of Arkansas at Little Rock for providing resources and a location for working.

References

- Barthelemy, L.: A brief survey of fully homomorphic encryption, computing on encrypted data, Quarkslab's blog (2016). <http://blog.quarkslab.com/a-brief-survey-of-fully-homomorphic-encryption-computing-on-encrypted-data.html>
- CIA, the world factbook, world (2016). <https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html>
- Gentry, C.: A fully homomorphic encryption scheme, Ph.D. Dissertation, Department of Computer Science, Stanford University, September 2009a
- Gentry, C.: Fully homomorphic encryption using ideal lattices. In: STOC 2009, Proceedings of the 41st Annual ACM Symposium on Theory of Computing, pp. 169–178 (2009b)
- Gentry, C., Halevi, S., Smart, N.P.: Homomorphic evaluation of the AES circuit. In: Cryptography ePrint Archives, Report 2012/099 (2012)

- Howard, R: Parallel processing of fully homomorphic encryption for a cloud environment. In: A thesis Submitted to the Graduate School of University of Arkansas at Little Rock, p. 38, May 2013
- Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the 18th EUROCRYPT, pp. 223–238, Prague, Czech Republic, May 1999
- Pfleeger, C., Pfleeger, S.: Security in Computing. Prentice Hall PTR, Upper Saddle River (2006)
- Lauter, K., Naehrig, M., Vaikuntanathan, V.: Can homomorphic encryption be practical? In: Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop, pp. 113–124 (2011)
- Rivest, R.L., Adleman, L., Dertouzos, M.L.: On data banks and privacy homomorphism. Found. Secure Comput. **4**(11), 169–180 (1978)
- Warren, L.V.: AddSubMulDivia: Volume 5 – Homomorphic Shape Encryption, Kindle edn. Amazon Digital Services LLC, Seattle (2016)
- Wooldridge, Nivens, Chiang: Securing databases with a combination of encryption and information hiding. Computer Science Department at the University of Arkansas at Little Rock (2016)

k-Depth Mimicry Attack to Secretly Embed Shellcode into PDF Files

Jaewoo Park and Hyoungshick Kim^(✉)

Department of Software, Sungkyunkwan University,
Seobu-ro, Suwon 2066, Republic of Korea
{bluereaper,hyoung}@skku.edu

Abstract. This paper revisits the shellcode embedding problem for PDF files. We found that a popularly used shellcode embedding technique called *reverse mimicry* attack has not been shown to be effective against well-trained state-of-the-art detectors. To overcome the limitation of the *reverse mimicry* method against existing shellcode detectors, we extend the idea of *reverse mimicry* attack to a more generalized one by applying the *k*-depth mimicry method to PDF files. We implement a proof-of-concept tool for the *k*-depth mimicry attack and show its feasibility by generating shellcode-embedded PDF files to evade the best known shellcode detector (PDFrate) with three classifiers. The experimental results show that all tested classifiers failed to effectively detect the shellcode embedded by the *k*-depth mimicry method when $k \geq 20$.

Keywords: Security · Malware · PDF · Shellcode · Mimicry attack

1 Introduction

Portable Document Format (PDF) based on “ISO 32000-1:2008 14.8” [7], originally created by Adobe [5], has become the de-facto standard for document files in the web environment, which accounts for over 75% of all online document files [8]. Naturally, the popularity of PDF files made them attractive targets [13] for cyber criminals who want to distribute malware through online document files. Many previous studies have tried to develop efficient methods to embed shellcode into a PDF document and defense mechanisms [6,9,12] to detect the embedded shellcode in PDF documents.

Mimicry attack [17] is a widely used general concept to cloak an attacker’s intrusion to avoid detection by security solutions. Mimicry attack has also been applied to secretly embed shellcode into a document file (e.g., PDF and Microsoft Office files) against shellcode detection techniques by blending shellcode with *normal* document objects. Maiorca [9] particularly proposed an automatic shellcode embedding technique called *reverse mimicry* by hiding a given malicious code as an object linked to the parent object through an *indirect reference*. This technique might be effective against the detectors based on PDF structure analysis (e.g., PJSscan [1]). However, the effectiveness of *reverse mimicry* is

still questionable against the-state-of-art detection technologies (e.g., **Malware Slayer** [2] and **PDFrate** [3]) that were developed based on machine learning to statistically distinguish the patterns of malicious PDF documents from those of benign PDF documents. In this paper, we show that the *reverse mimicry* method is not effective against **PDFrate**. Instead, we extend the idea of conventional *reverse mimicry* attack into a more generalized attack method against PDF files called *k*-depth mimicry attack by using *k* *indirect references* repeatedly. Through the *k*-depth mimicry attack method, we show the clear limitation of existing PDF shellcode detection tools. Our key contributions are as follows:

- **Design of a new PDF shellcode embedding technique.** We developed a new PDF shellcode embedding technique called *k*-depth mimicry attack, which can be used to secretly embed a JavaScript code into a PDF file.
- **Evaluation of *k*-depth mimicry attack against PDF shellcode detection tools.** We showed the effectiveness of the *k*-depth mimicry attack through experiments with popularly used PDF shellcode detection tool (**PDFrate**). Our experimental results demonstrate that the *k*-depth mimicry attack is significantly more effective in hiding the embedded shellcode in a PDF file than the conventional PDF shellcode embedding technique such as “reverse mimicry” [9].

The rest of this paper is organized as follows. In Sect. 2, we provide the background information about PDF files to understand the shellcode embedding attack and defense techniques. In Sect. 3, we present *k*-depth mimicry attack. In Sect. 4, we evaluate the effectiveness of *k*-depth mimicry attack compared with the performance of traditional mimicry attack. Finally, we conclude in Sect. 5.

2 Structure of PDF

In this section, we provide the basic information about the structure of PDF file. In general, a PDF file consists of a sequence of *objects* (e.g., fonts, pages and images) representing components of a document [4]. The description of PDF objects is explained in the ISO 32000-1 standard [7].

2.1 Object, Categorized by Types

Basically, objects in a PDF document are categorized into the following types:

- **Array.** This object is an elements list, enclosed in square brackets (i.e., [~]). This object is generally used to store the reference information of multiple objects.
- **Boolean.** This object represents a logical value (i.e., TRUE or FALSE).
- **Dictionary.** This object contains one or more entries that are enclosed in double angle brackets (i.e., << ~ >>). This is a collection of key and value pairs. Each key is always a name object while the value may be any other type of object, including another dictionary or even null. The dictionary object is one of the most common objects in PDF files.

- **Name**. This object is a sequence of characters starting with a slash character “/” for a fixed value set. This is mainly used for an entry’s attribute set such as “/Name”, “/Type” and “/Filter”.
- **Null**. The string `null` represents an empty object.
- **Numeric**. This object represents an integer or a real number.
- **Stream**. This object is a sequence of bytes to store large blobs of data that are in some other standardized format, such as XML grammars, font files, and image data. This starts with a specific string of `0x73747265616D` to indicate the start position of a stream object, and ends with another specific string of `0x656E6473747265616D` to indicate the end position of the stream object.
- **String**. This object represents a text string, which is enclosed in parentheses or angle brackets (i.e., (\sim) , $\langle \sim \rangle$).

In general, each object could have attributes called **entry** to specify the properties of the object. There are many different types of entries. **OpenAction** and **Annots** are popularly used. **OpenAction** entry is used to define an action to be taken after opening a PDF file. **Annots** is used to indirectly refer to another object such as text note, link, or embedded file. Naturally, those entries could be misused for injecting shellcode.

Objects could be labeled so that they can be referred to by other objects. A labeled object is called an *indirect object*.

2.2 File Structure

In general, a PDF file consists of four primary sections: header, body, Cross-reference table and Trailer. The header section contains the basic information (e.g., format version) about PDF file. The body section consists of a set of objects. The cross-reference (also known as ‘Xref’) table section is a large dictionary for the indexes by which all of the indirect objects, in the PDF file, are located. The trailer section contains the offset and object number information of **root**, the Xref’s starting point offset, etc. When these irremovable parts are complete, End-of-File marker “`%%EOF`” is used to indicate the end of a file.

2.3 Document Structure

In a PDF file, all objects are organized into a tree structure. This stems from the primary objects (i.e., **root** or **catalog**). A PDF document consists of objects contained in the body section of the PDF file. Each page of the document is represented by a **page** object, which is a dictionary that includes references to the page’s contents. Page objects are connected together and form a page tree, which is declared with an indirect reference in the document **catalog**. Each **page** object has its children objects called **Kids** that specifies all the children objects directly accessible from the current node. A PDF viewer application generally parses this tree structure of objects and renders the objects visible to the user.

2.4 Common Shellcode Types

Shellcode is a sequence of machine language instructions that can be injected into a running application. For PDF files, three different file types have been popularly used as shellcode that can be embedded inside a PDF file: (1) an executable file, (2) another PDF file and a JavaScript code.

3 k-depth Mimicry Attack

In this paper, we introduce a new shellcode injection method for PDF files called “k-depth mimicry” attack. We designed this sophisticated attack to secretly hide shellcode against existing shellcode detection techniques.

In a shellcode injection method, an exploit is crafted to fool the targeted application (e.g., Microsoft Office and Adobe PDF reader) into executing a malicious code, which is hidden within a document (e.g., office file) as shellcode. Several defense solutions [11, 14, 18] have been developed to prevent the shellcode injection by analyzing the difference between malware and normal applications because this technique was very popularly used. However, many advanced shellcode injection techniques have also been proposed to make it more difficult to detect the presence of shellcode [10]. In a PDF file, the indirect reference could be used to avoid shellcode detection methods, by scattering the embedded shellcode across objects [15]. Figure 1 shows how the reverse mimicry attack works with an indirect reference to a JavaScript code within the PDF document.

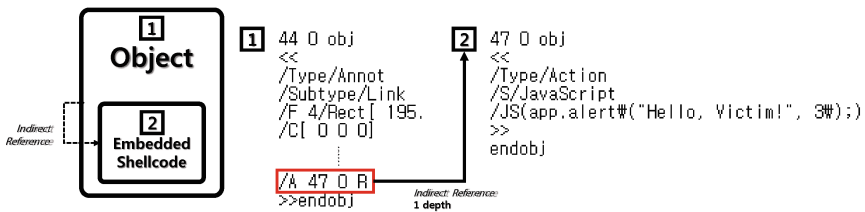


Fig. 1. Overview of reverse mimicry.

As depicted in Fig. 1, the reverse mimicry method modifies an object to have an indirect reference link to another object. The indirectly linked object is the carrier of the embedded shellcode. In the example of Fig. 1, the reverse mimicry method uses a seemingly harmless object as an intermediate, hiding the malicious shellcode underneath that object with an indirect reference to the “47 0 obj” object. When the “44 0 obj” object is rendered to show its contents, its indirect reference object “47 0 obj” would be sequentially launched through the object “44 0 obj”.

We extend the conventional reverse mimicry attack into a more generalized one by hiding shellcode with an indirect object with depth k. We call this technique “k-depth mimicry attack”. Here, depth is defined as the number of indirect

references from a referencing object to a referenced object. In the reverse mimicry method, one indirection reference is only used to hide the malicious shellcode, and thus, reverse mimicry can be regarded as the “1-depth” mimicry.

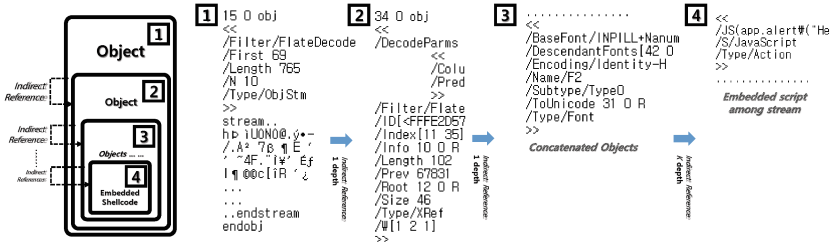


Fig. 2. Overview of k -depth mimicry.

Figure 2 shows how the k -depth mimicry attack works. As depicted in the Fig. 2, basically, each object has an indirect reference to the next object. In this chain of indirect references, the last object is the carrier with shellcode. We note that the main idea of the k -depth mimicry attack is to use multiple layers of indirect references to make it harder to extract the injected shellcode.

In the example of Fig. 2, “15 0 obj” object is indirectly linked to “34 0 obj” object, “34 0 obj” holds the indirect reference to another object, and so on. After k times of such indirect references among objects (e.g., $k = 4$ in Fig. 2), the last object with the JavaScript shellcode can be reached. This concatenation is launched when the PDF document is opened and its contents are rendered. The k -depth mimicry is similar to the reverse mimicry method except that k subsequent indirect references are used. However, we found that it is much harder to detect the presence of shellcode in practice when a reasonable number k of indirect references is used for hiding the shellcode.

In the following section, we will show that the k -depth mimicry attack is more effective than the reverse mimicry attack against even machine-learning based detectors through several experiments.

4 Experiments

For experiments, we created a normal PDF file (benign) using the k -depth mimicry technique to embed a JavaScript code (as shellcode) into the PDF file and then analyzed the file with the PDF shellcode detector called PDFrate [3] that is one of the most popular tools for malicious PDF detectors. PDFrate has achieved a high detection rate and a low false positive rate. Moreover, its performance has improved over time because it is publicly available and its training datasets are continuously updated by online users.

To find the optimal k for the k -depth mimicry method, we created several test PDF files with varying depth k and analyzed their classification scores

in PDFrate. We used the three different classifiers deployed by PDFrate (i.e., the classifiers trained on the **Contagio**, George Mason University (GMU), and PDFrate community (**Community**) datasets [16]). Figure 3 shows how these scores are changed with k .

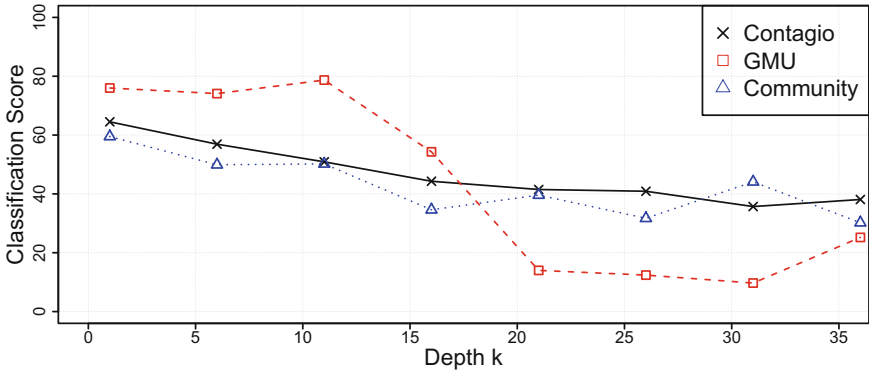


Fig. 3. Changes in the classification score with depth k .

From this figure, as expected, the performance of all classification scores overall reduced as k decreased. The classification scores of GMU dramatically decreased from $k = 10$ to 20 while the score curves of the other classifiers commonly had a gentle slope. Based on those experimental results, we recommend using a reasonably large $k \geq 20$ for the k -depth mimicry method to significantly reduce the classification scores of PDF files’ maliciousness.

To show the effectiveness of the k -depth mimicry attack on PDF file, we created the PDF file with the depth of 21 and compared its maliciousness on PDFrate with the PDF file created by the reverse mimicry technique. Table 1 shows the experimental results.

Table 1. Analysis results of PDFrate with Benign PDF, PDF by reverse mimicry, PDF by k -depth mimicry (Each percentage in parenthesis indicates a classification score of the tested PDF file’s maliciousness).

Classifier	Benign PDF	PDF by reverse mimicry	PDF by k -depth mimicry
Contagio	-(20.4%)	Detected (50.9%)	Evaded (41.5%)
GMU	-(5.6%)	Detected (78.7%)	Evaded (14.0%)
Community	-(27.6%)	Detected (50.2%)	Evaded (39.6%)

In Table 1, we can see that all those classifiers failed to successfully detect the shellcode embedded by the k -depth mimicry technique. Even though the classification scores returned by PDFrate for the PDF file by the k -depth mimicry

technique were overall higher than those scores for the benign PDF file with no JavaScript code, the *k*-depth mimicry attacks significantly reduce the classification scores of PDFrate from the mean of about 59.9% for the reverse mimicry technique to the mean of about 31.7%. Those experimental results demonstrate that the *k*-depth mimicry technique is practically effective for hiding shellcode against existing PDF shellcode detectors such as PDFrate.

To make matters worse, even for the latest version of Adobe Reader DC (v15.023.20053), the PDF document containing the JavaScript embedded by the *k*-depth mimicry method was successfully opened with no security warning whereas the PDF document was not opened when the reverse mimicry method was used to embed the same JavaScript code (See Fig. 4).

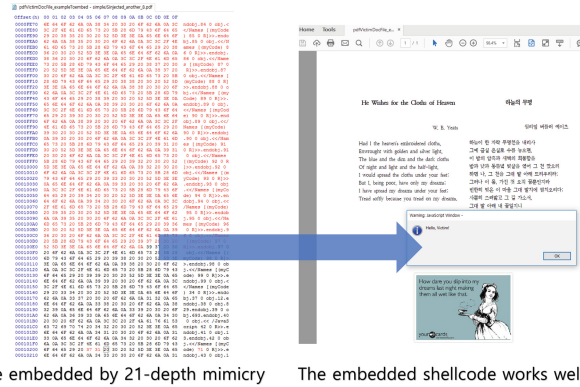


Fig. 4. Example of the embedded shellcode using the *k*-depth mimicry technique (with 21-depth).

5 Conclusion

We presented a novel shellcode embedding technique for PDF documents to successfully bypass PDF shellcode detectors. Our experimental results showed the proposed technique can be used to easily create a PDF file with shellcode against the state-of-the-art detectors such as PDFrate. However, our current results are not enough to generalize our observations because a single PDF file was tested. Therefore, as an extension to this paper, we plan to conduct intensive experiments with a large sample of PDF files.

In future work, we plan to develop defense mechanisms so as to make the *k*-depth mimicry technique ineffective. It would also be interesting to evaluate the performance of the defense mechanisms on real malicious PDF files.

Acknowledgement. This work was supported in part by the NRF Korea (No. 2014R1A1A1003707), the ITRC (IITP-2016-R0992-16-1006), and the MSIP/IITP (No. R0166-15-1041, R-20160222-002755). The first author’s research was mainly funded by the MSIP, under the “Employment Contract based Master’s Degree Program for Information Security” (H2101-16-1001) supervised by KISA. The contents of this article do not necessarily express the views of KISA.

References

1. PJSscan (2013). <https://sourceforge.net/p/pjscan/home/Home>
2. Malware Slayer (2014). <https://pralab.diee.unica.it/en/Slayer>
3. PDFrate (2016). <https://csmutz.com/pdfrate>
4. Adobe Systems Incorporated: PDF reference-adobe portable document format (2006). https://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf
5. Adobe Systems Incorporated: What is PDF? (2016). <https://acrobat.adobe.com/kr/ko/why-adobe/about-adobe-pdf.html>
6. Fratantonio, Y., Kruegel, C., Vigna, G.: Shellzler: a tool for the dynamic analysis of malicious shellcode. In: Proceedings of International Workshop on Recent Advances in Intrusion Detection (2011)
7. International Organization for Standardization: PDF (portable document format), version 1.7, base level (ISO 32000-1: 2008) (2008). <http://www.digitalpreservation.gov/formats/fdd/fdd000277.shtml>
8. Johnson, D.: PDF still dominates electronic documents online (2015). <http://duff-johnson.com/2015/10/07/pdf-still-dominates-electronic-documents-online>
9. Maiorca, D., Corona, I., Giacinto, G.: Looking at the bag is not enough to find the bomb: an evasion of structural methods for malicious PDF files detection. In: Proceedings of the 8th ACM Symposium on Information, Computer and Communications Security (2013)
10. Mason, J., Small, S., Monroe, F., MacManus, G.: English shellcode. In: Proceedings of the 16th ACM Symposium on Information, Computer and Communications Security (2009)
11. Polychronakis, M., Anagnostakis, K.G., Markatos, E.P.: Emulation-based detection of non-self-contained polymorphic shellcode. In: Kruegel, C., Lippmann, R., Clark, A. (eds.) RAID 2007. LNCS, vol. 4637, pp. 87–106. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-74320-0_5](https://doi.org/10.1007/978-3-540-74320-0_5)
12. Schmitt, F., Gassen, J., Gerhards-Padilla, E.: PDF scrutinizer: detecting javascript-based attacks in PDF documents. In: Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (2012)
13. Symantec: ISTR: Internet security threat report. In: Trend Report, vol. 21 (2016)
14. Toth, T., Kruegel, C.: Accurate buffer overflow detection via abstract pay load execution. In: Wespi, A., Vigna, G., Deri, L. (eds.) RAID 2002. LNCS, vol. 2516, pp. 274–291. Springer, Heidelberg (2002). doi:[10.1007/3-540-36084-0_15](https://doi.org/10.1007/3-540-36084-0_15)
15. Tzermias, Z., Sykiotakis, G., Polychronakis, M., Markatos, E.P.: Combining static and dynamic analysis for the detection of malicious documents. In: Proceedings of the 4th European Workshop on System Security (2011)
16. Šrđić, N., Laskov, P.: Practical evasion of a learning-based classifier: a case study. In: Proceedings of the IEEE Symposium on Security and Privacy (2014)
17. Wagner, D., Soto, P.: Mimicry attacks on host-based intrusion detection systems. In: Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security (2002)
18. Zhang, Q., Reeves, D.S., Ning, P., Iyer, S.P.: Analyzing network traffic to detect self-decrypting exploit code. In: Proceedings of the 2nd ACM Symposium on Information, Computer and Communications Security (2007)

Reconstruction of Task Lists from Android Applications

Xingmin Cui¹(✉), Ruiyi He¹, Lucas C.K. Hui¹, S.M. Yiu¹, Gang Zhou²,
and Eric Ke Wang³

¹ Department of Computer Science,
The University of Hong Kong, Pokfulam, Hong Kong
{xmcui,hui,smyiu}@cs.hku.hk, reillyhekazusa@gmail.com

² Peking University, Beijing, China
garretzhou@163.com

³ Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
wk_hit@hit.edu.cn

Abstract. The popularity of Android devices has made Android apps attractive targets for attackers. Some static checkers have been proposed to check whether an Android app is vulnerable to privacy leakage and other attacks. However, these checkers model the control flows in the app following the ICC events, ignoring the intrinsic purpose of users' interaction with mobile devices. In fact, users carry out various tasks using mobile apps, e.g. online shopping. An Android *task* consists of one or more Activities, which are organized in the back stack of the task. By extracting the task lists among Activities in Android apps, we can capture all control flow transitions between them, including those bring by ICC events and back button events. We design and implement a system, which leverages the combination of static and dynamic analysis to extract the task lists. Our system can be used to detect task related attacks and help static checkers construct more complete call graphs.

Keywords: Android applications · Task · Program analysis

1 Introduction

Android devices are widely used to handle private data; therefore, they have become attractive targets for malicious attackers. The security of Android apps has drawn the attention of researchers from both the academy and industry.

From the users' point of view, an Android app is used to fulfill tasks, such as on-line shopping and instant messaging. Therefore in practice users interact with the app based on the **Task** conception [1]. Among the four types of components in an Android app (Activity, Service, Broadcast Receiver, and Content Provider), users can only interact with Activities. Other components are launched by Activities directly or indirectly. Each Activity is designed around a specific action the user can perform or start other Activities. A *task* is a collection of Activities that users interact with when performing a certain job.

The activities are arranged in a stack (the *back stack*), in the order in which each activity is opened.

Control flow transitions among Android Activities are useful and prevalent. We summarize that the control flow transitions among Android components mainly result from the following two reasons:

1. Inter-component communication using ICC methods (e.g. `startActivity`, `startService`, etc.). Android allows components from the same application or different applications to interact with each other using Inter-Component Communication (ICC) methods. Although ICC enables function reuse and application cooperation, it also opens the door for malicious apps to attack vulnerable apps [10].
2. Back button events that are invoked when users press the back button. In this case, which activity will come to the foreground when the user presses the back button depends on which *task* this activity resides in and how the *back stack* of this task is organized.

Suppose Activity A starts Activity B using the ICC method `startActivity`. The control flow transits from A to B. When the user presses the *Back* button, the control flow transits back to A, along with implicit data flow (e.g. modification of static data in B would result in different behavior in A). Existing static checkers [4, 9] only considered the first type of transitions. Works dedicated to inter-component communication such as Epicc [7] also only consider control flow transition using ICC methods by spotting these methods and resolving their parameters to get the intent sender and receiver components. However, by ignoring the second type of transitions, we cannot gain a complete picture of inter-component communication, therefore fail to discover certain vulnerabilities (see our motivating example in Sect. 2.2).

The notion of *task* represents the essence of user interactions with the mobile device, which in turn determines the control flow and data flow in the app and the system. When activity A starts activity B, B will be added to the same or different task with A (which task it is added to depends on B's configuration in the manifest file, by default it is added to the same task with the activity which starts it). Suppose B and A are in the same task, when the user presses the back button, the `onBackPressed()` method of B will be invoked which destroys B and resumes A since in this case A is the activity next to B in the back stack.

With the concept of *task*, we can capture a more accurate control flow. We design and implement a system to extract the task lists among the Activities in a set of Android apps. The resulted task lists can be used to detect task related attacks and aid static checkers to construct complete call graphs.

The rest of the paper is organized as follows: In Sect. 2, we will give some background knowledge of Android tasks along with the motivating example. The design of the system and its application scenarios will be given in Sects. 3 and 4, respectively. Section 5 concludes this work and gives directions for future work.

2 Background and Motivating Example

2.1 Tasks and Back Stack

A *task* is a collection of activities that users interact with when performing a certain job [1]. The activities are arranged in a stack (the back stack), according to the order in which each activity is opened. When the user starts an application, that application's task comes to the foreground. If no task exists for the application, then a new task is created. Normally the launcher Activity is the root activity of the task and be put at the top of the back stack. If not specified, the task affinity is the package name. When the current activity starts another activity, the new activity will be added to the current task or create a new task depending on its configuration. The developer can determine how to manage the task by manipulating the attributes in the `<activity>` manifest element and the flags in the intent that pass to `startActivity`. By default, the new Activity is pushed on the top of the existing stack and becomes the foreground Activity. When the user presses the *Back* button, the current Activity is popped from the top of the back stack (i.e. the Activity is destroyed) and the previous Activity resumes. If the user continues to press Back, then each Activity in the stack would be popped out until the user returns to the Home screen or to whichever Activity was running when the task began. When all Activities are removed from the stack, the task no longer exists.

A task can move to the background as a cohesive unit when users begin a new task or go to the home screen. When users start the application later, this task can return to the foreground so users can pick up where they left off.

```

1  public class Act1 extends Activity {
2  public static String data;
3  protected void onCreate(Bundle savedInstanceState) {
4  data = "untainted data";
5  Button button1 = (Button) findViewById(R.id.button1);
6  button1.setOnClickListener(new OnClickListener() {
7  public void onClick(View arg0) {
8  Intent intent = new Intent(Act1.this, Act2.class);
9  startActivity(intent);
10 } }); }
11 protected void onResume(Bundle savedInstanceState) {
12 sink(data); }
13 }
14 public class Act2 extends Activity {
15 protected void onCreate(Bundle savedInstanceState) {...}
16 protected void onBackPressed(Bundle savedInstanceState){
17 Act1.data = source(); }
18 }

```

Fig. 1. Motivating example

2.2 Motivating Example

We give an example to show the necessity to consider the control flow transitions bring by both ICC methods and back button events. Taint analysis checks whether tainted variables can reach sink methods and is widely used in static

checkers [4,5]. If any tainted variables reach a sink method, an alarm will be raised. The example in Fig. 1 implements an application which consists of two Activities *Act1* and *Act2*. *Act1* is the launcher Activity. Therefore, when the user starts this app, *Act1* will be launched. *Act1* declares a global variable *data* which is shared between different components. When the user clicks *button1*, *Act2* would come to the foreground and *Act1* will be stored in the back stack next to *Act2*. When user clicks the back button, method *onBackPressed* would be invoked, which makes *data* a tainted variable. After that *Act1* will be resumed from the back stack. The *onResume* method of *Act1* would be invoked and *sink* on line 12 would raise an alarm.

If the checker fails to reconstruct the task $\{Act1 \rightarrow Act2 \rightarrow Act1\}$, it would miss the leak on line 12. To avoid such kind of false negatives, we propose a system to reconstruct all the task lists among Activities in Android apps and capture possible control flow transitions.

3 System Design

3.1 System Overview

Our system takes as input a set of Android apps, and prints out all possible task lists in these apps. It leverages a hybrid of static and dynamic analysis. The static preprocessor extracts the widget information related to ICC invocations to guide the subsequent dynamic runner. The dynamic runner runs the given set of apps dynamically, aiming to get a record of the list of tasks among Activities in these apps. The system overview is given in Fig. 2.

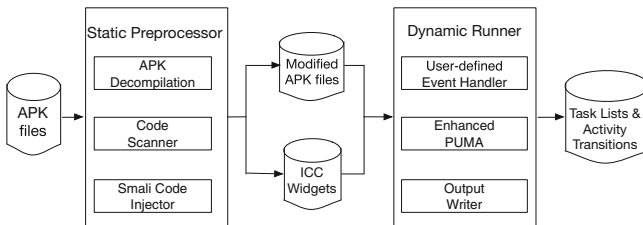


Fig. 2. System overview

3.2 Static Preprocessor

The static preprocessor aims to achieve two goals: Firstly, it records the list of widgets that have registered event listener methods. Secondly, it identifies the inter-component communication methods (i.e. `startActivity`, `startActivityForResult`) and performs backward reachability analysis to find out whether these methods are triggered by a widget event listener method.

If so, the execution path from the event listener method to the ICC method is recorded to guide the dynamic analysis.

During the pre-process, the apk files of Android apps are decompiled to smali files using APK Tool. After that the smali files are passed to the *code scanner* and *smali code injector*. Code scanner scans the smali files to identify ICC methods and extracts the widgets which trigger these methods. Once the scanner comes across an ICC invocation, it will launch the smali code injector to inject log information, including the current Activity and the ICC receiver Activity. Besides, for each layout file, it injects the information of the corresponding Activity into the root view group's content description. The apk file with injected code is then recompiled to a new apk file.

3.3 Dynamic Runner

As introduced in Sect. 2.1, developers can manipulate the task information in various ways. Relying on static analysis to emulate this process is not accurate, therefore we retrieve task lists with the help of dynamic analysis.

After the ICC widget extraction and smali code injection, we can use the dynamic runner to run the recompiled apk file following the guidance of the ICC widget information. Our runner is built on top of PUMA [6], which is programmable automatic tool that runs the tested app following the instructions in the so-called PUMA script.

In PUMA, a state is defined as follows:

$$state = \{Activity_1 \rightarrow Widget_1 : (WidgetState) \rightarrow Widget_2 : (WidgetState) \dots\}$$

It contains the current Activity plus the state of the widgets included in this Activity. A state transition denotes any kind of UI modification to the current state, e.g., inputting text to a TextFiled or clicking a button.

The exploration algorithm in PUMA is DFS (Depth First Search) based. We can define several event handlers to guide the DFS exploration process. One shortcoming of PUMA is that it can only handle one Android app at a time, we improved by allowing it to deal with a set of apps. Our runner provides 8 event handlers to support the exploration of multiple apps. These handlers include *getAppExecutionSequence*, *onNewAppLaunched*, *onReachingDiffApp*, *getClickingPolicy*, *onStateEquivalent*, *getNextClickItem*, *explorationDone*, *onResultAnalysis*. The usage of these event handlers are explained by their names.

The execution sequence of the apps is provided by the user, defined in the configuration file. The *getClickingPolicy* handler allows the user to define the types of widgets (e.g. Button, ImageView) to be explored, aiming to reduce the number of states. If this list is not provided, the runner will use the *Clickable* widgets defined in `AccessibilityService` as the default list.

In our runner, two states are equivalent if and only if all the widget states in them are the same. For the *getNextClickItem* event, we always choose the currently unexplored ICC widget. If user input is required, we decide the data

type based on the attributes of the widgets as in [3]. The content description is used as our first trial. If content description is not provided, we will look for the widget `hint` to predicate the data type. Otherwise, we use the `TextView` which follows the `EditText` to tell what to input. We match the text on the `TextView` with our pre-set key (e.g. phone, number, email, e-mail, etc.) to get the default value (e.g. *test@gmail.com* for email) and use it as the input to the `EditText`. At present, we set the state threshold as 1,500 to terminate the exploration process. The original PUMA fails at system levels above 5.0. We solved this problem and our system can be successfully launched in levels above 2.3. Our tool also gives users the flexibility to define event handlers to guide the state searching process.

After successful exploration of the target apps, we can easily extract the list of tasks represented by control flow transitions among Activities in these apps.

4 Application Scenarios

Our system outputs the possible task lists through Activities among a set of apps, which capture the intrinsic purpose of users' interaction with the mobile device. Our system can be used in several scenarios. For example, the extracted control flow transitions help static analysis tools construct complete call graphs, which reduces false negatives (e.g. the leak in Fig. 1). Besides, it can be used to detect task related attacks. Next we will show how our results can be utilized to detect these attacks.

Ren et al. [8] discovered *task hijacking* attacks, which result from the design flaws of Android multitasking. By launching these attacks, the attacker may steal private information, implement ransomware and spy on user's activities. Their analysis shows that this kind of attack is prevalent. They also show that due to the dynamic nature of Android tasks, statically detecting such attacks is not easy.

We take the UI spoofing attack as an example to show our system can detect task related attacks. We use two apps to illustrate the attack: *com.android.mms*, an app released together with the Android system; and *hku.cs.spoofing*, a home-made malicious app. The task affinity of the main Activity in *hku.cs.spoofing* is set as "android.task.mms", which is the same with the task affinity of *com.android.mms*. Besides, the malicious Activity's `allowTaskReparenting` is set as `true`. In this way, if the malicious app is launched before the *mms* app, the main Activity of *mms* will be put in the existing task created by the malicious app, redirecting the user to the malicious Activity instead of the real *mms* main Activity. The visualized task lists of these two apps are given in Fig. 3.

In Fig. 3, each node represents an Activity. Activities with the same shape belong to the same app. The node with label `-1` represents the system launcher "com.android.launcher". The filled arrowhead gives the direction of ICC event control flow transitions and the block arrowhead gives the direction of back button event control flow transitions. In the *mms* app, the launcher can launch two Activities, including one that does not belong to this app. This indicates the risk of UI spoofing.

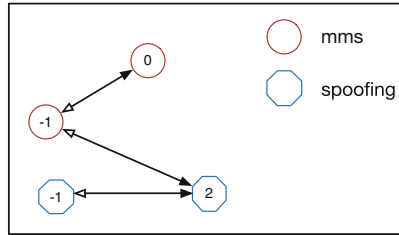


Fig. 3. Visualized task lists

5 Conclusion and Future Work

In this work, we propose a system to extract the task lists among Activities within a set of apps, which captures the intrinsic purpose of users' interaction with the mobile device. We show that our system can detect task related attacks. For future work, we will try to combine our results with existing static checkers [2, 4, 9] to detect vulnerabilities in a set of Android apps.

Acknowledgement. The work described in this paper was partially supported by the HKU Seed Fundings for Applied Research 201409160030; HKU Seed Fundings for Basic Research 201311159149, 201411159122, 201411159142, and 201511159034; National Natural Science Foundation of China (No.61572157), grant No.2016A030313660 from Guangdong Province Natural Science Foundation No.JCYJ20150617155357681, JCYJ20160428092427867 from Shenzhen Municipal Science and Technology Innovation Project.

References

1. Tasks and Back Stack. <http://developer.android.com/guide/components/tasks-and-back-stack.html>
2. Arzt, S., Rasthofer, S., Fritz, C., Bodden, E., Bartel, A., Klein, J., Le Traon, Y., Octeau, D., McDaniel, P.: Flowdroid: precise context, flow, field, object-sensitive and -aware taint analysis for android apps. In: Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation, p. 29. ACM (2014)
3. Carter, P., Mulliner, C., Lindorfer, M., Robertson, W., Kirda, E.: CuriousDroid: automated user interface interaction for android application analysis sandboxes. In: Financial Cryptography and Data Security (FC), February 2016
4. Cui, X., Wang, J., Hui, L.C., Xie, Z., Zeng, T., Yiu, S.: WeChecker: efficient and precise detection of privilege escalation vulnerabilities in android apps. In: Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks, p. 25. ACM (2015)
5. Cui, X., Yu, D., Chan, P., Hui, L.C.K., Yiu, S.M., Qing, S.: CoChecker: detecting capability and sensitive data leaks from component chains in android. In: Susilo, W., Mu, Y. (eds.) ACISP 2014. LNCS, vol. 8544, pp. 446–453. Springer, Cham (2014). doi:10.1007/978-3-319-08344-5_31

6. Hao, S., Liu, B., Nath, S., Halfond, W.G., Govindan, R.: PUMA: programmable UI-automation for large-scale dynamic analysis of mobile apps. In: Proceedings of the 12th Annual International Conference on Mobile Systems (2014)
7. Outeau, D., McDaniel, P., Jha, S., Bartel, A., Boddien, E., Klein, J., Le Traon, Y.: Effective inter-component communication mapping in android with epicc: an essential step towards holistic security analysis. In: Proceedings of the 22nd USENIX Security Symposium (2013)
8. Ren, C., Zhang, Y., Xue, H., Wei, T., Liu, P.: Towards discovering and understanding task hijacking in android. In: Proceedings of the 24th USENIX Security Symposium (2015)
9. Wei, F., Roy, S., Ou, X., Robby.: Amandroid: a precise and general inter-component data flow analysis framework for security vetting of android apps. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS 2014, pp. 1329–1341, New York. ACM (2014)
10. Wu, D., Luo, X., Chang, R.K.: A sink-driven approach to detecting exposed component vulnerabilities in android apps. arXiv preprint [arXiv:1405.6282](https://arxiv.org/abs/1405.6282) (2014)

Design and Evaluation of Chaotic Iterations Based Keyed Hash Function

Zhuosheng Lin¹(✉), Christophe Guyeux², Simin Yu¹, and Qianxue Wang¹

¹ College of Automation, Guangdong University of Technology, Guangzhou, China
zhuoshenglin@163.com, siminyu@163.com, wangqianxue@gdut.edu.cn

² Femto-st Institute, University of Bourgogne Franche-Comté, Besançon, France
christophe.guyeux@univ-fcomte.fr

Abstract. Investigating how to construct a secure hash algorithm needs in-depth study, as various existing hash functions like the MD5 algorithm have recently exposed their security flaws. At the same time, hash function based on chaotic theory has become an emerging research in the field of nonlinear information security. As an extension of our previous research works, a new chaotic iterations keyed hash function is proposed in this article. Chaotic iterations are used both to construct strategies with pseudorandom number generator and to calculate new hash values using classical hash functions. It is shown that, by doing so, it is possible to apply a kind of post-treatment on existing hash algorithms, which preserves their security properties while adding Devaney's chaos. Security performance analysis of such a post-treatment are finally provided.

Keywords: Chaotic iterations · Keyed hash function · Security performance analysis

1 Introduction

A hash function is any function that can be used to map data of arbitrary size to data of fixed size, which has many information-security applications. Rivest designed the first famous hash function called MD4 (Message Digest 4) in 1990, which is based on Merkle-Damgard iterative structure [6]. Later, various hash functions with an improved but similar design have been proposed. The latter encompass the well-known MD5 [7] and SHAs secure hash algorithm series [2]. However, recent researches have shown that security flaws exist too in these widely used standard hash functions. For instance, Lenstra, cooperated with Xiaoyun Wang, forged a digital certificate with different keys [8]. Then they improved the MD5 collision course and constructed an effective certificate [9]. This research result shocked cryptologists.

Some relationships can be emphasized between chaos properties and some targeted aims in cryptology. Thus it may be a good idea to investigate the use of chaos to enrich the design of cryptographic systems. In our previous work, we have proven that discrete chaotic iterations (CIs) produce topological chaos

as described by Devaney [1]. This topological chaos is a well studied framework and we have applied it in hash function, pseudorandom number generation, data hiding, and so on. However, all of them are used separately. In this research work, we intend to combine pseudorandom number generation and hash functions using CIs. Then we will check if this combination can improve the security performance of standard hash functions. More precisely, we will apply chaotic iterations on classical hash functions, adding by doing so provable chaos while preserving security properties like the collision. With such chaos, our desire is to reinforce diffusion and confusion of the inputted hash functions. In the meantime, General Formulation of the Chaotic Iterations (GFCIs) will be introduced and used, to deal with the output of standard hash functions and to construct chaotic strategies.

The remainder of this article is organized as follows. The first next section is devoted to some basic recalls on the general form of chaotic iterations. The third section introduces pseudorandom number generator with CIs. Our CI-based hash function is proposed and reformulated in this section too. Experimental evaluation is shown in the fourth section. This article ends by a conclusion section, in which our research work is summarized.

2 General Formulation of the Chaotic Iterations

In this section, we focus on the general formulation of chaotic iterations. Let us first define some notations. \mathbb{N} is the set of natural (non-negative) numbers. The domain $\mathbb{N}^* = \{1, 2, 3, \dots\}$ is the set of positive integers and $\mathbb{B} = \{0, 1\}$. $\llbracket 1; N \rrbracket = \{1, 2, 3, \dots, N\}$. A sequence which elements belong in $\llbracket 1; N \rrbracket$ is called a strategy. The set of all strategies is denoted by S . S^n denotes the n^{th} term of a sequence S , X_i stands for the i^{th} components of a vector X .

In here a new kind of strategies is introduced, namely a sequence of subsets of $\llbracket 1, N \rrbracket$, that is, a sequence of $\mathcal{P}(\llbracket 1, N \rrbracket)^{\mathbb{N}}$, where $\mathcal{P}(X)$ is for the powerset of the set X (i.e., $Y \in \mathcal{P}(X) \iff Y \subset X$). So we can now change multiple bits between two adjacent outputs, as follows.

The general form of the discrete dynamical system in chaotic iterations is

$$x^0 \in \mathbb{B}^N, (S^n)_{n \in \mathbb{N}} \in \mathcal{P}(\llbracket 1, N \rrbracket)^{\mathbb{N}}$$

$$\forall n \in \mathbb{N}^*, \forall i \in \llbracket 1; N \rrbracket, x_i^n = \begin{cases} x_i^{n-1}, & \text{if } i \notin S^n \\ (f(x^{n-1}))_{S^n}, & \text{if } i \in S^n \end{cases} \quad (1)$$

In other words, at the n^{th} iteration, only the cells whose id is contained into the set S^n are iterated.

Let us now rewrite these general chaotic iterations as usual discrete dynamical system of the form $X^{n+1} = f(X^n)$ on an *ad hoc* metric space. Such a formulation is required in order to study the topological behavior of the system.

Let us introduce the following function:

$$\psi : \llbracket 1; N \rrbracket \times \mathcal{P}(\llbracket 1; N \rrbracket) \longrightarrow \mathbb{B}$$

$$(i, X) \longleftarrow \begin{cases} 0 & \text{if } i \notin X, \\ 1 & \text{if } i \in X. \end{cases} \quad (2)$$

Given a function $f : \mathbb{B}^N \rightarrow \mathbb{B}^N$, define the function:

$$F_f : \mathcal{P}(\llbracket 1; N \rrbracket) \times \mathbb{B}^N \rightarrow \mathbb{B}^N, \tag{3}$$

$$(P, E) \mapsto \left(E_j \cdot \psi(j, P) + f(E)_j \cdot \overline{\psi(j, P)} \right)_{j \in \llbracket 1; N \rrbracket}.$$

Consider the phase space:

$$\mathcal{X} = \mathcal{P}(\llbracket 1; N \rrbracket)^{\mathbb{N}} \times \mathbb{B}^N, \tag{4}$$

and the map defined on \mathcal{X} :

$$G_f(S, E) = (\sigma(S), F_f(i(S), E)), \tag{5}$$

where, in a similar formulation than previously, σ is the *shift* function defined by $\sigma : (S^n)_{n \in \mathbb{N}} \in \mathcal{P}(\llbracket 1; N \rrbracket)^{\mathbb{N}} \rightarrow (S^{n+1})_{n \in \mathbb{N}} \in \mathcal{P}(\llbracket 1; N \rrbracket)^{\mathbb{N}}$ and i is the *initial function* $i : (S^n)_{n \in \mathbb{N}} \in \mathcal{P}(\llbracket 1; N \rrbracket)^{\mathbb{N}} \rightarrow S^0 \in \mathcal{P}(\llbracket 1; N \rrbracket)$. Then the general chaotic iterations defined in Equ.6 can be described by the following discrete dynamical system:

$$\begin{cases} X^0 \in \mathcal{X} \\ X^{k+1} = G_f(X^k). \end{cases} \tag{6}$$

To study the Devaney’s chaos property, a relevant distance between two points $X = (S, E), Y = (\check{S}, \check{E})$ of \mathcal{X} must be defined. Let us introduce:

$$d(X, Y) = d_e(E, \check{E}) + d_s(S, \check{S}), \tag{7}$$

where

$$\begin{cases} d_e(E, \check{E}) = \sum_{k=1}^{\mathbb{N}} \delta(E_k, \check{E}_k) \text{ is once again the Hamming distance,} \\ d_s(S, \check{S}) = \frac{9}{\mathbb{N}} \sum_{k=1}^{\infty} \frac{|S^k \Delta \check{S}^k|}{10^k}. \end{cases} \tag{8}$$

where $|X|$ is the cardinality of a set X and $A \Delta B$ is for the symmetric difference, defined for sets A, B as $A \Delta B = (A \setminus B) \cup (B \setminus A)$.

It has been proven in [4] that:

Theorem 1. *The general chaotic iterations defined on Eq.1 satisfy the Devaney’s property of chaos.*

3 Security Tools Based on CIs

We now investigate how to apply chaotic iterations on existing security tools. By such kind of post-treatment, we will add chaos to these tools, hoping by doing so to improve them in practice (increasing the entropy of random generators, the diffusion and confusion of hash functions, etc.) Such improvement must be such that existing security properties are preserved through iterations.

3.1 Pseudorandom Number Generator with CIs

In this section, we consider that the strategy $(S^n)_{n \in \mathbb{N}}$ is provided by a pseudorandom number generator, leading to a collection of so-called CIPRNGs [3]. The XOR CIPRNGs, for instance, is defined as follows [4]:

$$\begin{cases} x^0 \in \mathbb{B}^N \\ \forall n \in \mathbb{N}^*, x^{n+1} = x^n \oplus S^n, \end{cases} \tag{9}$$

where $N \in \mathbb{N}^*$ and \oplus stands for the bitwise exclusive or (xor) operation between the binary decomposition of x^n and S^n . In the formulation above, chaotic strategy $(S^n)_{n \in \mathbb{N}^*} \in \llbracket 1; N \rrbracket^N$ is a sequence produced by any standard pseudorandom number generator, which can be the well-known Blum Blum Shub (B.B.S.), Linear Congruential Generator (LCG), Mersenne Twister (MT), XORshift, RC4, or the Linear-Feedback Shift Register (LFSR). XOR CIPRNGs, which can be written as general chaotic iterations using the vectorial negation (see [4]), have been proven chaotic. They are able to pass all the most stringent statistical batteries of test, for well-chosen inputted generators.

3.2 CIs-Based Hash Function

Let us now present our hash function $H_h : K \times \mathbb{B}^* \rightarrow \mathbb{B}^N$ that is based on GFCIs. The key $k = \{k_1, k_2, k_3\}$ is in key space K , where k_1, k_2 , and k_3 are parameters of the function. The proposed hash function H_h is realized as follows.

The first step of the algorithm is to choose the traditional hash function h that it will be used in the proposed hash function. For our implementations, we have chosen MD5, SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512. The selected hash function determines the length N of the output hash value.

Then, the input message x is needed to transform into a normalized bits sequence of length a multiple of N , by applying the SHA-1 normalization stage. After this initialization, the length of the treated sequence X is L .

In the third step, k_1 is used as seed to generate k_2 pseudorandom binary vectors of length N , with XOR CIPRNGs. This sequence is the chaotic strategy $S = \{S^0 S^1 S^2 \dots S^{k_2-1}\}$.

In the fourth step, k_3 is considered as a binary vector of length N . $S^i \in S$ is then combined with k_3 using exclusive-or operation. After that we get a N bit binary output. Then we split X into $X = \{X^0 X^1 \dots\}$. Each X^i will be combined with the output of S and K_3 using exclusive-or operation. After that, we use this result as the input of traditional hash function h .

Lastly, to construct the digest, chaotic iteration of G_f are realized with the traditional hash function output $h(k_3, m, X)$ and strategy S as defined above. The result of these iterations is a N bits vector. It is translated into hexadecimal numbers to finally obtain the hash value.

So, the keyed hash function $H_h : K \times \mathbb{B}^* \rightarrow \mathbb{B}^N$ is described as Algorithm 1.

H_h is thus a chaotic iteration based post-treatment on the inputted hash function. If h satisfies the collision resistance property, then it is the case too

Algorithm 1. The proposed hash function H_h

Input:

The key $k = (k_1, k_2, k_3) \in K$;
 The input message $x \in \mathbb{B}^*$;
 The standard hash function $h()$;

Output:

Hash value H_h ;

- 1: Transform x to sequence X which length is L ;
 - 2: Use XOR CIPRNGs to generate m using k_1 as a seed and construct strategy $S = \{S^0 S^1 \dots S^{k_2-1}\}$ with m
 - 3: Use standard hash function to generate hash value $H = h(k_3, m, X)$;
 - 4: **for** $i = 0, 1, 2, \dots, k_2 - 1$ **do**
 - 5: Use GFCIs to generate hash value: $H_h = G_f(S^i, H)$;
 - 6: **end for**
 - 7: **return** H_h ;
-

for H_h . Moreover, if h satisfies the second-preimage resistance property, then it is the case too for H_h , as proven in [5]. With this post-treatment, we can thus preserve security while adding chaos: the latter may be useful to improve both confusion and diffusion.

4 Experimental Evaluation

In this section, experimental evaluations are provided including hash values, diffusion and confusion, and crash analysis. Let us consider that the input message is the poem “Ulalume” (E.A.Poe) and the selected pseudorandom number generator is B.B.S.

4.1 Hash Values

The standard hash function that we use here is MD5. To give illustration of the key properties, we will use the proposed hash function to generate hash values in the following cases:

- Case 1. $k_1 = 50, k_2 = 2, k_3 = 50$, and B.B.S.
- Case 2. $k_1 = 51, k_2 = 2, k_3 = 50$, and B.B.S.
- Case 3. $k_1 = 50, k_2 = 3, k_3 = 50$, and B.B.S.
- Case 4. $k_1 = 50, k_2 = 2, k_3 = 51$, and B.B.S.

The corresponding hash values in hexadecimal format are:

- Case 1. F69C3F042ABA1139FF443C278FDF3F7F.
- Case 2. C31BFBDD43273913C7CC845EC5E3D1EE.
- Case 3. 43353FA45B9560413C059F7FD4F485FB.
- Case 4. BEA4CAD480333117292F421BFA401BEB.

From simulation results, we can see that any little change in key space K can cause a substantial modification in the final hash value, which is coherent with the topological properties of chaos. In other words, it seems to be extremely sensitive to initial parameters.

A secured hash function should not only be sensitive to initial parameters, but also to initial values. This is why we test now our hash function with some changes in the input message, and observe the distribution of hash values. The key we use here is $k_1 = 50, k_2 = 2, k_3 = 50$, and standard hash function is MD5.

- Case 1. The input message is the poem "Ulalume" (E.A.Poe).
- Case 2. We replace the last point '.' with a coma ','.
- Case 3. In "The skies they were ashen and sober", 'The' become 'the'.
- Case 4. In "The skies they were ashen and sober", 'The' become 'Th'.
- Case 5. We add a space at the end of the poem.

The corresponding hash values in binary format are shown in Fig. 1. Through this experiment, we can check that the proposed hash function is sensitive to any alteration in the input message, which will cause the modification of the hash value.

For a secured hash function, the repartition of its hash values should be uniform. In other words, the algorithm should make full use of cryptogram space to make that the hash values are evenly distributed across the cryptogram space. The cases here are the same as discussed above. In Fig. 2(a), the ASCII codes of input message are localized within a small area, whereas in Fig. 2(b), the hexadecimal numbers of the hash value are uniformly distributed in the area of cryptogram space.

4.2 Diffusion and Confusion

We now focus on the illustration of the diffusion and confusion properties. Let us recall that diffusion means that the redundancy of the plain text must be dispersed into the space of cryptograms so as to hide the statistics of plain text. Confusion refers to the desire to make the statistical relationship between plain text, ciphertext, and keys as complex as possible, which makes attackers difficult to get relation about keys from ciphertext. So under the situation of that when the plain text is changed by only one bit, it leads to a modification of hash values that can be described by the following statistics:

- Mean changed bit number: $\bar{B} = \frac{1}{N} \sum_{i=1}^N B_i$;
- Mean changed probability: $P = \frac{\bar{B}}{L} \times 100\%$;
- Mean square error of B: $\Delta B = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (B_i - \bar{B})^2}$;
- Mean square error of P: $\Delta P = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\frac{B_i}{L} - P)^2} \times 100\%$;

where N denotes the statistical times, and B_i denotes the changed bits of hash value in i^{th} test, while L denotes the bits of hash value in binary format.

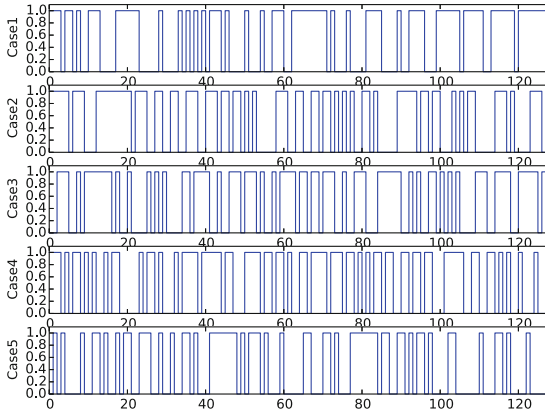


Fig. 1. 128 bit hash values in various cases

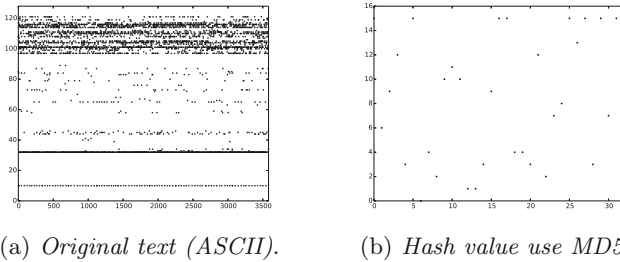
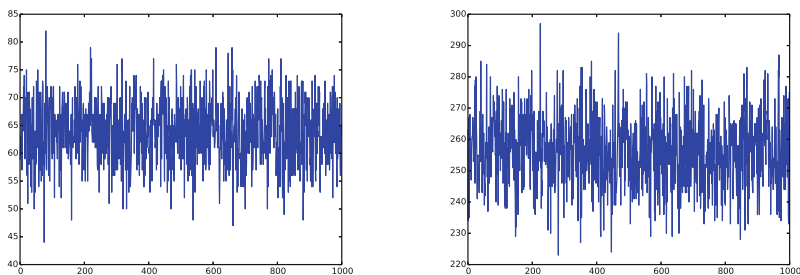


Fig. 2. Spread of input message and the corresponding hash value

For a secured hash function, the desired value of \bar{B} should be $L/2$. The desired distribution of hash algorithm should be that small toggle in plain text cause 50% change of hash value. ΔB and ΔP show the stability of diffusion and confusion properties. The hash algorithm is more stable if these two values are close to 0.

Let us check the diffusion and confusion of the proposed hash function. The test procedure is described below. Firstly, we obtain the original hash value of plain text. Then at each time, only one bit is changed in it. The hash values of these modified plain texts are used to compare with the original hash value. After N tests, \bar{B} , P , ΔB , and ΔP are calculated. The key used here is $k_1 = 50$, $k_2 = 2$, $k_3 = 50$, $N = 1000$. As shown in Fig. 3, we check the distribution of changed bits of proposed hash function with MD5 and SHA-512. From these figures, we can see that one bit changed in the plain text can cause about $L/2$ modifications in B_i .

Observing Table 1, even the iteration times are small, the mean changed bit numbers \bar{B} and the mean changed probabilities P are close to the desired values $L/2$ and 50%. ΔB and ΔP are quite small. In other words, the proposed hash function achieves desired values for such properties. These results illustrate the



(a) Standard hash function is MD5.

(b) Standard hash function is SHA-512.

Fig. 3. Distribution of changed bit numbers B_i with different standard hash functions

diffusion and confusion of the proposed hash function H_h , and these capabilities are quite stable. This feature is attributed to GFCIs that can change multi bits in one time. To sum up, due to the fact that computational complexity can be reduced here, we think it is better to apply it into practical applications. Furthermore, compared with the performance of standard hash functions, which is shown in Table 2, the proposed hash function shows better results.

4.3 Collision Analysis

We now consider the analysis of impact resistance attacks. If hash function's ability to face collision is stronger, then the hash function is more security. Through experiments can be quantitatively tested the collision resistance ability of the proposed hash function. Firstly, we obtain the original hash value of plain text and transform it to ASCII code. Then the plain text one bit modification is applied and we obtain a new hash value in ASCII code. By comparing these two hash values, we can get the positions where they have the same character. The absolute coefficient between these two hash values can be described as follows:

$$d = \sum_{i=1}^N |t(e_i) - t(\check{e}_i)|, \quad (10)$$

where N denotes the number of ASCII characters in the hash value, e_i and \check{e}_i are the i^{th} character in former and new hash value separately. Function $t(\cdot)$ is used to transform e_i and \check{e}_i to decimal format. the theoretical value of average absolute distance per character is 85.3333.

The key used here is $k_1 = 50, k_2 = 2, k_3 = 50$, while testing times is 2048. The experiment results are shown in Table 3. The second column shows the number of hits, in which the fist component is the number of hits to zero, the second component is to one, the third component is to two, the forth one is to three, and the last one is to four. We can see that the maximum number of hits

Table 1. Statistical performance of the proposed hash function (B.B.S.)

<i>hash_type</i>	Iterate times	\bar{B}	$P(\%)$	ΔB	$\Delta P(\%)$
MD5	1	63.906	49.926	5.819	4.546
	2	63.845	49.879	5.656	4.419
	10	63.846	49.880	5.845	4.566
SHA-1	1	79.774	49.859	6.446	4.029
	2	80.355	50.222	6.329	3.956
	10	79.779	49.862	6.131	3.832
SHA-224	1	112.087	50.039	7.619	3.401
	2	112.038	50.017	7.297	3.257
	10	111.883	49.948	7.268	3.244
SHA-256	1	128.075	50.029	7.845	3.064
	2	127.72	49.891	8.002	3.126
	10	127.806	49.924	8.215	3.209
SHA-384	1	192.098	50.255	9.579	2.495
	2	192.193	50.050	9.693	2.524
	10	191.843	49.959	9.704	2.527
SHA-512	1	256.043	50.008	10.867	2.122
	2	256.062	50.012	11.376	2.222
	10	256.032	50.006	11.438	2.234

Table 2. Statistical performance of the standard hash function

<i>hash_type</i>	\bar{B}	$P(\%)$	ΔB	$\Delta P(\%)$
MD5	63.893	49.916	5.437	4.248
SHA-1	79.770	49.856	6.359	3.975
SHA-224	112.284	50.127	7.324	3.270
SHA-256	127.746	49.901	8.405	3.283
SHA-384	191.81	49.951	10.036	2.613
SHA-512	256.084	50.016	11.232	2.194

is four with small probability. It is mainly in the number of collision to zero and one. On the other hand, the average absolute difference d of the two hash values per character, which is shown in the fifth column, is close to the desired value 85.3333. Based on these results, the collision resistance capability of the proposed hash algorithm is strong.

Table 3. Collision performance

<i>hash_type</i>	Number of hits	Sum of d	Avg d per character
MD5	(1931, 114, 3, 0, 0)	2956780	90.234
SHA-1	(1880, 159, 8, 1, 0)	3472244	84.772
SHA-224	(1837, 204, 7, 0, 0)	4629075	80.725
SHA-256	(1817, 214, 17,0, 0)	5348270	81.608
SHA-384	(1690, 328, 27,3, 0)	8398092	85.430
SHA-512	(1601, 392, 47,6, 2)	11042728	84.250

5 Conclusion

In this article, a chaotic iteration based hash function has been presented. When constructing strategies, pseudorandom number generator is used. Then the general formulation of chaotic iterations is exploited to obtain hash values. Through the experimental evaluation of hash values, we can see that the proposed hash function is highly sensitive to initial parameters, initial values, and keys. The statistical performances show that the proposed hash function has better features of diffusion and confusion even if the iteration times are small, which can be considered for practical applications. And the proposed hash algorithm has better performance of collision performance. To sum up, the proposed scheme is believed a good application example for constructing secure keyed one-way hash function.

References

1. Bahi, J.M., Guyeux, C.: Discrete Dynamical Systems and Chaotic Machines: Theory and Applications. CRC Press, Boca Raton (2013)
2. US Department of Commerce/National Institute of Standards and Technology: Secure Hash Standard (SHS). Fips Publication (1995)
3. Fang, X., Guyeux, C., Wang, Q., Bahi, J.: Randomness and disorder of chaotic iterations. Applications in information security field. In: NOLTA 2015, International Symposium on Nonlinear Theory and its Applications, Hong Kong, China, December 2015, pp. 1–4 (2015)
4. Guyeux, C., Couturier, R., Héam, P.C., Bahi, J.M.: Efficient and cryptographically secure generation of chaotic pseudorandom numbers on GPU. *J. Supercomputing* **71**(10), 3877–3903 (2015)
5. Guyeux, C., Wang, Q., Fang, X., Bahi, J.M.: Introducing the truly chaotic finite state machines and theirs applications in security field. In: Nolta 2014, International Symposium on Nonlinear Theory and ITS Applications (2014)
6. Rivest, R.: The MD4 Message Digest Algorithm. Springer, Heidelberg (1990)
7. Rivest, R.: The MD5 Message-Digest Algorithm. RFC Editor (1992)

8. Stevens, M., Lenstra, A., Weger, B.: Chosen-prefix collisions for MD5 and colliding X.509 certificates for different identities. In: Naor, M. (ed.) EUROCRYPT 2007. LNCS, vol. 4515, pp. 1–22. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-72540-4_1](https://doi.org/10.1007/978-3-540-72540-4_1)
9. Stevens, M., Sotirov, A., Appelbaum, J., Lenstra, A., Molnar, D., Osvik, D.A., Weger, B.D.: Short Chosen-Prefix Collisions for MD5 and the Creation of a Rogue CA Certificate. Springer, Heidelberg (2009)

Data Mining and Artificial Intelligence

Intellectual Overall Evaluation of Power Quality Including System Cost

Buhm Lee, Dohee Sohn, and Kyoung Min Kim^(✉)

Department of Electrical and Semiconductor Engineering, Chonnam National University,
50, Daehak-ro, Yeosu, Jeollanam-Do 59626, Korea
{buhmlee, kkm}@chonnam.ac.kr, sondoh@khnp.co.kr

Abstract. This paper presents a new methodology to evaluate power quality for a distribution system. Instead traditional evaluation methodology can evaluate a certain face of power quality, such as SAIFI and SAIDI for reliability, SARFI for voltage sag/swell, and THD for harmonics, by using IEEE Standard Indices, newly present evaluation methodology can overall evaluate for power quality items, maintenance cost, and system losses. This methodology uses AHP model and newly developed Ideal AHP. First, AHP model was employed and implemented for power quality overall evaluation. Second, Ideal AHP was developed to overcome different unit problems of power quality. This paper applied the method for a distribution system, and showed the process of the method, and obtained overall evaluation of the system.

Keywords: Overall evaluation · Power quality · AHP · Ideal AHP

1 Introduction

Every electric power company wants to supply reasonable quality electric power with reasonable price, instead every electric power customer desires high quality electric power at cheap price. Because electric power customers have to invest a lot of money to be supplied high quality power, decision maker of the system needs to find optimal point between numbers of distribution system alternatives. So, decision maker needs to know the overall power quality of each alternative.

Traditionally, power quality was evaluated by using IEEE Standards, such as SAIFI, SAIDI, CAIDI, CTAIDI, CAIFI, ASAI, ASIFI, ASIDI, $CEMI_n$, MAIFI, MAIFI_E, CEMSMI_n for Reliability [1, 2], SARFI_x, SEI, ASEI for Voltage Sags/Swells [3], THD, TDD, TIF for Harmonics [4]. Because these methodologies were developed for electric engineers which want to evaluate certain faces of power quality, decision maker needs to know overall power quality of the system for current system and a future system. To assume power quality of the system which is in planning, probabilistic model was developed [5], and this paper employed the method.

To unify above indices from IEEE Standards, Analytic Hierarchy Process (AHP) model [6, 7] was employed and authors developed AHP model [8] which can evaluate total power quality as an unified Index. To overcome unit different problem of indices, authors developed Ideal AHP model [9, 11] and applied it to the distribution system.

Another way to evaluate power quality and consider maintenance cost and losses, authors employed value-based methodology [10]. This methodology encounter conflicts between power companies and customers, because of their viewpoints.

This paper presents a combined traditional AHP model and Ideal AHP model which can overall evaluate for a distribution system. To overcome unit different problem of indices, Ideal AHP model which was developed by authors was employed. This model has three states, such as IDEAL, ACTUAL, and POSSIBLE, which reflects that electric power customer can feel ideal power quality and possible power quality. By setting ACTUAL between IDEAL and POSSIBLE, absolute power quality as competitiveness was obtained. To obtain overall power quality, AHP model was employed. By using combining AHP model and Ideal AHP model, overall power quality can be obtained as competitiveness. Biggest merit of this methodology is power quality can be shown as an absolute value which means competitiveness. This paper applied the method for a distribution system, and showed the process of the method, and obtained overall evaluation of the system.

2 Units of Power Quality

2.1 Indices of Power Quality, Operation, etc.

Indices of power quality can be calculated by IEEE standards, operation cost can be calculated by investment, and loss can be calculated by loadflow of the system.

(1) Reliability

Reliability and its indices are specified based on IEEE Std. 1366 [1]. These indices have no unit. Reliability indices are shown as follows:

- (a) System Average Interruption Frequency Index (SAIFI)
- (b) System Average Interruption Duration Index (SAIDI)
- (c) Customer average interruption duration index (CAIDI)
- (d) Customer total average interruption duration index (CTAIDI)
- (e) Customer average interruption frequency index (CAIFI)
- (f) Average service availability index(ASAI)
- (g) Average System Interruption Frequency Index (ASIFI)
- (h) Average System Interruption Duration Index (ASIDI)
- (i) Customers experiencing multiple interruptions ($CEMI_n$)
- (k) Momentary Average Interruption Frequency Index (MAIFI)
- (l) Momentary average interruption event frequency index ($MAIFI_E$)
- (l) Customers experiencing multiple sustained interruptions and momentary interruptions events ($CEMSMI_n$)

(2) Voltage Sags/Swells

Voltage Sags/Swells and its indices are specified based on IEEE Std. P1564 [2]. These indices have no unit. Voltage Sags/Swells indices are shown as follows:

- (a) System Average RMS variation Frequency Index ($SARFI_x$)
- (b) Voltage Sag Energy Index (SEI)
- (c) Average Sag Energy Index (ASEI)

(3) Harmonics

Harmonics and its indices are specified based on IEEE Std. 519 [1]. These indices have no unit. Harmonic indices are shown as follows:

- (a) Total Harmonic Distortion (THD)
- (b) Total Demand Distortion (TDD)
- (c) Telephone Interference (TIF)

(4) Maintenance, Loss

Every distribution system needs maintenance, so operators of the system have to prepare maintenance cost. Every system has losses to operate the system. Unit of maintenance cost is in dollar and unit of loss is in kWh.

- (a) Maintenance Cost [\$]
- (b) Loss [kWh]

2.2 Units of Power Quality Item and Their Mismatch

Power quality of a system has number of units. Reliability, Voltage Sag/Swell, and Harmonics indices have no unit, instead unit of Maintenance cost is money and unit of Loss is kWh. Even though indices of Reliability, Voltage Sag/Swell, and Harmonics have no unit, every index has different standard. For example, any of them are between 0 and 1, another of them are 0 to 100, the other of them is 0 to 10000.

So, Overall evaluation of power quality is not so easy.

2.3 Case Study: Application to a Sample System

This study shows an application to a sample system, and obtained power quality indices, maintenance cost and loss of the system.

(1) Distribution System

Current system has three 154 kV substations, such as node G, node A, and node B, and their 22 kV subsystems. This distribution system has plan to extend as node C and 22 kV subsystems (purple figure). Node C can be powered by node G (green line) and node A (blue line).

Diagram of distribution system is shown at Fig. 1, and loads of each load point are shown at Table 1. Failure frequencies and their durations are employed from IEEE data [12]. Because loads of the system are varying, authors select loads reflect maximum loads.

(2) Result of Evaluation

From Fig. 1, Reliability indices, Voltage Sags/Swells indices, Harmonic indices, Maintenance cost, and Loss are obtained as follows:

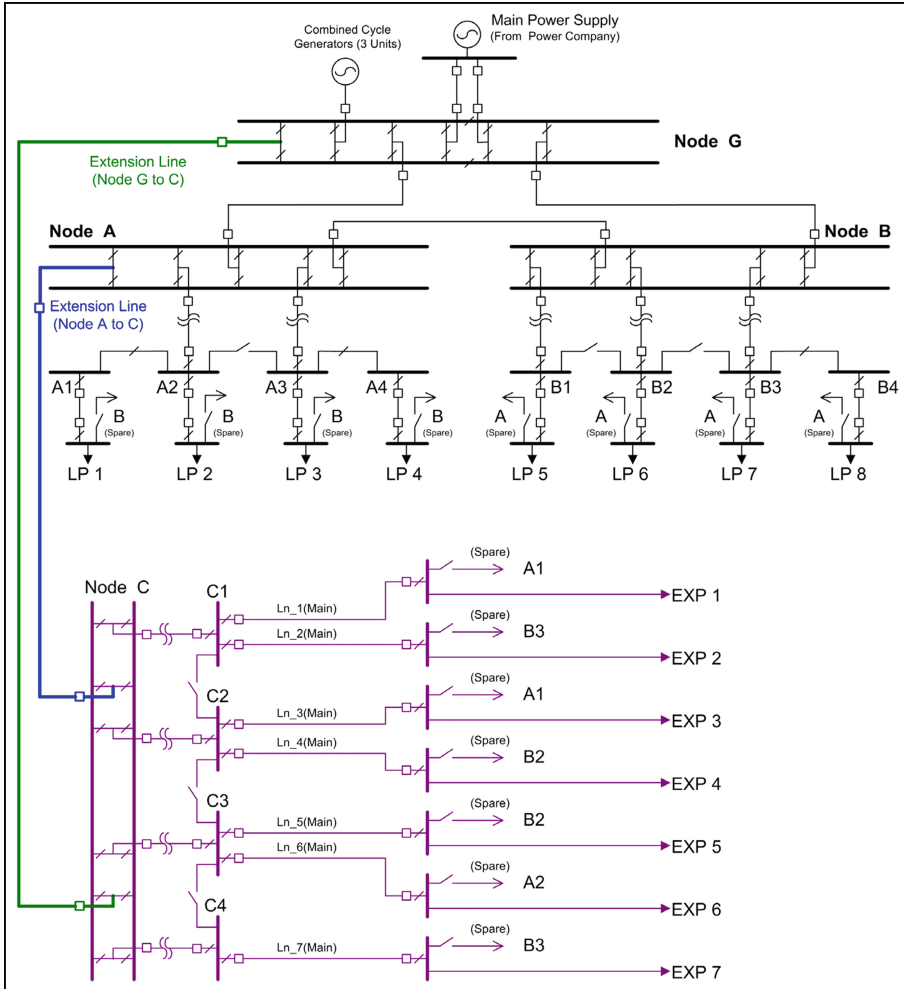


Fig. 1. Sample distribution system

(a) Reliability indices

SAIFI: 0.1142432	SAIDI: 20.8522665	ASIFI: 0.1141458
ASIDI: 20.8485466	MAIFI: 0.2856079	CAIDI: 182.5252693
ASAI: 0.9976196	ENS: 10226.2	AENS: 1733.3

(b) Voltage Sags/Swells indices

SARFI: 0.4569727

(c) Harmonic indices

THD: 0.036

(d) Maintenance Cost

Cable main:	\$260,000 (Annual, 30yr Lifetime)
Cable back up:	\$90,000 (Annual, 30yr Lifetime)
Switch:	\$290,000 (Annual, 30yr Lifetime)
Main transformer:	\$260,000 (Annual, 20yr Lifetime)
Accessory:	\$120,000 (Annual, 50yr Lifetime)
Maintenance:	\$180,000 (Annual)

Table 1. Loads at each load point

Load point	Load				Load point	Load			
	Size [kVA]	Num. [ea]	Harm [A]	Type		Size [kVA]	Num. [ea]	Harm [A]	Type
LP 1	23,000	100	120	Ind.	EXP 1	43,000	50	130	Ind.
LP 2	7,500	150	40	Office	EXP 2	15,000	120	70	Office
LP 3	42,000	320	200	Ind.	EXP 3	43,000	60	130	Ind.
LP 4	30,000	180	150	Ind.	EXP 4	36,000	80	110	Ind.
LP 5	68,000	1800	290	Ind.	EXP 5	29,000	150	140	Office
LP 6	39,000	1000	170	Ind.	EXP 6	50,000	50	150	Ind.
LP 7	16,000	800	80	Office	EXP 7	18,000	240	90	Office
LP 8	31,000	800	110	Ind.					

(e) Loss

Loss for active power:	1,200 kW (Annual)
Loss for harmonics:	50 kW (Annual)

3 Ideal AHP

3.1 Ahp

AHP model [6] can obtain competitiveness among alternatives, especially useful to obtain weighting such as determining priority. Authors applied this model to obtain the best power quality alternative [8] among number of alternatives in planning stage.

3.2 Ideal AHP

Ideal AHP model was developed by authors [9, 11]. This model has 3-state, such as IDEAL, ACTUAL, and POSSIBLE states, instead alternatives in traditional AHP model. Here, IDEAL is the ideal values that each customer feels as ideal, POSSIBLE is the possible values that each customer feels as extremely challenging because of power quality, and ACTUAL is calculated values that reflects current state. ACTUAL of index can vary from 0 to infinitive. Even though, index can vary 0 to infinitive, operators can only accept 0 to 100 (for example). Here, IDEAL and POSSIBLE are set to 0 and 100. Concept of traditional AHP and Ideal AHP model is shown at Fig. 2.

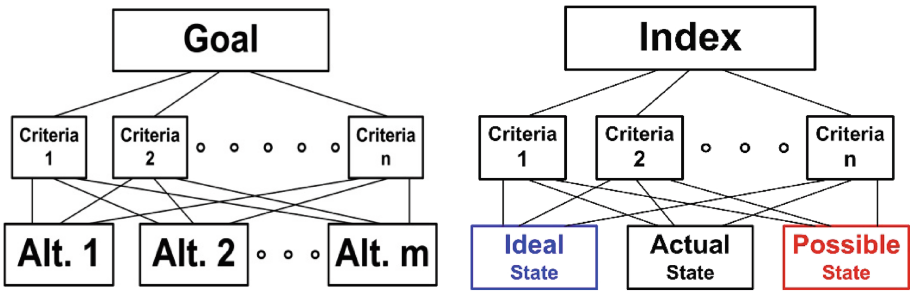


Fig. 2. Traditional AHP model vs. Ideal AHP model

4 Overall Evaluation Model

4.1 Overall Evaluation Model for Distribution System

Overall evaluation model for distribution system is shown at Fig. 3.

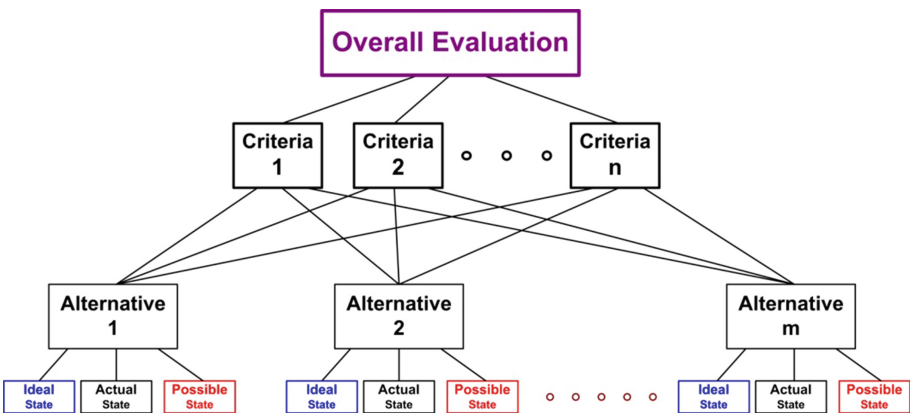


Fig. 3. AHP model for overall evaluation

At Fig. 3, lower hierarchy is consists of Ideal AHP and upper hierarchy is consist of traditional AHP. This model can overcome unit different problem, and obtain overall power quality evaluation.

4.2 Case Study: Application to a Sample System

(1) Re-indexing by using Ideal AHP (lower hierarchy)

Even though IEEE indices have no unit, these indices have different standards. By using Ideal AHP, these indices can be converted as a same standard. ACTUAL of power quality are shown from Tables 2, 3, 4, 5 and 6.

(a) Reliability

Table 2. Re-indexing for reliability indices by using ideal AHP

Reliability	IDEAL	ACTUAL	POSSIBLE
SAIFI	0.00	0.11424	1.00
SAIDI	0.00	0.10426	2.00
ASIFI	0.00	0.11415	1.00
ASIDI	0.00	0.10424	2.00
MAIF	0.00	0.14280	2.00

(b) Voltage Sags/Swells

Table 3. Re-indexing for voltage Sags/Swells index by using ideal AHP

Voltage Sag/Swell	IDEAL	ACTUAL	POSSIBLE
SARFI	0.00	0.15232	3.00

(c) Harmonics

Table 4. Re-indexing for harmonic index by using ideal AHP

Harmonics	IDEAL	ACTUAL	POSSIBLE
THD	0.00	0.12000	0.30

(d) Maintenance Cost

Table 5. Re-indexing for maintenance cost by using ideal AHP

Maintenance cost	IDEAL	ACTUAL	POSSIBLE
Cost [\$]	0.00	0.12000	10,000,000

(e) Loss

Table 6. Re-indexing for loss by using ideal AHP

Loss	IDEAL	ACTUAL	POSSIBLE
Loss [kW]	0.00	0.17000	10,000

(2) One to one matrix for upper hierarchy

One to one matrix for upper hierarchy is shown at Table 7. This matrix consider 3 category of power quality indices, cost, and loss. Among them, reliability consider 5 items. Here, elements of the matrix are arbitrarily.

Table 7. 1:1 matrix for upper hierarchy of developed AHP model

	Reliability					V. Sag	Harm.	Maint.	Loss
	SAIFI	SAIDI	ASIFI	ASIDI	MAIFI	SARFI	THD	Cost	Loss
SAIFI	1.000	0.952	1.000	0.952	2.000	2.000	1.000	0.750	0.500
SAIDI	1.050	1.000	1.050	1.000	1.667	2.500	1.250	0.800	0.600
ASIFI	1.000	0.952	1.000	0.952	2.000	2.000	2.000	0.750	0.500
ASIDI	1.050	1.000	1.050	1.000	1.667	2.500	1.250	0.800	0.600
MAIFI	0.500	0.600	0.500	0.600	1.000	3.000	3.000	0.500	0.400
SARFI	0.500	0.400	0.500	0.400	0.333	1.000	0.500	0.400	0.400
THD	1.000	0.800	0.500	0.800	0.333	2.000	1.000	0.300	0.300
Cost	1.333	1.250	1.333	1.250	2.000	2.500	3.333	1.000	0.800
Loss	2.000	1.667	2.000	1.667	2.500	2.500	3.333	1.250	1.000

By using Table 7, authors obtain Eigenvalues for upper hierarchy, and shown at Table 8. Obtained Overall evaluation of power quality is shown at Table 9.

Table 8. Eigenvalues for upper hierarchy of developed AHP model

	Reliability					V. Sag	Harm.	Maint.	Loss
	SAIFI	SAIDI	ASIFI	ASIDI	MAIFI	SARFI	THD	Cost	Loss
Weighting factor	0.1202	0.1174	0.1202	0.1175	0.0652	0.0434	0.0699	0.1459	0.2003

Table 9. Overall evaluation of power quality

Overall evaluation of power quality	0.12780905
-------------------------------------	------------

Obtained value of Overall evaluation of power quality is between ‘0’ which means IDEAL and ‘1’ which means POSSIBLE. If evaluation value is low, this value reflects high electric power quality.

5 Conclusion

This paper presents an overall evaluation model for distribution system. The contributions of the paper are:

- (1) Authors propose a new AHP model which combine traditional AHP model and newly developed Ideal AHP model.
- (2) By applying Ideal AHP to power quality indices, unit different problem of indices can be overcome.
- (3) By applying AHP to power quality evaluation, overall evaluation can be obtained.
- (4) By applying developed model for a sample distribution system, authors demonstrate the process of evaluation of power quality.

References

1. IEEE: IEEE Guide for Electric Power Distribution Reliability Indices, IEEE Std 1366 (2001)
2. IEEE: IEEE Recommended Practice for Monitoring Electric Power Quality, IEEE Std 1159 (1995)
3. Math, H.J., et al.: Voltage-Sag Indices – Recent Developments in IEEE P1564 Task Force, pp. 34–41 (2003)
4. IEEE: IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems, IEEE Std 519 (1992)
5. Billinton, R., et al.: Reliability Evaluation of Power Systems. Plenum Press, New York (1984)
6. Satty, T.L.: The Analytic Network Process. RWS Publications, Pittsburgh (1996)
7. Ilic, M.D., et al.: Hierarchical Power Systems Control – Its Value in a Changing Industry. Springer, Heidelberg (1996)
8. Lee, B., Choi, C.-H., Choi, N.-S., Kim, K.M., Kim, Y.-H., Choi, S.-K., Meliopoulos, Sakis, A.: Distribution system evaluation algorithm using analytic hierarchy process. In: Ali, M., Dapoigny, R. (eds.) IEA/AIE 2006. LNCS (LNAI), vol. 4031, pp. 177–186. Springer, Heidelberg (2006). doi:[10.1007/11779568_21](https://doi.org/10.1007/11779568_21)
9. Lee, B., et al.: Unified power quality index using ideal AHP. In: 13th International Conference on Harmonics and Quality of Power, Wollongong, Australia (2008)
10. Lee, B., et al.: Unified power quality index based on value-based methodology. In: IEEE Power & Energy Society General Meeting 2009, Calgary, Canada (2009)
11. Lee, B., Sohn, D., Kim, K.M.: Development of power quality index using ideal analytic hierarchy process. In: Kim, K., Joukov, N. (eds.) Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 783–793. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_75](https://doi.org/10.1007/978-981-10-0557-2_75)
12. IEEE. IEEE Recommended Practice for the Design of Reliable Industrial and Commercial Power Systems. IEEE ANSI (1991)

A Data-Driven Decision Making with Big Data Analysis on DNS Log

Euihyun Jung^(✉)

Department of Computer Science, Anyang University,
Jungang-ro Buleun-myeon, Incheon, Ganghwa-gun 602-14, Korea
jung@anyang.ac.kr

Abstract. Domain Name System (DNS) log has been considered as a great source of valuable information for the decision making on government policy or business strategy because querying DNS is the first step of all Internet activities. Due to the size of DNS log, Hadoop is considered as a prominent solution, but the geographical dispersal of DNS log hinders to adopt it in an ordinary way. Hadoop assumes all data source should be located on a single Hadoop File System (HDFS), but DNS log is stored on DNS servers dispersed all over the world. To resolve this issue, a new method named “Localized Analysis & Merge (LAM)” is proposed in this paper. The proposed method enables Hadoop to analyze DNS log on the dispersed DNS servers and it reduced the whole processing time dramatically. Also, the LAM method showed that DNS log can be used to extract a lot of valuable information such as a malware detection, the access frequency over countries, etc.

Keywords: DNS · Big Data · Data mining · Malwares · Decision making

1 Introduction

Domain Name System (DNS) was developed to translate a host’s domain name into the host’s network address, and vice versa [1]. In IP networking, every peer has to find the target peer’s IP address before making a session by querying DNS for the peer’s domain name. Since querying DNS is the first step of all activities on Internet, researchers in data science and computer networks have expected to extract valuable information from DNS log [2].

Although there have been several studies in analyzing DNS log, most researchers have focused on the security patterns watched in DNS log [3–5]. In theory, it is possible to detect malwares by monitoring DNS log because the patterns of the domain queries related to the malwares will be changed before an attack starts. Whyte et al. [3] suggested the method for scanning worms by investigating DNS. Korea University research team [4] forecasted the behaviors of botnets through the simulation of DNS queries. EXPOSURE [5] tried to find patterns in particular attacks by applying several characteristics to DNS log.

The previous studies more or less succeeded to find useful patterns and various methods to adopt DNS in finding security holes, but these studies have some limitations.

First of all, the studies were mainly confined to the security issue. Needless to say, the monitoring DNS is to be an effective way in detecting anomalies in Internet traffic, but DNS log has more possibility to be a gold mine of valuable information. The information extracted from DNS log can be used to determine government policies or to plan business strategies. Secondly, the results of the previous studies were drawn from the analysis of the small fraction of DNS log or limited simulations. Since it is very difficult to gather large amount of DNS log due to some political issues, it can be unfair the previous studies are criticized for this reason. However, it is undeniable that the analysis of huge DNS log can provide various viewpoints that the previous studies did not give.

To get the enough size of DNS log, researchers should need the cooperation of Top Level Domain (TLD) managing organizations instead of operating recursive servers by themselves [6]. In Korea, Korea Internet & Security Agency (KISA) takes charge of managing the national DNS which resolves the queries on the Korea country TLD domain, “.kr”. In 2013, the national DNS processes about 1.5 billion queries a day and records the results of the queries in the size of 300 GB. Unlike ordinary DNS servers, the national DNS servers resolve the whole DNS queries for the top-level country code domain, therefore researchers can look into the patterns of the national-level host accesses from all over the world.

However, even if DNS log has been obtained from the organization, there was another problem in processing them. In general, DNS log is too huge to handle with the conventional tools because the tools have the limitation of physical memory and they cannot use the power of parallel programming. Fortunately, the emerging Big Data technology especially Hadoop can resolve it. Since Hadoop adopts the horizontal scalability [7], it is expected to process huge amounts of data in a fast and efficient way. However, it is not enough to simply adopt Hadoop in this problem because Hadoop assumes all the data in the localized file system named Hadoop File System (HDFS) [8]. The assumption conflicts with the fact that DNS logs are stored on the geographically dispersed servers. In order to use Hadoop, the dispersed DNS logs should be gathered into a single central HDFS, but it is impractical because the procedure takes too much time.

To resolve the issue, we suggested “Localized Analysis & Merge (LAM)” method to be used for the analysis on the Korean national DNS log. The LAM method enables Hadoop to analyze DNS logs separately on the dispersed DNS servers and it saves the whole processing time a lot. With the LAM method, the several analyses were conducted to get valuable information such as a detection of malwares. In this paper, the analysis results showed DNS log can be used to help people to do a better decision making if they were analyzed properly.

The rest of this paper is organized as follows. Section 2 describes the data set used in the analysis and the proposed method named LAM. In Sect. 3, the results of the analysis are discussed and Sect. 4 concludes the paper.

2 Analysis

2.1 The Data Set

An ordinary organization can run its DNS server called a recursive DNS server [6]. The recursive DNS server takes charge of resolving a query for the domain belonging to the organization. When the recursive DNS server cannot resolve the query, it hands over the query to the upper level recursive DNS server along the hierarchy of the DNS domain. This kind of delegation keeps going until the query is arrived to Top Level Domain (TLD) DNS servers or country TLD DNS servers [6]. The Korea national DNS servers as a country TLD server were deployed on 15 sites around the world. In 2013, the national DNS servers served about 1.5 billion queries a day and they recorded the DNS log in the size of 300 GB.

The data format of the Korea DNS log is the pcap (packet capture) format for capturing network traffic. This is useful to network analysis tools but not suitable for data mining. Since Hadoop usually needs text-based data, the transformation was required. Also, the DNS log lacks the additional information except the network related data. Therefore, we appended the geolocation information according to the IP from a DNS packet to make a new data set, which contains both DNS features and IP-based geolocation.

2.2 The Proposed Method

The Korea national DNS log is too huge for the conventional in-memory analysis tools such as R and it even also takes unacceptable time to load the log into the memory. Therefore, the only solution for the analysis of DNS log is Big Data technology such as Hadoop. However, Hadoop assumes all data sources are located on a single HDFS which consists of computers connected with high-speed network. In theory, the geographically dispersed DNS servers can be combined into a single HDFS because they are connected on the Internet. However, it is impractical to combine 15 Korea national DNS servers dispersed around the world into a single HDFS because the network speed of HDFS is assumed to be 1 Gbps at least. Therefore, in the data mining of DNS log, it can be the most challenging task to gather the logs into the one central analyzing server.

In order to resolve this issue, we proposed a new method named “Localized Analysis & Merge (LAM)”. The LAM method borrows its idea from Hadoop’s philosophy [7], “Code moves near data for computation”. Hadoop puts a map-reduce job’s code to the HDFS nodes which hold data instead of loading data from the nodes to the node running the code. Similarly, in the LAM method, a Hadoop analysis is locally performed on each national DNS server and then the results of each analysis are merged. The LAM method not only resolves the problem of the need for the one single HDFS, but also reduces the whole processing time dramatically. The method has two saving points in processing time during the process. First, it does not need to deliver raw DNS logs from the dispersed DNS servers to the central system. Second, the analysis is performed in parallel on each local DNS server, so the burden of the analysis is shared and it results in reducing the processing time.

Although the LAM method is powerful, it can only be applied to associative operations. The associative operations do not matter the order in which the operations performed, so they can be divided and distributed to the process nodes [9]. The associative operations are sum, average, frequency, max, min, etc. An Eq. (1) represents this concept mathematically and “Operation” in the Eq. (1) can be replaced with the actual operations such as sum.

$$Operation_{total} = \sum Operation_{local} \quad (1)$$

For example, since the operation for the access frequency with various conditions is an associative operation, it can be represented like the Eq. (2). It means that the access frequency can be divided and analyzed on each DNS server and the summed result of each operation is same with the result from the access frequency on whole data.

$$Access_Freq.(condition)_{total} = \sum Access_Freq.(condition)_{local} \quad (2)$$

3 The Results

3.1 The Most Visited Domains Per Country

In the decision making based on the analysis of DNS log, the most visited domain is an important factor. To get the result for this, a Pig code is performed on the DNS log by changing the condition of the country code. The below code is the Pig code for the most 2000 visited domain from Brazil.

```
dns = load '0816' using PigStorage(',') as
(server_name:chararray, day:int, hour:int, min:int,
sec:int, msec:int, src_ip:chararray, length:int,
dst_port:int, flag:int, question:int, answer_rr:int,
auth_rr:int, add_rr:int, domain:chararray, d_type:int,
d_class:int, cc:chararray, cont:chararray,
tzone:chararray, region:chararray, isp:chararray,
city:chararray, lat:float, lon:float);

day_group = foreach dns generate d_type, domain, cc;
query_dns = filter day_group by d_type == 1;
brdomain = filter query_dns by cc matches 'BR';
br_grpd = group brdomain by domain;
br_cnt = foreach br_grpd generate group,COUNT(brdomain)
as br_domain_count;
order_br = order br_cnt by br_domain_count desc;
limit_br = limit order_br 2000;
store limit_br into '0816-out/nation-br';
```

By changing the country code, we got the most visited and noticeable domains as shown in Table 1. From the result, “ns.hardware.co.kr” ranked as the fifth in China. Since the domain was belonged to the game company, “Com2US”, it was easily guessed that the company’s game was popular in China. Therefore, the game company could get more profits by doing advertising campaign intensely in China. In Brazil, “www.style.co.kr” was popular, but the domain was rarely accessed in Korea. Since this

domain was the domain of a fashion site, it implied Brazilians were interested in Korea fashion. Lastly, we found a strange domain, “smartfind.co.kr” because the access rate was abnormally high in US and Korea, but the site did not have any corresponding web page. The domain seemed a Command and Control (C&C) server and we found it from the malware reporting site [10].

Table 1. The most visited domains per country. “-” means the rank is below 2000.

Domain	US	Brazil	Japan	China	France	Korea
www.style.co.kr	564	29	761	595	174	-
ns.hardware.co.kr	14	41	11	5	21	246
www.auction.co.kr	18	169	321	260	4	-
www.koreatimes.co.kr	38	331	146	345	74	121
www.alba.co.kr	650	-	883	1204	145	7
smarfind.co.kr	9	43	732	593	-	5

3.2 A Detection of Hidden Malwares

In order to check why “smartfind.co.kr” is so popular even though it is not linked to any web page, more analyses were needed. First, the access pattern over time was watched. Figure 1 showed the access pattern of the “smartfind.co.kr” domain depended on human working hours. The access rate rose from 8 AM rapidly and dropped down since 6 PM. It meant that the domain access was related to human activity.

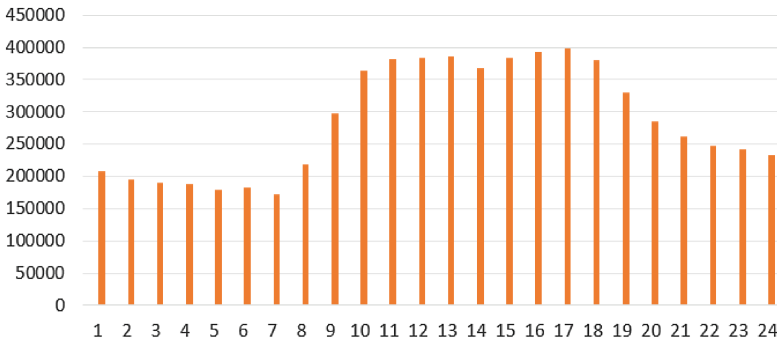


Fig. 1. The access frequency pattern of “smartfind.co.kr” over time

Secondly, the most querying IPs for the domain were found as shown in Table 2. A particular host queried more than two hundred thousand and some hosts looked like they were in the same company because they shared common class C type IP address. Using the “nslookup”, we found the hosts having IP addresses started from 211.235.32 belonged to the same company. Due to the privacy issue, we cannot specify the company’s name. However, from two results, we guessed that the hosts in the company were booted when workers came to the office and the hosts acted as malware zombie PCs for working hours.

Table 2. The most querying IP for the “smarfind.co.kr” domain

Rank	IP	Number of accesses
1	203.251.153.x	276,237
2	58.87.60.x	184,763
3	211.235.32.x	150,371
4	211.235.32.x	135,916

3.3 A Geographical Visualization

Infographics greatly helps people to understand situations and to make better decision. For this purpose, a geographical map was adopted to represent the access pattern in the research. Since the DNS log is closely related to the region information, this kind of infographic enables people to figure out the access pattern from particular region at one view.

Simply thinking, it seems very easy to show the locations of DNS queries on a map because the datasets already had the location information. However, the problem is that there is too much location information on the dataset. As shown in the Fig. 2(a), a map is easily covered with location pins. Therefore, in the research, the heat map visualization [11] was adopted to avoid this problem. It allows to display the data on the map as color-coded areas based on the density of locations on the map. The heat map from the access pattern according to countries is shown in Fig. 2(b). In the heat map, it easily guessed that the users in US, China, and Japan visited to Korea domain more than other countries.

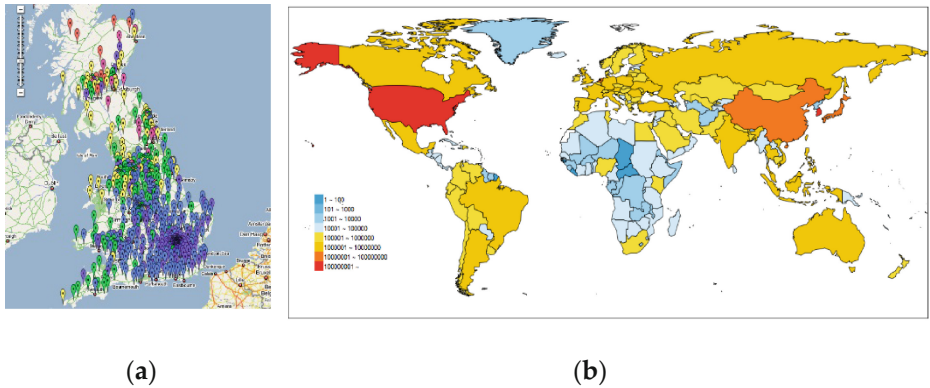


Fig. 2. (a) A map covered with location pins. (b) A heat map representing the access frequency according to country.

4 Conclusion

DNS log has been considered as a gold mine of valuable information to be used for the decision making on government policy or business plan because every Internet activity needs a DNS query before its session. Since DNS log is too huge for the conventional

analysis tools, Hadoop should be adopted for the analysis of DNS log. Unfortunately, the geographical dispersal of DNS conflicts with the Hadoop's basic assumption in which all data should be located on a single HDFS connected with high speed network. In this paper, a new method named LAM was proposed to solve this conflict. In the emulation, the LAM method proved that it enabled Hadoop to independently analyze DNS logs on the geographically dispersed DNS servers and it greatly reduced whole processing time. Several data mining results such as a malware detection were found and the results showed the data mining for DNS log can help people to do a better decision making.

References

1. Mockapetris, P., Dunlap, K.J.: Development of the domain name system. *ACM SIGCOMM Comput. Commun. Rev.* **18**(4), 123–133 (1988)
2. Snyder, M.E., Sundaram, R., Thakur, M.: Preprocessing DNS log data for effective data mining. In: *IEEE International Conference on Communications*, pp. 1–5 (2009)
3. Whyte, D., Kranakis, E., van Oorschot, P.C.: DNS-based detection of scanning worms in an enterprise network. In: *NDSS* (2005)
4. Choi, H., Lee, H., Lee, H., Kim, H.: Botnet detection by monitoring group activities in DNS traffic. *Comput. Inf. Technol.*, 715–720 (2007)
5. Bilge, L., Kirda, E., Kruegel, C.: EXPOSURE: finding malicious domains using passive DNS analysis. In: *NDSS* (2011)
6. Postel, J.: RFC-1591: domain name system structure and delegation. In: *IETF*, March 1994
7. White, T.: *Hadoop: The Definitive Guide*. O'Reilly, Sebastopol (2012)
8. Borhkur, D.: *The Hadoop distributed file system: architecture and design*. Hadoop Proj. Website (2007)
9. Thomas, W.H.: *Algebra*, 1st edn, p. 24. Springer, New York (1974)
10. McAfee Inc.: *Virus Profile & Definition* (2016). <https://home.mcafee.com/virusinfo/virusprofile.aspx?key=610775#none>
11. Nathan, Y.: *Visualize This: The Flowing Data Guide to Design Visualization, and Statistics*, 1st edn. Wiley, Hoboken (2011)

A Case-Based Approach to Colorectal Cancer Detection

Pedro Morgado¹, Henrique Vicente^{1,2}, António Abelha¹, José Machado¹,
João Neves³, and José Neves^{1(✉)}

¹ Centro Algoritmi, Universidade do Minho, Braga, Portugal
pedrommcs@hotmail.com, {abelha,jmac,jneves}@di.uminho.pt

² Departamento de Química, Escola de Ciências e Tecnologia, Universidade de Évora,
Évora, Portugal
hvicente@uevora.pt

³ Mediclinic Arabian Ranches, PO Box 282602, Dubai, United Arab Emirates
joaocpneves@gmail.com

Abstract. Colorectal cancer is one of the most common malignancies in developed countries. Although it is not well known what causes this type of cancer, studies have showed that there are certain risk factors associated that may increase the likelihood of developing such malignancy. These factors comprise, among others, individual's age, lifestyle habits, personal disease history, and genetic syndromes. Despite its high mortality, colorectal cancer may be prevented with an early diagnosis. Thus, this work aims at the development of Artificial Intelligence based decision support system to assess the risk of developing colorectal cancer. The framework is built on top of a Logic Programming approach to Knowledge Representation and Reasoning, complemented with a Case-based approach to computing that caters for the handling of incomplete, unknown, or even self-contradictory data, information or knowledge.

Keywords: Colorectal cancer · Knowledge representation and reasoning · Logic programming · Case-based reasoning · Decision support systems

1 Introduction

ColoRectal Cancer (CRC) refers to a type of cancer that starts in either the colon or the rectum. Colon cancer and rectal cancer have many features in common. Indeed, the main difference between them is in the anatomical location and the treatment procedure. Nevertheless, the tumour biology is exactly the same. In the majority of cases, colorectal cancers start with the development of a polyp, i.e., an abnormal growth that originates from the epithelial cells lining of the colon or rectum. Although most of these polyps are not carcinogenic, the adenomas may turning into adenocarcinomas and become cancer. Nowadays, *CRC* may be detected in an early stage since it is possible to find and remove precancerous polyps before turning into cancer [1].

CRC is the third most common cancer in Portugal, being only overtaken by breast and prostate cancer. Its incidence reaches over 30 cases per 100,000 inhabitants per year, and the mortality rate has been gradually increasing over the years. Indeed, throughout the last 3 decades the mortality rate of this disease has registered an increase of 80%.

In average, there are 3,300 deaths each year, corresponding to 10 deaths per day [2]. It occurs in all cultures, with evidences that certain races, such as black people, have a higher incidence of this disease than any other racial group. Male gender also shows a slightly higher risk of developing colorectal cancer than the female one [3].

Several risk factors may be associated to this illness, namely patient's age (more than 90% of cases occur in people over 50 years old), obesity, lifestyle issues (e.g., cigarette smoking, lack of physical activity), insulin resistance [3], genetic syndromes (e.g., familial adenomatous polyposis or Lynch syndrome) [4, 5], inflammatory intestinal conditions (e.g., ulcerative colitis or Crohn's disease) [6], or personal history of colorectal cancer [7], and even low levels of *HDL* cholesterol (e.g., values above 60 mg/dL convey some protection against *CRC*) [8]. The knowledge of the colorectal cancer risk factors may support preventive behaviours in order to reduce the likelihood of developing the disease. Despite the presence of one or more risk factors does not guarantee that an individual will develop rectal cancer, but increases the expectation of such happening. Solving problems related to the early detection of *CRC* requires a proactive strategy able to take into account all these factors, where the available data can be incomplete, self-contradictory and even unknown. Thus, this work is focused on the development of a hybrid methodology for problem solving, aiming at the elaboration of a clinical decision support systems to detect *CRC*, according to a historical dataset, under a *Case Based Reasoning* (*CBR*) approach to computing [9, 10]. Indeed, *CBR* provides the ability of solving new problems by reusing knowledge acquired from past experiences [9], i.e., *CBR* is used especially when similar cases have similar terms and solutions, even when they have different backgrounds [10]. Indeed, its use may be found in different arenas, namely in *Online Dispute Resolution* [11], or Medicine [12, 13], just to name a few.

2 Knowledge Representation and Reasoning

In specific judgments the available data, information or knowledge is not always exact in the sense that it can be estimated values, probabilistic measures, or degrees of uncertainty. Furthermore, knowledge and belief are generally incomplete, self-contradictory, or even error sensitive, being desirable to use formal tools to deal with the problems that arise from the use of these types of information [14, 15]. Many approaches to Knowledge Representation and Reasoning have been proposed using the *Logic Programming* (*LP*) epitome, namely in the area of *Model Theory* [16, 17] and *Proof Theory* [14, 15]. In the present work the proof theoretical approach is followed in terms of an extension to *LP*. An *Extended Logic Program* is a finite set of clauses in the form:

$$\begin{aligned}
 &\{\neg p \leftarrow \text{not } p, \text{not exception}_p \\
 &p \leftarrow p_1, \dots, p_n, \text{not } q_1, \dots, \text{not } q_m \\
 &?(p_1, \dots, p_n, \text{not } q_1, \dots, \text{not } q_m) \\
 &\text{exception}_{p_1} \quad \text{exception}_{p_j} (0 \leq j \leq k) \text{ being } k \text{ an integer number} \\
 &\}::\text{scoring}_{\text{value}}
 \end{aligned}$$

where “?” is a domain atom denoting *falsity*, the p_i, q_j , and p are classical ground literals, i.e., either positive atoms or atoms preceded by the classical negation sign \neg [14]. Indeed, \neg stands for a strong declaration that speaks for itself, and *not* denotes *negation-by-failure*, or in other words, a flop in proving a given statement, once it was not declared explicitly. Under this framework every program is associated with a set of *abducibles* [16, 17], given here in the form of exceptions to the extensions of the predicates that make the program, i.e., clauses of the form:

$$exception_{p_1}, \dots, exception_{p_j} (0 \leq j \leq k), \text{ being } k \text{ an integer number}$$

that stand for data, information or knowledge that cannot be ruled out. On the other hand, clauses of the type:

$$?(p_1, \dots, p_n, notq_1, \dots, notq_m) (n, m \geq 0)$$

also named invariants or restrictions, allows us to set the context under which the universe of discourse has to be understood. The term *scoring_{value}* stands for the relative weight of the extension of a specific *predicate* with respect to the extensions of peers ones that make the inclusive or global program.

2.1 Quantitative Knowledge

Aiming to set one’s approach to knowledge representation, two metrics will be set, namely the *Quality-of-Information (QoI)* and the *Degree-of-Confidence (DoC)*. The *QoI* of a logic program should be understood as a mathematical function that will return a truth-value ranging between 0 and 1, once it is fed with the extension of a given predicate, i.e., $QoI_i = 1$ when the information is *known (positive)* or *false (negative)* and $QoI_i = 0$ if the information is *unknown*, where the “*i*” denotes the term or clause “*i*” in a predicate’s extension. For situations where the extensions of the predicates that make the program also include *abducible* sets, its terms (or clauses) present a $QoI_i \in]0, 1[$ [18].

The *DoC*, in turn, stands for one’s confidence that the argument values or attributes of the terms that make the extension of a given predicate, having into consideration their domains, are in a given interval [19]. The *DoC* is figured using $DoC = \sqrt{1 - \Delta l^2}$, where Δl stands for the argument interval length, which was set to the interval [0, 1], since the ranges of attributes values for a given predicate and respective domains were normalized using $(Y - Y_{min}) / (Y_{max} - Y_{min})$, where the Y_s stand for themselves [19].

Thus, the universe of discourse is engendered according to the information presented in the extensions of such predicates, according to productions of the type:

$$predicate_i - \bigcup_{1 \leq j \leq m} clause_j(((A_{x_1}, B_{x_1})(QoI_{x_1}, DoC_{x_1}))), \dots, ((A_{x_j}, B_{x_j})(QoI_{x_j}, DoC_{x_j})) :: QoI_j :: DoC_j \tag{1}$$

where \mathcal{U} , m and l stand, respectively, for *set union*, the *cardinality* of the extension of *predicate* _{i} and the number of attributes of each clause [19]. The subscripts of *QoIs* and *DoCs*, x_1, \dots, x_p , stand for the attributes values ranges.

2.2 Qualitative Knowledge

In present study both qualitative and quantitative data/information/knowledge are present. Aiming at the quantification of the qualitative part and in order to make easy the understanding of the process, it will be presented in a graphical form. Taking as an example a set of n issues regarding a particular subject (where there are k possible choices (e.g., *absence*, *low*, ..., *high* and *very high*), let us itemized an unitary area circle split into n slices (Fig. 1). The marks in the axis correspond to each of the possible options.

If the answer to issue 1 is *high* the area correspondent is $\pi \times \left(\sqrt{\frac{k-1}{k \times \pi}} \right)^2 / n$, i.e., $(k-1)/(k \times n)$ (Fig. 1(a)).

Assuming that in the issue 2 are chosen the alternatives *high* and *very high*, the correspondent area ranges between $\left[\pi \times \left(\sqrt{\frac{k-1}{k \times \pi}} \right)^2 / n, \pi \times \left(\sqrt{\frac{k}{k \times \pi}} \right)^2 / n \right]$,

i.e., $[(k-1)/(k \times n), k/(k \times n)]$ (Fig. 1(b)). Finally, in issue n if no alternative is ticked, all the hypotheses should be considered and the area varies in the interval $\left[0, \pi \times \left(\sqrt{\frac{k}{k \times \pi}} \right)^2 / n \right]$, i.e., $[0, k/k \times n]$ (Fig. 1(c)).

The total area is the sum of the partial ones, i.e., $[(2 \times k - 2)/(k \times n), (3 \times k - 1)/(k \times n)]$ (Fig. 1(d)). In some situations similar responses to different issues have opposing impact in the subject in consideration. For example the assessment of healthy lifestyle includes issues like physical exercise practices and smoking status. The response *high* to the former issue has a positive contribution for healthy lifestyle, while the same response to smoking status has a negative one. Thus, the contribution of the items with negative impact on the subject in analysis is set as $k/(k \times n)$ minus the correspondent area, i.e., $(k/(k \times n) - (k-1)/(k \times n)) = 1/(k \times n)$ for issue 1, $[0, 1/(k \times n)]$ for issue 2 and $[0, k/k \times n]$ for issue 3.

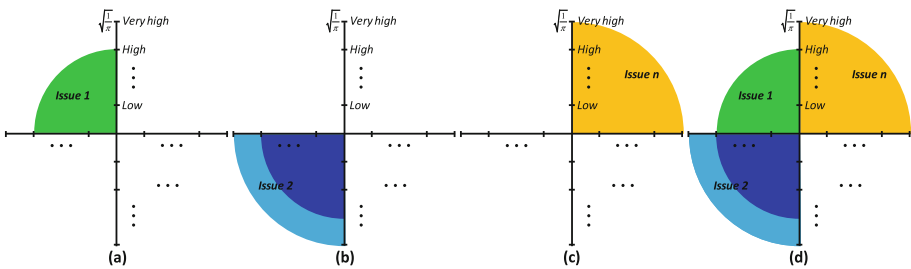


Fig. 1. A view of the qualitative data/information/knowledge processing.

3 A Case Based Methodology for Problem Solving

The *CB* methodology for problem solving stands for an act of finding and justifying a solution to a given problem based on the consideration of similar past ones, by reprocessing and/or adapting their data/knowledge [9, 10]. In *CB* the *cases* are stored in a *Case Base*, and those cases that are similar (or close) to a new one are used in the problem solving process. The typical *CB* cycle presents the mechanism that should be followed, where the former stage entails an initial description of the problem. The new case is defined and it is used to retrieve one or more cases from the *Case Base*.

Despite promising results, the current *CB* systems are neither complete nor adaptable enough for all domains. In some cases, the user cannot choose the similarity(ies) method(s) and is required to follow the system defined one(s), even if they do not meet their needs. Moreover, in real problems, the access to all necessary information is not always possible, since existent *CB* systems have limitations related to the capability of dealing, explicitly, with unknown, incomplete, and even self-contradictory information. Neves *et al.* [13, 20] induced a different *CB* cycle which takes into consideration the case's *QoI* and *DoC* metrics. It also contemplates a cases optimization process present in the *Case Base*, whenever they do not comply with the terms under which a given problem as to be addressed (e.g., the expected *DoC* on a prediction was not attained) [13, 20]. The optimization process can use *Artificial Neural Networks*, *Particle Swarm Optimization* or *Genetic Algorithms* generating a set of new cases which must be in conformity with the invariant:

$$\bigcap_{i=1}^n (B_i, E_i) \neq \emptyset \quad (2)$$

that states that the intersection of the attribute's values ranges for cases' set that make the *Case Base* or their optimized counterparts (B_i) (being n its cardinality), and the ones that were object of a process of optimization (E_i), cannot be empty.

4 Case Study

Aiming to develop a predictive model for early detection of *CRC*, a database was set, built on the health records of patients at a major health care institution in the north of Portugal. This section demonstrates how the information comes together and how it is processed.

4.1 A Logic Programming Approach to Data Processing

After having collected the data it is possible to build up a knowledge database given in terms of the extensions of the relations depicted in Fig. 2, which stand for a situation where one has to manage information aiming to detect *CRC*. Under this scenario some incomplete and/or unknown data is also available. For instance, in case 1, the *Previous*

CRC Episodes are unknown, which is depicted by the symbol \perp , while the HDL Cholesterol ranges in the interval [45, 55].

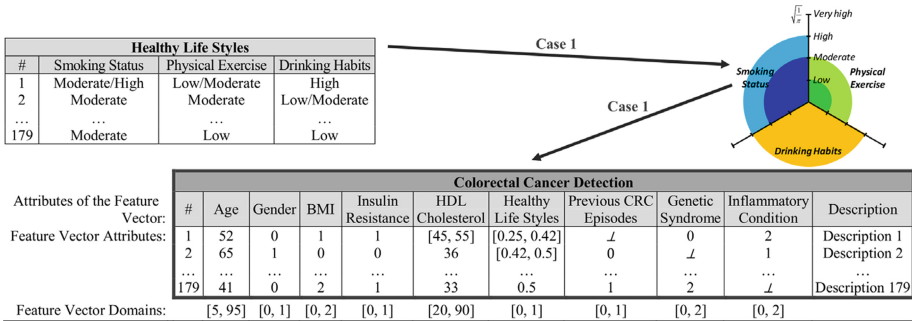


Fig. 2. A fragment of the knowledge base aiming at the early detection of colorectal cancer.

The *Insulin Resistance* and *Previous CRC Episodes* columns are filled with 0 and 1 denoting, respectively, *No* and *Yes*, while in *Gender* column 0 and 1 stand for *Female* and *Male*. The *BMI*, *Genetic Syndrome* and *Inflammatory Condition* columns are populated with 0, 1 and 2, standing, respectively, for *BMI < 25*, *BMI ranging in interval [25, 30[*, and *BMI ≥ 30* in the former case, *absence, familial adenomatous polyposis* and *Lynch syndrome*, in the second case and, finally, *absence, ulcerative colitis* and *Crohn’s disease* in the last one.

In order to quantify the information present in the *Healthy Life Style* tables the procedures already described in Sect. 2.2 were followed. Applying the algorithm presented in [19] to the fields that make the knowledge base for CRC detection (Fig. 2), excluding of such a process the *Description* ones, which will not be object of attention in this work, and looking to the *DoC_s* as described in [19], it is possible to set the arguments of the predicate *detection (detec)* referred to below, that also denotes the objective function with respect to the problem under analyze:

$$\begin{aligned}
 \text{detec: } & A_{ge}, G_{ender}, B_{ody}M_{ass}I_{ndex}, Ins_{uline}R_{esistance}, HDL_{Cholesterol}, L_{ife}S_{tyle} \\
 & H_{abits}, P_{revious}CRCE_{pisodes}, G_{enetic}S_{yndrome}, I_{nflammatory}C_{onditions} \rightarrow \{0, 1\}
 \end{aligned}$$

where 0 (zero) and 1 (one) denote, respectively, the truth values false and true.

The algorithm presented in [19] includes different phases. In the former one the clauses or terms that make extension of the predicate under study are established. In the subsequent stage the arguments of each clause are set as continuous intervals. In a third step the boundaries of the attributes intervals are set in the interval [0, 1] according to a normalization process given by the expression $(Y - Y_{min}) / (Y_{max} - Y_{min})$, where the Y_s stand for themselves. Finally, the *DoC* is evaluated. As an example considers a term (patient) that presents the feature vector $A_{ge} = 57, G_{ender} = 0, B_{ody}M_{ass}I_{ndex} = 1, I_{nsulin}R_{esistance} = 0, HDL_{Cholesterol} = [48, 62], L_{ife}S_{tyle}H_{abits} = [0, 0.25], P_{revious}CRCE_{pisodes} = 0, G_{enetic}S_{yndrome} = 0, I_{nflammatory}C_{onditions} = \perp$, one may obtain:

modulus of the arithmetic difference between the arguments of each case of the selected cluster and those of the new one. Thus, one may have:

$$\begin{aligned}
 &detec_1(\dots, ((0, 0.5)(1, 0.87))) :: 1::0.70 \\
 &detec_2(\dots, ((0, 0)(1, 1))) :: 1::0.76 \\
 &\quad \vdots \\
 &detec_j(\dots, ((0, 0)(1, 1))) :: 1::0.74
 \end{aligned}$$

$\underbrace{\hspace{15em}}_{\text{normalized cases from retrieved cluster}}$

Assuming that every attribute has equal weight, for the sake of presentation, the dissimilarity between $detec_{new}$ and the $detec_j$ may be computed as follows:

$$detec_{new \rightarrow 1}^{DoC} = \frac{\|1 - 1\| + \dots + \|1 - 0\| + \|1 - 0.97\| + \|1 - 0.87\|}{9} = 0.19$$

Thus, the similarity for $detec_{new \rightarrow 1}^{DoC}$ is set as $1 - 0.19 = 0.81$. Regarding QoI the procedure is similar, returning $detec_{new \rightarrow 1}^{QoI} = 1$. Thus, one may have:

$$sim_{newcase \rightarrow 1}^{QoI, DoC} = 1 \times 0.81 = 0.81$$

i.e., the product of two measurements is a new type of measurement. For instance, multiplying the lengths of the two sides of a rectangle gives its area, which is the subject of dimensional analysis. In this work the mentioned outcome gives the overall similarity between the new case and the retrieved ones. These procedures should be applied to the remaining cases of the retrieved clusters in order to obtain the most similar ones, which may stand for the possible solutions to the problem. This approach allows users to define the most appropriate similarity methods to address the problem (i.e., it gives the user the possibility to narrow the number of selected cases with the increase of the similarity threshold).

The proposed model was tested on a real data set with 179 examples. In order to evaluate the performance of the proposed model the dataset was divided in exclusive subsets through the ten-folds cross validation. In the implementation of the respective dividing procedures, ten executions were performed for each one of them. To ensure statistical significance of the attained results, 30 experiments were applied in all tests. The model accuracy was 91.6% (i.e., 164 instances correctly classified in 179). Thus, the predictions made by the *CB* model are satisfactory, attaining accuracies higher than 90%. The sensitivity and specificity of the model were 90.5% and 72.2%, while *Positive* and *Negative Predictive Values* were 86.4% and 94.7%, denoting that the model exhibits a good performance in the *CRC* detection. The methodology presented in this work is a generic one, and therefore may be applied in different grounds. Indeed, some interesting results have been obtained, namely to model the organizational efficiency in training corporations [18] and in health care context [13, 20].

5 Conclusions

This work presents an *Artificial Intelligence* based *Decision Support System* to detect *CRC* centred on a formal framework based on *LP* for knowledge representation and reasoning, complemented with a *CB* approach to computing that caters for the handling of incomplete, unknown, or even self-contradictory information. The proposed model is able to provide adequate responses once the overall accuracy is higher than 90%. Indeed, it has also the potential to be disseminated across other prospective areas, therefore validating a universal attitude. Additionally, it gives the user the possibility to narrow the search space for similar cases at runtime by choosing the most appropriate strategies to address the problem. In fact, the added values of the presented approach arises from the complementarity between *Logic Programming* (for knowledge representation and reasoning) and the computational process based on *Case Based* approach.

Acknowledgments. This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

References

1. Chehade, R., Robertson, D.: Screening and surveillance guidelines. In: Anderson, J., Kahi, C. (eds.) *Colorectal Cancer Screening*, pp. 43–55. Humana Press, New York (2011)
2. Directorate-General of Health: PORTUGAL Oncological Diseases in Numbers – 2015. Directorate-General of Health, Lisbon (2016)
3. Anderson, J.: Risk factors and screening for colorectal cancer. In: Anderson, J., Kahi, C. (eds.) *Colorectal Cancer Screening*, pp. 7–23. Humana Press, New York (2011)
4. Lynch, H., de la Chapelle, A.: Genetic susceptibility to non-polyposis colorectal cancer. *J. Med. Genet.* **36**, 801–818 (1999)
5. Khan, M., Burke, C.: Hereditary adenomatous colorectal cancer syndromes. In: Anderson, J., Kahi, C. (eds.) *Colorectal Cancer Screening*, pp. 25–41. Humana Press, New York (2011)
6. Eaden, J., Mayberry, J.: Guidelines for screening and surveillance of asymptomatic colorectal cancer in patients with inflammatory bowel disease. *Gut* **51**, V10–V12 (2002)
7. Haggard, F., Boushey, R.: Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin. Colon Rectal Surg.* **22**, 191–197 (2009)
8. van Duijnhoven, F., Bueno-de-Mesquita, H., Calligaro, M., Jenab, M., Pischon, T., Jansen, E., et al.: Blood lipid and lipoprotein concentrations and colorectal cancer risk in the European prospective investigation into cancer and nutrition. *Gut* **60**, 1094–1102 (2011)
9. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Commun.* **7**, 39–59 (1994)
10. Richter, M.M., Weber, R.O.: *Case-Based Reasoning: A Textbook*. Springer, Berlin (2013)
11. Carneiro, D., Novais, P., Andrade, F., Zeleznikow, J., Neves, J.: Using case-based reasoning and principled negotiation to provide decision support for dispute resolution. *Knowl. Inf. Syst.* **36**, 789–826 (2013)
12. Blanco, X., Rodríguez, S., Corchado, Juan, M., Zato, C.: Case-based reasoning applied to medical diagnosis and treatment. In: Omatu, S., Neves, J., Rodríguez, Juan M., Corchado, Paz Santana, J.F., Gonzalez, S.R. (eds.) *AISC*, vol. 217, pp. 137–146 Springer, Heidelberg (2013). doi:[10.1007/978-3-319-00551-5_17](https://doi.org/10.1007/978-3-319-00551-5_17)

13. Quintas, A., Vicente, H., Novais, P., Abelha, A., Santos, M.F., Machado, J., Neves, J.: A case based approach to assess waiting time prediction at an intensive care unity. In: Arezes, P. (ed.) *Advances in Safety Management and Human Factors. Advances in Intelligent Systems and Computing*, vol. 491, pp. 29–39. Springer International Publishing, Cham (2016)
14. Neves, J.: A logic interpreter to handle time and negation in logic databases. In: Muller, R., Pottmyer, J. (eds.) *Proceedings of the 1984 Annual Conference of the ACM on the 5th Generation Challenge*, pp. 50–54. Association for Computing Machinery, New York (1984)
15. Neves, J., Machado, J., Analide, C., Abelha, A., Brito, L.: The halt condition in genetic programming. In: Neves, J., Santos, M.F., Machado, J.M. (eds.) *EPIA 2007. LNCS (LNAI)*, vol. 4874, pp. 160–169. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-77002-2_14](https://doi.org/10.1007/978-3-540-77002-2_14)
16. Kakas, A., Kowalski, R., Toni, F.: The role of abduction in logic programming. In: Gabbay, D., Hogger, C., Robinson, I. (eds.) *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 5, pp. 235–324. Oxford University Press, Oxford (1998)
17. Pereira, L., Anh, H.: Evolution prospection. In: Nakamatsu, K. (ed.) *Studies in Computational Intelligence*, vol. 199, pp. 51–64. Springer, Berlin (2009)
18. Fernandes, A., Vicente, H., Figueiredo, M., Neves, M., Neves, J.: An adaptive and evolutionary model to assess the organizational efficiency in training corporations. In: Dang, T.K., Wagner, R., Küng, J., Thoai, N., Takizawa, M., Neuhold, E. (eds.) *Future Data and Security Engineering. Lecture Notes on Computer Science*, vol. 10018, pp. 415–428. Springer International Publishing, Cham (2016)
19. Fernandes, F., Vicente, H., Abelha, A., Machado, J., Novais, P., Neves, J.: Artificial neural networks in diabetes control. In *Proceedings of the 2015 Science and Information Conference (SAI 2015)*, pp. 362–370. IEEE Edition (2015)
20. Silva, A., Vicente, H., Abelha, A., Santos, M.F., Machado, J., Neves, J., Neves, J.: Length of stay in intensive care units – a case base evaluation. In: Fujita, H., Papadopoulos, G.A. (eds.) *New Trends in Software Methodologies, Tools and Techniques, Frontiers in Artificial Intelligence and Applications*, vol. 286, pp. 191–202. IOS Press, Amsterdam (2016)
21. Figueiredo, M., Esteves, L., Neves, J., Vicente, H.: A data mining approach to study the impact of the methodology followed in chemistry lab classes on the weight attributed by the students to the lab work on learning and motivation. *Chem. Educ. Res. Pract.* **17**, 156–171 (2016)

Multi-Modes Cascade SVMs: Fast Support Vector Machines in Distributed System

Lijuan Cui^{1(✉)}, Changjian Wang¹, Wanli Li², Ludan Tan¹, and Yuxing Peng¹

¹ Science and Technology on Parallel and Distributed Laboratory,
National University of Defense Technology, Changsha, China
cuilijuan2010@163.com, wangchangjian@163.com, tanludan@163.com,
pengyuxing@163.com

² College of Computer Science and Electronic Engineering,
Hunan University Changsha, Changsha 410073, Hunan, China
liwanli@163.com

Abstract. Machine learning is one field of Artificial Intelligence (AI) to help machines solve problems. Support Vector Machines (SVMs) are classic methods in machine learning field and are also used in many other AI fields. However, the model training is very time-consuming when meeting large scale data sets. Some efforts have been devoted to develop it for distributed memory clusters. Their bottleneck is the training phase, where the structure is immobile. In this paper, we propose Multi-Modes Cascade SVMs (MMCascadeSVMs) to adaptively reshape the structure. MMCascadeSVMs employs analytical hierarchy process to qualitatively analyse the similarity between adjacent hierarchies. Furthermore, MMCascadeSVMs leverages a two-stage algorithm: the first stage is to compute the similarity between two adjacent models, and the similarity is built for halting criterion. The second stage is to predict new samples based on multi models. MMCascadeSVMs can modify the structure of SVMs in distributed systems and reduce training time. Experiments show that our approach significantly reduces the total computation cost.

Keywords: Cascade SVMs · Analytical hierarchy process

1 Introduction

In many real world problems such as text classification, image recognition, financial markets, Support Vector Machines (SVMs) is a popular classification tool [2, 4, 5]. SVMs establishes the separating hyperplane by discovering the support vectors (SVs). However, the training of SVMs requires to solve a quadratic programming (QP) problem with n inequality constraints and one equality constraint [8–10], where n is the training set size. The iterative process can't be easily paralleled for its recurrence relation.

To parallelize SVMs, some “cascade methods” [6, 7, 11] are proposed. They divide the data into several subproblems, then train local models on their own nodes. They pass the local models to the higher layer and combine them in a “tree” way. The process repeats until a single model remains. These works only differ from the way dividing data

sets and the concept passing to higher layers. The computational process is still slow when the data set is large, which narrows the applicable scope of cascade SVMs.

We notice the fact that: (i) cascade SVMs are time-consuming in higher layers but identify few support vectors. We call those layers as ineffective structure; (ii) the model in the highest layer is not always the best one; (iii) the nodes utilization rate is low in higher layers. Motivated by this insight, we aim at adaptively reshaping the structure to reduce training time without losing accuracy. The intuitive idea is to prune the ineffective structure. However, we still need to face the following challenges: (i) how to define the ineffective structure? (ii) how to automatically prune the ineffective structure without losing accuracy when there is no prior knowledge? (iii) how to predict new instances in the new structure of cascade SVMs?

In this paper, we respond to these challenges with Multi-Modes Cascade SVMs (MMCascadeSVMs), a novel training method adaptively changing the structure of cascade SVMs. In order to measure the similarity between adjacent hierarchical models, MMCascadeSVMs constructs the hierarchical model based on local models in the same layer. Then hierarchical similarity is proposed to qualitatively analyse how much the model is modified between adjacent hierarchies. MMCascadeSVMs exploits the hierarchical similarity to serve as the stop condition of cascade SVMs. Since the structure has changed, there exist multi models when the process halts. These models work cooperatively when dealing with new instances in the testing phase. Our approach efficiently reduces the computation cost compared to the existing approaches.

To summary up, we list our contributions as follows:

1. We propose hierarchical similarity in cascade SVMs to measure the similarity between adjacent hierarchical models. MMCascadeSVMs computes the hierarchical similarity to measure the degree of modification.
2. We propose Adaptively Reshaping Cascade SVMs (ARC) method applying hierarchical similarity. ARC can automatically prune the ineffective structure in cascade SVMs.
3. We propose MMP method to improve the detection accuracy in the testing phase. MMP ensures that new instances can be predicted even there exist multi models.

MMCascadeSVMs reshapes the structure of cascade SVMs and the number of total layers is not more than the latter one. Therefore, it performs less computation than the existing works. Extensive experiments have been conducted on four benchmark data sets (ijcnn1, a7a, shuttle, covtype) [1] to evaluate the classification performance. MMCascadeSVMs significantly reduces the training time and keeps good accuracy.

2 Related Works

Related works of paralleling SVMs on large-scale data sets can be roughly classified into two categories as follows.

The first category is to parallel the internal algorithms. For example, Zanghirati proposes a parallel implementation of SMO [12]. It splits the problem into a sequence of subproblems which are parallely solved. PSVM parallels “Inner-Point method” [3]. Through

approximating extensive matrix manipulations using parallel computing, the method only loads essential data in each iteration and can reduce a lot of training time. P-pack SVM parallels stochastic gradient descent [13]. It can pre-calculate the predicted labels and kernel matrix, and update the distributed hash table after iterations finishing. But there are intense communication between nodes.

Another approach is cascade SVMs which are popular multi-levels approaches in distributed environment. Figure 1 shows the training way. The main idea is that they divide the data sets into p pieces, where p is the number of nodes. They obtain local models based on the local data. They combine the local models in a “tree” structure until there is a single model. Based on the way dividing the data and the content passing, the methods are classified into Cascade SVM and DCSVM. Cascade SVM divides the data evenly while DCSVM divides it using K-means. Cascade SVM passes the candidate support vectors while DCSVM passes all the data using the lower models to initialize the higher model. Since the data in higher layer is the support vectors of the lower layer models in Cascade SVM, the number is fewer than that in DCSVM. The data in higher layer can correct the error that occurs in the lower layers of DCSVM, the detecting accuracy is higher than that in Cascade SVM.

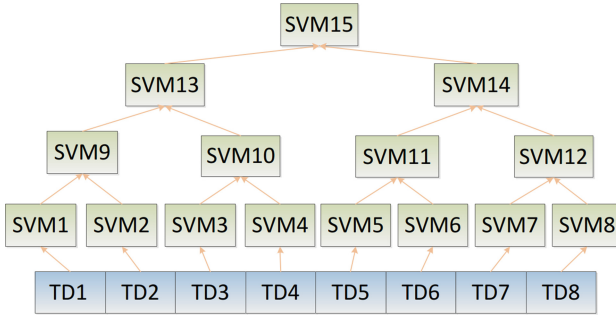


Fig. 1. The structure of cascade SVMs

3 ARC: Adaptively Reshaping Cascade SVMs

In binary-class SVMs, training set with n samples is represented as (x_i, y_i) , $i = 1, 2, 3, \dots, n$, where $x_i \in \mathbf{R}^d$ is the i-th input data point and $y_i \in \{-1, 1\}$ is the i-th output. The decision classification rule for a data point x becomes:

$$f(x) = \text{sign}\left(\sum_{i=1}^N \alpha_i y_i K(x_i, x)\right) \tag{1}$$

Where α_i are Lagrange multipliers. Each $\alpha_i \geq 0$ is associated with the sample (x_i, y_i) , $K(x_i, x_j)$ is the kernel function between sample x_i and sample x_j . The support vectors are those samples whose coefficients α_i are non-zero. All the coefficients are expressed as $\alpha^* = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$. The fact that only the support vectors have effect on the construction of hyperplane, can be observed from Eq. (1). In SVMs, if the coefficients of two models are same when dealing with the same data sets, the separating hyperplanes are same. Similarly,

if the coefficients of two models are semblable, the separating hyperplanes are alike. Therefore, if the solutions of the problem α^* are similar, the separating hyperplanes are similar.

We evaluate the performance of every layer in cascade SVMs. We observe that cascade SVMs are time-consuming in higher layers but identify few support vectors. In this paper, we call those layers as ineffective structures. In order to reduce the training time by pruning the ineffective structures, we propose the hierarchical similarity to measure the similarity between adjacent hierarchical models. Assuming the hierarchical models α and α' , a given threshold $\tau \geq 0$, if cosine similarity between α and α' is greater than τ , the models associated with α and α' are similar.

Note that the ineffective structures of cascade SVMs have little effect on the final model. The basic thinking is to identify those structures and remove them in the training phase, thus it will reduce the computation time. The key point is how to identify ineffective structures and screen out the “inactive” structures. We investigate the similarity between hierarchical layers and motivate our identifying rules. In MMCascadeSVMs, the hierarchical similarity can help identify the ineffective layers.

In the training phase, if the hierarchical similarity is high, the degree of model modified little by the higher level. The process of the training is long, but the number of support vectors identified in the higher layer is small. If the hierarchical similarity between layer l_{t-1} and layer l_t is higher than the threshold τ , then the l_t layer and layers higher than l_t are deemed to the ineffective structures. Therefore, the ineffective models are screened out. The cascade SVMs will stop at l_t layer. The process is summarized in Algorithm 1.

Algorithm 1. The algorithm of ARC

Input: Training Sets $\{(x_i, y_i)\} i = \{1, 2, \dots, n\}$, constant m , threshold τ .

- 1: for $l = l^{max}, \dots, 1$ do
- 2: if $l = l^{max}$ then
- 3: Sample m instances from the training set.
- 4: else
- 5: Sample m instances whose coefficient $\alpha_i^{(l+1)} > 0$
- 6: end if
- 7: Run kernel k-means on the m selected instances.
- 8: Identify the clusters for the rest of the data and obtain the data partition v_1, v_2, \dots, v_{kl} .
- 9: for $c = 1, \dots, k'$ do
- 10: Obtain the local solution α_c
- 11: end for
- 12: Concatenate the local solution to get the approximate layer solution α
- 13: if $l = l^{max}$ then
- 14: if $\frac{last \cdot \alpha}{|last| \cdot |\alpha|} > \tau$ then
- 15: BREAK;
- 16: end if
- 17: end if
- 18: last $\leftarrow \alpha$
- 19: end for

The algorithm 1 builds the halting criterion. Based on the rule, we can get the layer MMCascadeSVMs stops.

4 MMP: Predicting Algorithm Based on Multi-Modes Cascade SVMs

Multi-modes have two different meanings. The first dimension is multi training modes in the same layer; and the second dimension is multi models when the training process halts.

In MMCascadeSVMs, the data is split based on the clustering in the same way as DCSVM. Imbalance may occur in the process. There is a risk that the instances in one node are a single class of data. If this case happens, the treatment in the node is different from others. Those samples are likely to be non-support vectors which locate far from the separating hyperplane. We set the solution of the node α_i^* is $\mathbf{0}$ or we can use one class SVMs to train the model. We use original SVMs classifiers to train models in other nodes. The models and data are passed to the higher layer. If there exist some support vectors, they will be identified in the higher layer. There exist different SVMs training modes in the same layer of MMCascadeSVMs.

There exist multi-models when the training process halts. In the predicting phase, the new instances are usually predicted based on the final one model instead of multi models. To cope with the problem, three strategies are proposed for multi models to predict new instances.

Strategy 1: The practical method is to collect the data and models and then pass them to the higher layer. We concatenate the models to form an approximate solution and use it as the initial value of the highest layer. It trains the final model using all the data in the highest layer. There is only one model finally, and the final model can be used to predict new instances. Since some ineffective layers are removed from the structure, and it reduces the training time and improves the training efficiency.

Strategy 2: When MMCascadeSVMs halts, a multitude of sub-SVMs models remain to predict new instances. Each model is a binary-class classifier. MMCascadeSVMs constructs an ensemble classifier that consists all the retained models and outputs the class that is the mode of the class's output by every individual model. Since the models store in multi nodes, the predicting processes are non-interfering in the predicting phase. The support vectors in each model are fewer than all the support vectors, so the predicting time using sub-model is shorter. In this way, it can reduce the predicting time.

Strategy 3: Each sub-model represents the hyperplane constructed by the local data. New instances can be classified based on the local models. Therefore, MMCascadeSVMs can select the model whose cluster is closest to the new instance in predicting phase. After MMCascadeSVMs halting, the predicting label depends on the particular cluster where we are trying to evaluate x . To make predictions using local model, we need to keep the entire models around to compute the nearest cluster.

5 Experiments

We adopt four benchmark data sets for experiments: *ijcnn1*, *a7a*, *shuttle*, *covtype* [1]. Figure 2(a) shows the statistical result of the hierarchical similarity in MMCascadeSVMs. The influence of layer to hierarchical similarity is considered. The difference between layers is smaller when the training layer is higher.

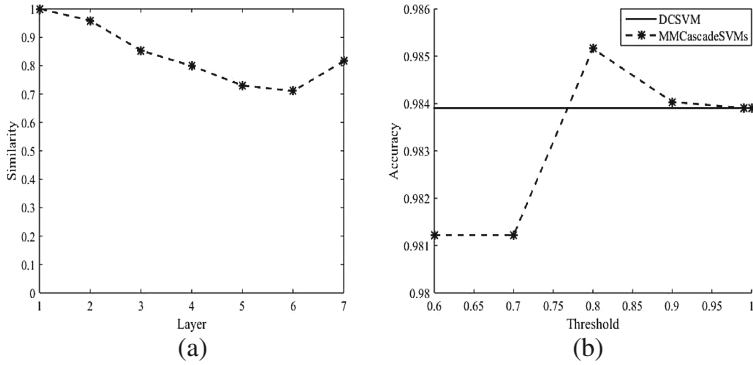


Fig. 2. (a) the hierarchical similarity on *ijcnn1*. The layer 1 is the highest layer while layer 7 is the lowest one; (b) the relationship between the threshold τ and accuracy.

The threshold τ we set in Algorithm 1 is the basis we halt the process of cascade SVMs training. When the training process stops, we employ the first strategy to evaluate the performance of accuracy and efficiency. Figure 2(b) shows the relationship between accuracy and the threshold τ . Moreover, Fig. 3 shows the relationship between the threshold τ and the training time on different data sets. From these experimental results, we have the following observation:

- (1) The similarities are near when τ ranges within 0.6 and 1. The model in the highest level is not the best one. Therefore, MMCascadeSVMs can halt the training process before it arrives the highest layer.
- (2) Training time increases with τ . Combining the rule above, MMCascadeSVMs can get higher accuracy than the original method in the predicting phase.

Table 1. Training accuracy and training time of the proposed three strategy on data set *ijcnn1*.

	Second layer		Third layer		Fourth layer	
	Accuracy	Time	Accuracy	Time	Accuracy	Time
Strategy 1	0.9839	15.1477	0.9835	17.6125	0.9835	17.6125
Strategy 2	0.9237	5.1324	0.9244	4.7736	0.9082	2.5428
Strategy 3	0.9839	14.7733	0.9835	14.1181	0.9829	13.8841

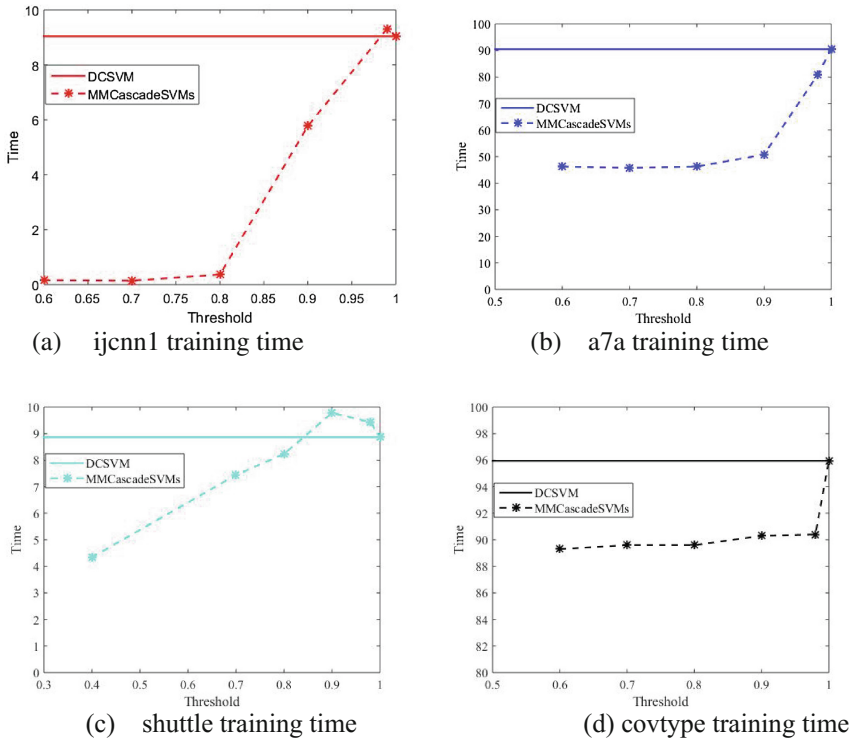


Fig. 3. Classification performance measured in training time on the four benchmark data sets using RBF kernel function.

To cope with the problem of multi models in the predicting phase, we propose three strategies. Table 1 shows the experimental results.

We have the following observation:

- (1) The accuracy of strategy 1 is the highest while strategy 2 is the lowest. The weak independence models affect the predicting result of strategy 2, leading poor accuracy.
- (2) The predicting time of strategy 2 is one-third of that in strategy 1. In strategy 2, each model stores in its own node and it owns fewer support vector than that of the total model. The predicting time is less than other strategies.

6 Conclusion

In this paper, we propose Multi-Modes Cascade SVMs (MMCascadeSVMs) to accelerate cascade SVMs. MMCascadeSVMs identifies the ineffective layers in the structure. This is achieved by computing the hierarchical similarity between adjacent layers, removing the ineffective structures. To cope with the multi models in the predicting phase, the predicting algorithm based on Multi-Models Cascade SVMs is proposed. We conduct extensive experiments on large-scale data sets to demonstrate the efficiency of MMCascadeSVMs.

Acknowledgment. The work was supported by the National Basic Research Program of China (project No. 2014CB340303) and the National Natural Science Foundation of China (project No. 61402514).

References

1. <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>
2. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**(2), 121–167 (1998)
3. Chang, E.Y., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., Cui, H.: PSVM: parallelizing support vector machines on distributed computers. In: *Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December, pp. 213–230 (2007)
4. Chien, L.J., Chang, C.C., Lee, Y.J.: Variant methods of reduced set selection for reduced support vector machines. *J. Inf. Sci. Eng.* **26**(1), 183–196 (2010)
5. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods*. China Machine Press, Beijing (2005)
6. Graf, H.P., Cosatto, E., Bottou, L.: Parallel support vector machines: the cascade SVM. In: *Advance in Neural Information Processing Systems* (2008)
7. Hsieh, C.J., Si, S., Dhillon, I.S.: A divide-and-conquer solver for kernel support vector machines. In: *International Conference on Machine Learning*, pp. 566–574 (2014)
8. Menon, A.K.: *Large-scale support vector machines: algorithms and theory*. Research Exam University of California, San Diego, pp. 1–17 (2009)
9. Platt, J.C.: *Sequential minimal optimization: a fast algorithm for training support vector machines* (1998)
10. Tsang, I.W., Kwok, J.T., Cheung, P.M.: Core vector machines: fast SVM training on very large data sets. *J. Mach. Learn. Res.* **6**(1), 363–392 (2005)
11. You, Y., Demmel, J., Czechowski, K., Song, L.: CA-SVM: communication avoiding support vector machines on distributed systems, pp. 847–859 (2015)
12. Zanghirati, G., Zanni, L.: A parallel solver for large quadratic programs in training support vector machines. *Parallel Comput.* **29**(4), 535–551 (2002)
13. Zhu, Z.A., Chen, W., Wang, G., Zhu, C., Chen, Z.: P-packSVM: parallel primal gradient descent kernel SVM, pp. 677–686 (2009)

Deep Learning Based Recommendation: A Survey

Juntao Liu¹ and Caihua Wu²

¹ 709th Research Institute, China Shipbuilding Industry Corporation, Wuhan, China
prolay@163.com

² Section of Automatic Command, Huang Pi NCO School, Air Force Early-Warning Academy,
Wuhan, China
wucaihua2009@163.com

Abstract. Due to the great success, deep learning gains much attentions in the research field of recommendation. In this paper, we review the deep learning based recommendation approaches and propose a classification framework, by which the deep learning based recommendation approaches are divided according to the input and output of the approaches. We also give the possible research directions in the future.

Keywords: Deep learning · Recommendation · Neural network

1 Introduction

In recent years, deep learning, as a kind of machine learning approach, is applied in many different research domains successfully, such as computer vision, speech recognition, natural language processing and so on. In these fields, compared with traditional approaches, deep learning based approaches improve the performance remarkably. Due to the great success of deep learning, some researchers try to use deep learning in recommendation systems [1–11, 13–15], and wish these newly proposed model can improve the performance of the recommendation systems, just as the deep learning models do in other research fields.

In this paper, the deep learning based recommendation approaches are reviewed under the proposed classification framework for this kind of approaches. In this framework, the deep learning based recommendation approaches are classified by the input and the output of the approaches. Introducing deep learning in the recommendation systems is a related new research direction. Several approaches are proposed, but few is used in practice. Scalability is a problem in the recommendation systems, in which there are huge number of items and users. According to this fact, we also point out the future research direction of the deep learning based recommendation.

This paper is organized as follows. First, the background of recommendation and deep learning are introduced in Sect. 2. In Sect. 3, the classification framework for the deep learning approaches are introduced, and these approaches are also reviewed. The future research directions are given in Sect. 4 followed by the conclusion in Sect. 5.

2 Background

2.1 Recommendation

Traditionally, recommendation systems are divided into three kinds: content-based, collaborative and hybrid recommendation approaches, in the view of the recommendation approaches [17]. Recommendation systems are also classified according to the application fields [18]. In this paper, we review the recommendation systems in the view of the underlying recommendation problems. In other word, we divide the recommendation systems according to the input and output of the recommendation problems addressed in the systems.

Recommender systems usually collect the users' activities in the systems, including rating, clicking, buying, comment and so on, which are the input of the recommendation approaches. Rating is the most used input in recommendation systems, in which users are allowed to rate the item by a k -point integer. Recommendation results are generated according to the ratings. The typical approaches are PMF [19] and its extensions. Usually, rating is regarded as a kind of explicit feedback. Some approaches take implicit feedback, such as clicks, view and so on, as input. Compared with explicit feedback, implicit feedback can be collected easily by the recommendation systems. BPR [20] infers the binary preference relation of a user between items from the implicit feedback of this user, but it can handle only one kind of implicit feedbacks. Instead of learning from only one kind of feedback, some recommender systems lavage multi-kind of feedback [6, 21]. Wu et al. [6] also use the time information of the feedback in their approach.

The output of recommender systems is the recommendation result given to the users. Different recommender systems give different kinds of recommendation results. Some systems predict the ratings that the users have not issued. Matrix factorization based approaches are the most popular approaches resolving the rating prediction problem. Some recommendation systems predict the preference order of the uses among the items, which are called ranking based recommendation approaches. This kind of approaches include BPR [20], ListPMF [22] and QPMF [23]. In some cases, users may expect a combination of products, such as a jacket and the matched pant. It is raised the combination recommendation problem. [24, 25] address this problem in the field of cloth recommendation.

2.2 Deep Learning

Nowadays, deep learning refers to class of machine learning algorithm. Usually, the model of deep learning contains a cascade of nonlinear transformation layers. The parameters in the models are learned by end-to-end optimization.

Several kinds of deep learning models are proposed. One of the most used model is the feedforward neural network. A typical feedforward neural network with only one hidden layer is shown in Fig. 1(a). The feedforward neural networks used in deep learning usually have several hidden layers with different dimensions in order to encode the input in high level abstractions. The output O in Fig. 1(a) is computed as follows,

$$\begin{aligned}
 h &= \sigma(Wx) \\
 O &= \delta(Vh)
 \end{aligned}
 \tag{1}$$

where V and W are the weight matrices for hidden layer to output layer and input layer to hidden layer. Functions δ and σ are the nonlinear transformation functions, such as tanh and sigmoid function.

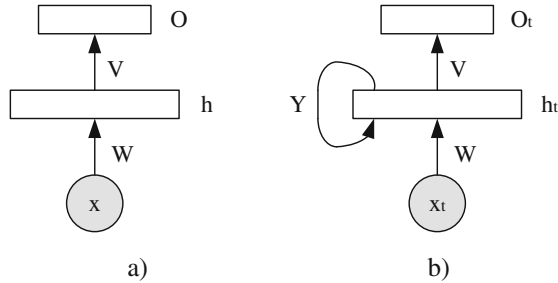


Fig. 1. Structure of feedforward neural network (a) and recurrent neural network (b).

Because the feedforward neural network can't analyze sequence data, such as natural language, recurrent neural network is designed. In this model, the hidden layers are connected recurrently to the input layers, as shown in Fig. 1. Structure of feedforward neural network (a) and recurrent neural network (b). The output at time t , O_t , is calculated as follows,

$$\begin{aligned}
 h_t &= \sigma(Wx_t + Yh_{t-1}) \\
 O_t &= \delta(Vh_t)
 \end{aligned}
 \tag{2}$$

where Y is the weight matrix for previous hidden layer to current hidden layer.

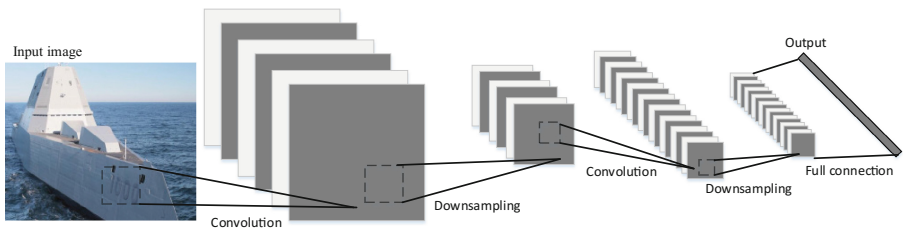


Fig. 2. Typical convolutional neural network for image encoding.

Convolutional neural networks are widely used for image recognition, and it contains one or more convolutional layers where the neurons are tiled in such a way that they respond to overlapping regions in the visual field. Between two neighbor convolutional layers there is always a pooling layer for subsampling. A typical convolutional neural network is shown in Fig. 2. Typical convolutional neural network for image encoding..

Because the parameters in CNN are much fewer than other deep learning model, it is easier to train. This makes them a highly attractive architecture.

3 Classification of Deep Learning Based Recommendation Methods

3.1 Classification Framework

In this paper, the deep learning based recommendation approaches are classification by the input and output of the approaches. In the aspect of input, some approaches take the content information into consideration and some approaches do not using this kind of information. In the view of output, user ratings and user activity sequence are predicted, respectively. Our classification frame work is shown in Table 1.

Table 1. Classification framework of deep learning based recommendation

Output	Input	
	Approaches using content information	Approaches without content information
Rating	[1, 2, 14]	[7, 8, 10, 15]
Order	[3, 4, 11, 13]	[5, 6]

3.2 Classification by Input

Approaches Using Content Information. Oord et al. [1] propose a music recommendation method based on deep neural network. In this method, the user and item latent feature vectors is learned first by weighted matrix factorization (WMF) algorithm [16]. And then the item latent feature vectors are learned by a deep convolutional network further, in which the already learned user latent feature vectors and the music signal are taken as input. The objective functions used to train the neural network are the weighted rating prediction error (WPE) just as that in WMF and the difference between the item latent feature vectors learned by WMF and those learned by the neural network. Finally, recommendation is made as in standard MF i.e. inner product between user and item latent factors. Because the item latent feature vectors are learned from item content, this method is specifically useful in Item cold start situation, where no feedback on target item is available. This method was later used at Spotify in an experiment [2].

Matrix factorization and deep belief network (DBN) are integrated in [14]. In this model, the item (music) latent feature vector is the output of DBN, which is trained previously according to the content information of the music.

Also in the music domain, Hamel et al. [3] designed two neural network models to predict music tags. The predicted tags are used for music recommendation. The input of these models is the preprocessed music feature, which is obtained by discrete Fourier transform (DFT), mel-compression and principal component analysis whitening (PCA). The author investigate the performance of several pool functions and find that combining several pool functions can improve the performance.

Elkahky et al. [4] propose a Deep Learning approach to map users and items to a latent space where the similarity between users and their preferred items is maximized. They learn the item feature learn from different domains. And the user feature is learned user features by a multi-view Deep Learning model. The proposed method is test for Windows Apps, news and Movie/TV recommendation.

Wang et al. [11] propose Bayesian stacked denoising autoencoder (SDAE) [12], and integrate this model with Bayesian probabilistic matrix factorization (BPMF), which is called collaborative deep learning (CDL), to address the problem of implicit feedback recommendation. In their method, the latent item feature is first generated by a previous trained SDAE model according to the item content information. And then, user latent feature vector and rating issued by this user are generated in the way similar to that in BPMF.

There are two neural network models in the recommender system of YouTube [13]. One model is used to generate hundreds of candidate items from huge number of items according to the historical behaviors of the users. The candidate items are then ranked by the other model according to the historical user behavior, context information and item feature.

Approaches Without Content Information. Some approaches do not use content information. In these method, only the feedback of the users, such as ratings, clicking and so on, are used to generate the recommendation results. For example, [7, 8] use the observed rating to predict the ratings of the items that the user have not accessed. In [5], implicit feedback is used to generate the ranked item orders. And [6] take multiple feedback as the input of the deep learning model. The input of [15] is the user-tag matrix.

3.3 Classification by Output

Item Order Prediction Methods. As mentioned above, the approaches using content information, such as [3, 4, 11, 13], train deep learning models and rank the items. The ranked item orders are recommended to the users.

Some approaches do not use content information. Hidasi et al. [5] propose a session-based recommendation method using recurrent neural network (RNN). In their method, the user accessed items are treated as sequences. The predicted item sequences that the users may accessed is generated by the trained RNN model in the end-to-end manner. There are embedding layer, feed forward layer and several GRU layers in the proposed model. The authors find that pair-wise loss function is better than point-wise loss function. And the model with single GRU layer is better than that with several GRU layers.

Wu et al. [6] propose a recurrent neural network based recommendation approach (RNNRec) to address the problem of time heterogeneous feedback recommendation. In this work, historical feedback activities with time stamps of the users are treated as sequences. And a recurrent neural network is trained using these feedback sequences. It is reported that the recommendation results generated RNNRec are more accurate than those generated by the traditional recommendation methods.

Rating Prediction Methods. Content information are used in [1, 2, 14] to predict the ratings.

There are approaches only using user feedback, such as rating and tag, to generate recommendation. Salakhutdinov et al. [10] use RBMs for collaborative filtering. RBMs can be used as a fundamental units of Deep Neural Networks. But in [10], there is only a single layer in RBM. Additionally, Edwin Chen walks through a more basic use of RBM's for collaborative filtering in this blog post: Introduction to Restricted Boltzmann Machines.

Zhang et al. [7] propose a deep learning model to predict the ratings. The input of the model is the concatenation of the embedding feature vectors of user and item. There is only one hidden layer in the model, and output of the model is the predicted rating. Zheng et al. [8] use Neural Autoregressive Distribution Estimator (NADE) [9] model to address recommendation problem. The model is modified to share parameters among ratings. And to scale to large dataset, a factorizing version model is proposed inspired by RBM [10]. In this work, authors also present a list-wise loss function. Zou et al. [15] use stacked autoencoders in tag-aware recommender systems. The user latent feature vectors are generated by the stacked autoencoders according to the user-tag matrix. Recommendation results are obtained through aggregating the user latent feature vectors and item-user rating matrix.

4 Future Research Directions

One of the problems of the deep learning based recommendation approaches is scalability for recommendation systems, in which there are huge number of items and users. And user feedback are collected every second. The performance of the recommendation approaches is importance in this circumstance. In other hand, training the deep learning model is time-consuming. So how to improve the scalability of the deep learning based recommendation approaches is an importance issue in the future research.

Another possible research direction is to design new kinds of deep learning model to solve special problems in recommendation. RNN and feedforward neural network are used in the existing deep learning recommendation approaches. Convolutional neural network (CNN) is seldom used in recommendation. Maybe it can get good results for some recommendation problems.

5 Conclusion

Recently, due to the great success of deep learning, several researchers propose to use deep learning approach in recommendation systems. In this paper, the deep learning based recommendation approaches are reviewed and classified by the input and output of the approaches. It is found that the most used deep learning models are feedforward neural network and recurrent neural network. Convolutional neural network is seldom used. It is prompted that CNN based recommendation approaches is the possible research direction. Deep learning can improve the accuracy of the recommendation systems, but scalability is a critical problems for the huge number of items and users in

the systems. So improving efficiency of the deep learning based recommendation approaches is the main work in this field.

Acknowledgments. This work is supported by National Natural Science Foundation of China (No. 61403350 and No. 61401228).

References

1. van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: NIPS (2013)
2. Recommending music on Spotify with deep learning (2014). <http://benanne.github.io/2014/08/05/spotify-cnns.html>
3. Hamel, P., Lemieux, S., Bengio, Y., Eck, D.: Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), pp. 729–734 (2011)
4. Elkahky, A., Song, Y., He, X.: A multi-view deep learning approach for cross domain user modeling in recommendation systems. In: Proceedings of the 24th International Conference on World Wide Web (WWW 2015), pp. 278–288 (2015)
5. Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D.: Session-based recommendations with recurrent neural networks. In: ICLR 2016 (2016)
6. Caihua, W., Wang, J., Liu, J., Liu, W.: Recurrent neural network based recommendation for time heterogeneous feedback. *Knowl.-Based Syst.* **109**, 90–103 (2016)
7. Junlin, Z., Heng, C., Tongwen, H., Huiping, X.: A Distributional Representation Model For Collaborative Filtering (2015). <https://arxiv.org/abs/1502.04163>
8. Zheng, Y., Tang, B., Ding, W., Zhou, H.: A neural autoregressive approach to collaborative filtering. In: Proceedings of the 33rd International Conference on Machine Learning (2016)
9. Larochelle, H., Murray, I.: The neural autoregressive distribution estimator. In: International Conference on Artificial Intelligence and Statistics, pp. 29–37 (2011)
10. Salakhutdinov, R., Mnih, A., Hinton, G.: Restricted boltzmann machines for collaborative filtering. In: Proceedings of the 24th International Conference on Machine Learning, pp. 791–798. ACM (2007)
11. Wang, H., Wang, N., Yeung, D.-Y.: Collaborative deep learning for recommender systems. In: KDD 2015 (2015)
12. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *JMLR* **11**, 3371–3408 (2010)
13. Covington, P., Adams, J., Sargin, E.: Deep neural networks for YouTube recommendations. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 191–198 (2016)
14. Wang, X., Wang, Y.: Improving content-based and hybrid music recommendation using deep learning. In: Proceeding of ACM MM 2014 (2014)
15. Zuo, Y., Zeng, J., Gong, M., Jiao, L.: Tag-aware recommender systems based on deep neural networks. *Neurocomputing* **204**, 51–60 (2016)
16. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (2008)
17. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **17**(6), 734–749 (2005)

18. Park, D.H., Kim, H.K., Choi, I.Y., Kim, J.K.: A literature review and classification of recommender systems research. *Expert Syst. Appl.* **39**(11), 10059–10072 (2012)
19. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: *Proceedings of Advances in Neural Information Processing Systems, NIPS 2008*, pp. 1257–1264 (2008)
20. Gantner, Z., Drumond, L., Freudenthaler, C., Schmidt-Thieme, L.: Bayesian personalized ranking for non-uniformly sampled items. In: *Proceedings of Knowledge Discovery and Data Mining (KDD) Cup and Workshop 2011* (2011)
21. Pan, W., Zhong, H., Xu, C., Ming, Z.: Adaptive Bayesian personalized ranking for heterogeneous implicit feedbacks. *Knowl.-Based Syst.* **73**, 173–180 (2015)
22. Liu, J., Wu, C., Xiong, Y., Liu, W.: List-wise probabilistic matrix factorization for recommendation. *Inf. Sci.* **278**, 434–447 (2014)
23. Liu, W., Wu, C., Feng, B., Liu, J.: Conditional preference in recommender systems. *Expert Syst. Appl.* **42**(2), 774–788 (2015)
24. McAuley, J., Targett, C., Shi, Q., Hengel, A.: Image-based recommendations on styles and substitutes. In: *SIGIR 2015* (2015)
25. Jagadeesh, V., Piramuthu, R., Bhardwaj, A., Di, W., Sundaresan, N.: Large scale visual recommendations from street fashion images. In: *KDD 2014* (2014)

Towards Collaborative Data Analytics for Smart Buildings

Sanja Lazarova-Molnar^{1(✉)} and Nader Mohamed²

¹ Center for Energy Informatics, University of Southern Denmark, Campusvej 55,
5230 Odense, Denmark
slmo@mmmi.sdu.dk

² Middleware Technologies Lab., P.O. Box 33186, Isa Town, Bahrain
nader@middleware-tech.net

Abstract. Smart buildings are buildings equipped with the latest technological and architectural solutions, controlled by Building Management Systems (BMS), operating in fulfillment of the typical goals of increasing occupants' comfort and reducing buildings' energy consumption. We witness a slow, but steadily increasing trend in the number of buildings that become smart. The increase in availability and the decrease in prices of sensors and meters, have made them almost standard elements in buildings; both in newly built and existing ones. Sensors and meters enable growing collections of data from buildings that is available for further analytics to support meeting BMS' performance goals. For a single building to benefit from this data-based analytics, it will take a long time. Collaboration of BMS in their data analytics processes can significantly shorten this time period. This paper makes two contributions: one, a careful examination of the potential of buildings for collaborative data analytics; and two, description of models for collaborative data analytics.

1 Introduction

Smart buildings are buildings that incorporate advanced building intelligence technologies and latest architectural solutions, usually with the goals of enhancing energy performance and occupants' comfort. Smart buildings are commonly controlled by Building Management Systems (BMS) that are Cyber-Physical Systems (CPS), embedding a multitude of sensors and meters, as well as other hardware and software. Meters and sensors facilitate collection of big data that is utilized for different purposes, such as fault discovery and diagnosis (FDD) or monitoring energy consumption.

The benefits and synergetic effects of sharing and collaborating in data processing are apparent. BMS, or systems in general, can also effectively gain through sharing data and performing collaborative data processing and analytics. In this way, buildings can benefit from better quality and apt decision support in achieving their performance goals. Cloud computing and Internet of Things (IoT) platforms can support such collaboration among BMS by providing mechanisms and applications to support the access and sharing of data, as well as the flows of data.

In this paper we aim to explore and identify the opportunities for collaborative data processing and analytics for smart buildings. The paper is structured as follows. In

Sect. 2 we provide the state-of-the-art on common collaborative data processing, further focusing on building management systems. In Sect. 3 we elaborate on the opportunities in this area with respect to BMS, as well as describe the different models of collaboration. In Sect. 4 we provide analysis of the effects of collaboration based on a simple example, and finally, in Sect. 5 we conclude the paper.

2 Overview of Collaborative Data Processing for Buildings

Collaborative data processing, as usually referred to, is processing of data through collaboration of humans. One popular and successful example of this is Wikipedia [1], where the contents is developed by a large-scale contribution of people. Thus, as currently understood, collaborative data processing has a “human” dimension. This is a constraint that we aim to relax in this paper to refer to more than just humans.

We define *collaborative data processing* as *joint processing of data from multiple and diverse sources, by multiple and diverse processors (human or hardware), for achieving synergetic effect*. Sources can be human/expert knowledge, but also data gathered through various means, including sensors and meters.

Collaborative data processing for buildings has not been mainstreamed and systematized. Therefore, there are only a small number of approaches that target this beneficial manner of handling and processing buildings’ data, and usually they don’t go further than sharing of data. In the following, we first provide a brief overview on the advances in the collaborative data processing in common terms, further focusing on the research on collaborative data processing for buildings.

2.1 Collaborative Data Processing

Collaborative data processing is the combination of an entire collection of data into a joint analysis to synergistically derive considerably more knowledge as compared to their sole and separate analysis. Collaborative data processing usually refers to approaches that support humans to collaboratively process or analyze data, statistically or otherwise. There is a wide range of tools that support the collaboration of humans, e.g. in brainstorming, videoconferencing, or collaborative writing of documents. We expand this understanding and definition to apply to any entities, including systems, or in our case Building Management Systems.

In a research presented in [2] Frenklach et al. demonstrate that collaborative data processing leads to systematic development of predictive models and the joint analysis extracts substantially more of the information content of the data, thus supporting our claims. In a similar attempt [3], Seiler et al. present a framework for collaborative numerical data processing among researchers. The subjects are again humans (researchers), and the approach relies on data from many researchers within a community. The conclusion again emphasizes that the collaborative algorithm extracts more information than the non-collaborative one.

In a recent book on the challenges of collaborative data processing by humans [4], Noël et al. identify the following ingredients for successful collaborative data analysis:

- Sharing the data, along with its privacy challenges and its advantages,
- Task specialization, meaning that each contributor has a precisely defined task,
- Credit, ensuring that contributors' work would not be credited to others,
- Access, precisely defining access to ensure privacy wherever needed,
- Expertise, enough expertise should be available to approach and solve problems,
- Concurrency, meaning controlling the concurrent work,
- Usability, as it is humans that need to work with these tools,
- Flexible semantics, to compensate for people's disagreements on data semantics,
- Motivation, contributors should be able to see collaboration's benefit.

Apparently, some of these ingredients would be invalid when collaborators are not humans, but systems. In Sect. 3 we reflect back on these ingredients and elaborate on them in the context of systems.

2.2 Collaborative Data Processing for Smart Buildings: State of the Art

Most of the collaborative data processing approaches for buildings go as far as enabling data sharing, but typically no further than that. Among the more relevant works in the area [5], Wang et al. describe an experimental Internet-enabled system that integrates various BMS, which has also been implemented and tested. The research was guided by the need for sharing information among BMS and accessing BMS databases remotely to make it convenient for both use and development. There was no further joint data analytics for enhanced decision support. Working towards a similar goal of supporting intelligent data-driven methods, Agarwal et al. [6], elaborate on an architecture for buildings' data storage, access and sharing.

Related to the subject of sharing BMS data [7], Granzer et al. focus on the security aspects of doing it, and state that sharing the data of just one sensor to multiple applications can reduce investment and maintenance costs as well as facilitate management and configuration of the integrated BMS, as a multitude of different management solutions can be substituted for a unified view and a single central configuration access point.

3 Collaborative Data Analytics for Smart Buildings

Collaborative Data Analytics for Smart Buildings is emerging to become necessity for when high quality models are needed, along with highly accurate results and higher quality decisions. The ease of collecting relevant data, along with the possibility to store it remotely in the cloud together with the processing applications, makes the collaborative data processing almost inevitable.

The ingredients necessary for successful data collaborative processing from Sect. 2.1, when the main subjects are systems, need to be revisited, as follows.

- Sharing the data – also a relevant issue that needs addressing,
- Task specialization – depends on collaboration model, elaborated in Sect. 3.3,
- Credit – irrelevant for systems' collaboration due to their technical nature,
- Access – also relevant for systems, access needs to be precisely defined,

- Expertise – need for relevant tools/resources to satisfy BMS’s requirements,
- Concurrency – it is also relevant to be considered and handled,
- Usability – irrelevant as the collaborative tools communicate with BMS,
- Flexible semantics – irrelevant if there are no people in the middle of the process,
- Motivation – in BMS case, buildings’ owners should be motivated to participate.

Therefore, the ingredients for collaboration of systems are slightly related to those for collaboration of humans, except that we can exclude those ingredients that are related to the human nature and psychology. For instance, while in collaborative data analytics of humans, tools should motivate participants to collaborate; in the BMS case buildings’ owners need to clearly see the incentive to participate. In the following we elaborate on the potential areas where collaborative processing can enhance performance of buildings.

3.1 Data Analytics Needs of Smart Buildings

Data that is being collected through BMS can serve various purposes. More significant BMS data analytics processes are the following: Establishing ground truth data, Data validation, Fault detection and diagnosis, Energy consumption prediction, Generation of maintenance schedules, Estimation of occupant comfort level, Building simulation, etc. All of these processes can benefit from collaborative data processing, as all of them depend on high quality and large quantity of data.

Establishing ground truth data is obtaining data that represents correct operation of the building management system and labeling it as such. It is an incredibly important step for any further data analytics. Ground truth data is very hard to ensure given the high complexity of BMS, and its obtaining can be supported by the various performance and commissioning tests. However, if a set of buildings that collaborate have been classified as similar, that would increase the confidence of stating that a building operates correctly (e.g. if there are n similar buildings, and one building has different operation patterns from the rest $n - 1$ buildings, most likely the different one is the odd one, and therefore other $n - 1$ buildings’ data can be labeled as ground truth). This would be very difficult to state with confidence if there is only one building. The similarity does not necessarily need to be defined at the building level, it can also be defined at subsystem/room/zone/component level.

Another important data process is data validation, which is ensuring that the data collected from data sources is clean and error-free. It is a process that is very tightly intertwined with fault detection and diagnosis, as faults can also result in incorrect data [8]. Apparently, faults are rare events, so it will take a long time before a building can have accurate and exhaustive fault models. Therefore, having a high number of participating buildings can significantly enhance these data-based model-building processes. Remaining smart buildings data analytics processes can share the conclusions on the benefits of collaborative data processing, especially that most of them are data-based. In the following two subsections we detail the opportunities and models of collaboration.

3.2 Opportunities for Collaborative Data Processing for Smart Buildings

Collaborative data processing and analytics can immensely enhance the decision support and control of BMS, as well as achieving their corresponding predefined goals. As stated before, typical goals of BMS are increasing occupants' comfort and reducing energy consumption, but also reducing maintenance cost, increasing reliability, increasing safety and security, as dependent on building's purpose [9, 10]. Reaching most of these goals can be enhanced through having more and better quality data, i.e. both quantity and quality of data are important. It is apparent that the more BMS participate and collaborate, the more data can be obtained in a shorter time, and through peer-checking, a better quality could be achieved. To gain real benefit from large amounts of data, it needs to be structured and addressed adequately. Clustering can be used to group similar BMS and similar buildings, as well as similar subsystems, zones and all phenomena that are relevant to be grouped. Clustering can be performed based on physical characteristics, purpose, consumption patterns, etc. Clustering can be also utilized for peer-observation of BMS for various purposes. To illustrate this, we can take school buildings; one would imagine that similarly sized school buildings' BMS would belong to the same cluster. Therefore, a deviation of the behavior of one building would be easily detected, as it would be compared against a number of similar buildings. The occupancy patterns of these similar buildings would be also expected to be similar, so any deviations in occupancy would also be noticed in a timely manner (this can also tackle on the aspect of safety). Furthermore, there would be a much larger pool of data on faults that would significantly enhance the FDD processes for all buildings in a cluster.

One of the questions is how to most efficiently and most effectively share the knowledge obtained through the collaborative data analytics. One way to could be through generation of rules that have been learned from the behavior of all participating BMS on the cloud, and shared back to the BMS. An example would be a new and unseen fault that has been discovered on a new component. Once it is matched to the circumstances that have led to it and labeled adequately, this knowledge in form of rules can be shared to all other participating BMS, thus enabling them to timely and accurately diagnose it when it occurs. The advantage of collaboration as opposed to having a solitary BMS in this scenario is very obvious. From the noted example, we have seen that collaboration of BMS can be through different collaboration models, as we elaborate in the following.

3.3 Collaboration Models

Collaborative data analytics for smart buildings can be classified as either offline or online collaborative analytics. With offline collaborative analytics, each smart building data will be individually collected periodically such as monthly or quarterly. This data is stored locally within the BMS of the building. After a specified time, the collected data is transferred using a portable storage unit or uploaded to a place where collaborative data analytics for all participating smart buildings is performed. On the other hand, online collaborative analytics continually collect data from the participating smart buildings. In this model, collaborative data analytics can tune the data collection process for better results and find interesting insights at any time. The tuning process can include

new collected data and new data collection rates. This allows having adaptive collaborative data analytics for better results, as well as better overall control. However, this model requires using a special type of BMS that can be connected to external networks and systems such as the Internet and the Cloud [11].

Collaborative data analytics for smart buildings can be offered for different scenarios. It can be provided within private, community, or public scenarios. In private scenarios, one owner of multiple smart buildings utilizes collaborative data analytics to improve energy consumption in these buildings. In the community scenarios, multiple organizations or owners of similar buildings can share their buildings information for the analytics processes to benefit all owners or organizations. Examples of similar buildings can be schools, hospitals, or commercial buildings. In public scenarios, different buildings owners and organizations share their buildings information for the benefit of all participants. The analytics process of all three scenarios can be provided as offline or online analytics. In addition, buildings' owners can perform their own analytics processes by having their own software and hardware for that purpose. Alternatively, they can use special analytics services available on the Cloud [12]. In this case, a third party company can provide the needed services for a fee. This company will be responsible for all hardware, software, storage, technical support, and collaborative data analytics for smart buildings. The second approach of using a cloud-based solution can provide a more cost-effective solution compared to the first one as the cloud company that provides the analytics services will serve multiple smart buildings and the cost will be significantly reduced. In addition, the analytics results can be better as more data is collected from more buildings, as we will show in the analysis section.

4 Advantages Analysis Through an Illustrative Example

In this section, we analyze the advantages of collaboration among smart buildings through a case study example. Let us assume that there is a set of b school buildings with the same/similar design in a large city. These schools have an average of r classrooms each. Each classroom is equipped with a temperature sensor to monitor the temperature and a controller that links with the classroom's schedule and the number of students in each classroom to turn air conditioning on or off to maintain a convenient temperature in the classroom, while at the same time optimize energy consumption. There is a p probability of each sensor to have a persistent fault per year in measuring the correct temperature. Such fault may cause unnoticeable errors or fluctuations in the automation process for air conditioning. This can result in extra consumed energy cost averaging x Euros per year. Let us assume that there is a smart system that discovers common faults and it is individually installed in each building. This smart system can detect a fault after having f occurrences of the same new fault. By this, we can find the expected number of years, y_1 , that the system will detect this type of fault in a single building. We have $y_1 = f/(r \times p)$ years. For simplicity, let us assume that the fault occurs in uniform time distribution, then the expected total extra consumed energy cost of each building from this fault before it is discovered is $c_1 = (f \times x \times y_1)/2$ Euros, i.e. $c_1 = (f^2 \times x)/(2 \times r \times p)$. The total extra consumed energy cost for all school buildings without collaboration is $c_{wc} = b \times c_1$ Euros.

Now, if we use the same smart diagnosis system for all buildings, the fault will be discovered after y_b years, where $y_b = f/(b \times r \times p)$. In this case, the expected total extra consumed energy cost of all buildings from this type of fault before discovering the fault is $c_b = (b \times f \times x \times y_b)/2$ Euros which equals $(f^2 \times x)/(2 \times r \times p)$ Euros. This is the same as the extra cost of a single building regardless of the number of the buildings. Now, to get the saving factor of using the collaborative solution in energy saving, we need to divide c_b over c_{wc} , and we get $1/b$. This means that the collaborative solution will only cost $1/b$ of the total extra consumed energy cost for all school buildings without collaboration. In addition, the saving factor of the collaborative solution in the extra energy consumption will be $(1 - 1/b)$ or $(b - 1)/b$ percent of the extra cost of that without collaboration as shown in Fig. 1. As we can see, the saving factor will increase with the increase in the number of participating buildings.

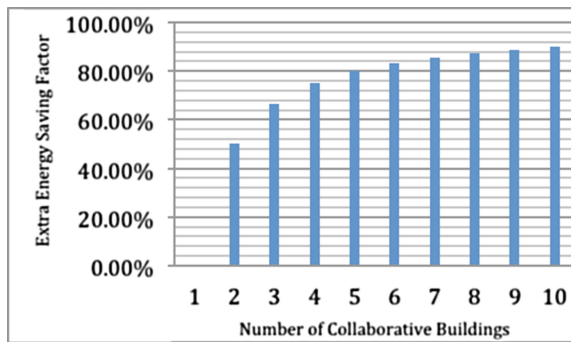


Fig. 1. Collaborative solution saving factor in energy consumption.

5 Summary and Outlook

Benefits from collaborative and cloud supported data processing for smart buildings seem to be apparent since more data with a high diversity can also mean timely models with higher accuracy, when developed using data-based approaches. What remains is the matter of gathering participating BMS. This implies that there is a need for more assertive education, as well as a more enhanced addressing of privacy and security issues, to inform and convince stakeholders about the benefits of sharing data and its collaborative processing. Furthermore, the availability and advancements of the enhanced data sharing platforms, such as Cloud Computing and Internet of Things, make the emergence of the collaborative data analytics highly feasible. Such BMS that support collaborative data processing and analytics can be also considered as more prepared and mature for the smart grid, as some of the smart grid challenges will only become feasible to solve through collaborative data analytics.

Acknowledgements. This work is supported by the Innovation Fund Denmark for the project COORDICY.

References

1. Wilkinson, D.M., Huberman, B.A.: Cooperation and quality in Wikipedia, pp. 157–164. ACM
2. Frenklach, M., Packard, A., Seiler, P., Feeley, R.: Collaborative data processing in developing predictive models of complex reaction systems. *Int. J. Chem. Kinet.* **36**, 57–66 (2004)
3. Seiler, P., Frenklach, M., Packard, A., Feeley, R.: Numerical approaches for collaborative data processing. *Optim. Eng.* **7**, 459–478 (2006)
4. Noël, S., Lemire, D.: On the challenges of collaborative data processing. In: Collaborative Information Behavior: User Engagement and Communication Sharing: User Engagement and Communication Sharing, vol. 55 (2010)
5. Wang, S., Xie, J.: Integrating building management system and facilities management on the Internet. *Autom. Constr.* **11**, 707–715 (2002)
6. Agarwal, Y., Gupta, R., Komaki, D., Weng, T.: BuildingDepot: an extensible and distributed architecture for building data storage, access and sharing. In: Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, pp. 64–71. ACM (2012)
7. Granzer, W., Praus, F., Kastner, W.: Security in building automation systems. *IEEE Trans. Industr. Electron.* **57**, 3622–3630 (2010)
8. Lazarova-Molnar, S., Shaker, H.R., Mohamed, N.: Fault detection and diagnosis for smart buildings: state of the art, trends and challenges. In: 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1–7. IEEE (2016)
9. Lazarova-Molnar, S., Kjærgaard, M.B., Shaker, H.R., Jørgensen, B.N.: Commercial buildings energy performance within context: occupants in spotlight. In: SmartGreens 2015 (2015)
10. Lazarova-Molnar, S., Shaker, H.R., Mohamed, N.: Reliability of cyber physical systems with focus on building management systems. In: IEEE International Workshop on Communication, Computing, and Networking in Cyber Physical Systems (CCN-CPS) in Association with IEEE International Performance Computing and Communications Conference (IPCCC 2016). IEEE, Las Vegas (2016)
11. Mohamed, N., Lazarova-Molnar, S., Al-Jaroodi, J.: CE-BEMS: a cloud-enabled building energy management system. In: 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC), pp. 1–6. IEEE (2016)
12. Mohamed, N., Lazarova-Molnar, S., Al-Jaroodi, J.: SBDaaS: smart building diagnostics as a service on the cloud. In: 2016 2nd International Conference on Intelligent Green Building and Smart Grid (IGBSG), pp. 1–6. IEEE (2016)

English and Malay Cross-lingual Sentiment Lexicon Acquisition and Analysis

Nurul Amelina Nasharuddin¹, Muhamad Taufik Abdullah¹(✉), Azreen Azman¹,
and Rabiah Abdul Kadir²

¹ Department of Multimedia, Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
{nurulamelina,mta,azreenazman}@upm.edu.my

² Institute of Visual Informatics, Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor, Malaysia
rabiahivi@ukm.edu.my

Abstract. Sentiment analysis finds opinions, sentiments or emotions in user-generated contents. Most efforts are focusing on the English language, for which a large amount of sources and tools for sentiment analysis are available. The objective of this paper is to introduce a cross-lingual sentiment lexicon acquisition method for the Malay and English languages and further being test on a set of news test collections. Several part of speech tags are being experimented using the Word Score Summation technique in order to classify the sentiment of the news articles. This method records up to 50% as experimental accuracy result and works better for verbs and negations in both the English and Malay news articles.

1 Introduction

Many existing research on textual information processing has focused on mining and retrieval of factual information. But recently with the growth of the Web, there is a lot of user-generated information available such as users' views and opinions on products on merchant sites, forums, blogs and discussion groups. Text processing focusing on opinions and sentiments has become increasingly important, not only to individuals but also for organisations. Sentiment analysis or opinion mining may be defined as a general method to find opinions, sentiments and emotions in text. Sentences in texts can be classified as objective or subjective where a subjective sentence may contains a positive, negative or neutral opinion. Most efforts are focusing on the English language, where a large amount of sources and tools are available especially for sentiment analysis. There are two major problems exist in the under-resourced sentiment analysis work: (1) limited number of formally standardised sentiment lexicon and (2) few number of sentiment classifier that are publicly available.

Hence, a better solution is to have a sentiment orientation analysis based on existing linguistics resources in highly-resourced languages, such as the English language. This work proposed an automatic cross-lingual sentiment lexicon acquisition for the English and Malay languages where Malay is being considered as an under-resourced language.

Then an analysis on the sentiment for both English and Malay languages was experimented using the developed sentiment lexicon.

2 Related Work

2.1 Sentiment Analysis

There are two main approaches in extracting sentiment from texts namely the lexicon-based (an unsupervised approach) and the statistical or machine-learning (a supervised approach). Lexicon-based unsupervised learning approach uses sentiment dictionaries. Dictionaries or lexicons which are being used for this approach can either be constructed manually, using seed words, utilising existing dictionaries or making use of other lexical resources like WordNet [1, 2]. The pre-built dictionaries such as SentiWordNet [3] contain words along with their associated sentiment polarity (positive or negative) and strength. The SentiWordNet was developed following the structure of the English WordNet built by Princeton University [4]. The word class (nouns, verbs, adjectives and adverbs) are grouped into synonyms sets (or synsets) linked by semantic relations. Lexicon-based approach does not require storing a large data corpus and training, so the whole process is much faster. Classifiers built using statistical or supervised approaches usually perform very well in detecting the polarity of a text as they are generally trained on a large annotated corpus [5]. However, the performance is usually not good when the same classifier is applied on different domain that they have been trained on.

2.2 Malay Sentiment Analysis

Research in sentiment analysis for the Malay languages is scarce compared to the English language. Two studies focused on finding sentiments in Malay news document using Artificial Immune System technique inspired by the biological immune system responding toward foreign antigen [6, 7]. This supervised learning algorithm is widely adapted by other research such as in computer and network security. Zamani and his colleagues studied on how to analyse sentiments in English and Malay words in Facebook [8]. Their work focused on quantifying Facebook sentiments using a lexicon-based approach for both the English and Malay texts. Liao and Tan [9] also used a lexicon-based approach to study the customer's satisfaction level towards low-cost airlines in Malaysia, in order to understand the consumer's needs. These two researches created a small-scale lexicon and employed a very simple classifier to calculate the polarity of the sentiments.

One of a Malay lexical databases widely used in Malay language research study is the Wordnet Bahasa. It is a combination of lexical semantics from three different resources which are the Malay Wordnet, the Indonesian Wordnet and the Wordnet Bahasa [10]. It contains over 45,000 Malay words and 58,000 Indonesian words. The Wordnet Bahasa was also developed based on the English WordNet. Thus, a cross-lingual sentiment lexicon for English and Malay language can be developed by mapping both the Wordnet Bahasa and the English SentiWordNet with the English WordNet as they were built using the same structures. This mapping method is independent of

translation using a bilingual dictionary as both lexical databases have the same unique synsets value for each word entry.

3 Methodology

Figure 1 shows the methodology performed in this work where it comprised of four main phases namely, (i) pre-processing of test documents, (ii) cross-lingual sentiment lexicon acquisition, (iii) sentiment processor and finally (iv) document sentiment categorisation, in order to classify the documents into positive, negative or objective classes.

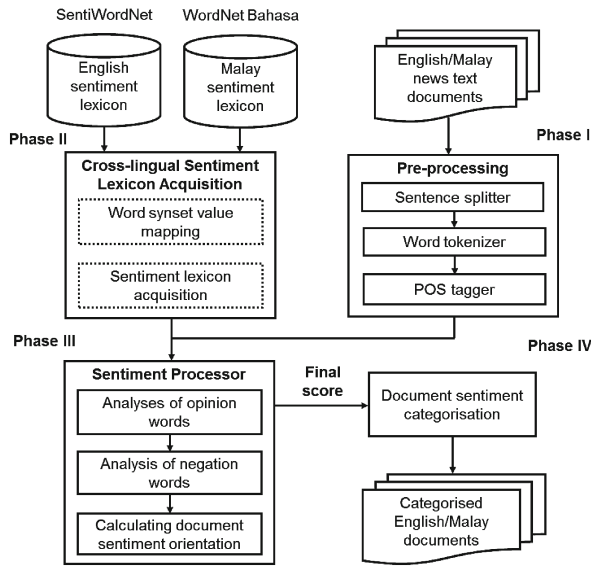


Fig. 1. The proposed methodology for the English and Malay sentiment analysis.

3.1 Test Documents Collection

In this work, a total of 883 of an unstructured English and Malay news articles were selected as the test documents. These articles which published in the year of 2005 were from a national news agency in Malaysia, the Bernama. The documents contain various contents which include politics, business, economy, executive reports and sports. This will avoid the domain-specific limitation during the experiments.

3.2 Test Documents Pre-processing

The English and Malay test documents were first annotated with the XML tags describing the author, publication date, headline as well as the contents. Then the documents were being pre-processed. Three main tasks of the pre-processing for both English and Malay documents were:

Lexical Analysis. The operations being done in this step were treating digits, hyphens and punctuation marks. All the digits in the documents were removed and the hyphens were replaced by space. All the capital letters were left unchanged as they were important indicators during the part-of-speech (POS) tagging.

Stopwords Removal. Next was to remove stopwords - the most frequent words that often do not carry much meaning. This work used the SMART stopword list of 571 words for the English documents and a list of 321 stopwords for the Malay documents.

Stemming and POS Tagging. The English test documents were stemmed using the Stanford CoreNLP Java library, a natural language analysis tool in order to remove the inflectional endings and return the base of the words in the documents. For extracting the POS of the English documents, the work employed the Stanford POS Tagger [11]. As for the Malay documents, this work implemented the stemmer from Abdullah, Ahmad, Mahmud and Tengku Sembok [12] and a rule-based POS tagger by Alfred and his colleagues [13] to tag the words into their POS.

3.3 English-Malay Cross-lingual Sentiment Lexicon Acquisition

A cross-lingual sentiment lexicon is needed to enable sentiment analysis to be performed on both English and Malay test documents. Thus, an automatic acquisition method was proposed to create a cross-lingual sentiment lexicon for English and Malay languages using the similarity structures of the English SentiWordNet and Wordnet Bahasa. The suggested approach consists of two main processing steps, explained in details as below:

Automatic Mapping of the Cross-lingual Synonyms. The Wordnet Bahasa was mapped to the English SentiWordNet using the synset value and POS value as the key to build the cross-lingual sentiment lexicon. Then, the synonyms for both languages were matched with each other. The mapping includes all types of POS in the Wordnet Bahasa which were nouns, verbs, adjectives and adverbs. As the Word Sense Disambiguation (WSD) was not one of the objectives of this work, word's prior polarity was being used to address the problem on how to compute the sentiment orientation of one word. Though the technique was less precise, it was guaranteed that the same score was given to the same word in different contexts. To find the prior polarity of a word, the weighted mean formula as in Eq. (1) was adopted

$$posScore = \frac{\sum_{i=1}^n \left(\frac{1}{i} \times posScore_i \right)}{\sum_{i=1}^n \left(\frac{1}{i} \right)} \quad (1)$$

where *posScore* represents the positive value of a word, *n* represents the total number of senses of the word and *posScore_i* represents the absolute value of the positive value of the *i*-th sense of that word [14]. The *objScore* (the objective or neutral value of a word) and *negScore* (the negative value of a word) were calculated using the same steps. In this technique, each sense weight was chosen according to a harmonic series according

to their frequency of being used. The choice was based on the assumption that more frequent senses should bear more “affective weight” than very rare senses.

Cross-lingual Sentiment Lexicon Acquisition. An English-Malay sentiment lexicon was generated from the previous step. The advantage of this method is it does not require translation unlike other cross-lingual lexicon-based approaches. The construction of this lexicon was based on the following assumptions; (1) the senses of the words in both Malay and English language were the same and (2) the sentiment score for each English word was similar to the matched Malay word.

3.4 English and Malay Sentiment Processor

In this step, the relevant words which used to find the sentiment of the document were extracted from each document. The relevant words include the adjectives, adverbs, verbs and negation words. The positivity, negativity and neutrality (or objectivity) values from the previous step were combined to determine the documents’ sentiment orientations. For each document in both English and Malay languages, the method first finds the *posScore*, *negScore* and *objScore* values of the related word. The method then checks if there were negations used before the related word and if yes, the three sentiment values of the words will be updated. After all the related words have been checked, each of the sentiment values was aggregated to get the overall *posScore*, *negScore* and *objScore* values for the documents. Four different combinations of POS were being experimented in this work were *Adj + Neg*, *Verb + Neg*, *Adv + Adj + Neg* and *Adv + Verb + Neg*.

For the aggregation method of the documents’ sentiment scores, this work enhanced the Word Score Summation method by Hamouda and Rohaim by including the objective scores of the documents where these scores were important in a general domain text collection [15]. In this method, the sentiment scores for each positive, negative and objective word in a document were added up and negated if a negation word appeared. The overall sentiment orientation of the document was chosen based on which from these three overall scores had the highest value. It was assumed that the document’s author sentiment was correlated to the choice and number of relevant words presented in the document.

3.5 Document Sentiment Categorisation

The results from the previous step were used to automatically categorise the documents into the three orientation categories, which are positive, negative and neutral. Then for each of the document, the overall document’s orientation will be classified as positive if the score bigger than zero. If the overall score is less than zero, the document’s orientation will be classified as negative and if equal to zero, the orientation is neutral.

4 Experiments

4.1 Experimental Settings

The automatic categorisation of the documents sentiments were then assessed manually by three language experts in both Malay and English languages. A contingency table was being used to compare the performances of the method and the manual assessments by the experts. Two different tables were created for the English and Malay test documents. The three-class contingency table representing the positive, negative and objective categories as in Table 1.

Table 1. Contingency table for three sentiment categories to compare between experts' assessments and proposed method

		Proposed method		
		Positive	Objective	Negative
Expert	Positive	T_{Pos}	E_{PosObj}	E_{PosNeg}
	Objective	E_{ObjPos}	T_{Obj}	E_{ObjNeg}
	Negative	E_{NegPos}	E_{NegObj}	T_{Neg}

For example, T_{Pos} is the total number of documents that were correctly categorised as positive and E_{ObjNeg} is when an objective document is wrongly categorised to negative category by the proposed method. From this table, the performance of the proposed method was measured using four commonly-used measurements which are the accuracy, precision, recall and F1. For example in calculating the performance of the positive category, the equations are as follows:

$$accuracy = \frac{T_{Pos} + T_{Obj} + T_{Neg}}{\text{all instances in the table}} \tag{2}$$

$$precision_{Pos} = \frac{T_{Pos}}{T_{Pos} + E_{ObjPos} + E_{NegPos}} \tag{3}$$

$$recall_{Pos} = \frac{T_{Pos}}{T_{Pos} + E_{PosObj} + E_{PosNeg}} \tag{4}$$

$$F1 = \frac{2 * precision_{Pos} * recall_{Pos}}{precision_{Pos} + recall_{Pos}} \tag{5}$$

4.2 Experimental Results

Table 2 shows the accuracy results for the proposed sentiment method for the English and Malay test documents. It can be observed that the POS combination of *Verb + Neg* shows the highest accuracy for both English and Malay documents at 50.2% and 47.9%, respectively. However, the worst performance is obtained by the *Adv + Adj + Neg*

combination with only 24.3% for English documents and 21.6% for Malay documents. The macro-averaged precision, recall and F1 results for English and Malay documents that are obtained are shown in Table 3. Macro-averaged measurements are the average values of the three categories of sentiment, being experimented.

Table 2. Accuracy comparison between different POS combinations for the English and Malay test documents.

	English	Malay
Adj + Neg	44.1	44.2
Verb + Neg	50.2	47.9
Adv + Adj + Neg	24.3	21.6
Adv + Verb + Neg	33.1	39.2

Table 3. Macro-averaged precision, recall and F1 comparisons between different POS combinations for the English and Malay test documents.

POS	Precision		Recall		F1	
	English	Malay	English	Malay	English	Malay
Adj + Neg	0.30	0.31	0.34	0.32	0.32	0.31
Verb + Neg	0.31	0.32	0.34	0.36	0.32	0.34
Adv + Adj + Neg	0.18	0.34	0.34	0.34	0.24	0.34
Adv + Verb + Neg	0.34	0.35	0.36	0.33	0.35	0.34

None of the macro-averaging F1 results of the proposed method are statistically significant. The best performance (0.35 and 0.34) of the proposed method is achieved when the POS combination of *Adv + Verb + Neg* are used for both English and Malay documents and also the combinations of *Verb + Neg* and *Adv + Adj + Neg* for Malay documents. In general, both accuracy and F1 scores for the English and Malay test documents are somewhat low. This finding was unexpected and suggests that rigorous steps need to be done in improving the scores. A possible explanation for this might be that the adoption of prior polarity which totally removes the senses of each words. Another possible explanation for this is that the lack of established and reliable pre-processing tools on the Malay language might poses problems to the overall experiments.

5 Conclusion and Future Work

This paper presents an extensive work on the development of a cross-lingual English and Malay sentiment lexicon and further used the lexicon in sentiment analysis task. The main contribution of this work is to effectively develop an automatic method for the cross-lingual sentiment lexicon construction which in turn will help the Malay language sentiment analysis in future. This paper attempts to determine which POS combinations perform best to detect sentiments on English and Malay documents. The results indicate that the best POS combinations are the *Adv + Verb + Neg* and *Adv + Adj + Neg* for both English and Malay test documents. Future work will be

focusing on improving the pre-processing phase of the Malay documents especially on the stemming task. The word sense disambiguation should also be included in future research to address the word sense problem for the Malay sentiment classification. Different kind of test collection could also be experimented in order to confirm the method robustness. Despite these limitations, it is believed that this study has contributed to this subject especially in the Malay sentiment analysis research.

Acknowledgments. This work was supported by the Fundamental Research Grant Scheme (FRGS), MOHE Malaysia. We thank Bernama for their assistance and advise during the data collection process.

References

1. Taboada, M., Brooke, J., Tofilofski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. *Comput. Linguis.* **32**(2), 267–307 (2011)
2. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of HTML documents. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1075–1083 (2007)
3. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. European Language Resources Association, pp. 2200–2204 (2010)
4. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
5. Boiy, E., Hens, P., Deschacht, K. Moens, M.-F.: Automatic sentiment analysis in on-line text. In: Leslie, C., Martens, B. (eds.) *Proceedings of the 11th International Conference on Electronic Publishing*, pp. 349–360 (2007)
6. Isa, N., Puteh, M., Raja Mohamad Hafiz, R.K.: Sentiment classification of Malay newspaper using immune network (SCIN). In: *Proceedings of the World Congress on Engineering* (2013)
7. Puteh, M., Isa, N., Puteh, S., Redzuan, N.A.: Sentiment mining of Malay newspaper (SAMNews) using artificial immune system. In: *Proceedings of the World Congress on Engineering* (2013)
8. Zamani, N.A.M., Abidin, S.Z.Z., Omar, N., Abiden, M.Z.Z.: Sentiment analysis: determining people's emotions in Facebook. In: Zaharim, A., Sopian, K., Psarris, K., Margenstern, M. (eds.) *Proceedings of the 13th International Conference on Applied Computer and Applied Computational Science*, pp. 111–116 (2014)
9. Liau, B.Y., Tan, P.P.: Gaining customer knowledge in low cost airlines through text mining. *Ind. Manag. Data Syst.* **114**(9), 1344–1359 (2014)
10. Bond, F., Lim, L.T., Tang, E.K., Riza, H.: the combined Wordnet Bahasa. In: Chung, S.-F., Nomoto, H. (eds.) *Current Trends in Malay Linguistics*, vol. 57, pp. 83–100 (2014)
11. Toutanova, K., Dan, K., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 173–180 (2003)
12. Abdullah, M., Ahmad, F., Mahmud, R., Tengku Sembok, T.: Rules frequency order stemmer for Malay language. *Int. J. Comput. Sci. Netw. Secur.* **9**(2), 433–438 (2009)

13. Alfred, R., Mujat, A., Obit, J.H.: A ruled-based part of speech (RPOS) tagger for malay text articles. In: Selamat, A., Nguyen, N.T., Haron, H. (eds.) ACIIDS 2013. LNCS (LNAI), vol. 7803, pp. 50–59. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-36543-0_6](https://doi.org/10.1007/978-3-642-36543-0_6)
14. Gatti, L., Marco, G.: Assessing sentiment strength in words prior polarities. In: Proceedings of COLING 2012: Posters, COLING 2012, pp. 361–370 (2012)
15. Hamouda, A., Rohaim, M.: Reviews classification using SentiWordNet lexicon. Online J. Comput. Sci. Inf. Technol. **2**(1), 120–123 (2012)

A Novel Natural Language Processing (NLP) Approach to Automatically Generate Conceptual Class Model from Initial Software Requirements

Mudassar Adeel Ahmed, Wasi Haider Butt, Imran Ahsan,
Muhammad Waseem Anwar^(✉), Muhammad Latif, and Farooque Azam

Department of Computer Engineering, College of Electrical and Mechanical Engineering,
National University of Sciences and Technology, H-12, Islamabad, Pakistan
{mudassar.adeel14, imran.ahsan14}@ce.ceme.edu.pk,
{wasi, waseemanwar, mlatif, farooq}@ceme.nust.edu.pk

Abstract. Conceptual class model is an essential design artifact of Software Development Life Cycle (SDLC). The involvement of several resources and additional time is required to generate the class model from early software requirements. On the other hand, Natural Language Processing (NLP) is a knowledge discovery approach to automatically extract elements of concern from initial plain text documents. Consequently, it is frequently utilized to generate various SDLC artifacts like class model from the early software requirements. However, it is usually required to perform few manual processing on textual requirements before applying NLP techniques that makes the whole process semi-automatic. This article presents a novel fully automated NLP approach to generate conceptual class model from initial software requirements. As a part of research, Automated Requirements 2 Design Transformation (AR2DT) tool is developed. The validation is performed through three benchmark case studies. The experimental results prove that the proposed NLP approach is fully automated and considerably improved as compared to the other state-of-the-art approaches.

Keywords: NLP · AR2DT · Class diagram · Software requirements · Natural language processing

1 Introduction

Getting significant information from preliminary set of requirements in the analysis phase is inherently a crucial task and requires more manual intervention that leads to massive data processing time. Moreover, these manual interventions can cause crucial data processing errors. Natural Language Processing (NLP) shows some propitious and more encouraging results to overcome such issues, especially in bio-medical domain [1]. NLP allows automated data processing features and is applied to various software development phases to generate the requirement specifications [2], design artifacts and test cases [3] in an automated manner. A lot of research is done over design phase which include class diagram generation [4], use case generation [5], collaboration diagram generation [6] and so on.

Although there is a noticeable research available that deals with the generation of class model from initial plain text software requirements, the existing studies usually requires few manual processing on textual requirements before generating the class model that makes the whole process semi-automatic. This deviates the actual spirit of true automation. Therefore, in this article, we propose a novel and fully automated NLP approach to generate the class model from early software requirements. The overview of this study is shown in Fig. 1.

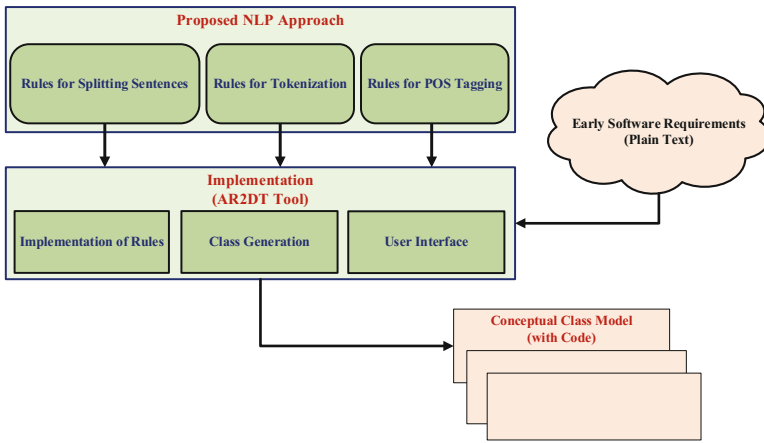


Fig. 1. Overview of research

Firstly, we defined the novel and improved rules for splitting sentences, tokenization and POS tagging (Sect. 2). Secondly, we implement the defined rules in AR2DT tool (Sect. 2.1). There are three components of AR2DT tool i.e. Implementation of Rules, Class generation and User Interface. It takes early software requirements as a plain text and generate conceptual class model with code as shown in Fig. 1. Finally, we utilize three case studies for the validation of proposed approach (Sect. 3). The comparative analysis with state-of-the-art is given in Sect. 4. The paper is concluded in Sect. 5.

2 Proposed Methodology and Implementation

The proposed NLP approach comprises the novel rules of sentence splitting, tokenization and POS tagging as shown in Fig. 2. The defined rules are applied to the initial plain text software requirements to generate conceptual class model. Our proposal mainly concerns with the extraction of Noun Plural (NNS), Proper Noun Singular (NNP) and Proper Noun Plural (NNPS) by using matching nouns. The summary of rules is as follows:

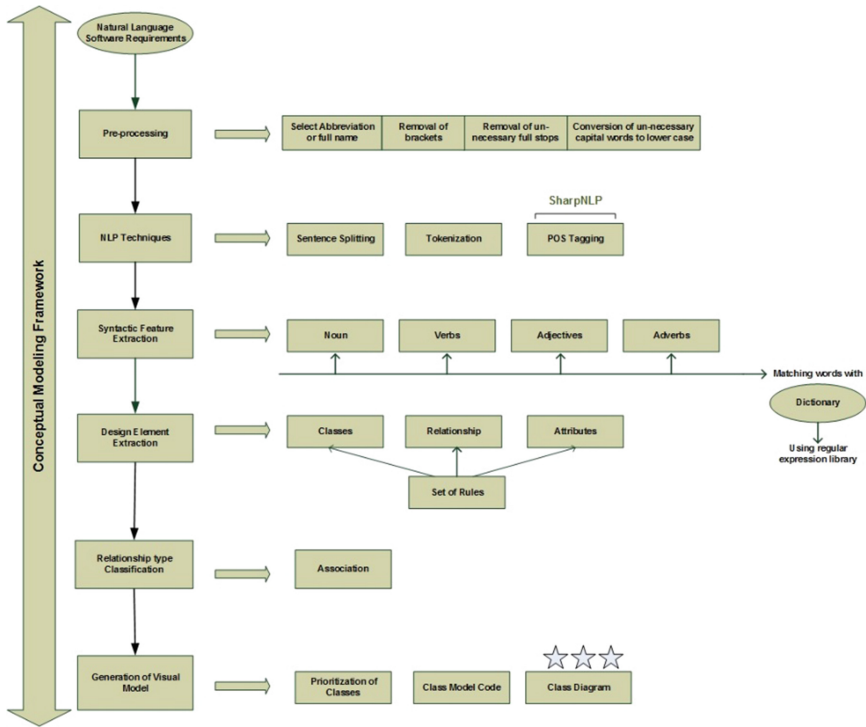


Fig. 2. Proposed architecture of AR2DT

Nouns or Classes Identification: It can be concluded from the state-of-the-art (Sect. 4) that researchers usually consider all types of nouns as classes. However, we propose to only consider NNS, NNP and NNPS as classes.

Conversion of Plural to Singular: We are converting the plural nouns to singular e.g. convert books to book.

Remove Redundant Classes: Repeated classes are only considered once. This concept is implemented by defining a dictionary which includes all the irrelevant glossary words e.g. user, software, number etc. The special set of the standard guidelines are defined while developing the glossary of the dictionary in order to avoid any sort of biasness.

The identification of classes from plain text are performed on the basis of pre-defined rules. A class can be described by this equation: $C: \epsilon \{[C, A, O, R]\}$ Where C is the candidate class, A belongs to the attribute of this class, O is the operation or function of the class and R represents the relationship of the class. The relationships between the classes can be expressed as follow: $R: \epsilon \{[rT, Cr, Rc]\}$ Where R belongs to relationship, rT is the relationship type i.e. association, Cr is the cardinality and Rc is the related class.

2.1 Implementation

AR2DT is developed in Visual Studio 2010 and written in C# with 1500 line of codes. The SQL Server 2012 has been used for the storage. In AR2DT, the rules are implemented through *SharpNLP-1.0.2529* [14] library. Subsequently, *Regular Expression* library is used to match classes by utilizing the concept of dictionary. The interface of AR2DT implementing ATM case study (Sect. 3) is shown in Fig. 3.

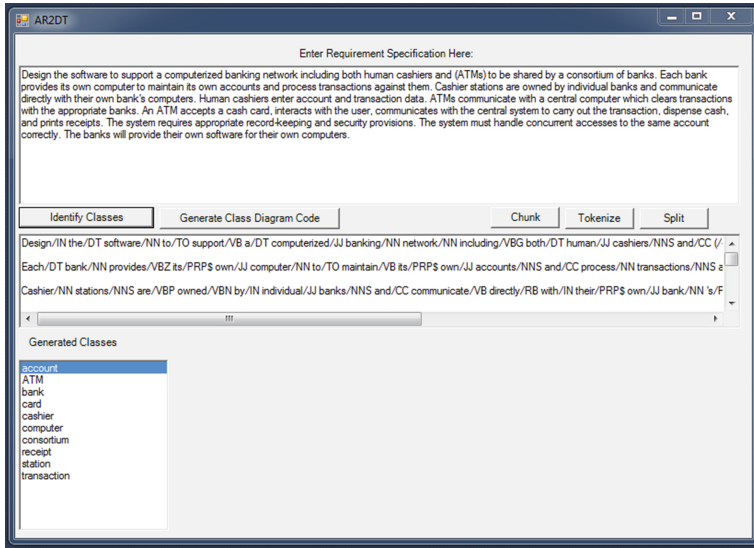


Fig. 3. AR2DT user interface

The text area is provided to write and copy/paste the desired case study. The classes can be identified by pressing *Identify Classes* button where the business logic for the rules of sentence splitting, tokenization and POS tagging has been implemented. The *Generate Class Diagram Code* button creates the code of the class diagram. The generated classes can be viewed in a grid view. The operations like tokenization and spitting can be performed separately (without the generation of classes) as shown in Fig. 3. The details about AR2DT tool like installation/user manual, executable file, source code and sample case studies can be found at [20].

3 Validation

Automatic Teller Machine (ATM) Case Study: Rumbaugh et al. [19] first analyzed the automatic teller machine case study by using OMT methodology. We took the same problem statement to present the results of analysis. The initial software requirements of ATM, expressed as a plain text, are shown in Fig. 4.

Design the software to support a computerized banking network including both human cashiers and automatic teller machines (ATMs) to be shared by a consortium of banks. Each bank provides its own computer to maintain its own accounts and process transactions against them. Cashier stations are owned by individual banks and communicate directly with their own bank's computers. Human cashiers enter account and transaction data. Automatic teller machines communicate with a central computer which clears transactions with the appropriate banks. An automatic teller machine accepts a cash card, interacts with the user, communicates with the central system to carry out the transaction, dispenses cash, and prints receipts. The system requires appropriate record-keeping and security provisions. The system must handle concurrent accesses to the same account correctly. The banks will provide their own software for their own computers.

Fig. 4. Automatic teller machine problem statement

Rumbaugh et al. [19] took all the nouns and created a list of classes from the case study. The set of classes are 23: Software, Consortium, Cash receipt, Cash card, Account data, Baking network, Bank computer, Bank, Traction, Access, Cashier station, Central computer, Transaction, ATM, Cashier., Transaction data, Security provision, Record keeping provision, System, Cost, Receipt, Account, and Customer. In our case, AR2DT generate 10 classes for ATM case study as shown in the Fig. 5.

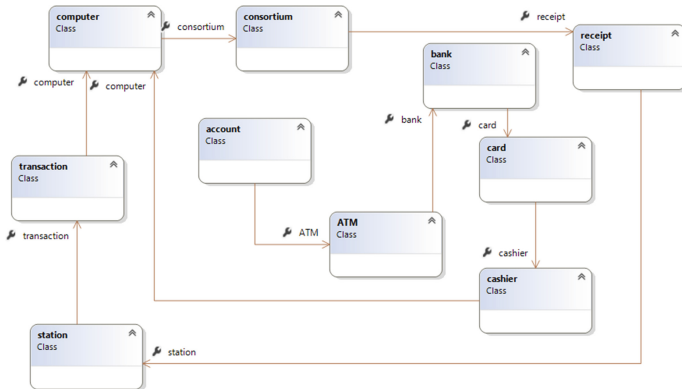


Fig. 5. Conceptual class model of ATM by AR2DT

3.1 Evaluation of Results with the State-of-the-Art

It is assumed in this paper that models given in the object oriented books are correct so we took them all as our answer key for matching our results. For evaluation purpose, we considered three type of measures i.e. precision, recall and over specification. Precision shows how much the information was correct and present in the answer key. Following equations are used to calculate precision, recall and over specification:

$$\begin{aligned}
 \textit{Precision} &= N_{correct} / (N_{correct} + N_{incorrect}) \\
 \textit{Recall} &= N_{correct} / (N_{correct} + N_{missing}) \\
 \textit{Over - specification} &= N_{extra} / (N_{correct} + N_{missing})
 \end{aligned}$$

We evaluate the performance of AR2DT tool against three case studies i.e. ATM, Electronic Filling Program (EFP) and Local Hospital Problem (LHP). However, due to space limitations, we only provide the details of ATM case study and further details can be found at [20]. We compare the results of AR2DT tool with high impact journal research study (i.e. Class-Gen [8]). The results are summarized in Table 1.

Table 1. Evaluation of results with Class-Gen [8]

Sr. #	Case study	Correct classes	Incorrect classes	Missing	Extra	Precision	Recall	OS
1	ATM	9	0	1	1	100%	90%	10%
2	EFP	6	1	3	2	85.7%	66.7%	22.2%
3	LHP	5	0	0	0	100%	100%	0%
(AR2DT)	Avg.					94.9%	85.56%	10.73%
Class-Gen	Avg.					82.6%	83.3%	42%

It can be seen from the Table 1 that the results of AR2DT for precision, recall and over specification are significantly improved as compared to Class-Gen [8].

4 Comparative Analysis with the State-of-the-Art

In this section, we compare our proposed approach with state-of-the-art approaches. We believe it is important to first highlight few studies relevant with the subject of automatic conceptual class model generation from NL software requirements using natural language processing. We considered latest paper ranging from 2003-14 from well-known repositories Springer [15], ACM [16], Elsevier [17] and IEEE [18].

Ibrahim and Ahmad [4] suggested a methodology for the automation of analysis process for class diagram generation from natural language text using NLP. They developed a RACE tool to extract the classes and relationships for class diagram generation. Kumar and Sanyal [5] analyze the natural language text and generated class model and use case model from the SUGAR tool.

Deeptimahanti and Sanyal [7] used Stanford parser, JavaRAP and WordNet for the conversion of NL requirements to UML models semi-automatically. Elbendak et al. [8] developed class-gen tool to generate class diagram from use case descriptions through semi-automated approach. Sharma et al. [9] developed FCDT tool and used RSA algorithm for production of functional design.

Viney et al. [10] developed R-tool to analyze the NL requirements for identification of classes, attributes, methods and relationships which serves as the basis for the creation of class diagram. Author used tokenization as NLP technique. Alkhaider et al. [11] suggested a framework for class diagram generation from the NL requirements by using MIMB and GATE tool. Tripathy and Rath [12] developed a methodology for the identification of class name from the SRS documents in automated manner. Harmain and Gaizauskas [13] developed CM-Builder for the creation of class diagram from NL requirements in semi-automated way.

4.1 Comparative Analysis

To this point, we present existing state-of-the-art approaches in the given context. Now, we compare significant studies with our proposed approach to highlight the strengths and weaknesses. We use three parameters to perform this comparison as follows (1) *Input* define the format of requirements which has been used to generate class model. (2) *Coverage* describes the coverage area of the selected research study i.e. whether the research study covers the generation of a Class (C), Relationship (R), Attribute (A) and Operation (O). (3) *Automated* evaluates the involvement of manual steps required on the textual requirements before apply NLP approach. It can be evaluated as Automatic and Semi-Automatic (in case some manual processing is required). The summary of comparison is given in Table 2.

Table 2. Comparative analysis of proposed approach with state-of-the-art

Paper	Input	Coverage				Automated
		C	R	A	O	
Ibrahim and Ahmad [4]	Plain text	Yes	Yes	Yes	No	Semi-automatic
Kumar and Sanyal [5]	Plain text	Yes	No	Yes	Yes	Semi-automatic
Deeptimahanti and Sanyal [7]	Plain text	Yes	No	No	No	Semi-automatic
Elbendak et al. [8]	Plain text	Yes	Yes	Yes	No	Semi-automatic
Sharma et al. [9]	RS	Yes	Yes	No	Yes	Semi-automatic
Viney et al. [10]	Plain text	Yes	Yes	Yes	Yes	Semi-automatic
Alkhader et al. [11]	Plain text	Yes	Yes	Yes	No	Semi-automatic
Tripathy and Rath [12]	RS	Yes	Yes	Yes	Yes	Semi-automatic
Harmain and Gaizauskas [13]	RS	Yes	Yes	Yes	No	Semi-automatic
AR2DT	Plain text	Yes	Yes	No	No	Automatic

It can be seen from the Table 2 that our approach fully automate the requirement to design automation process which is a significant contribution. Furthermore, our experimental results (Sect. 3.1) are more encouraging as compared to other studies. However, we are not dealing with the generation of association and operation. We intend to include such missing features in AR2DT in near future.

5 Conclusions and Future Work

This article presents a novel Natural Language Processing (NLP) approach to automatically generate conceptual class model from early software requirements. Particular, the new sentence splitting, tokenization and POS tagging rules are defined to avoid the manual processing which is usually required on textual requirements before the generation of class model. As a part of research, Automated Requirement 2 Design Transformation (AR2DT) tool has been developed to automatically generate class model with code from initial plain

text requirements. The application of AR2DT is validated through three benchmark case studies. Experimental results prove that the recall, precision and over specification of AR2DT tool are significantly improved as compared to the state-of-the-art. Furthermore, AR2DT is fully automated and does not require any manual processing on textual requirements.

Currently, AR2DT does not deal with the generation of class relationships like aggregation, composition and inheritance. Furthermore, the generation of methods and cardinalities are also missing. We intend to include such missing features in AR2DT in our future article.

References

1. Meter, M., Borukhov, B., Crivaro, M., Shafir, M., Thamrongattanarit, A.: MedLingMap: a growing resource mapping the bio-medical NLP field. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), Montreal, Canada, 8 June 2012, pp. 140–145 (2012)
2. UMBER, A., Bajwa, I.S., Asif Naeem, M.: NL-based automated software requirements elicitation and specification. In: Abraham, A., Lloret Mauri, J., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) ACC 2011. CCIS, vol. 191, pp. 30–39. Springer, Heidelberg (2011). doi: [10.1007/978-3-642-22714-1_4](https://doi.org/10.1007/978-3-642-22714-1_4)
3. Sneed, H.M.: Testing against natural language requirements. In: Seventh International Conference on Quality Software. IEEE (2007)
4. Ibrahim, M., Ahmad, R.: Class diagram extraction from textual requirements using natural language processing (NLP) techniques. In: Second International Conference on Computer Research and Development, pp. 200–204. IEEE Computer Society, IEEE (2010)
5. Kumar, D.D., Sanyal, R.: Static UML model generator from analysis of requirements (SUGAR). In: Advanced Software Engineering and Its Applications, pp. 77–84. IEEE (2008)
6. Liu, D., Subramaniam, K., Eberlein, A., Far, Behrouz, H.: Natural language requirements analysis and class model generation using UCDA. In: Orchard, B., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS (LNAI), vol. 3029, pp. 295–304. Springer, Heidelberg (2004). doi: [10.1007/978-3-540-24677-0_31](https://doi.org/10.1007/978-3-540-24677-0_31)
7. Deeptimahanti, D.K., Sanyal, R.: Semi-automatic generation of UML models from natural language requirements. In: ISEC 2011, pp. 165–174. ACM (2011)
8. Elbendak, M., Vickers, P., Rossiter, N.: Parsed use case descriptions as a basis for object-oriented class model generation. *J. Syst. Softw.* **84**, 1209–1223 (2011). 2011 Published by Elsevier Inc.
9. Sharma, V.S., Sarkar, S., Verma, K., Panayappan, A., Kass, A.: Extracting high-level functional design from software requirements. In: 16th Asia-Pacific Software Engineering Conference. IEEE (2009)
10. Vinay, S., Aithal, S., Desai, P.: An NLP based requirements analysis tool. In: International Advance Computing Conference. IEEE (2009)
11. Alkhader, Y., Hudaib, A., Hammo, B.: Experimenting with extracting software requirements using NLP approach. In: ICIA. IEEE (2006)
12. Tripathy, A., Rath, S.K.: Application of natural language processing in object oriented software development. In: International Conference on Recent Trends in Information Technology. IEEE (2014)
13. Harmain, H.M., Gaizauskas, R.: CM-builder: a natural language-based CASE tool for object-oriented analysis. *Autom. Softw. Eng.* **10**, 157–181 (2003). Springer

14. <https://sharpnlp.codeplex.com/>. Accessed 12 Sept 2016
15. Springer. <http://www.springer.com/in/>. Accessed Sept 2016
16. ACM. <http://dl.acm.org>. Accessed Sept 2016
17. Elsevier. <https://www.elsevier.com>. Accessed Sept 2016
18. IEEE Scientific database. <http://ieeexplore.ieee.org/Xplore/home.jsp>. Accessed Sept 2016
19. Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., Lorensen, W.: Object-Oriented Modeling and Design. Pearson Education, Upper Saddle River (1991)
20. AR2DT Tool. <http://ceme.nust.edu.pk/ISEGROUP/Resources/ar2dt/ar2dt.html>

The Applications of Natural Language Processing (NLP) for Software Requirement Engineering - A Systematic Literature Review

Farhana Nazir, Wasi Haider Butt, Muhammad Waseem Anwar^(✉),
and Muazzam A. Khan Khattak

College of Electrical and Mechanical Engineering (CEME),
National University of Sciences and Technology (NUST), H-12, Islamabad, Pakistan
farhana.nazir14@ce.ceme.edu.pk,
{wasi,waseemanwar,muazzamak}@ceme.nust.edu.pk

Abstract. Natural Language Processing (NLP) is a well-known technique of artificial intelligence to extract the elements of concerns from raw plain text information. It can be utilized to process the early software requirements in order to achieve the goals like requirement prioritization and classification (functional and non-functional). To the best of our knowledge, no research work is available yet to examine and summarize the utilization of NLP in the domain of Software Requirement Engineering (SRE). Therefore, in this paper, we investigate the applications of NLP in the context of SRE. A Systematic Literature Review (SLR) is carried out to select 27 studies published during 2002–2016. Consequently, 6 NLP techniques and 14 existing tools are identified. Furthermore, 9 tools and 2 algorithms, proposed by the researchers, are presented. It has been concluded that the NLP techniques and tools are highly supportive to accelerate the SRE process. However, some manual operations are still required on initial plain text software requirements before applying the desired NLP techniques.

Keywords: NLP · SRE · NLP tools · Software requirements

1 Introduction

Software requirements are the foremost attributes of the system under development. These are usually classified into four major groups i.e. Business Requirements, Functional Requirements (FR), Non-Functional Requirements (NFR) and Domain Requirements. Initially, the software requirements are gathered and expressed in human readable natural language as a plain text. However, such textual requirements are of least use for technical stake holders. Therefore, it is essential to refine the early requirements for appropriate further utilization. However, manual enhancement of initial software requirement is laborious and time-consuming activity. On the other hand, Natural Language Processing (NLP) is a knowledge discovery approach to automatically extract the elements of concerns from raw plain text documents. Consequently, it is utilized to polish and extract desired software requirements from initial natural language artifacts.

As NLP provides sophisticated text mining features, it is commonly used in various software engineering areas [1–5]. For example, NLP is utilized to transform the functional software requirements into design artifacts [6, 7]. It is also used to refine the ambiguities from initial textual requirements [8–10]. Although NLP techniques have been practiced in several software engineering areas [11–18], there is no study available yet to the best of our knowledge that investigate and summarize the applications of NLP in software requirement engineering domain. Therefore, in this article, we investigate the application of NLP techniques in software requirement engineering to get the answers of the following research questions:

RQ1: What are the leading software requirements areas where NLP techniques are frequently practiced?

RQ2: What are the primary NLP activities in the context of software requirement engineering?

RQ3: What are the leading tools, proposed/used by the researchers, for software requirement engineering?

We develop a review protocol (Sect. 2.2) that contains selection and rejection criteria. We define six categories (Sect. 2.1) for the classification of selected 27 studies as shown in Fig. 1. We investigate the selected studies to identify 6 NLP techniques (Sect. 3.1) that are frequently practiced independently as well as jointly in the area of software requirement engineering. Furthermore, we also identified 9 proposed tools, 14 utilized tools and 2 algorithms (Sect. 3.2).

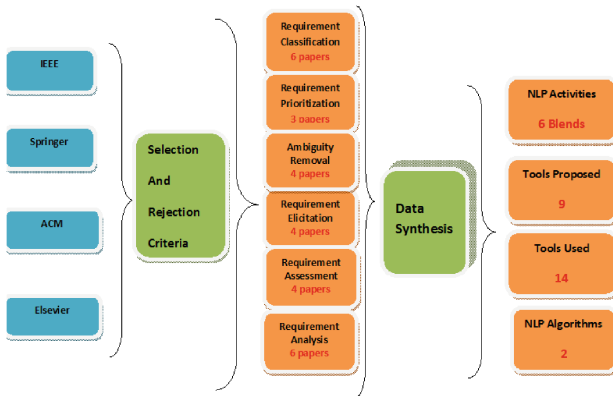


Fig. 1. Overview of research

2 Research Methodology

Systematic Literature Review (SLR) [20] is used to perform this study. The research methodology consists of five stages: (i) Category Definition (ii) Selection Rejection Criteria (iii) Search process (iv) Quality Assessment (v) Data Extraction.

2.1 Category Definition

We define six significant categories for the selection of studies as follows:

Classification: The initial requirements are further classified on the basis of functionality like functional requirements, Non-Functional Requirements, etc. All research work dealing with such classification of requirements are included in this category.

Prioritization: In software system, priority is given to requirements according to the importance of their impact on the system. All the researches that deal with requirement prioritization are included in this category.

Ambiguity Removal: All research works that deal with the removal of ambiguous requirements from initial text are included in this category.

Requirement Elicitation: All research works that deal with the requirement elicitation from initial text by utilizing NLP techniques are placed in this category.

Requirement Assessment: The research works that deal with the evaluation of the impact of the requirements from initial plain text by employing NLP techniques are placed in this category.

Requirement Analysis: The studies that perform analysis on the initial textual requirements to get the desired features are placed in this category.

General: It is possible that some studies belong to more than one above-mentioned categories. All such studies are placed in this category.

2.2 Review Protocol

Review protocol has been set by maintaining the rules and regulations of SLR [20]. The RQ's and background are already covered in Sect. 1. The details of further review protocol stages have been described in subsequent sections.

Selection and Rejection Criteria. The selection and rejection of research papers is based on following parameters:

- Select only those researches which are relevant to our research questions i.e. the study must utilize NLP approach for software requirement engineering.

- We only choose studies that must be published in one of these databases i.e. IEEE [21], ACM [22], Springer [23] and Elsevier [24] during 2002–2016.
- The papers with same research contents should be selected once.

Search Process. The given selection and rejection criteria show that we just used four scientific databases (IEEE, ACM, Springer, and Elsevier) for our research process. We use specific terms related to our topic for the search and the results of these search are shown in Table 1. We apply various filters (e.g. publication year, etc.) to shorten the number of search results.

Table 1. Search terms and results

Sr.#	Terms	Operator	No. of search results			
			IEEE	ACM	Springer	Elsevier
1	NLP		1925	1750	222	3,465
2	Software requirement classification	AND	213	401	61	1810
		OR	717	7,732	23	5,623
3	Software requirement categorization	AND	18	177	26	90
		OR	67	2,300	16	4,433
4	Software requirement classification + NLP	AND	11762	3	0	10
		OR	3,321	5,231	3,345	4,321

Quality Assessment. We assess the quality of selected studies with following parameters: (1) The data assessment from selected studies is based on the solid facts and theoretical perspective (2) The selected studies have been properly validated through case studies and experiments (3) The search we choose is most important factor therefore we use four most genuine and globally accepted scientific databases, i.e. IEEE, SPRINGER, ELSEVIER, ACM.

Data Collection and Synthesis. The elements of data extraction/synthesis are shown in Table 2. We use this template for each selected study to get desired data.

Table 2. Details of data collection and synthesis

Sr.#	Description	Details
1	References information	Title, author, publication year, publisher detail
Extraction of data		
2	Overview	The main proposal/objective of study
3	Results	Results acquire from the study
Synthesis of data		
4	Classification	According to defined categories Table 3
5	Techniques	NLP techniques used in the studies Table 4
6	Tools	Tools used and proposed in studies (Tables 5 and 6)

3 Results

We have identified overall 27 research papers i.e. 6 Journal and 21 Conference. The selected studies are classified into six categories (Sect. 2.1) as shown in Table 3.

Table 3. Classification of studies

Sr.#	Category	Total	Corresponding studies
1	Classification	6	[19, 26–30]
2	Requirement prioritization	3	[31–33]
3	Ambiguity removal	4	[34–37]
4	Requirement elicitation	4	[38–41]
5	Quality assessment	4	[42–45]
6	Requirement analysis	6	[46–51]

3.1 NLP Techniques

We identify 6 main NLP techniques that have been utilized independently as well as jointly in the domain of software requirement engineering as shown in Table 4.

Table 4. Identification of NLP techniques

Sr.#	NLP techniques	Studies
1	Tokenization	[19, 27, 30, 31, 36, 37, 43, 44]
2	POS tagger	[19, 27–32, 34–36, 42, 44, 50]
3	Text chunking	[19, 27, 38–40, 51]
4	Parsing	[19, 30, 42, 43, 46, 48]
5	VSM	[19]
6	TF-IDF	[19, 49, 51]

There are studies that utilized more than one NLP activities e.g. [19] utilized all six NLP activities. Consequently, we place it against each technique as shown in Table 4. Similar is the case with other studies e.g. [28] etc.

3.2 Tools

We identified 14 existing NLP related tools that have been used by the researchers as shown in Table 5.

Table 5. Tools utilized in the given research context

Sr. No	Tool	Studies
1	SharpNLP [25], SMT solver	[31]
2	C4.5 decision tree algorithm ReqSAC, rational XDE	[42]
3	NLP engine	[19]
4	NARCIA (Natural Language Requirements Change Impact Analyzer)	[38]
5	Stanford tagger, NER.	[29]
6	NLTK (Natural Language ToolKit), pyEnchant	[43]
7	Drools expert	[26]
8	Antiword, jauman	[34]
9	C4.5 algorithm	[45]
10	Stanford parser	[30]
11	QUARS, ARM, WSD, RESI, SREE and NAI	[36]
12	RegeX parser	[43]
13	GATE tool	[37]
14	OpenNLP	[51]

We identify 9 tools and 2 algorithms as given in Table 6.

Table 6. Proposed tools and algorithms

Sr. No.	Tool/algorithms proposed	Reference
1	SNIPR	[31]
2	Text classifier, FEATURE XTRACTOR	[42]
3	NARCIA	[40]
4	NLARE (Natural Lang. Automat. Requ. Evaluator)	[43]
5	WordNET	[34]
6	Model for NLP	[45]
7	MUPRET	[30]
8	ReqAligner	[44]
9	RUBRIC	[39]
Algorithms		
10	Unnamed algorithm	[26]
11	LSAN Bayes classifier	[41]

4 Discussion and Limitations

It has been analysed that NLP techniques show encouraging outcomes while extracting relevant elements from initial plain text software requirements. However, it is usually

required to perform few manual steps at lower level NLP activities e.g. tokenization and POS tagging. Therefore, it cannot be said that NLP fully automate the process of requirement refinement from initial plain text. However, the proposal of latest tools in this regard is highly beneficial. For example, [40] proposed a tool NARCIA for analysing the impact of change in natural language requirements. Although we utilized renowned scientific repositories, there is a chance that we might miss few studies from other scientific resources e.g. Google scholar etc.

5 Conclusion and Future Work

This article investigates the applications of Natural Language Processing (NLP) for Software Requirement Engineering (SRE). A Systematic Literature Review (SLR) has been carried out to select 27 studies published during 2002–2016. As a result, 6 NLP techniques are identified that can be applied alone as well as jointly. Moreover, 14 existing tools are presented. Furthermore, the 9 tools and 2 algorithms, proposed by the researchers, are also identified. It has been concluded that the NLP techniques and tools certainly accelerate the SRE process. However, few manual operations are usually required on the initial plain text requirements before applying desired NLP approach. The tools and techniques, presented in this SLR, provide the platform for the SRE researchers. For example, this research can be extended to examine the application of NLP for the whole Software Development Life Cycle (SDLC) e.g. design and testing phases etc.

References

1. Ryan, K.: The role of natural language in requirement engineering. In: IEEE International System Requirement Engineering (1993)
2. Binkley, D., Lawrie, D.: Information Retrieval Applications in Software Maintenance and Evolution (2010)
3. Funke, M.H., Sauer, S., Engels, G., Guldali, B.: Semi-automated test planning for e-ID systems by using requirements clustering. In: Automated Software Engineering (2009)
4. Frakes, W.B., Nejme, B.A.: SIGIR Forum, vol. 21, pp. 30–36 (1987)
5. Berry, D.M., Kaiser, G.E., Maarek, Y.S.: An information retrieval approach for automatically constructing software libraries. IEEE Trans. Softw. Eng. **17**, 800–813 (1991)
6. Yue, T., Briand, L.C., Labiche, Y.: A systematic review of transformation approaches between user requirements and analysis models. Requir. Eng. **16**, 75–99 (2010)
7. Paech, B., Martell, C. (eds.): Innovations for Requirement Analysis. From Stakeholders' Needs to Formal Designs. Springer, Berlin, Heidelberg (2008)
8. Kiyavitskaya, N., Zeni, N., Mich, L., Berry, D.M.: Requirements for tools for ambiguity identification and measurement in natural language requirements specifications. Requir. Eng. **13**, 207–239 (2008)
9. Nuseibeh, B., de Roeck, A., Francis, A.W.C.: Identifying nocuous ambiguities in natural language requirements. In: 14th IEEE International Conference Requirements Engineering, pp. 59–68 (2006)

10. Yang, H., Willis, A., De Roeck, A., Nuseibeh, B.: Automatic detection of nocuous coordination ambiguities in natural language requirements. In: Proceedings of the IEEE/ACM International Conference on Automated Software Engineering (2010)
11. Weber-Jahnke, J.H., Onabajo, A.: Finding defects in natural language confidentiality requirements. In: Proceedings of the 17th IEEE International Conference on Requirements Engineering (2009)
12. De Lucia, A., Oliveto, R., Sgueglia, P.: Incremental approach and user feedbacks: a silver bullet for traceability recovery. In: Proceedings of the 22nd IEEE International Conference on Software Maintenance (2006)
13. De Lucia, A., Oliveto, R., Tortora, G.: Assessing IR-based traceability recovery tools through controlled experiments. *Empirical Softw. Eng.* **14**, 57–92 (2009)
14. Sultanov, H., Hayes, J.H., Kong, W.-K.: Application of swarm techniques to requirements engineering: requirements tracing. In: Proceedings of the IEEE International Conference on Requirements Engineering (2010)
15. Sundaram, S.K., Hayes, J.H., Dekhtyar, A., Holbrook, E.A.: Assessing traceability of software engineering artifacts. *Requir. Eng. J.* **15**, 313–335 (2010)
16. Duan, C., Cleland-Huang, J.: Clustering support for automated tracing. In: Proceedings of the 22nd IEEE/ACM International Conference on Automated Software Engineering (2007)
17. Zou, X., Settimi, R., Cleland-Huang, J.: Term-based enhancement factors for improving automated requirement trace retrieval. In: Proceedings of the International Symposium on Grand Challenge in Traceability (2007)
18. Lormans, M., van Deursen, A.: Can LSI help reconstructing requirements traceability in design and test? In: Proceedings of the International Conference on Software Maintenance and Reengineering (2006)
19. Falessi, D., Cantone, G., Canfora, G.: Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques. *IEEE Trans. Softw. Eng.* **39**, 18–44 (2013)
20. Kitechenhem, B.: Procedures for Performing Systematic Reviews. Keele University (2004)
21. IEEE (2016). <http://ieeexplore.ieee.org/>
22. ACM. <http://dl.acm.org/>
23. Springer Scientific Database (2016). <http://www.springer.com/>
24. Elsevier (2016). <https://www.elsevier.com>
25. CodePlex. <https://sharpenlp.codeplex.com/>
26. Roshni, R., Sharma, S.V.: A framework for identifying and analyzing non-functional requirements from text. In: TwinPeaks Proceedings of Requirements, pp. 1–8 (2014)
27. Al-Zahgoul, F.A., Hudaib, A., Abushariah, M., Alqudah, A.: A suggested framework for software requirement classification. In: IEEE 17th UKSIM-AMSS (2015)
28. Ilieva, M., G., Ormandjieva, O.: Automatic transition of natural language software requirements specification into formal presentation. In: Montoyo, A., Muñoz, R., Métails, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 392–397. Springer, Heidelberg (2005). doi: [10.1007/11428817_45](https://doi.org/10.1007/11428817_45)
29. Sharma, A., Singh, D.: Natural Language based Component Extraction from Requirement Engineering Document and its Complexity Analysis (2011)
30. Thanwadee, C.A.: Ontology-based multiperspective requirements traceability framework. *Knowl. Info. Syst.* **25**(3), 493–522 (2009)
31. Sarkani, S., Holzer, T., McZara, J.: Software requirements prioritization and selection using linguistic tools and constraint solvers a controlled experiment. *Empir. Softw. Eng.* **20**, 1721–1766 (2014). Springer

32. Nordin, A., Ng, K.-Y., Lau, K.-K.: Extracting Elements of Component-Based Systems from Natural Language Requirements (2011)
33. Ramzan, M., Ghayyur, S.A.K., Babar, M.I.: Challenges and Future Trends in Software Requirements Prioritization (2011)
34. Matsuoka, J., Lepage, Y.: Ambiguity spotting using WordNet semantic similarity in support to recommended practice for software requirements specifications. In: NLP-KE (2011)
35. Shah, U.S., Jinwala, D.C.: Resolving ambiguities in natural language software requirements: a comprehensive survey. In: ACM SIGSOFT (2015)
36. Daniel, M., Berry, D., Kamsties, E., Denger, C.: Higher quality requirements specifications through natural language patterns. In: Proceedings of SwSTE (2003)
37. Chug, A., Hayrapetian, A., Raje, R., Malhotra, R.: Analyzing and evaluating security features in software requirements. In: ICICCS 2016 (2016)
38. Sabetzadeh, M., Goknil, A., Briand, L.C., Zimmer, F., Arora, C.: Change impact analysis for natural language requirements: an NLP approach. In: Requirements Engineering Conference (2015)
39. Sabetzadeh, M., Frank, Z., Gnaga, R., Briand, L., Arora, C.: RUBRIC: a flexible tool for automated checking of conformance to requirement boilerplates. In: Proceedings of Software Engineering (2013)
40. Arora, C., Sabetzadeh, M., Goknil, A., Briand, L.C., Zimmer, F.: NARCIA: an automated tool for change impact analysis in natural language requirements. In: Proceedings of Software Engineering (2015)
41. Feng, G., Guo, J., Jing, B.-Y., Sun, T.: Feature subset selection using naive Bayes for text classification. *J. Pattern Recogn. Lett.* **65**, 109–115 (2015)
42. Ormandjieva, O., Kosseim, L., Hussain, I.: Automatic quality assessment of SRS text by means of a decision-tree-based text classifier. In: Conference on Quality Software(2007)
43. Huertas, C., Juárez-Ramírez, R.: NLARE, a natural language processing tool for automatic requirements evaluation. In: CUBE 2012, pp. 371–378 (2012)
44. Rago, A., Marcos, C., Diaz-Pace, J.A.: Identifying duplicate functionality in textual use cases by aligning semantic actions. *Softw. Syst. Model.* **15**, 579–603 (2014)
45. Ormandjieva, O., Hussain, I., Kosseim, L.: Toward a Text Classification System for the Quality Assessment of Software Requirements Written in Natural Language (2007)
46. MacDonell, S.G., Min, K., Connor, A.M.: Autonomous Requirements Specification Processing Using Natural Language Processing
47. Nattoch Dag, J., Regnell, B., Carlshamre, P.: A feasibility study of automated natural language requirements analysis in market-driven development. *Requir. Eng.* **7**, 20–33 (2002)
48. Casagrande, E., Woldeamlak, S., Woon, W.L., Zeineldin, H.H., Svetinovic, D.: NLP-KAOS for systems goal elicitation: smart metering system case study. *IEEE Trans. Softw. Eng.* **40**(10), 941 (2014)
49. Ferrari, A., Spagnolo, G.O., Dell’Orletta, F.: Mining commonalities and variabilities from natural language documents. In: Proceedings of the 17th International Software Product Line Conference (2013)
50. Fatwanto, A.: Software requirements specification analysis using natural language processing. In: IEEE 2013 International Conference in QiR (Quality in Research) (2013)
51. Hayes, J.H., Dekhtyar, A., Holbrook, E.A.: A study of methods for textual satisfaction assessment. *Empir. Softw. Eng.* **18**(1), 139–176 (2012)

Smart Fetal Monitoring

Jane You¹(✉), Qin Li², Zhenhua Guo³, and Ruohan Zhao¹

¹ Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
{csyjia, csrzhao}@comp.polyu.edu.hk

² School of Software Engineering, Shenzhen Institute of Information Technology,
Shenzhen, Guangdong, China
kenneth_lee_qin@qq.com

³ Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guangdong, China
zhenhua.guo@sz.tsinghua.edu.cn

Abstract. Fetal movement is an important index of fetal well-being. The absence or a reduction in fetal movement is a symptom or an alarming sign of fetal compromise or even death. The timely detection of abnormalities in fetal movement is vital to reduce the incidence of fetal loss, perinatal morbidity and maternal distress. This paper presents a smart fetal monitoring system to detect fetal movement and monitor movement pattern safely and reliably by using a new fabric sensor belt. The monitoring belt is wearable, non-intrusive, radiation free and washable. The new algorithms are robust for automated analysis, detection and assessment of fetal condition, which include effective noise removal, feature extraction, time sequence data analysis and decision support. Both the design of fabric sensor and functionality implementation of the belt are original and unique. The results of preliminary clinical trials demonstrate the feasibility of our prototype. There are no such similar products available in the market.

1 Introduction

Fetal health assessment aims to identify any possible problems during pregnancy and at labor. The timely detection of fetal abnormalities is very important to prevent fetal death. Electronic fetal monitoring (EFM) is widely adopted in practice. In general, clinical approaches to fetal monitoring include monitoring of fetal movements, symphysial–fundal height (SFH) measurement, auscultation of fetal heart rate (FHR), Doppler assessment of fetal heart rate, ultrasound assessment of fetal growth (and interpretation with both standardized growth charts and customized growth charts), amniotic fluid volume measurements, ultrasound assessment of the fetal biophysical profile, cardiotocography (CTG), Doppler ultrasound recording of blood flow in the fetal umbilical artery, Doppler recording blood flow in other fetal vessels (e.g. middle cerebral artery) and placental grading.

Fetal movements have been considered an important indicator of fetus well-being. Clinically decreased fetal movements or an absence of fetal movements for an extended period of time are regarded as alarming sign of risk – a fetus at risk for fetal compromise or associated with an increased risk of intrauterine fetal death. Although mothers can

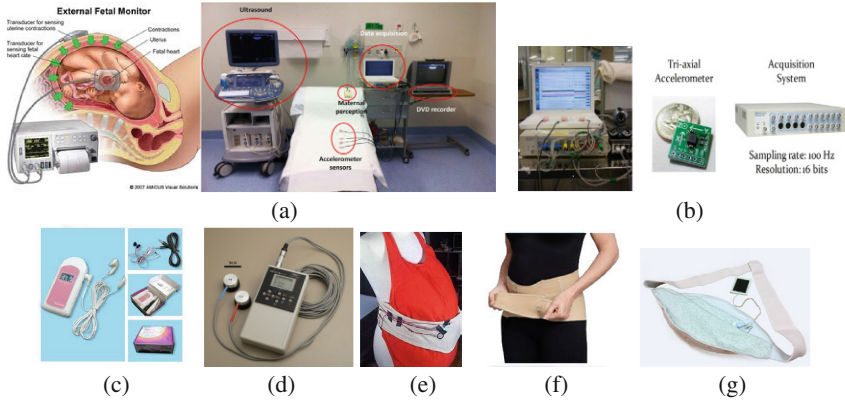


Fig. 1. Different samples for fetal movement detection.

feel fetal movements, the use of “kick-charts” based on fetal movement counting is subjective. The outcome of the detection will not be accurate because the counting is based on the quantifying the number of movements felt by the woman over a set period of time. Studies show that there are wide variations in practices with respect to the fetal movement counting and the way how pregnant women feel and count the movement. It is essential to detect true fetal movements and count the number of movement patterns.

Conventional Doppler fetal monitor uses ultrasound to project the fetus’ heart rate and detect fetal movement for prenatal care. Although the advances in ultrasound technology have led to the development of different models of ultrasonic fetal monitors to capture fetal movement, the non-stress test using Doppler ultrasonic monitor is subject to two major limitations: (1) the frequent exposure to ultrasound by pregnant women for regular medical check-up may not be healthy to pregnancy; (2) not practical for home-based objective monitoring of fetal movement by Doppler device. Therefore it is highly desired to develop a safe, convenient, reliable and easy way to monitor fetal health. Figure 1 shows the selected samples of different methods for fetal movement monitoring with the following features: (a) Doppler fetal monitor – use of ultrasound technology for detection of fetal heart rates and movement; (b) A data acquisition system for fetal movement in [8] – a sophisticated data acquisition system at hospital (ultrasound); (c) New models of Doppler ultrasound monitor – portable, but ultrasound used; (d) Fetal movement acceleration recorder in [7] – new fetal movement detector, portable device for data recording only, no monitoring functions; (e) Fetal belt reported in news [9] – detect fetal movement only, no monitoring functions, sensitive to noise, entertainment oriented (mobile connection, electromagnetic wave, not healthy); (f) Pregnancy support belt – Support belt only without any functions; (g) The proposed fetal monitoring belt – a special monitoring belt with pressure sensors to detect the changes of fetal movement for fetal health assessment (NEW).

Based on the comprehensive literature survey, it is concluded that so far there have been no practical products available in the market to meet all of the requirements of safe, reliable, non-intrusive, wearable, waterproof, radiation free, comfortable, convenient, easy to use, automated detection and interactive data sharing. The innovation of our

approach is characterized on the aspects of both hardware and software. The hardware components of our system include: (1) a new design of fabric pad embedded with an array of fabric sensors to detect the movement of fetus within the proper force range for data acquisition; (2) control card with embedded DSP circuits for robust movement detection with functions including noise removal, fusion of data from the array of fabric sensors, movement counting, data recording and pattern assessment. The software consists of: (1) data fusion of multi-channel signals and noise removal; (2) computer-aided analysis of fetal movement patterns for robust individual monitoring by ensemble learning; (3) interactive communication & information sharing with privacy and security protection.

This paper is organized in the following sections. The system structure and its hardware components are briefly described in Sect. 2 while the new algorithms for robust noise removal and data analysis are highlighted in Sect. 3. Section 4 summarizes the testing results of preliminary clinical trials. Finally the conclusion is presented in Sect. 5.

2 The System Design of Smart Fetal Monitoring

The major functions offered by our fetal monitoring system is illustrated in Fig. 2, where the fetal monitoring belt is embedded with a soft sensor pad to capture signals of fetal movement, the control card processes the data in a real-time fashion to serve different purposes – data recording and packaging for a series of fetal movements (kicks); automated fetal health monitoring by robust analysis of movement patterns via ensemble learning; warning under abnormal conditions. To avoid any radiation exposure to mother and fetus, no online data transmission is performed. Instead, the recorded data is transmitted to the eligible parties offline via internet or mobile network for information sharing. Our control card provides reliable online service for real-time monitoring and issuing warning signals if any abnormalities occur.

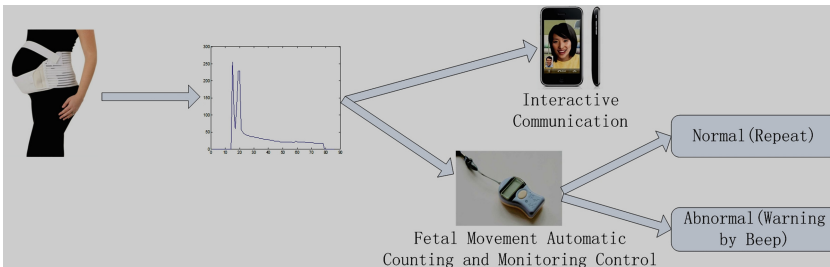


Fig. 2. System structure of the smart fetal monitoring.

One of the major contributions and achievements of our smart fetal monitoring system is the development of a fetal monitoring belt with our own design of a soft sensor pad which consists of an array of fabric sensors and a control card for data acquisition and analysis. More specifically, the relevant technology issues are centered on the development of reliable sensors which not only detect the body changes caused by fetus

kicks reliably with robustness to disturbing noises but also do no harm to the health of both mom and fetus. The major features of the new monitoring belt are summarized as follows:

- *Radiation free*: no radio frequency, very low electromagnetic signals under low power consumption and CPU frequency mode;
- *Safe*: powered by USB with special low power consumption design (low voltage <5 V, low currency <60 mA, sleep mode for idle);
- *Robust to disturbances*: array of fabric sensors with control card embedded with DSP board for optimization;
- *Reliable and easy to use*: stable, wearable with comfort and waterproof for easy cleaning

2.1 Soft Sensor Pad

Our new design fabric sensor pad extends the one-dimension fabric strain sensor string to a two-dimension array of fabric pressure sensor strings as shown in Fig. 3(a). The connection of different sensor components is illustrated in Fig. 3(b), to make the sensor pad robust to disturbances with the expected signal output shown in Fig. 3(c), where a strong signal output for the true fetal kicks and very low signal responses to disturbances. To identify the signals caused by fetal kicks, we will apply competitive algorithm to determine the node which contributes most to the force change among its neighboring nodes. A sigmoid function is used to map the sum of the neighboring signals to the final output which satisfies linearity within a specified range. Figure 4 demonstrates the data process procedure of signals from fabric sensor pad. The innovative aspect of our new sensor pad is to enhance the system computing power and performance by a control card with an embedded DSP board for data processing. The structure of control card is shown in Fig. 5.

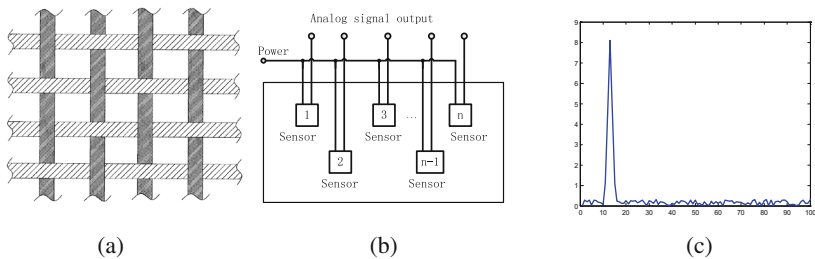


Fig. 3. The structure of monitoring belt with a fabric pressure sensor pad. (a) An array of fabric pressure sensors; (b) The connection of sensor components; (c) The expected signal output.

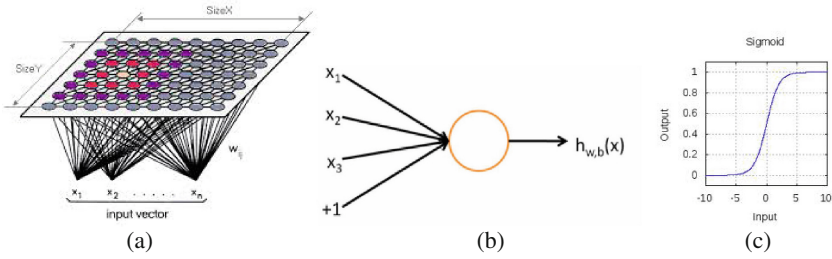


Fig. 4. The data processing procedure of fabric sensor signals. (a) Competitive node signal detection; (b) Fusion of multiple node signals; (c) Mapping function for output.

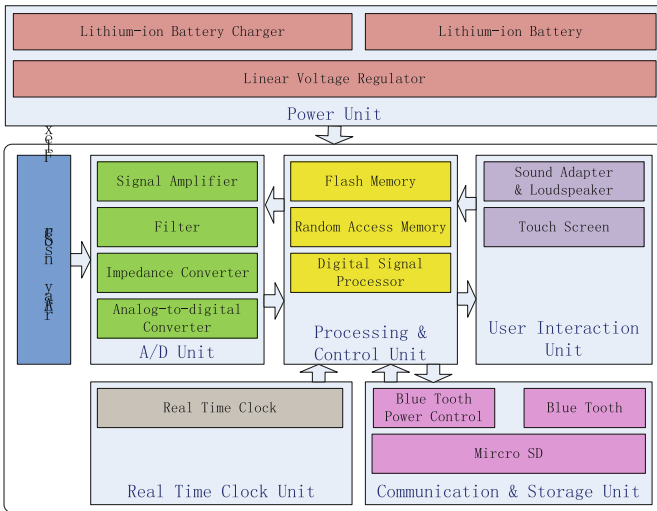


Fig. 5. The structure of control card.

2.2 Control Card

The purpose to develop a special control card with embedded DSP board is for system control and performance enhancement. In contrast to the current fabric sensors [10,11] which respond to the input signals passively without any capacity for data processing, our control card with embedded DSP board will not only control the process for data acquisition, but also enhance performance with data pre-processing including noise filtering, fusion of signals from sensor pad and output signal mapping. Figure 5 illustrates the structure of the control card with components for data acquisition, filtering, storage and processing.

3 The Implementation of Smart Fetal Monitoring

The software development of smart fetal monitoring resolves around the new methods to handle the three fundamental issues: *multiple feature extraction and fusion, detection of movement, classification for abnormalities detection*. The following highlights the relevant details of our approach on the aspects of signal enhancement, feature extraction, classification and performance evaluation.

3.1 Detection of Fetal Movement by Adaptive Signal Enhancement

It is noted that the bandwidth of fetal movement signal is very wide. It is very important to reduce noise of the captured signal of fetal kicks to achieve reliable classification of moving patterns. An adaptive band-pass filtering algorithm is proposed for robust noise removal and signal enhancement.

Considering the performance of Gabor filters which can serve as excellent band-pass filters for signal enhancement, we proposed an adaptive band-pass filtering algorithm by extending 1D Gabor filters with an adaptive band-width selection scheme. The conventional 1D Gabor filters are defined as

$$g(t) = g_e(t) + ig_o(t) \tag{1}$$

$$g_e(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{t^2}{2\sigma^2}} \cos(2\pi f_0 t) \quad g_o(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{t^2}{2\sigma^2}} \sin(2\pi f_0 t) \tag{2}$$

The imaginary part of Gabor filter out put can be used as band-pass filter, which can enhance the signals within the specified bandwidth and smooth the signals over the bandwidth frequencies. Figure 6 shows the imaginary outputs at different frequencies with respect to the same standard deviation σ of Gaussian.

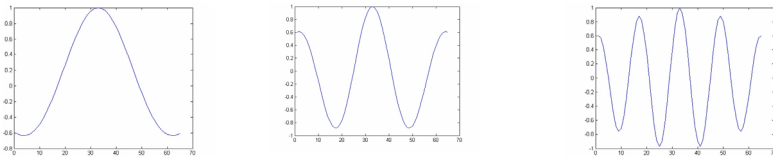


Fig. 6. Gabor filtering with different bandwidth

$R^i(t) = g_r^i(t) * f(t)$. It is important to select the proper bandwidth for band-pass filtering. In contrast to the conventional algorithm which a pre-defined bandwidth, we developed an adaptive scheme to determine an optimized bandwidth by introducing our previous work on scale product of matched filters. The response output of Gabor filter at the specified scale i is defined as:

$P^{ij}(t) = R^i(t) \cdot R^j(t)$. The scale product is defined as the product of the response outputs from the two Gabor filters of two adjacent frequencies i and j ; Fig. 7 illustrates

the effect of the scale product to improve signal noise ratio (SNR), where the signal in each row represents the following items:

- Row 1: input signal s for testing;
- Row 2: noise signal f added to the input signal for simulation testing;
- Row 3, 4 and 5: response outputs R_1 , R_2 and R_3 from the matched filter at different scales;
- Row 6: max plot of maximum values among R_1 , R_2 and R_3 ;
- Row 7: $P_{1,2}$ is the plot of scale product of R_1 and R_2 ;
- Row 8: $P_{2,3}$ is the plot of scale product of R_2 and R_3 .

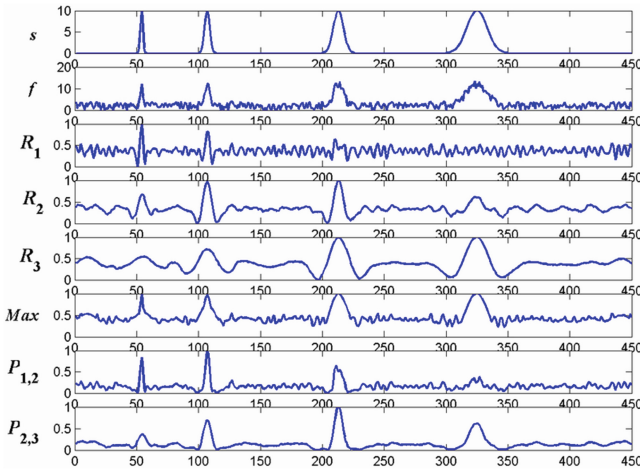


Fig. 7. The effect of scale product for signal enhancement.

3.2 Multiple Feature Fusion and Ensemble Classifier

This process is to integrate multiple fetal movement features in terms of their ranking for robust change detection. The fusion techniques used include: (1) dominant single feature based fusion; (2) likelihood distribution normalization (PPDN)-based fusion; (3) multithreshold fusion; (4) supervised fusion and (5) combination of supervised and PPDN system. Our previous testing results on DR analysis shows that the combination of supervised and PPDN based approach outperforms other techniques. In this project, we will conduct comprehensive comparative studies on the effect of different fusion techniques for multi-feature based fetal movement pattern classification.

Change analysis is an important research topic in various areas and there have been many algorithms developed for change detection. However, the performance of the direct use of these algorithms to detect the changes in a time series of fetal movements is not satisfactory because of the variations of disturbances and time duration of movement period. Traditional approaches often apply single classifier to detect changes, which is more sensitive to the change of parameters or the fluctuation of training data, resulting in low stabilization and

poor robustness. Thus, we adopt ensemble classifier to identify movement sequences with different types of pattern.

4 Testing Results of Clinical Trials

A preliminary clinical trial was conducted at Guangzhou Women & Children Medical Center and Guangzhou No. 3 Hospital, China. Four groups of volunteer subjects participated in the trial. The testing includes both the automatic counting of fetal movement by signal analysis and the manual counting by over 80 volunteer subjects over 3-month data tracking and comparison studies.

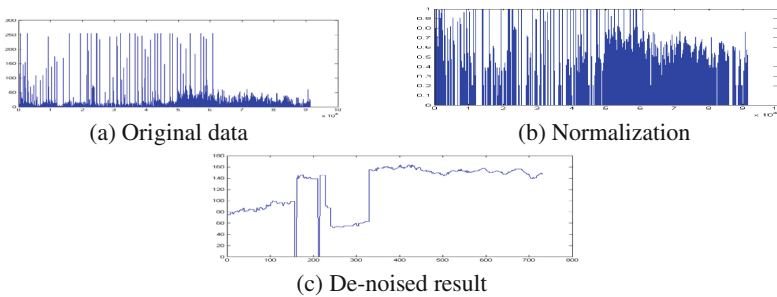


Fig. 8. Noise removal and signal enhancement testing.

Table 1. SNR record.

	Original signal	De-noised signal
SNR (dB)	12.6	140.1

Table 2. Performance matrix.

Data set (no. of recorded data)	Accuracy (%)	FPR (%)	Sensitivity (%)	Specificity (%)
120600	81.9–97.3	2.1–18.3	75.1–87.2	81.6–99.7
252739	78.2–83.7	8.1–17.2	59.3–68.2	61.4–65.3
26777	85.6–95.6	1.9–14.5	71.4–82.1	75.6–95.8
40924	78.9–97.7	1.8–21.3	37.0–92.6	78.6–98.1

The fetal movement and other interference signals (human movement, breathing, etc.) are simulated using a specially designed water channel. Gaussian noises are generated using the largest amplitude of the fetal movement. Figure 8 illustrates the effect of noise removal on a sequence of signals. The SNRs are recorded in Table 1. The signal power is calculated including the interference signals. Figure 9 shows the data distribution, where the fetal movement was captured and recorded in the 7th and 8th signal channel while the marker set

by the volunteer subjects is recorded in the 9th row. Figure 10 is the ROC curve and Table 2 presents some detail performance values. The average accuracy rate of the movement counting is 85.7%. Better results are achieved by adopting the competitive algorithm and ensemble learning algorithm.

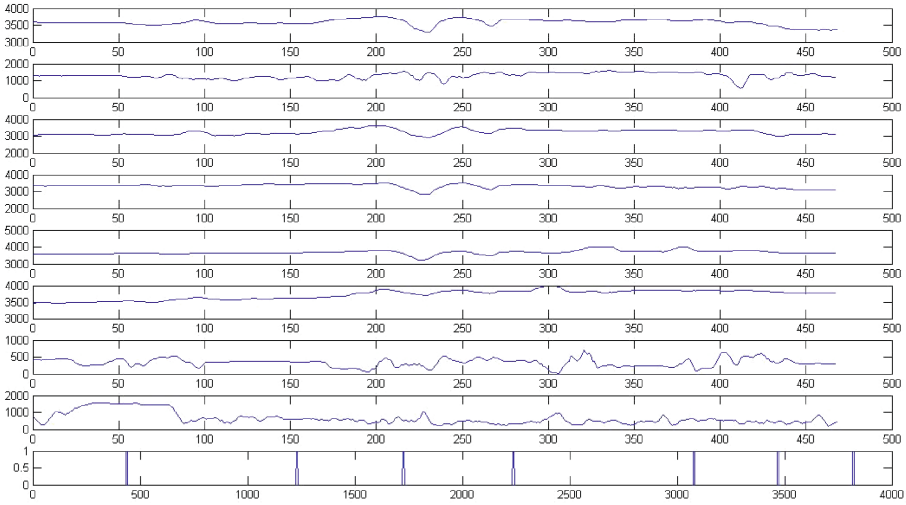


Fig. 9. Fetal movement detection.

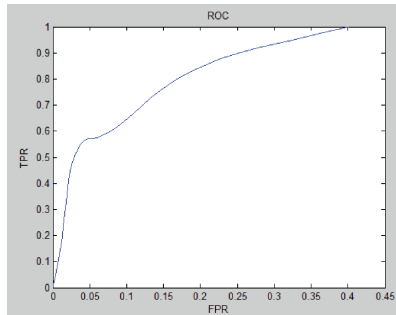


Fig. 10. ROC curve.

5 Conclusions

We conclude that our smart fetal monitoring belt which can detect fetal movement and monitor fetal movement safely and reliably by integrating a new soft sensor pad with robust algorithms. Both the design and functionality implementation of the belt are original and unique. The experimental results provide the basis for further development with excellent market potentials.

Acknowledgement. The authors would like to thank for the partial support of the research grants from Hong Kong Government Innovation Technology Fund (ITF) under its HKRITA Scheme (ITP 025/14TP) and General Research Fund (GRF 152202/14E).

References

1. Sontag, L.W., Wallace, R.F.: An apparatus for recording fetal movement. *Am. J. Psychol.* **45**, 517–519 (1933)
2. Sadosky, E., et al.: Correlation between electromagnetic recording and maternal assessment of fetal movement. *Lancet* **26**, 1141–1143 (1973)
3. Valentin, L., et al.: Recording of fetal movement: a comparison of three methods. *J. Med. Eng. Technol.* **10**, 239–247 (1986)
4. Freda, M.C., et al.: Fetal movement counting: which method? *Am. J. Matern. Child Nurs.* **18**, 314–321 (1993)
5. Froen, J.F., et al.: Fetal movement assessment. In: *Seminars in Perinatal Medicine*, vol. 32, pp. 13–24 (2004)
6. Patrelli, T.S., et al.: Correlation between fetal movement revealed in actography and fetal-neonatal wellbeing. *Clin. Exp. Obstet. Gynaecol.* **38**(4), 382 (2011)
7. Ryo, E., et al.: A new method for long-term home monitoring of fetal movement by pregnant women themselves. *Med. Eng. Phys.* **34**, 566–572 (2012)
8. Boashash, B., et al.: Passive detection of accelerometer-recorded fetal movements using a time-frequency signal processing approach. *Digit. Sig. Process.* **25**, 134–155 (2014)
9. <http://health.dbw.cn/system/2016/11/30/057458900.shtml>

A Network-Based Approach on Big Data for the Comorbidities of Urticaria

Yi-Horng Lai^{1(✉)}, Chih-Chiang Ho², and Piao-Yi Chiou³

¹ Department of Healthcare Management, Oriental Institute of Technology,
New Taipei City 22061, Taiwan
FL006@mail.oit.edu.tw

² Medical Affairs Office, West Garden Hospital,
Taipei City 10864, Taiwan
duke5988@gmail.com

³ School of Nursing, National Taipei University of Nursing and Health Sciences,
Taipei City 11219, Taiwan
piaoyi@ntunhs.edu.tw

Abstract. This study investigates the network properties of urticaria comorbidity. Comorbidities are the presence of one or more additional disorders or diseases that co-occur with a primary disease or disorder. The purpose of this study is to identify diseases that co-occur with urticaria. Research data was collected from 1,154,534 urticaria outpatient department medical records out of 163,141,270 outpatient department medical records from 1997 to 2010 in Taiwan. Through the phenotypic disease network (PDN), this study has identified the diseases that are associated with urticaria. It has been discovered that the PDN has a complex structure where some diseases are highly connected while others are barely connected at all. While not conclusive, these findings can explain that the more connected the diseases are, the higher the mortality rate is, as patients developing highly connected diseases are more likely to be diagnosed at an advanced stage of the disease, which can be reached through multiple paths in the PDN.

Keywords: Medical records · Big data · Urticaria · Comorbidities · The human phenotypic disease network (PDN)

1 Introduction

The medical record is a systematic file that provides a chronicle of a patient's medical history and care. Physicians, nurses and other members of the health care team may make entries in the medical record. Hospitals, over the years, have generated large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient care. To meet the mandatory requirements and the goal to improve the quality of healthcare service, these massive quantities of big data hold the promise of supporting a wide range of medical and healthcare functions, including clinical decision support, disease monitor, and public health management.

For many diseases, there are no definite boundaries, as diseases can have multiple causes and can be related in different dimensions. From a genetic point of view, a pair of diseases can be related because they have both been associated with the same gene, whereas from a proteomic perspective, diseases can be related because disease-associated proteins act on the same pathway [1].

Over the past half-decade, several resources have been constructed to help understand the entangled origins of many diseases. Many of these resources have been presented as networks in which interactions among disease-associated genes, proteins, and expression patterns have been summarized. Goh et al. built a network of Mendelian gene-disease associations by connecting diseases that have been associated with the same genes [2]. Besides, more and more researches have applied the network approach to diseases, such as neurodegenerative diseases [3], infertility etiologies [4], diabetes mellitus [5], and HIV/AIDS [6, 7].

A comorbidity relationship exists between two diseases when they affect the same individual substantially more than chance alone. Over the years, comorbidities have been used extensively to construct synthetic scales for mortality prediction [8, 9], yet their utility exceed their current use. Studying the structure defined by entire sets of comorbidities can help the understanding of many biological and medical questions from a perspective that is complementary to other approaches. For example, a recent study created a comorbidity network in an attempt to elucidate neurological diseases' common genetic origins [10].

Network-based approach can be utilized to analyze high-throughput big data, such as medical records. In this present study, the big data of all diseases recorded in the medical claims is presented in the form of phenotypic disease network (PDN). In order to guide urticaria-related diseases prevention program, this study conducted the PDN of urticaria to explore the relationship between urticaria and other diseases. The objective of this study is to identify diseases that are highly correlated with urticaria.

2 Materials and Research Method

2.1 Data Source

The National Health Insurance (NHI) program was launched in Taiwan in 1995 and covers nearly all residents. In 1999, the Bureau of NHI began to release all claims data in electronic form under the National Health Insurance Research Database (NHIRD) project. The structure of the claim files is described in detail on the NHIRD website and in other publications [11]. NHIRD offers reliable, systematic, and complete data for disease detection. The datasets contained only the visit files, including dates, medical care facilities and specialties, patients' genders, dates of birth, and the four major diagnoses coded in the International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) format [11, 12]. In total, the ICD-9-CM classification consists of 17 different categories at the 3 digit level and 16459 categories at 5 digits [12]. To protect privacy, the data on patients' identities and their medical institutions had been scrambled cryptographically.

In ICD-9-CM, urticaria is coded 708 as urticaria, 708.0 as allergic urticarial, 708.1 as idiopathic urticarial, 708.2 as urticaria due to cold and heat, 708.3 as dermatographic urticarial, 708.4 as vibratory urticarial, 708.5 as cholinergic urticarial, 708.8 as other specified urticarial, 708.9 as urticaria, unspecified, and 995.1 as angioneurotic edema.

The visit files in this study represent 163,141,270 outpatient department data within the NHI from 1997 to 2010. Demographically, the data set consists of 1,154,534 outpatient department medical records with urticarial record from 1,000,000 patients. Of all these patients, 55.71% were females, 44.21% were males, 23.91% were over 60 years of age, and 0.71% of these cases were diagnosed with urticaria (Table 1).

Table 1. Data characteristics of outpatient department medical records in this study.

Variable		Cases	%
Gender	Female	90,885,007	55.71
	Male	72,124,235	44.21
	Unknown	132,028	0.08
Age	–19	38,685,969	23.71
	20–21	18,163,880	11.13
	30–39	21,266,487	13.04
	40–49	24,014,487	14.72
	50–59	22,003,145	13.49
	60–	39,007,302	23.91
Year	1997	1,810,544	1.11
	1998	2,763,207	1.69
	1999	5,788,164	3.55
	2000	12,134,948	7.44
	2001	12,279,135	7.53
	2002	13,209,636	8.10
	2003	13,284,444	8.14
	2004	14,789,420	9.07
	2005	15,029,401	9.21
	2006	14,269,844	8.75
	2007	14,319,129	8.78
	2008	14,208,231	8.71
	2009	14,662,769	8.99
2010	14,592,398	8.94	
Urticaria	Yes	1,154,534	0.71
	No	161,986,736	99.29
Total		163,141,270	100.00

These urticaria outpatient department medical records included 272,258 medical records with allergic urticarial (708.0), 43,116 medical records with idiopathic urticarial (708.1), 4,874 medical records with urticaria due to cold and heat (708.2), 957 medical records with vibratory urticarial (708.4), 1,018 medical records with cholinergic urticarial (708.5), 102,716 medical records with other specified urticarial (708.8), 726,754

medical records with urticaria, unspecified (708.9), and 2,841 medical records with angioneurotic edema (995.1) (Table 2).

Table 2. Data characteristics statistics of urticaria outpatient department medical records in this study.

Variable		Cases	%
Gender	Female	645,244	55.89
	Male	507,983	44.00
	Unknown	1,307	0.11
Age	–19	241,683	20.93
	20–21	162,458	14.07
	30–39	181,774	15.74
	40–49	201,765	17.48
	50–59	150,705	13.05
	60–	216,149	18.72
Year	1997	1,750	0.15
	1998	6,959	0.60
	1999	20,620	1.79
	2000	75,592	6.55
	2001	82,966	7.19
	2002	87,160	7.55
	2003	93,395	8.09
	2004	105,520	9.14
	2005	111,321	9.64
	2006	112,306	9.73
	2007	113,524	9.83
	2008	113,259	9.81
	2009	113,426	9.82
Urticaria	708.0	272,258	23.58
	708.1	43,116	3.73
	708.2	4,874	0.42
	708.4	957	0.08
	708.5	1,018	0.09
	708.8	102,716	8.90
	708.9	726,754	62.95
	995.1	2,841	0.25
Total		1,154,534	100.00

2.2 Measure of the Strength of Comorbidity Relationships

To measure the comorbidity relationships, it is necessary to quantify the strength of comorbidities by introducing a notion of distance between two diseases. However, a

drawback of this approach is that different statistical measures may result in over- or under-estimate of the relationships between rare or prevalent diseases.

In this study, the ϕ -correlation is used to quantify the distance between two diseases. The ϕ -correlation, which is Pearson’s correlation for binary variables, can be expressed mathematically as Eq. 1 [1, 13]. The ϕ -coefficient is given by

$$r_{\phi} = \frac{p_a - p_X p_Y}{(p_X p_Y (1 - p_X)(1 - p_Y))^{1/2}} \tag{1}$$

if $0 < p_X, p_Y < 1$, and $r_{\phi} = 0$ otherwise.

The distribution of r_{ϕ} values representing all disease pairs is presented in Fig. 1.

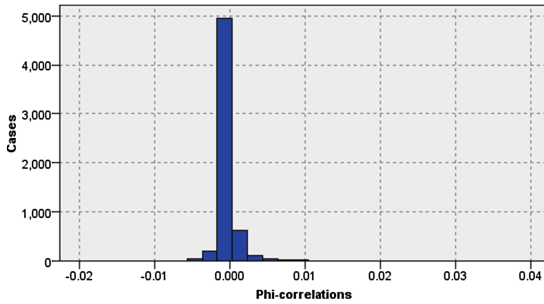


Fig. 1. Distribution of the ϕ -correlation between all disease pairs/groups.

2.3 Network Approach

The data from NHIRD used in this study contained the five major diagnoses codes of all patients. This study calculated ϕ -correlation with Eq. 1. R 3.3.1 with Spark 2.1.0 (with SparkR) was the main software used for data linkage and processing. Descriptive data, including frequencies, percentage and means, are presented. SparkR is an R package that provides a light-weight frontend to use Apache Spark from R. SparkR provides a distributed data frame implementation that supports operations such as selection, filtering, and aggregation with big data [14]. Pajek 4.10 program [15] was used to compute the degree of centrality and betweenness of each node and the path value (ϕ -correlation).

3 Results

The set of all comorbidity associations among all diseases in the study population can be constructed through PDN. In the PDN, nodes are disease phenotypes are identified by unique ICD-9-CM codes, and links phenotypes that show significant comorbidity according to the measures introduced above.

In principle, the number of disease-disease associations in the PDN is proportional to the square of the number of phenotypes. However, many of these associations are

either not strong or not statistically significant [1]. The structure of the PDN can be explored by focusing on the strongest and the most significant of these associations. The PDN can be viewed as a network of the phenotypic space. This network allows people to understand the relationship among illnesses.

The distribution of ϕ -values representing all disease pairs/groups is presented in Fig. 1. Most of them are between 0 and 0.005. A discussion on the confidence interval and statistical significance of these measures can be found in Hidalgo et al.'s study, and ϕ -correlation > 0.01 is statistically significant in this study [1].

Figure 2 shows that there is a high correlation between/among the following pairs/groups of diseases:

- (1) Allergic urticarial (708.0), contact dermatitis and other eczema, unspecified cause (692.9), contact dermatitis and other eczema, due to oils and greases (692.1), and contact dermatitis and other eczema, due to other specified agents (692.8).
- (2) Idiopathic urticarial (708.1), contact dermatitis and other eczema, unspecified cause (692.9), contact dermatitis and other eczema, due to solvents (692.2), and contact dermatitis and other eczema, due to other specified agents (692.8).
- (3) Urticaria due to cold and heat (708.2) and cholinergic urticarial (708.5).
- (4) Vibratory urticarial (708.4) and malignant neoplasm of submandibular gland (142.1).
- (5) Other specified urticarial (708.8), other specified disorders of sweat glands (705.8), other acne (706.1), and seborrheic dermatitis (690.1).
- (6) Urticaria, unspecified (708.9), angioneurotic edema (995.1), and unspecified pruritic disorder (698.9).

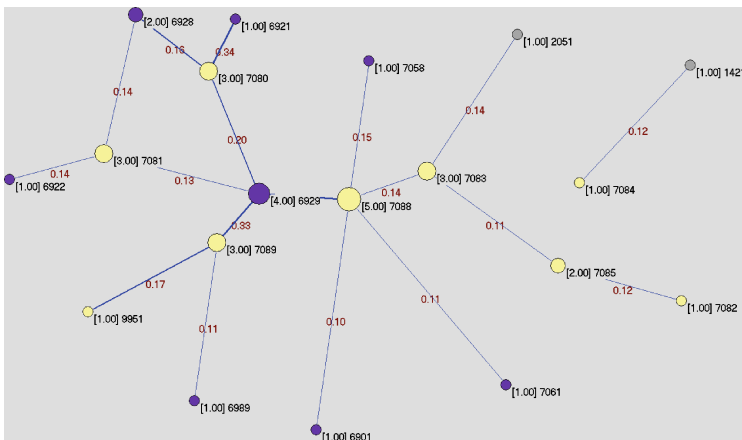


Fig. 2. The PDNs of urticaria, ϕ -correlations > 0.01. [] is the summary of ϕ -correlations with other diseases of this disease, and the size of the node is dependent on the value. For example, the summary of ϕ -correlations with other diseases of 477.9 is 21.00. Node color is based on the ICD9 category.

4 Conclusion

This study provides a comprehensive view of the network characteristics of urticaria. Through the PDN, this study has identified the diseases that are associated with urticaria. It shows that the PDN has a complex structure where some diseases are highly connected while others are barely connected at all. While not conclusive, these observations can explain that the more connected the diseases are, the higher the mortality rate is as patients developing highly connected diseases are more likely those at an advanced stage of disease, which can be reached through multiple paths in the PDN. Exploring comorbidities from a network perspective could help determine the risk between each disease, and give the medical team a reliable recommendation for healthcare.

Both big data and graph analytics can play an important role in future clinical practice and medical research [16]. PDN is a good tool for unsupervised learning in data mining in medical big data, such as medical record and health insurance data. With the result of PDN, future studies can help determine whether differences in the comorbidity patterns are resulted from differences in races, nationalities, or socioeconomic status. The PDN can be the starting point of researches exploring these and related questions.

Acknowledgments. This study is based in part on data from the National Health Insurance Research Database provided by the Bureau of National Health Insurance, Department of Health and managed by National Health Research Institutes (NHRI). The interpretation and conclusions contained herein do not represent those of Bureau of National Health Insurance, Department of Health or National Health Research Institutes.

References

1. Hidalgo, C.A., Blumm, N., Barabási, A.L., Christakis, N.A.: A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**(4), 1–11 (2009)
2. Goh, K.I., Cusick, M.E., Valle, D., Barton, C., Vidal, M., Barabási, A.L.: The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* **104**(21), 8685–8690 (2007)
3. Santiago, J.A., Potashkin, J.A.: A network approach to clinical intervention in neurodegenerative diseases. *Trends Mol. Med.* **20**(12), 694–703 (2014)
4. Tarín, J.J., García-Pérez, M.A., Hamatani, T., Cano, A.: Infertility etiologies are genetically and clinically linked with other diseases in single meta-diseases. *Reprod. Biol. Endocrinol.* **13** (2015). <http://www.rbej.com/content/13/1/31>. Accessed 20 April 2016
5. Lai, Y.H., Wang, T.Y., Yang, H.H.: Network-based analysis of comorbidities: case study of diabetes mellitus. *Commun. Comput. Inf. Sci.* **540**, 210–222 (2015)
6. Moni, M.A., Liò, P.: Network-based analysis of comorbidities risk during an infection: SARS and HIV case studies. *BMC Bioinform.* **15**, 1471–2105 (2014)
7. Lai, Y.H.: A network approach for the comorbidities of HIV/AIDS in Taiwan. *Technol. Health Care* **24**(s1), 377–383 (2016)
8. Iezzoni, L.I., Heeren, T., Foley, S.M., Daley, J., Hughes, J., Coffman, G.A.: Chronic conditions and risk of in-hospital death. *Health Serv. Res.* **29**(4), 435–460 (1994)
9. Schneeweiss, S., Wang, P.S., Avorn, J., Glynn, R.J.: Improved comorbidity adjustment for predicting mortality in medicare populations. *Health Serv. Res.* **38**(4), 1103–1120 (2003)

10. Rzhetsky, A., Wajngurt, D., Park, N., Zheng, T.: Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **104**(28), 11694–11699 (2007)
11. National Health Research Institutes: National Health Insurance Research Database (NHIRD) (2016). https://nhird.nhri.org.tw/date_01.html. Accessed 1 Nov 2016
12. Centers for Disease Control and Prevention: International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (2016). <https://www.cdc.gov/nchs/icd/icd9cm.htm>. Accessed 1 Nov 2016
13. Ekström, J.: The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule debate (2011). <http://statistics.ucla.edu/preprints/uclastat-preprint-2008:40>. Accessed 20 Feb 2016
14. Apache Software Foundation: SparkR 2.1.0 (2016). <http://spark.apache.org/docs/latest/sparkr.html>. Accessed 31 Dec 2016
15. De Nooy, W., Mrvar, A., Batagelj, V.: *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, Cambridge (2011)
16. Tschentlin, B.: *Big Data and Graph Analytics in a Health Care Setting*. Mayo Clinic (2012). www.graphanalysis.org/SC12/03_Tschentlin.pdf. Accessed 1 June 2015

Application of Automated Theorem-Proving to Philosophical Thought: Spinoza's *Ethics*

Maciej Janowicz, Luiza Ochnio, Leszek J. Chmielewski,
and Arkadiusz Orłowski^(✉)

Faculty of Applied Informatics and Mathematics (WZIM),
Warsaw University of Life Sciences (SGGW), ul. Nowoursynowska 159,
02-775 Warsaw, Poland

{maciej_janowicz, luiza_ochnio, arkadiusz_orlowski}@sggw.pl

<http://www.wzim.sggw.pl>

Abstract. We have applied the automatic theorem-prover Prover9 to prove first eleven theorems of Part 1 of Benedict de Spinoza's *Ethics*. We have used a previous formalization of that segment developed by Blum and Malinovich. We have found Prover9 to be very efficient, providing proofs in tens of miliseconds. It appears that the only, but fundamental, limitation for testing philosophical reasoning is related to the difficulty of unique formalization.

Keywords: Automated reasoning · Theorem proving · Spinoza's philosophy

1 Introduction

As one might expect, automated reasoning has found its most important applications in the fields of mathematical logic, foundations of mathematics, and computer science. To convince oneself about that, it is sufficient to browse recent issues of the leading magazine – the Journal of Automated Reasoning. However, one can readily conceive other applications. In fact, automatic theorem-provers can also be used to any field of human thought which admits formalization. By “formalization” we mean here (to avoid lengthy discussion) simply a procedure which results with a set of terms, definitions, and axioms added to the classical first-order logic with the inference rules inherited from the latter.

In particular, it should be possible to apply automatic reasoning programs to prove, or find counterexamples to, statements made by philosophers. This is especially so in the context of what is known as “analytic philosophy” (as contrasted with “continental philosophy”, usually associated with (post-)Kantian, (post-)Hegelian, or phenomenological/existential schools of thought). This is not because the “analytic” way of philosophical considerations is “better” or “more rigorous”, but because it is more oriented towards solving specific problems and closer, it seems, to the way the mathematical logic works.

We would like to argue that the philosophical thought has always been and remains relevant in everyday life as it can generate consistent worldviews and help us to take more rational and more moral decisions about ourselves. We also argue that automatic theorem proving can greatly help us to select a life philosophy which is logically consistent and morally viable. This is because automatic theorem proving can radically amend the time effectiveness and reduces costs in every situation where it can be applied.

As our first attempt to fulfill the program outlined above, we have decided to try and employ a well-known automated theorem-proving program Prover9 [1] written by William McCune to a part of *Ethics*, a fundamental work by Benedict de Spinoza [2]. This has been done for three reasons. Firstly, Spinoza belongs to the venerable grandsires of all the modern thought, invoked both by Hegel and the analytic school. Secondly, his *opus magnum*, the *Ethics*, has the structure very favorable from our point of view: it is made of definitions, axioms, and theorems. Thirdly, the hardest work, namely, the formalization of a segment of *Ethics*, has already been performed several times (see, e.g., [3,4]). Below, we shall use the formalization provided by Blum and Malinovich [4], who have also provided formal (human-made) derivations of the theorems. As far as we are aware, this is the first attempt to apply automated theorem proving to philosophy; in a sense, it constitutes the first real attempt to fulfill Leibniz's dream about such an application.

The main body of this work is organized as follows. In Sect. 2 we describe the Blum-Malinovich formalization of the relevant part of Spinoza's work. In Sect. 3 we list the theorems proved by both Blum and Malinovich and by Prover9 as well as an example of an output file from Prover9. Section 4 contains some concluding remarks.

2 Blum-Malinovich Formalization of Spinoza's *Ethics*

In [4] the terms necessary for the formalization of *Ethics* have been introduced (see Table 1).

The list appears to be rather lengthy, but all the concepts there are necessary to characterize the Spinoza thought. Actually, additional terms, not listed here, would have been introduced were we to check the whole body of *Ethics*. For instance, we have found the relations "is conceived by" and "having more reality than" painfully lacking in the list (Table 1).

What is more, let us provide the list of Spinoza's definitions in the Blum-Malinovich settings. We will use the notation of Prover9, that is, all the entries look almost exactly the same as in the Prover9 input file. "Almost" means that \rightarrow represents the sign " $_$ " followed by " $>$ " and \leftrightarrow by " $<$ " followed by " $_$ " and " $>$ " in the input file. We believe the notation is self-explanatory, except that all the formulae have to be ended with a dot.

1. all x (C(x, x) \leftrightarrow N(x)).
2. all x (Kf(x) \leftrightarrow exists y (K(x, y) & L(y, x) & \neg (x = y))).

Table 1. List of terms in the Blum-Malinovich formalization of *Ethics*.

Term	Meaning
$A(x, y)$	x is an attribute of y
$C(x, y)$	x causes y
$D(x, y)$	x depends on y
$Et(x)$	x is eternal
$E(x, y)$	x is essence of y
$F(x)$	x is finite
$H(x)$	x is absolutely infinite
$I(x, y)$	x is contained in y
$Kf(x)$	x is finite after its kind
$K(x, y)$	x and y are of the same kind
$L(x, y)$	x limits y
$M(x, y)$	x is a mode of y
$N(x)$	x has necessary existence
$P(x, y)$	x is prior to y
$Q(x)$	x is free
$S(x)$	x is a substance
$T(x, y)$	x is the effect of y
$U(x, y)$	x knows y
$W(x, y, z)$	x and y have z in common

3. all x ($S(x) \leftrightarrow I(x, x)$).
4. all x ($S(x) \leftrightarrow D(x, x)$).
5. all x ($S(x) \rightarrow \neg(\text{exists } y (D(x, y) \ \& \ \neg(y = x)))$).
6. all x all y ($A(x, y) \leftrightarrow (S(y) \ \& \ E(x, y))$).
7. all x (exists y ($S(x) \rightarrow A(y, x)$)).
8. all x all y ($M(x, y) \leftrightarrow (S(y) \ \& \ I(x, y) \ \& \ \neg(x = y)) \ \& \ D(x, y)$).
9. all x ($G(x) \rightarrow (S(x) \ \& \ H(x))$).
10. all x ($Q(x) \leftrightarrow (N(x) \ \& \ C(x, x))$).
11. all x ($Et(x) \leftrightarrow N(x)$).

Spinoza’s axioms have been written as follows:

1. all x ($I(x, x) \mid \text{exists } y (\neg(x = y) \ \& \ I(x, y))$).
2. all x all y (($\neg(x = y) \ \& \ \neg D(x, y)$) $\rightarrow D(x, x)$).
3. all x (exists y ($C(x, y) \rightarrow \text{exists } z (T(z, x))$)) $\ \& \$ all x (exists y ($T(x, y) \rightarrow \text{exists } z (C(z, x))$)).
4. all x all y ($C(x, y) \rightarrow \text{all } z (U(z, x) \rightarrow U(z, y))$).
5. all x all y ($\neg(\text{exists } z (W(x, y, z))) \rightarrow ((\text{exists } v (U(v, x) \ \& \ \neg(U(v, y)))) \ \& \ (\text{exists } v (U(v, y) \ \& \ \neg(U(v, x)))) \ \& \ \neg(D(x, y)) \ \& \ \neg(D(y, x)))$).
6. all x ($N(x) \rightarrow \text{all } y (E(y, x) \rightarrow \text{all } z (A(z, y) \rightarrow \text{exists } u (z = u)))$).

One additional axiom schema has been introduced by Spinoza (“A true idea must correspond with its ideate or object”) which has not used here. There have been, however, several concepts which Spinoza most likely considered as too obvious to define or characterize in the axiomatic form. And yet, their explicit introduction is necessary in the formal proofs. That fact have compelled the authors of [4] to introduce the following additional “postulates”:

1. all x all y ($(\neg(x = y) \ \& \ D(x, y)) \rightarrow P(y, x)$).
2. all x ($D(x, x) \leftrightarrow C(x, x)$).
3. all x all y ($(D(x, y) \mid D(y, x)) \leftrightarrow \text{exists } w (W(x, y, w))$).
4. all x all y all u all v ($(E(u, x) \ \& \ E(v, y)) \rightarrow ((x = y) \leftrightarrow (u = v))$).
5. all x ($Q(x) \leftrightarrow (\neg(\text{exists } y (L(y, x))))$).

We have had Prover9 found proofs for the first eight- and the eleventh theorem from *Ethics*:

1. all x all y ($(S(x) \ \& \ M(y, x)) \rightarrow P(x, y)$).
2. all x all y ($((S(x) \ \& \ S(y) \ \& \ \text{exists } z \ \text{exists } v (A(z, x) \ \& \ A(v, y) \ \& \ (z \neq v))) \rightarrow (\neg(\text{exists } w (W(x, y, w))))$).
3. all x all y ($(\neg(\text{exists } z (W(x, y, z)))) \rightarrow ((\neg C(x, y)) \ \& \ (\neg C(y, x)))$).
4. all x all y ($((S(x) \ \& \ S(y) \ \& \ (x \neq y)) \rightarrow ((\text{exists } z (A(z, x) \ \& \ \neg A(z, y))) \mid (\text{exists } v (M(v, x) \ \& \ \neg M(v, y))))$).
5. all x all y ($((S(x) \ \& \ S(y) \ \& \ (\neg(x=y))) \rightarrow (\neg(\text{exists } z (W(x, y, z))))$).
6. all x all y ($((S(x) \ \& \ S(y) \ \& \ (\neg(x=y))) \rightarrow (\neg(C(x, y) \mid C(y, x)))$).
7. all x all y ($S(x) \rightarrow N(x)$).
8. all x ($S(x) \rightarrow ((\neg Kf(x)) \ \& \ (\neg(\text{exists } y (L(y, x)))))$).
9. all x ($G(x) \rightarrow N(x)$).

As for the Theorem 9, “The more reality or being a thing has the greater the number of its attributes.”, we believe it is actually a definition of the relation $R(x, y)$, “ x have more reality or being than y ”. Prover9 has demonstrated, however, a proposition which remotely resembles Theorem 9 without using a new definition:

- all x all y ($((\text{all } z (E(z, x) \rightarrow E(z, y))) \rightarrow (\text{all } v (A(v, x) \rightarrow A(v, y))))$).

Similarly, it is difficult to convincingly formalize Theorem 10: “Each particular attribute of the one substance must be conceived through itself.”

We have played with several formulae which, again, very remotely resemble the meaning of the above proposition. Among other things, Prover9 has proved the following:

- all x all y ($((S(y) \ \& \ A(x, y)) \rightarrow (I(x, x)))$).

The difficulties in formalization of Theorems 9 and 10 have probably been the reason of their omission in [4].

3 Example of the Output from Prover9

After some testing, Prover9 has been used by us as a black box. We knew in advance that all the propositions presented to it have already been proved in [4].

The output files from Prover9 are rather lengthy as they provide quite detailed information about the proof search. In this section we just show the text of final part of the output file containing proof of the famous Theorem 11.

We do this mainly to demonstrate beyond any doubt that the Prover9 has really performed the required action!

```

===== PROOF =====
% Proof 1 at 0.02 (+ 0.00) seconds.
% Length of proof is 15.
% Level of proof is 4.
% Maximum clause weight is 6.000.
% Given clauses 45.

1 (all x (C(x,x) ↔ N(x))) # label(non_clause). [assumption].
4 (all x (S(x) ↔ D(x,x))) # label(non_clause). [assumption].
9 (all x (G(x) → S(x) & H(x))) # label(non_clause). [assumption].
20 (all x (D(x,x) ↔ C(x,x))) # label(non_clause). [assumption].
24 (all x (G(x) → N(x))) # label(non_clause) # label(goal). [goal].
31 ¬S(x) | D(x,x). [clausify(4)].
39 ¬G(x) | S(x). [clausify(9)].
40 G(c1). [deny(24)].
43 ¬G(x) | D(x,x). [resolve(39,b,31,a)].
64 ¬C(x,x) | N(x). [clausify(1)].
77 ¬D(x,x) | C(x,x). [clausify(20)].
81 ¬N(c1). [deny(24)].
108 D(c1,c1). [resolve(43,a,40,a)].
128 ¬C(c1,c1). [ur(64,b,81,a)].
135 $F. [resolve(108,a,77,a),unit_del(a,128)].

===== end of proof =====

===== STATISTICS =====

Given=45. Generated=95. Kept=70. proofs=1.
Usable=45. Sos=25. Demods=0. Limbo=0, Disabled=99. Hints=0.
Kept_by_rule=0, Deleted_by_rule=0.
Forward_subsumed=24. Back_subsumed=0.
Sos_limit_deleted=0. Sos_displaced=0. Sos_removed=0.
New_demodulators=0 (0 lex), Back_demodulated=0. Back_unit_deleted=0.
Demod_attempts=0. Demod_rewrites=0.
Res_instance_prunes=0. Para_instance_prunes=0. Basic_paramod_prunes=0.
Nonunit_fsub_feature_tests=24. Nonunit_bsub_feature_tests=92.

```

Megabytes=0.28.

User_CPU=0.02, System_CPU=0.00, Wall_clock=0.

===== end of statistics =====

===== end of search =====

THEOREM PROVED

4 Concluding Remarks

To our best knowledge, this work describes the first attempt to employ an automated-reasoning computer program to the philosophical thought. We have not (or, better to say, not yet) attempted to write our own software dedicated to that purpose. This is because the major difficulty in such applications is the proper formalization of a philosophical theory. Indeed, every formalization is at the same time an interpretation. And even in our case, Spinoza could have had complained that Blum-Malinovich formalization does not fully reflect his thought, regardless of the differences of historical epoch and in the language. For instance, one could compare the above symbolic Definition 1 with its original verbal counterpart: “By that which is self-caused, I mean that of which the essence involves existence, or that which the nature is only conceivable as existent”. Clearly, the formalization here does not only interpret the original in some way, but also trivializes it. The last statement can, of course, be used against any attempt to seek for computer-generated proofs. But we would like to conclude with a more positive statement: there are many levels of possible formalizations, and, by iterating, we can eventually approach the philosopher’s intentions arbitrarily close. And the endeavor itself is valuable as some thinkers sometimes have some important things to say to us.

Let us finally remark that automated theorem proving has recently found application in the analysis of optical system [5]. That interesting development possibly heralds new ways in which automated provers will be applied in both theoretical thought and technology.

We would like to dedicate this work to the memory of late Professor William McCune.

References

1. Prover9. <http://www.cs.unm.edu/~mccune/prover9/manual/2009-11A/>. Accessed 20 Dec 2016
2. Spinoza, B., Ethics, E.: Translation from Latin. In: Elwes, R.H.M. (ed.) *The Chief Works of Benedict De Spinoza*. Dover Publications, New York (1951)
3. Jarrett, C.: The logical structure of Spinoza’s Ethics. Part I Synth. **37**, 15–65 (1978)
4. Blum, A., Malinovich, S.: A formalization of a segment I of Spinoza’s Ethics, *Metalogicon*. **VI**, 1 (1993)

5. Hasan, O., Khan Afshar, S., Tahar, S.: Formal analysis of optical waveguides in HOL. In: Berghofer, S., Nipkow, T., Urban, C., Wenzel, M. (eds.) TPHOLs 2009. LNCS, vol. 5674, pp. 228–243. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-03359-9_17](https://doi.org/10.1007/978-3-642-03359-9_17)

Improving the B+-Tree Construction for Transaction Log Data in Bank System Using Hadoop

Cong Viet-Ngu Huynh^{1,2}, Jongmin Kim², and Jun-Ho Huh^{3(✉)}

¹ Troy University,

Ho Chi Minh Campus, Socialist Republic of Vietnam

hcvngu@pukyong.ac.kr

² Department of IT Convergence Engineering, Pukyong National University at Daeyeon,
Busan, Republic of Korea

{hcvngu, jmkim}@pukyong.ac.kr

³ Department of Software, Catholic University of Pusan, Busan, Republic of Korea
72networks@cup.ac.kr

Abstract. In Socialist Republic of Vietnam, applying the Big data to process any kind of data is still a challenge, especially in the banking sector. Until now, there is only one bank applied Big data to develop a data warehouse system has focused, consistent, can provide invaluable support to executives make immediate decisions, as well as planning long-term strategies, however, it still not able to solve any specific problem. Nowadays, from the fact large amounts of traditional data are still increasing significantly, if B-tree is considered as the standard data structure that manage and organize this kind of data, B+-tree is the most well-known variation of B-tree that is very suitable for applying bulk loading technique in case of data is available. However, it usually takes a lot of time to construct a B+-tree for a huge volume of data. In this paper, we propose a parallel B+-Tree construction scheme based on a Hadoop framework for Transaction log data. The proposed scheme divides the data into partitions, builds local B+-trees in parallel, and merges them to construct a B+-tree that covers the whole data set. While generating the partitions, it considers the data distribution so that each partitions have nearly equal amounts of data. Therefore the proposed scheme gives an efficient index structure while reducing the construction time.

Keywords: B-tree · B+-tree · Hadoop · Map-Reduce · Big data in Vietnam

1 Introduction

Nowadays, from the fact large amounts of traditional data are still increasing significantly, the B-tree is considered as an optimal index mechanism that will help retrieve data quickly. A B-tree [1] is a self-balancing tree data structure that keeps data sorted and allows searches, sequential access, insertions, and deletions in logarithmic time. The B-tree is a generalization of a binary search tree in that a node can have more than two children. Unlike self-balancing binary search trees, the B-tree is optimized for systems that read and write large blocks of data. If B-tree is a good example of a data structure for external memory, B+-tree [2] is the most well-known variation of the B-tree and it

is very suitable for applying bulk loading technique in case of bank system, where almost the history transactions are sorted and available. The B+-tree example is shown in Fig. 1, where branching factor equal 3.

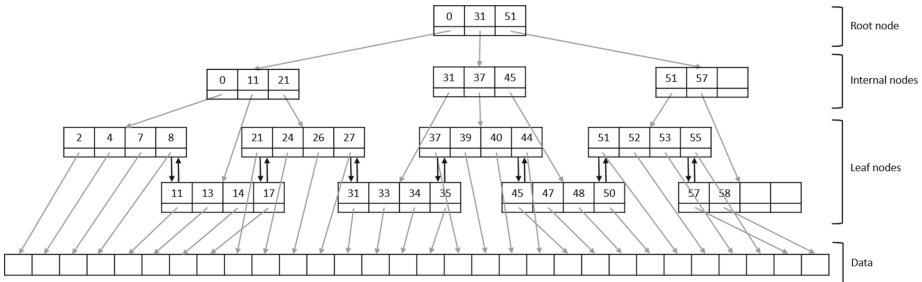


Fig. 1. Example of B+-tree with branching factor equal 3.

A B+-tree is an index structure that enables fast accesses to one dimensional data. However, it usually takes a lot of time to construct a B+-tree for a huge volume of sorted data. In this paper, we propose a parallel B+-Tree construction scheme based on a Hadoop framework. The proposed scheme divides the data into partitions, builds local B-trees in parallel using bulk-loading technique that can maximize the number of leaf node, so the height of B+-tree is minimized, and merges them to construct a B+-tree that covers a whole data set. While generating the partitions, it considers the data distribution so that each partition has nearly equal amounts of data. Therefore the proposed scheme gives an efficient index structure while reducing the construction time.

2 Related Research

2.1 Hadoop-MapReduce

A few years ago, to store or process data, most enterprises had a super computer to perform this task. Here data can be stored in an RDBMS such as Oracle Database, MS SQL Server or DB2. But with this approach, when it has to handle huge amounts of data, it faces many difficulties in processing such data through a traditional database server [2–6]. Facing those difficulties, in 2005, an Open Source Project called Hadoop was released. In order to handle a huge amounts of data, Hadoop runs all applications using the MapReduce algorithm, where the data is processed in the parallel way on different nodes. MapReduce is a programming model suited for parallel computation, it handles parallelism, fault tolerance and other level issues [5–10].

Furthermore, MapReduce consists of both a map and reduce function which are user-defined. The input data format is specified by the user and the output is a set of <key,value> pairs. As shown in Fig. 2, the mapper applies user-defined logic on every input key/value pair (k1,v1) and transforms it into a list of intermediate key/value pairs (k2,v2). Then the reducer will apply user-defined logic to all intermediate values (v2)

associated with the same k_2 and produces a list of final output key/value pair (k_3, v_3) . The data flow of the MapReduce framework is illustrated in Fig. 3.

	Input	Output
Map	$\langle k_1, v_1 \rangle$	$list(k_2, v_2)$
Reduce	$\langle k_2, list(v_2) \rangle$	$list(k_3, v_3)$

Fig. 2. Input and output in MapReduce

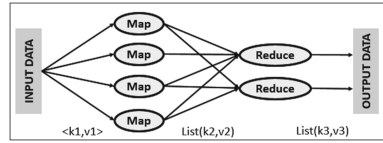


Fig. 3. MapReduce framework

2.2 B+-Tree Bulk Loading

As we discuss before, when data is available, it is very suitable to apply bulk loading technique for B+-tree. For bulk-loading technique of B+-tree, there are two primary way to implement: Top-down building and Bottom up building. As their name, in Top-down insertion, the data is sorted using External MergeSort, and inserted in a top down manner into the index.

In this approach, the nodes are always half full, as they get filled in the sorted order and split, and none of the new entries go to the old nodes. But this approach is fast in terms of creation time.

In Bottom up building, first the leaf node layer is created by populating nodes completely and connecting till all the entries are finished. Then a recursive function builds the internal node. It starts from a node which is designated as the root, and then recursively builds the subtrees corresponding to all it's children. The height at a point in the recursion is maintained. The base case is reached when the recursion reaches the node just above the leaf node level. There, the node is populated, and the pointed to the next leaf node is passed above. Then the recursion builds the rest of the internal nodes in the similar manner, while handling the keys and the pointers to the leaf nodes as the recursion passes up. This approach is very efficient in terms of space efficiency of the B+-tree, as almost all the nodes are completely filled. Hence, the height is smaller, and the accesses are faster.

3 Our Parallel B+-Tree Construction Scheme

Before we describe how to build B+-tree indexing on a MapReduce framework, there are some notations that we will use in the rest of the paper which are as follows.

Our parallel B+-tree construction has three phases, and two of them are implemented in the Hadoop environment using the MapReduce model:

- Partitioning phase – Partition boundary of big data set.
- B-tree construction – small B+-trees are built concurrently with bulk loading technique in bottom up fashion.
- B-tree consolidation – merge small B+-trees into the final B+-tree.

Firstly, let us start our description by defining the problem. The data set that we are using is a large CVS file where each line represents one transaction, it contains <o.d, o.t> where o.d is the time of the transaction, it is used as a unique identifier and o.t is the transaction information such as sender ID, transaction type, money, etc., as show in Table 1. All items in file are sorted by time.

Table 1. Transaction log data.

Date	Sender ID	Transaction type	Money	Receiver ID	Sender after transaction amount	Receiver after transaction amount
01/03/2015	15	Transfer	500	16	1000	3255
01/14/2015	52	Deposit	1200	52	1500	1500
02/05/2015	125	Transfer	250	168	2503	4456
04/22/2015	32	Transfer	200	88	1112	2284

Our scheme consists of three phases executed in sequence, as shown in Fig. 4. First, we determine the partition boundary based on the transaction time. Next, data is partitioned into the corresponding partition which creates small B+-Trees.

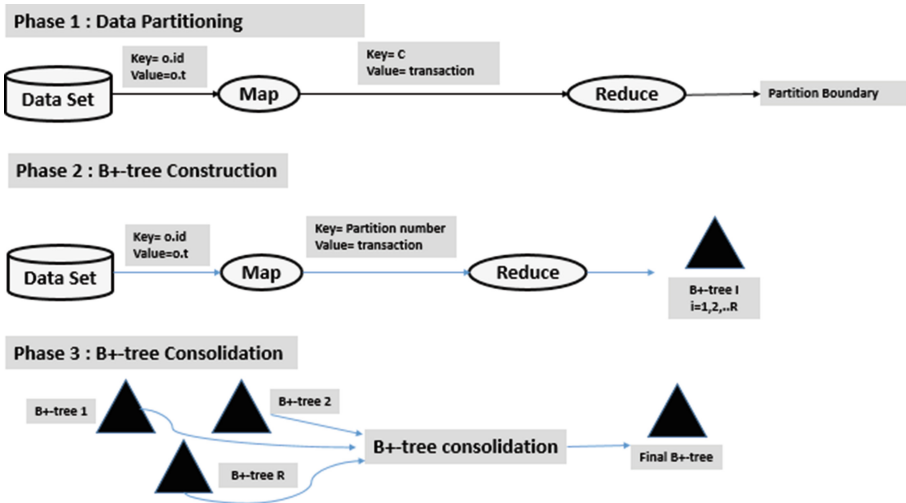


Fig. 4. Phases involved in building a B+-tree in MapReduce.

Finally, the small B+-Trees are merged into the final B+-Tree. The first two phases are executed in MapReduce, while the last phase does not require high computation, so it is executed outside of the cluster.

3.1 Data Partitioning

In this phase, to determine the partition boundary for data, our idea is to read random objects from the input file via data sampling with a default ratio of input data. The MapReduce algorithm runs M Mappers that take sample objects from the input file, then in each Mapper, it reads transaction, then takes the date of each item. Then a single Reducer, firstly, it sorts the list from Mapper' output, then it determines a new list K of $R-1$ partition boundary that split the list of sample into R nearly equal-size partition.

The specific MapReduce key/value input pairs are presented in Table 2. Mappers read the default ratio of data from input file and take the date of each item. The intermediate key is a constant that helps to send all the Mappers' outputs to a single Reducer. Then Reducer determines the list splitting point K base on the date of transaction.

Table 2. Map and Reduce inputs/outputs for data partitioning.

Function	Input	Output (key, value)
Map	(o.d, o.t)	(C, transaction date)
Reduce	(C, list (transaction date))	K

3.2 B+-Tree Construction

In this phase, an individual B+-tree is built concurrently. Mappers partition the input file into R groups, then every partition is executed by a different Reducer, each Reducer, B+-tree is built independently using bulk-loading technique with bottom up fashion. The output of every Reducer is a root node of their constructed B+-Trees, as shown in Table 3.

Table 3. Map and Reduce inputs/outputs for B+-tree construction.

Function	Input	Output (key, value)
Map	(o.d, o.t)	(Partition number, transaction)
Reduce	(Partition number, list (transactions))	B+-tree, root

In traditional B+-tree construction, the item is inserted sequentially as they arrive. But with this method, it will take a long time for B+-tree construction with a huge volume of sorted data. If the item is inserted one by one, the nodes that are not affected by the insertion procedure have a minimum order of keys, which causes the height of the tree increase. To apply the bulk-loading to build the B+-tree, we believe the height of B+-tree can be minimized, so the search performance also will be better.

3.3 B+-Tree Consolidation

In this phase, we are going to combine the R individual B+-tree, built in the second phase, under a single root. Because it's not computationally intensive and the logic to run this phase is fairly simple, it is executed outside the cluster as shown in Fig. 5.



Fig. 5. B+-tree consolidation.

4 Conclusion and Future Work

In this paper, we proposed a scheme for parallel B+-tree construction for Transaction log data on Hadoop environment using the MapReduce model. To enhance quality of B+-tree construction, we also use the bulk-loading technique in the second phase. With our scheme, we hope to contribute to the improvement of the data processing in general, and in particular, in Bank system.

Our scheme has three phases, in which, the first two phases are executed in parallel with MapReduce model, while the last phase is executed outside the cluster because it does not require the high computational. Nowadays, with the amount of data is increasing significantly, the availability of Big Data and commodity hardware, has opened many opportunities for analyzing astonishing data sets quickly and cost-effectively for the first time in history.

References

1. Douglas, C.: The ubiquitous B-Tree. *Comput. Surv.* **11**(2), 121–137 (1979). ACM
2. Cong, V.-N.H., et al.: Improving the quality of an R-tree using the Map-Reduce framework. In: CUTE 2016, LNEE. Springer (2017, accepted)
3. Huh, J.-H., Je, S.-M., Seo, K.: Design and configuration of avoidance technique for worst situation in zigbee communications using OPNET. In: Kim, K., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016*. LNEE, vol. 376, pp. 331–336. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_33](https://doi.org/10.1007/978-981-10-0557-2_33)
4. Huh, J.-H., Otgonchimeg, S., Seo, K.: Advanced metering infrastructure design and test bed experiment using intelligent agents: focusing on the PLC network base technology for Smart Grid system. *J. Supercomput.* **72**(5), 1862–1877 (2016). Springer
5. Kajioka, S., Mori, T., Uchiya, T., Takumi, I., Matsuo, H.: Experiment of indoor position presumption based on RSSI of Bluetooth LE beacon. In: *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*, pp. 337–339. IEEE (2014)

6. Huh, J.-H., Je, S.-M., Seo, K.: Communications-based technology for smart grid test bed using OPNET simulations. In: Kim, K., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016*. LNEE, vol. 376, pp. 227–233. Springer, Heidelberg (2016). doi: [10.1007/978-981-10-0557-2_23](https://doi.org/10.1007/978-981-10-0557-2_23)
7. Birkenmeier, G.F., Park, J.-K., Rizvi, S.T.: Principally quasi-Baer ring hulls. In: Van Huynh, D., López-Permouth, S.R. (eds.) *Advances in Ring Theory*. Trends in Mathematics, pp. 47–61. Springer, Basel (2010)
8. Birkenmeier, G.F., Park, J.-K., Rizvi, S.T.: Ring hulls of semiprime homomorphic images. In: Brzeziński, T., Gómez Pardo, J.L., Shestakov, I., Smith, P.F. (eds.) *Modules and Comodules*. Trends in Mathematics, pp. 101–111. Springer, Basel (2008)
9. Apache Hadoop: <http://hadoop.apache.org>
10. Prasad, S.K., McDermott, M., He, X.: GPGPU-based parallel R-tree construction and querying. In: *2015 IEEE International Conference (IPDPSW)*, pp. 619–627 (2015)
11. Li, Z., Pan, H., Liu, W., Xu, F., Cao, Z., Xiong, G.: A network attack forensic platform against HTTP evasive behavior. *J. Supercomput.* 1–12 (2016). Springer, USA
12. Sung, Y., Jeong, Y.-S., Park, J.-H.: Beacon-based active media control interface in indoor ubiquitous computing environment. *Clust. Comput.* **19**(1), 547–556 (2016). Springer, USA
13. Cheong, H., Eun, J., Kim, H., Kim, K.: Belief propagation decoding assisted on-the-fly Gaussian elimination for short LT codes. *Clust. Comput.* **19**(1), 309–314 (2016). Springer, USA

Calculate Deep Convolution Neural Network on Cell Unit

Haofang Lu¹✉, Ying Zhou², and Zi-Ke Zhang¹

¹ Alibaba Research Center for Complexity Sciences, Hangzhou, China
sd185115@163.com, zhangzike@gmail.com

² DataCastle, Chengdu, China
zhouying@datacastle.cn

Abstract. We introduce CACU, a new deep learning algorithm framework for CNN which using binary method to reduce the consumptions in convolution calculating. CACU introduces bit data flow to fit the CPU platform. Using binary-weights and xnor methods to speed-up the convolution's computation on CPU device. GPU is also supported in CACU. CUDA version is implemented for accelerating large scale models' training and inference. CACU is a C++ library with no dependencies except Boost. CACU is not only developed for the CNN's usage in application, it's helpful for researchers to take an inner investigation of bit method in CNN. It's a fully open-source platform which is available on GitHub.

Keywords: CNN · Framework · Binary network · XNOR

1 Introduction

The scale and complexity of deep learning (DL) algorithms are becoming increasingly large. Traditional CPU cannot return the convolution result with float in real-time which is frequent appearances in CNN. So high-powered parallel calculation device GPU is introduced into the deep learning algorithms frameworks [1, 2]. Undeniably GPU has largely boosted deep learning's development, it'll play a more important role in AI devices sooner or later. Although there are still wide gaps to hinder us using CNN on smaller devices like smart phones. High performed GPU haven't been extensively used on smart phones, besides the great power consumption for running Deep Neural Network (DNN) models. So how to setup DNN model that makes CPU running more fast is a significant way to promote the development of deep learning.

To realize DNN model running on smart device, there are several important facts need to be settled: (a) there must be a clean and highly portable runtime core that fits the background model training and device sample inference. (b) Fast algorithm level bit computation supported based on the CPU core architecture. (c) Small complied software size.

Squeeze 32-bit float into a single bit width which is easily calculated on CPU is a way. But till now almost all popular DL framework do not supply bit computation on dataflow. And most of them supply different interfaces for multi-language (e.g. C++/PYTHON/MATLAB/LUA). That makes the frameworks easily used for different

domain users (Caffe/Mxnet), but difficult for users to transplant the work to heterogeneous architectures. These frameworks are mainly used for large-scale distributed clusters or background-services.

Recently more and more works prove that bit method could take a well performance [3, 4] on CNN. Bit method framework is extremely needed. To address the problem we present CACU, a fully open-source framework that contains 32-bit and single bit data bottoms for deep learning architectures.

CACU is written by C++ with CUDA for GPU. It's very efficient for model training on GPU device and easily model transferring to lighter computation architecture unit. For the feature, CACU is more likely used as a dynamic library which is included by DL module in applications or other logical computation units. In research aspect, CACU provides an real bit computation platform for those who want to get a better understanding or promote works on binary methods.

In CACU, deep learning engineers and researchers can easily create models by C++ code. Training the model on GPU then include the well trained model to the application scene. CACU could speed-up the computation in convolution inference process which cost 90% time consumption. That makes the CNN model running on mobile devices becoming true.

While CACU is first designed for CNN models running on smart devices. We hope to see more bit level deep learning models appearance. CACU is fully open-source. We welcome open-source contributions at: <https://github.com/luhaofang/CACU>.

2 Features

CACU provides a complete toolkits for user who wants to train 32-bit or single bit models. For engineering users, CACU will be an even more suitable choice for its easily use.

2.1 Bit Computation High-Performance

CACU supports a true bit data bottom to maintain data which is fed to bit computation. So in real application scenes, CPU will perform well. In our experiments, CACU makes ~X10 speed-up on CPU architecture by XNOR methods [4].

2.2 Light Equipment Application Scenes Design

No-dependencies is a really important feature for a portable deep learning frameworks. CACU has no-dependencies except Boost. `Dynamic_bitset` is a bit container for C++, we use the library for bit unit computation.

CACU is a well portable framework. It's code programming oriented, there is no need to install CACU. We supply a header file for those who want to use CACU in their applications. After include the header file, all modules in CACU could be used. You can create a model on the GPU devices, training the model, then transfer it to your application. All these will be easily done on CACU.

2.3 Modularity

CACU is established by various of layer modules like Caffe. In CACU, it's easy to add a new layer module. There is no need to register new a layer, rather creating new layer header file, then implement layer code. Actually we just need to implement forward and backward functions at most of times.

Pre-trained model is provided in CACU. We provide well-trained models for different application usage. Fine-tuning models for new applications is supported in CACU.

3 Architecture

3.1 Data Bottom

In CACU, there are two types of data interfaces for DNN model. We use blob to load full precision data like Caffe, bin blob is a new data container in CACU for bit data flow technically.

Table 1. Comparison of popular deep learning frameworks. Bindings have an officially supported library interface for feature extraction, training, while bit Computation is the real bit data bottoms type. No dependencies is accompanied with bit computation that increase the framework's portability.

System	Banding language	Devices	Modeling	Bit computation	No dependencies	Pre-trained model
CACU	-	GPU/CPU/ mobile	C++ code	√	√	√
Caffe	Python, Matlab	GPU/CPU	CONFIG File			√
Torch7	-	GPU/CPU	Lua code			
Theano	-	GPU/CPU	Python code			
Tiny-Dnn	-	CPU/ mobile	C++ code		√	√

Blob type is similar with Blobs in Caffe. In CPU/GPU, it maintains an array of float data. While there are still something different. CACU does not load image data from channel by channel, but pixel by pixel.

Bin blob is an important container in CACU. It created with a dynamic_bitset data type. Bit flip op is usually needed in the BNN [3] algorithm. We keep the real bit data until they should be serialized. It's necessary for bit data transform to unsigned integer data. In fact that's why we confirm bit method could make algorithm speed-up on CPU. Nowadays, 32-bit float computation is the standard type in DNN algorithm, the time consumption between sum op and multi op in 32-bit float, really have no differences on most of CPU architectures. In XNOR-NET convolution computation with 32-bit float is transform by xnor computation and bit count, CACU use a 32-bit unsigned integer to

compute xnor. Theoretically, convolution computation speed-up by ~X32. Bin blob doesn't pass backward binary gradient, both blob and bin blob contains full precision gradient (Table 1).

Model squeeze is also achieved by binary weights. CACU provides 32-bit float weights on most of layers. Binary convolution layer's weights are stored in unsigned integer that is size of 32-bit for every block.

3.2 Memerys

CACU is written by C++, so we organize data by vector form STL on CPU, the similar form on GPU. In CACU each layer is defined by code and contains two lists of probe, one for maintaining the blob data spaces that passing forward the tensor data and taking backward gradient data. Another for maintaining bin blob bit data spaces. These blobs/bin blobs used in algorithm are manually added by the layer's creator. CACU provides a data container for intermediate data that created by the algorithm in each layer.

3.3 Layer Algorithms

Layers in DNN carry most of the logical computation. CACU provides foremost layers in CNN models: batch normalization layer [6], convolution layer, element wise add layer, max pooling layer, average pooling layer, inner product layer, softmax, cross entropy loss, etc. Additionally, CACU contains a complete interface for layer creators. New layers are easily created.

DNN layer's algorithm is always presented in forward and backward processes. CACU users just need to focus on the forward and backward functions.

3.4 XNOR Method

An typical form of NN structure like (1). Theoretically, we could activate the features by tanh. Furthermore, if we shape the activation function to hard-tanh, all 32-bit float features are binary to -1 or +1 that we expect. In BNN forward propagation, we change the model into (2), compare with the backward propagation (1). So every sign-function creates an error.

$$a_l = \tanh\left(\text{Batch_normalize}\left(\sum_i^{k_l}(w_l \cdot a_{l-1})\right)\right) \tag{1}$$

$$a_l = \text{hard_tanh}\left(\text{Batch_normalize}\left(\sum_i^{k_l}(w_l \cdot a_{l-1})\right)\right) \tag{2}$$

We consider that the neural network will output most of nearly -1 or +1 features if it has been well trained. Here we present an output feature distribution of the every convolution layer after activated by tanh in Fig. 1. It denotes that the estimated error is becoming larger when the net going deeper.

$$w \cdot x = N - 2 \times \text{bitcount}(\text{xnor}(x_b, w_b)), x_b, w_b \in \{0, 1\} \tag{3}$$

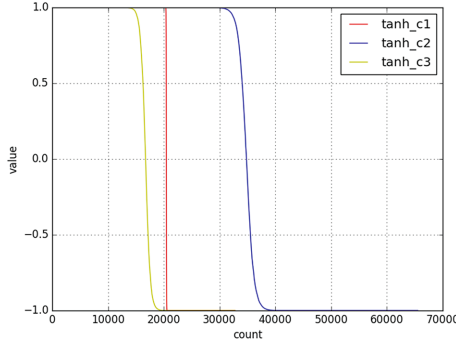


Fig. 1. The trained activation quantization for full precision cifar10 model. tanh_c1 denote the first activation.

To minimize the error accompanied by the formal change, Rastegari et al. [4] introduces hyper parameter alpha for weights estimation. 1/0 result transform into +1/-1, by function (3).

3.5 Network

Each network is created before the data spaces are allocated. We setup the architecture of the network in C++ code, then allocate the model spaces. Under this mechanism, CACU model can be transferred to different platforms, after compiling the code with the corresponding architecture.

3.6 Training and Testing

CACU uses standard SGD with momentum for network training. When we need to training the network, several data preprocesses are required. CACU supplies the pixel and channel mean center tools.

If we need to train a new model, setup the network first, create a new data blob for network to fetch the training data, turn network phrase to ‘train’, and get the training model after every snapshot steps. Testing the model is similar with training steps.

In the repository, we supplies two simple examples, one for cifar-10 [9], another for MNIST [10]. Figure 2 shows the models’ training performances and Fig. 3 shows the typical cifar10 network models.

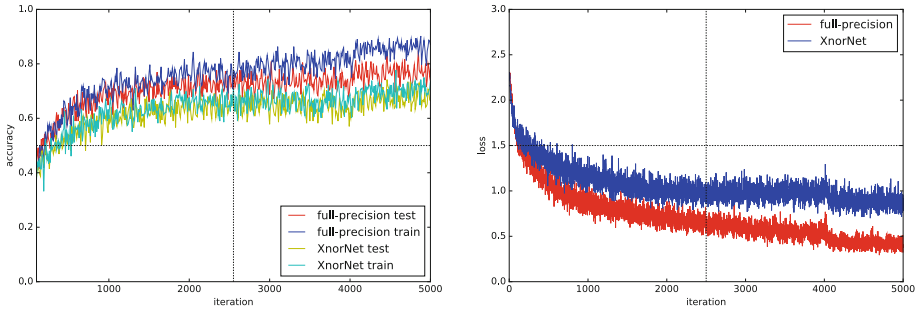


Fig. 2. The training performances for cifar10 model. (a) Demonstrate the comparison of accuracy between 32-bit full precision model and xnor model. (b) Is the loss curves for xnor model and full precision model.

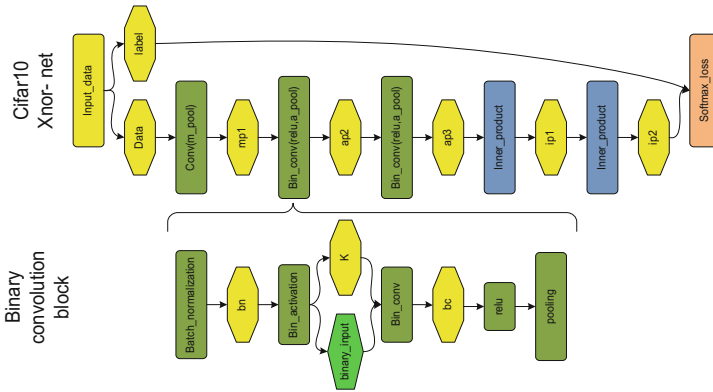


Fig. 3. The demo models for cifar10 which are presented in CACU. We change the C-B-A-P block into B-A-C-P xnor model. C is represent to convolution layer, B is batch normalization layer, A is activation layer and P demonstrates pooling layer. First convolution layer is corresponding to full precision type [3–5].

4 Evaluation

Recently, more and more works on deep learning are carried out for reducing the computational complexity of the network. There are two aspects reasons: (a) Network model is complex in computation. (b) The extension of a, we haven't figured out why does a complex model like DNN works well in detailed mechanism. But we see that BNN model is performed nearly well as the full-precision model. Maybe that's a real question need to be considered. Under this background we present CACU for researchers.

4.1 True Bit Accelerate

CACU is a new and light framework which supports real bit computation for deep learning algorithms, CACU achieve $\sim X10$ speed-up for CNN models on CPU. CACU provides a real bit platform for the people who want to have farther insight in this direction.

4.2 Nice Light Equip. Portability

CACU is written by clean C++ code. The main framework is just 2 MB after compiled. For CNN models, CACU reduce $\sim X32$ of models size. It's easy to transfer CACU and models to light equipments.

5 Conclusion

CACU provides real bit computation for layer algorithms. It's high efficient computation on CPU for convolution algorithm makes the large scale CNN model running on light equipment becoming true.

DNN model is becoming the tendency of the artificial intelligence. But with the great progress carried out on the models' precision, the model becoming more complex. Recently complexity is beginning to be valued by the academic community gradually. We hope CACU could help those who want to create more bit methods on deep learning algorithms. Clearly it has not been fully completed yet, we'll keep it on the progress.

Acknowledgments. This work was partially supported by Zhejiang Provincial Natural Science Foundation of China (Grant No. LY14A050001) and Natural Science Foundation of China (Grant No. 61673151).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
2. Jia, Y., Shelhamer, E., Donahue, J., et al.: Caffe: convolutional architecture for fast feature embedding (2014). arXiv preprint: [arXiv:1408.5093](https://arxiv.org/abs/1408.5093)
3. Courbariaux, M., Bengio, Y.: BinaryNet: training deep neural networks with weights and activations constrained to +1 or -1 (2016). arXiv preprint: [arXiv:1602.02830](https://arxiv.org/abs/1602.02830)
4. Rastegari, M., Ordonez, V., Redmon, J., et al.: XNOR-Net: ImageNet classification using binary convolutional neural networks (2016). arXiv preprint: [arXiv:1603.05279](https://arxiv.org/abs/1603.05279)
5. Rastegari, M. (2016). <https://github.com/allenai/XNOR-Net>
6. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015). arXiv preprint: [arXiv:1502.03167](https://arxiv.org/abs/1502.03167)
7. Courbariaux, M. (2016). <https://github.com/MatthieuCourbariaux/BinaryNet>
8. Tiny-Dnn Developers (2016). <https://github.com/tiny-dnn/tiny-dnn>
9. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
10. LeCun, Y., Cortes, C., Burges, C.J.C.: The MNIST database of handwritten digits (1998)

Stepwise Structure Learning Using Probabilistic Pruning for Bayesian Networks: Improving Efficiency and Comparing Characteristics

Godai Azuma¹(✉), Daisuke Kitakoshi², and Masato Suzuki²

¹ Department of Communication Engineering and Informatics,
Faculty of Informatics and Engineering, The University of Electro-Communications,
Chofu, Tokyo 182-8585, Japan
god.ai@uec.ac.jp

² Department of Computer Science, National Institute of Technology, Tokyo College,
Hachioji, Tokyo 193-0997, Japan
{kitakosi,suz}@tokyo-ct.ac.jp

Abstract. This paper evaluates a structure learning method for Bayesian networks called Stepwise Structure Learning with Probabilistic pruning (SSL-Pro). Probabilistic pruning allows this method to obtain appropriate network structures while reducing computational time for structure learning. Computer experiments were conducted to investigate the characteristics of the SSL-Pro. Results showed that the SSL-Pro generally provided favorable performance, and revealed several parameter-setting guidelines to ensure reasonable learning.

1 Introduction

Stochastic models have received much attention as a leading tool for efficiently dealing with various kinds of data. There have been many reports of research efforts to apply the Bayesian Network (BN), a stochastic model visually reflecting the features of data. A large number of structure learning [1, 2] and probabilistic inference algorithms [3] have been proposed for BNs. Our research group has proposed a structure learning method called the Stepwise Structure Learning (SSL), which can obtain an appropriate network structure in a relatively short time based on multivariate sample data. Furthermore, probabilistic pruning has been introduced to the SSL in order to further decrease learning time, after which its characteristics were evaluated.

This paper focuses on the parameters that trigger the probabilistic pruning in the improved version of SSL (SSL-Pro), and aims to empirically discuss desirable settings for the SSL-Pro to perform “high-speed structure learning,” while maintaining the validity of the obtained network. We verified the adequacy of obtained network structures by comparing with the structures obtained by conventional structure learning methods.

2 Preparation

This section describes Bayesian networks [4] and some typical structure learning algorithms, including Stepwise Structure Learning, the basis of the proposed algorithm.

2.1 Bayesian Networks (BNs)

BNs are well-known knowledge representation models representing stochastic dependences between random variables as directed acyclic graphs based on joint probability distribution $P(X_1, \dots, X_n)$. Each node in a BN corresponds to a random variable. Arcs drawn between nodes denote stochastic dependences between them. The joint probability distribution in an entire BN is given as:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \pi(X_i)), \tag{1}$$

where $\pi(X_i)$ is a set of random variables that are all initial nodes of an incident arc to X_i . The elements of $\pi(X_i)$ are called *parent nodes* of X_i . In addition, probabilities (i.e., $P(X_i | \pi(X_i)), \forall i \in \{1, \dots, n\}$) corresponding to each element on the right-hand side of Eq. (1) are listed as a *Conditional Probability Table (CPT)*. Figure 1 shows an example of a BN with 4 nodes in a Bernoulli distribution, and its CPTs. In this example, random variables take only two values (0 or 1).

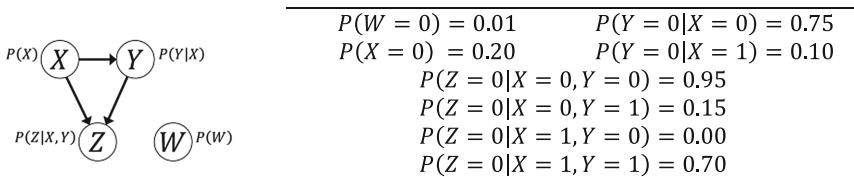


Fig. 1. An example of a Bayesian network and CPT.

2.2 Methods for Learning Structure of BN

A network structure and CPTs of BNs are generally decided based on multivariate data, so that the constructed BN can appropriately express the characteristics of the data. A variety of structure learning methods have been proposed.

Stepwise Structure Learning (SSL) [5], a platform for the proposed method, achieves learning by performing the following steps: (1) all nodes are *divided* into a certain number of small sets (called a *cluster*), and respective structures corresponding to clusters are learned (*inner-cluster learning*); (2) network structure between two clusters (i.e., layout of arcs connecting the nodes in one cluster and the ones in another) chosen at random is learned (*outer-cluster learning*); and (3) step (2) is repeated over and over until the number of clusters equals a desired value. The SSL can learn an appropriate network structure rapidly. Several experiments have revealed guidelines for determining

desirable settings for division and inner-cluster learning for the SSL [6]. Meanwhile, preferable settings for outer-cluster learning have not yet been clarified. Moreover, computational complexity for outer-cluster learning increases with the number of nodes. This leads to an increase in total computational time. To alleviate this problem, a revised version of SSL that introduces an approach to improve computational performance while maintaining adequacy of the BN has been proposed [7].

3 Stepwise Structure Learning with Probabilistic Pruning

To suppress the increase in computational time, an approach called *probabilistic pruning* has been proposed for the SSL. This aims to restrict the search space for BN structure learning. We call the method *Stepwise Structure Learning with Probabilistic Pruning (SSL-Pro)*. Probabilistic pruning is a process that pays attention to two clusters $C_a, C_b (\forall a, b \in \text{index set of clusters})$, which are targets of structure learning in SSL. To estimate significance of performing outer-cluster learning for the clusters, probabilistic pruning employs the following “*meta-dependence*” function $md(C_x, C_y)$, which considers each cluster as a superordinate scale of stochastic dependence between nodes.

$$md(C_x, C_y) = \frac{1}{|C_x||C_y|} \sum_{x \in C_x} \sum_{y \in C_y} I(X; Y), \quad (2)$$

where $I(X; Y)$ denotes mutual information between two nodes. We consider that if meta-dependence between two clusters is low, the possibility of arcs being added between respective nodes in the corresponding clusters would also be low. Outer-cluster learning between two such clusters would not be expected to have a desirable impact on constructing an appropriate structure. Skipping to learn between low meta-dependence clusters contributes to a decrease in computational time for outer-cluster learning while maintaining the adequacy of the structure for the data. An agglomerative hierarchical clustering approach based on a group-average method [8] is employed so as to perform probabilistic pruning effectively. Since Eq. (2) is defined from the viewpoint of distance in the group-average method, this equation can be used for agglomerative hierarchical clustering, as with probabilistic pruning. This approach divided all nodes into each cluster, and repetitively merges the two clusters having the highest meta-dependence into one cluster until the number of clusters reaches the desired value. The SSL-Pro excludes the possibilities between clusters with low meta-dependence by performing probabilistic pruning with agglomerative hierarchical clustering.

3.1 Procedure of SSL-Pro

This section explains the structure learning procedure in the SSL-Pro method. Before explaining this method, we must introduce several parameters. A^t is defined as a set of cluster indexes i corresponding to cluster C_i^t in t -th iteration, and S denotes a set of pairs of mergeable cluster indexes. The structure of a BN with N nodes $X_i (i = 1, \dots, N)$ is learned according to the following steps:

1. $t \leftarrow 0, C_i^0 \leftarrow \{X_i\}, A^0 \leftarrow \{1, \dots, N\}, S \leftarrow \{(i, j) | \forall i, j \in A^0, i \neq j\}$
2. Outer-cluster learning between two clusters $C_j^t, C_k^t ((j, k) \in S)$ is performed after the two clusters with the highest meta-dependence are selected as C_j^t, C_k^t
3. $C_j^{t+1} \leftarrow C_j^t \cup C_k^t, C_m^{t+1} \leftarrow C_m^t (\forall m \in A^t \setminus \{j, k\})$
4. Probabilistic pruning of C_j^{t+1} from C_m^{t+1} is performed based on three conditions as follows ($(\forall m \in A^t \setminus \{j, k\})$):
 - (a) if $(j, m) \in S \wedge (k, m) \in S, S \setminus \{(j, m)\}$ is substituted for S with probability p_1
 - (b) if $(j, m) \in S \vee (k, m) \in S, j^*$ is regarded as the new cluster composed of old clusters j and k , and the following are then applied:
 - (i) $S \setminus \{(j^*, m)\}$ is substituted for S with probability p_2 only if $(j, m) \in S$
 - (ii) $S \cup \{(j^*, m)\}$ is substituted for S with probability $1 - p_2$ only if $(k, m) \in S$
 - (c) otherwise, $S \setminus \{(j, m)\}$ is substituted for S
5. $A^{t+1} \leftarrow A^t \setminus \{k\}, S \leftarrow S \setminus \{(k, n) | \forall n \in A^t\}$
6. Step 2 to step 4 are iterated unless $|A^t| = 1$ or $S = \emptyset$

Pruning probabilities p_1 and p_2 in step 4 above, are defined as parameters computed according to the following equations using the meta-dependence:

$$(a) p_1 = \alpha \frac{md(C_j^{t+1}, C_m^{t+1})}{\bar{I}}, \quad (b) p_2 = \alpha \frac{md(C_z^t, C_m^t)}{md(C_j^t, C_k^t)}, \tag{3}$$

where \bar{I} is the average mutual information of all pairs of nodes, and α is a meta-parameter that suggests the possibility of performing probabilistic pruning ($\alpha \in [0, 1)$). Additionally, z equals j if $(j, m) \in S$, while it equals k if $(k, m) \in S$. Then p_2 is greatly affected only by $md(C_z^t, C_m^t)$, in which C_m^t is a merging candidate of C_z^t (i.e., either j or k , which is not equal to z , is ignored). This is because in Eq. (3)(b), the numerator of the exponent is $md(C_z^t, C_m^t)$, and the denominator of the exponent is fixed to $md(C_j^t, C_k^t)$, corresponding with the *highest* meta-dependence in the t -th iteration.

3.2 Improvement Efficiency of Probabilistic Pruning

When either of C_j^t or $C_k^t (C_j^{t+1} = C_j^t \cup C_k^t)$ can merge with $C_m^{t+1} (= C_m^t), C_m^{t+1}$ may be pruned from merging candidate of C_j^{t+1} with pruning probability p_2 . The exponent of p_2 is less than the exponent of p_1 in most cases, since p_2 uses the highest meta-dependence in each iteration as a calculation criterion. Those of p_2 thus tend to be larger than those of p_1 . This tendency shows that with probabilistic pruning, it is easy to omit merging possibilities between the clusters required to obtain an appropriate structure. In addition, since either $\{C_j^t, C_m^t\}$ or $\{C_k^t, C_m^t\}$ is ignored given the conditions of step 4(b) when calculating

p_2 , learning chances of arcs that should be connected between the ignored cluster and C_m^{t+1} can be lost. As a result, we confirmed problematic situations such that network structures between two clusters having significant stochastic dependences cannot be learned, while improving computational time, in the previous research.

To improve the problems described above, we introduce three kinds of p_2 settings, AVE, RMS, and EP1 (Table 1) to probabilistic pruning, and evaluate the characteristics of structure learning and adequacy of possible structures for each p_2 setting. Since the problem of the calculating equation is out of consideration of either C_j^t or $C_k^t (C_j^{t+1} = C_j^t \cup C_k^t)$, in AVE (Table 1), p_2 takes the average meta-dependence between both of them and C_m^t into account. RMS considers the average meta-dependence similarly to the AVE. However, RMS differs from AVE in the sense that it is expected to restrain excess pruning procedures due to the influence of the clusters with low meta-dependence. In those calculation formulae, the denominator of the exponent in α is also the highest meta-dependence $md(C_j^t, C_k^t)$ at the t -th iteration, and so the minimum value of p_2 is equal to α like the original p_2 (Eq. (3)(b)). We additionally prepare another improvement idea in which that p_2 is equal to p_1 in terms of the calculation method; we call this ‘‘Equivalent to p_1 ’’ (EP1), because p_1 adopting \bar{l} as a comparative basis of meta-dependence worked relatively well in the results of computer experiments [7].

Table 1. New options for parameter p_2 .

$\frac{md(C_j^t, C_m^t) + md(C_k^t, C_m^t)}{2}$	RMS: $\sqrt{\frac{\{md(C_j^t, C_m^t)\}^2 + \{md(C_k^t, C_m^t)\}^2}{2}}$
$md(C_j^t, C_k^t)$	
AVE: $p_2 = \alpha$	$p_2 = \alpha$
$\frac{md(C_j^{t+1}, C_m^{t+1})}{\bar{l}}$	
EP1: $p_2 = \alpha$	

4 Computer Experiments

The experiments investigated each parameter setting by comparing computational times and adequacy of obtained structures. The existing parameter setting (EXT) was used as the basis for comparison. In addition, to reviewing the characteristics of learning using the SSL-Pro itself, the experiments compared the characteristics of SSL-Pro with those of simulated annealing (SA) [9] as a conventional method. Six datasets (data size: 2,000,000) were used for targets of structure learning. These datasets were probabilistically generated from respective BN (Fig. 2). In this figure, the Sachs network, available from the Bayesian Network Repository [10], was used for research purposes. The Alarm network was available from the same source [10]; however, the node ‘‘INTUBATION’’ was removed from the original structure, and the corresponding CPTs were modified. The other network structures in Fig. 2 were manually designed in a manner similar to [10]. Table 2 lists the characteristics

of all the networks in Fig. 2. Both the SSL and SSL-Pro employ brute-force and hill-climbing algorithms for inner-cluster and outer-cluster learning, respectively, because the validity of using these algorithms was confirmed in previous research [6]. The annealing schedule of SA was as follows: $T = 10000\beta^\tau$ (T : temperature, β : rate of temperature reduction, τ : the number of steps). Unless $T < 0.1$, the SA repeats generation, removal, or reversal of an arc while increasing τ . The experiments evaluated the SSL-Pro using each value from 0.00 to 0.30 as the meta-parameter α , and SA using each value from 0.990 to 0.825 as the values of β . To evaluate the network structures, the minimum description length (MDL) criterion [11] was employed for each learning method.

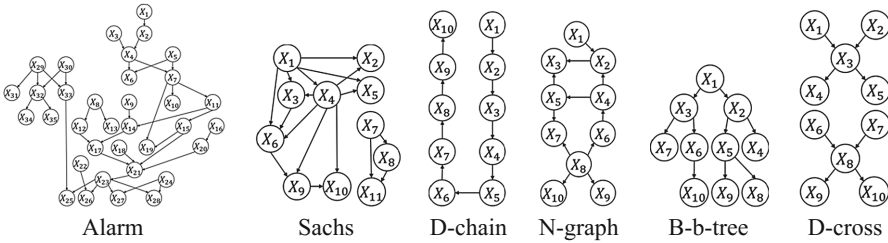


Fig. 2. Structures of BNs used to generate data for structure learning.

Table 2. Characteristics of BNs for generating data.

BN	Nodes	Arcs	Degree of freedom	BN	Nodes	Arcs	Degree of freedom
Alarm	35	39	1.9×10^{15}	N-graph	10	11	59049
Sachs	11	17	177147	B-b-tree	10	9	59049
D-chain	10	9	59049	D-cross	10	8	59049

In the experiments, after learning methods were applied to multivariate datasets just 20 times, obtained structures were compared with the original one. Simultaneously, an inference of $P(Q|E)$ was compared with its true value in order to evaluate the adequacy of obtained structures (Q : Query, E : Evidence). The node corresponding to Q and its value were randomly chosen, while the combination of nodes and their values in E were set similarly to those in Q . True values for probabilistic inference were computed by Likelihood Weighting [3], an approximate inference method, based on the respective original BNs. The number of samples generated in each iteration was 1,000,000, and the convergence condition was that difference of probability calculated at the t -th and $(t - 1)$ -th iteration becomes 10^{-4} or less. Inference of 50 probabilities was made for each of the obtained structures. Then, the difference in the absolute mean of inference error (mean absolute error: MAE) between inferences and true values were compared.

5 Results and Discussion

The experimental results are presented in this section. Unless otherwise stated, all results represented in tables except MAEs are the arithmetic means for 20 obtained structures. The values of MAE in Tables 3 and 4 are the arithmetic means of probability inferences performed 50 times for each of 20 obtained structures. The discussion is roughly divided as follows: (i) detailed investigation of the characteristics of SSL-Pro; and (ii) evaluation of the performance depending on parameters and data features.

Table 3. Learning performance and network features of SSL-Pro and SA.

BN	SSL-Pro (EXT)						SA					
	α	Time (s)	MAE	DA	AA	VMI	β	Time(s)	MAE	DA	AA	VMI
Alarm	0.00	4370.79	0.031	1.45	62.7	+4.73	0.990	12165.5	0.070	22.9	44.5	-3.23
	0.10	762.279	0.056	11.9	29.4	+3.67	0.950	1260.10	0.095	35.0	17.4	-8.50
	0.20	635.303	0.057	12.3	25.6	+3.10	0.925	737.665	0.099	36.3	11.6	-9.37
	0.30	566.063	0.058	14.2	25.7	+2.95	0.825	257.984	0.102	37.9	6.65	-10.2
Sachs	0.00	7.30	0.018	0.00	9.55	+1.10	0.990	287	0.020	0.10	0.60	+0.08
	0.10	3.89	0.049	1.85	9.65	+0.95	0.950	44.4	0.067	4.75	8.60	-0.08
	0.20	3.53	0.055	2.45	8.70	+0.71	0.925	24.5	0.095	7.75	7.20	-0.87
	0.30	3.06	0.074	3.85	8.05	+0.44	0.825	7.35	0.155	12.3	4.85	-1.95
D-chain	0.00	2.16	0.016	0.00	6.65	+0.36	0.990	96.0	0.017	0.05	0.10	-0.01
	0.10	0.56	0.038	2.60	2.05	+0.16	0.950	18.3	0.031	2.10	4.60	-0.46
	0.20	0.57	0.037	2.55	2.20	+0.18	0.925	11.8	0.045	3.35	6.15	-0.59
	0.30	0.52	0.040	2.90	2.20	+0.24	0.825	3.79	0.069	6.00	5.45	-1.76
D-cross	0.00	2.07	0.012	0.00	5.90	+0.40	0.990	94.6	0.011	0.05	0.40	+0.00
	0.10	0.60	0.041	2.15	3.55	+0.34	0.950	17.1	0.029	1.70	4.95	-0.17
	0.20	0.56	0.042	2.30	3.25	+0.26	0.925	10.0	0.046	2.65	4.25	-0.54
	0.30	0.51	0.050	2.75	2.45	+0.10	0.825	3.18	0.088	5.10	3.20	-1.38

5.1 Outline of Characteristics of SSL-Pro

This section evaluates the basic characteristics of SSL-Pro. In Table 3, (a) deleted arcs (DA), and (b) added arcs (AA) denote the number of arcs which (a) *exist* in the original BN and do *not exist* in the obtained structure, and (b) do *not exist* in the original BN and *exist* in the obtained network structure, respectively. Neither (a) nor (b) take the arc direction into account. Variation in MI (VMI) is the difference between total mutual information (MI) between nodes in added arcs in the obtained network and that in original BN. We define total mutual information as the sum of all mutual information between two nodes that are adjacent by arcs.

Relationships Between Computational Time and MAEs. First, *trade-off* relationships between Time and MAE were evaluated (Table 3). In the SSL-Pro, Time tended to decrease, while MAE tended to increase with the value of meta-parameter α . Since computational time depends in large part on the search space, probabilistic pruning was able to produce a good effect from the perspective of time by the increase of α . As a side-effect of probabilistic pruning, an increase of MAE was simultaneously observed due to the descent of network appropriateness (SA showed a similar trend). We believe

that coordinating the (meta-)parameters depending on the respective users' purposes is the key to improving learning performance for the users.

Second, we consider the relationships between α and features of the obtained networks. Table 3 implies that AA or DA increased (decreased): (i) as α decreased; and (ii) as the number of nodes (i.e., random variables) increased. The above (ii) appears to be due to the increase of chances, which judge whether an arc should be added, with the number of nodes. Since extremely complex structures should be avoided for ease of observation, especially when the number of nodes is large, we believe the desirable setting for α is not too small (e.g., greater than 0.10).

Comparison of SSL-Pro with SA. Table 3 shows that the values of DA in SA were larger than those in the SSL-Pro in almost all datasets, despite the value of α . It was also confirmed that VMI was smaller (less than 0 in some cases) in SA than in the proposed method. For example, in the D-cross dataset, although the values of DA and AA in the SSL-Pro ($\alpha = 0.20$) were similar to those in SA ($\beta = 0.925$), VMI for the SSL-Pro was larger than that for SA. Moreover, the values of "AA - DA" in SA were smaller than those in the SSL-Pro. This means that decrement of MI because of arc deletion was greater than increment of MI because of arc addition. These results suggest that SA has a tendency to obtain "simple" BN structures in terms of the number of nodes, while some arcs between two nodes having strong stochastic dependences can be deleted.

Table 3 also indicates that the SSL-Pro obtained networks that could perform higher-accuracy probabilistic inference than SA in the same computational time. For example, under conditions of similar computational time, MAE for the SSL-Pro ($\alpha = 0.10$, time = 762 s) was smaller than that for SA ($\beta = 0.925$, time = 738 s). Moreover, although the SSL-Pro ($\alpha = 0.00$, time = 2.16 s) could perform learning for the D-chain dataset faster than SA ($\beta = 0.825$, time = 3.79 s), MAE for the SSL-Pro was smaller. These results indicate that the SSL-Pro can perform learning faster than SA with equal accuracy of probabilistic inference. We therefore concluded that the SSL-Pro is superior to SA in terms of both learning performance and possible network structures.

5.2 Characteristics of Parameter Settings in p_2

We summarize the characteristics of the SSL-Pro using different parameters. Tables 4 and 5 represent learning performance and features of network structures, respectively obtained by the proposed learning method ($\alpha = 0.10, 0.20$, and 0.30).

To evaluate the trade-off between computational time and MAE, we defined $RDA(x, y)$ as "the growth rate of MAE/the decreasing rate of computational time" when α is changing from x to y . A low $RDA(x, y)$ means that benefit of time-reduction is larger than risk of MAE-increment. As shown in Table 6, because $RDA(0.10, 0.30)$ in Alarm and B-b-tree was low, $RDA(0.10, 0.30)$ was generally expected to be low in networks having clear relationships between nodes. This means that the SSL-Pro would be able to maintain preferable learning efficiency in terms of RDA even if it performed probabilistic pruning more frequently when the target data for learning had the above-mentioned characteristics. Table 6 shows that the learning of the SSL-Pro with EXT can

be improved with α because RDA(0.10, 0.30) in EXT was the lowest of all parameter settings.

Table 4. Learning performances of SSL-Pro using various settings.

BN	Method	EXT			AVE			RMS		EPI	
	α	Time (s)	MAE		Time (s)	MAE		Time (s)	MAE	Time (s)	MAE
Alarm	0.10	762.279	0.056		718.057	0.055		711.245	0.055	951.066	0.054
	0.20	635.303	0.057		615.652	0.058		647.195	0.057	780.308	0.055
	0.30	566.063	0.058		566.482	0.060		627.292	0.057	753.281	0.056
Sachs	0.10	3.885	0.049		3.769	0.054		3.952	0.036	3.975	0.034
	0.20	3.533	0.055		3.550	0.057		3.223	0.069	4.080	0.037
	0.30	3.059	0.074		3.049	0.076		3.279	0.068	3.701	0.047
N-graph	0.10	0.408	0.043		0.352	0.055		0.394	0.040	0.363	0.041
	0.20	0.282	0.069		0.287	0.061		0.286	0.061	0.352	0.050
	0.30	0.250	0.071		0.243	0.070		0.281	0.065	0.324	0.053
D-cross	0.10	0.601	0.041		0.706	0.030		0.711	0.032	0.679	0.031
	0.20	0.559	0.042		0.626	0.039		0.619	0.039	0.622	0.036
	0.30	0.506	0.050		0.554	0.045		0.610	0.040	0.638	0.036

Table 5. Features of network structures obtained by SSL-Pro using various settings.

BN	Method	EXT			AVE			RMS			EPI		
	α	DA	AA	VMI	DA	AA	VMI	DA	AA	VMI	DA	AA	VMI
Alarm	0.10	11.9	29.4	+3.67	11.4	27.7	+3.30	11.4	27.8	+3.49	10.8	33.6	+3.80
	0.20	12.3	25.6	+3.10	12.5	23.0	+2.78	12.1	26.9	+3.46	11.4	27.2	+3.72
	0.30	14.2	25.7	+2.95	13.3	23.3	+2.79	12.9	24.1	+2.75	12.0	28.9	+3.68
Sachs	0.10	1.9	9.7	+0.95	2.1	9.4	+0.87	1.1	9.2	+0.98	1.0	9.8	+1.01
	0.20	2.5	8.7	+0.71	2.3	9.2	+0.77	3.7	7.8	+0.45	1.4	9.2	+0.97
	0.30	3.9	8.1	+0.44	3.9	7.6	+0.38	3.0	8.1	+0.58	2.0	9.6	+0.86
N-graph	0.10	3.8	4.5	+0.24	4.4	3.9	+0.14	4.0	4.9	+0.23	4.2	4.2	+0.26
	0.20	5.0	3.7	+0.12	4.6	3.1	+0.10	4.8	3.3	+0.06	4.3	4.0	+0.20
	0.30	5.0	3.2	+0.09	5.1	3.6	+0.14	4.8	3.6	+0.15	4.5	3.9	+0.21
D-cross	0.10	2.2	3.6	+0.34	1.5	4.1	+0.41	1.6	4.6	+0.32	1.6	4.1	+0.36
	0.20	2.3	3.3	+0.26	2.1	3.8	+0.29	2.0	3.7	+0.28	1.9	3.4	+0.29
	0.30	2.8	2.5	+0.10	2.4	4.0	+0.33	2.1	3.6	+0.33	1.9	3.7	+0.32

Table 6. Ratio of Degradation in Accuracy (RDA) for various settings of p_2 .

	(a) RDA(0.00, 0.30)					(b) RDA(0.00, 0.10)					(c) RDA(0.10, 0.30)				
	EXT	AVE	RMS	EPI	Ave.	EXT	AVE	RMS	EPI	Ave.	EXT	AVE	RMS	EPI	Ave.
Alarm	0.25	0.25	0.27	0.32	0.27	0.32	0.29	0.29	0.38	0.32	0.77	0.86	0.92	0.83	0.84
Sachs	1.76	1.80	1.73	1.35	1.66	1.48	1.58	1.10	1.04	1.30	1.19	1.14	1.58	1.30	1.30
D-chain	0.59	0.54	0.55	0.61	0.57	0.60	0.52	0.61	0.60	0.58	0.97	1.04	0.91	1.01	0.98
N-graph	0.76	0.72	0.78	0.73	0.75	0.74	0.83	0.66	0.63	0.71	1.03	0.87	1.18	1.17	1.06
B-b-tree	0.64	0.56	0.61	0.76	0.64	0.77	0.66	0.80	0.72	0.74	0.83	0.84	0.76	1.05	0.87
D-cross	0.98	0.96	0.94	0.89	0.94	0.96	0.81	0.88	0.83	0.87	1.02	1.18	1.07	1.08	1.09

Difference of RMS from AVE. We discussed the features of RMS designed to restrain the adverse effect in AVE. Table 4 revealed that, as an overall trend, learning with RMS requires more computational time than that with AVE. It is thought that computational time was increased because p_2 was prevented from becoming large due to the influence of the clusters having small meta-dependence, as stated in Sect. 3.2. However, RMS did not always provide smaller values of MAE than AVE. Therefore, we believe that employing AVE for parameter setting is better than employing RMS.

Comparison of AVE with the Others. Table 6 (a) and (b) indicate that the SSL-Pro with AVE was able to perform efficient learning for most datasets except Sachs, N-graph (in $\alpha \leq 0.1$), and D-cross (in $\alpha > 0.1$). For AVE in Sachs and N-graph, the RDA(0.00, 0.10) and RDA(0.10, 0.30) showed the same tendencies: RDA(0.00, 0.10) was the highest while RDA(0.10, 0.30) was the lowest.

Furthermore, in Table 5, we focused on the measure computed by “AA - DA” expressing a similarity between the obtained structure and the original BN in terms of the number of arcs added to the network. It was confirmed that the above-mentioned measures for AVE were smaller (i.e., closer to 0) than those for EXT and EP1 in almost all datasets and α settings. This means that the obtained network structures based on AVE had a lower number of arcs than those based on EXT and EP1 while maintaining favorable learning efficiency. The SSL-Pro with AVE thus tends to be able to obtain appropriate structures regardless of the values of α without requiring the addition of “superfluous arcs.”

Characteristics of EP1. The SSL-Pro with EP1 provided the best performance in terms of MAE except in the case of N-graph ($\alpha = 0.10$) and D-cross ($\alpha = 0.10$), although it occasionally took relatively longer computational time compared to the other settings (see Table 4). It was also able to perform efficient learning when $\alpha \leq 0.10$. This was demonstrated by the fact that the values of RDA(0.00, 0.10) for EP1 were the first or the second lowest in every parameter settings for all datasets except Alarm. As shown in Table 5, DA and AA through structure learning tended to be the smallest and the largest in EP1, respectively. This means that possible networks may have complex structures due to many added arcs. However, as mentioned earlier, the SSL-Pro with EP1 was able to perform well for the Alarm dataset, which has many nodes.

6 Conclusions and Future Work

We introduced new parameter settings in probabilistic pruning to the improved version of SSL (SSL-Pro) in order to refine both computational time and adequacy of obtained network structures. The experimental results indicated that the SSL-Pro provided superior performance to SA. We also discussed appropriate parameter settings to allow the SSL-Pro to perform appropriate structure learning. In the case of applying the SSL-Pro, it should be noted that the appropriate parameter settings depended on datasets used for structure learning and user’s preference. However, the number of datasets was insufficient to discuss appropriate settings. Consequently, it will be necessary to carry out more

detailed verification by performing further experiments using additional datasets on the basis of the tendencies found in this paper.

In the SSL-Pro, there is a problem in that computation of mutual information between all pairs of nodes involves the meta-dependence calculation. Therefore, an increase of random variables in a dataset of learning targets causes exponential increase of computational time. We will therefore discuss a new learning method based on probabilistic pruning that employs an information criterion.

References

1. Cheng, J., Greiner, R., Kelly, J., Bell, D., Liu, W.: Learning Bayesian networks from data: an information-theory based approach. *Artif. Intell.* **137**, 43–90 (2002)
2. Larranaga, P., Poza, M., Yurramendi, Y., Murga, R.H., Kuijpers, C.M.H.: Structure learning of bayesian networks by genetic algorithms: a performance analysis of control parameters. *IEEE TPAMI* **18**, 912–926 (1996)
3. Fung, R., Chang, K.: Weighing and integrating evidence for stochastic simulation in bayesian networks. In: 5th Annual Conference on Uncertainty in Artificial Intelligence (UAI 1989), pp. 209–219. Elsevier Science (1989)
4. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, Burlingto (1988)
5. Fukui, H., Kitakoshi, D.: Prior knowledge-based stepwise structure learning of bayesian networks (in japanese). *IEICE Technical report*, vol. 108, pp. 55–60 (2009)
6. Nishiyama, H., Kitakoshi, D., Suzuki, M.: A study on appropriate parameter settings in a stepwise structure learning for bayesian networks (in japanese). In: Proceeding 38th SICE Symposium on Intelligent Systems, pp. 79–84 (2011)
7. Kitakoshi, D., Azuma, G., Suzuki, M.: Improving learning speed in stepwise structure learning method for bayesian networks by using probabilistic pruning (in japanese). *IPSI SIG Technical report*, vol. 2016-ICS-182, pp. 1–6 (2016)
8. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Upper Saddle River (1988)
9. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. *Science* **220**, 671–680 (1983)
10. Marco, S.: bnlearn (2016). www.bnlearn.com/bnrepository/
11. Rissanen, J.: Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Co., Singapore (1989)

Differential-Weighted Global Optimum of BP Neural Network on Image Classification

Lin Ma¹, Xiao Lin^{1,3}, and Linhua Jiang^{1,2(✉)}

¹ Shanghai Key Lab of Modern Optical Systems,
University of Shanghai for Science and Technology, Shanghai, People's Republic of China
honorsir@yandex.com

² CCSR, Stanford University, 269 Campus Drive, Stanford, CA 94305, USA

³ The College of Information, Mechanical and Electrical Engineering,
Shanghai Normal University, Shanghai, People's Republic of China

Abstract. This paper investigates the problem of image classification with limited or no annotations, but abundant unlabelled data. We propose the DBP (Differential-weighted Global Optimum of BP Neural Network) to make the performance of the BP Neural Network to become more stable. In details, the optimal weights will be saved as potential global optimum during the process of iteration and then we combine the BP Neural Network with the potential global weights to adjust parameters in the backward feedback process for the first time. As the model has fallen into local optimization, we replace the present parameters with the potential global optimal weights to optimize our model. Besides, we consider EP, CNN, SIFT image features and conduct several experiments on eight standard datasets. The results show that DBP mostly outperforms other supervised and semi-supervised learning methods in the state of the art.

Keywords: Image classification · BP neural network · Global optimum · Learning with limited supervision

1 Introduction

Image set classification has occupied an important status in computer vision. Image classification almost could be divided into several approaches in recent years such as unsupervised feature learning [1–4], semi-supervised learning [5–7], supervised learning and image clustering [8, 9]. Semi-supervised learning [10] means taking both labelled and unlabelled data into account when training machine learning models. That means that semi-supervised learning merely need limited labelled instances. This is very convenient and adaptable to the trend. Even the semi-supervised learning, most of them usually make strong assumptions such as single Gaussian, subspace or mixture of subspaces to represent image sets. In many complex databases, these assumptions may not be held.

In this paper, we propose a semi-supervised image classification strategy which exploits the BP Neural Network optimized by the differential-motivated global optimum. The key idea of the proposed approach is shown in Fig. 1. Given each image

feature, we first model it as a nonlinear manifold [11] can effectively describe the geometrical and structural information. Then we can obtain the possibilities that the image feature belongs to each class from the output layer and adjust the parameters through the feedback acquiring from the predictable labels and real labels. Although the BP Neural Network indeed works in some degree, it would result in the instability of the parameters and the decline of classification effectiveness if the margin of error is too large. Motivated by the factor that the BP Neural Network may fall into the local optimization and not be robust to image classification, we propose a DBP method to optimize the parameters of the BP Neural Network though storing potential global parameters and replacing the parameters when the performance is not good. Experimental results on eight datasets validate the effectiveness of the proposed method.

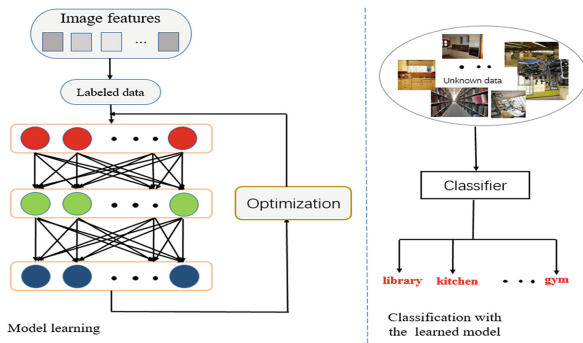


Fig. 1. Schema of the proposed strategy. For each labeled data, we sampled it from image features as training samples and pass it into multiple layers of deep neural network to nonlinearly map each sample into another feature space. After each iteration, we replace the ineffective parameters with the optimal values stored previously. The top of our network represents the possibilities that belongs to each class.

2 Related Work

Our method is generally relevant to BP Neural Network, image features and datasets.

BP Neural Network: Recently, the BP Neural Network [12–14], a deep learning method [12], has attracted increasing interest in computer vision and machine learning. The BP Neural Network has three layers which is generally divided into the input layer, hidden layer and output layer. For image classification, we typically use image feature as the input of the first layer and every element in the image feature as a node will be non-linear mapped to hidden layer and output layer. As a result, the output layer represents the possibility that the image feature belongs to each class. Then we select the number which possibility is the highest and the number is predictable label.

Image Features and Datasets: In this work, we utilize three image features which are CNN, EP [15] and SIFT features to evaluate our method. CNN feature [16–18] as a mature image representation is widely used in image classification field. While EP

feature which is relatively novel captures not only the information of individual images, but also the relationships among images. At last, SIFT feature has the invariance for sizes or angles of images. Eight datasets cover a variety of scenarios, including complex and single background, outdoor and indoor scenes and so on.

3 Proposed Approach

Figure 1 shows the basic idea of our proposed DBP method, and the following subsections present the details of the proposed method.

3.1 DBP

Let $X = [X_1, \dots, X_c, \dots, X_C]$ be the training set of C different classes, where $X_c = [x_{c1}, x_{c2}, \dots, x_{ci}, \dots, x_{cN_c}] \in \mathbb{R}^{d \times N_c}$ denotes the c th image set, $1 \leq c \leq C, N_c$ is the number of samples in this image set, x_{ci} is the i th image in this image set, and d is the feature dimension of each image. As shown in Fig. 1, we construct a deep neural network which has L layers in the work. Next we will introduce the mathematical details of BP Neural Network. The first step is the forward propagation, we assume that $\alpha^{x,1} = x_{ci}^1$ as the input activation of the first layer. For each training sample x , its output of $L-1$ layer in the network is computed as follows:

$$z_{ci}^l = w^l \alpha^{x,l-1} + b^l. \tag{1}$$

$$\alpha^{x,l} = \sigma(z_{ci}^l). \tag{2}$$

where $2 \leq l \leq L, w^l$ is the projection matrix and b^l is the bias vector to be learned in the l th layer of the network. σ is a nonlinear active function which applies component-wisely, which is widely used in previous deep learning algorithms. The second step is calculating the output error $\delta^{x,L}$, after finishing the forward propagation, the output error of the top layer is computed as follows:

$$\delta^{x,L} = \nabla_{\alpha} C_x \odot \sigma'(z^{x,L}). \tag{3}$$

where $\nabla_{\alpha} C_x$ is defined as a vector and the rate of output activation, $z^{x,L}$ is defined in (1) and $\sigma'(z^{x,L})$ represents derivative of $\sigma(\cdot)$. The last step is feedback, when the output error of the top layer is computed, we will compute the error of each layer as follows:

$$\delta^{x,l} = ((w^{l+1})^T \delta^{x,l+1}) \odot \delta'(z^{x,l}). \tag{4}$$

where $2 \leq l \leq L - 1, \delta^{x,l+1}$ is defined in (3).

Let $f_c = \{w^1, w^2, \dots, w^L, b^1, b^2, \dots, b^L\}$ be the parameters of the network, we formulate the following optimization problem to minimize the margin between the real labels and predictable labels:

$$\min_{f_c} \frac{1}{2} \|y - \alpha^L\|^2 \tag{5}$$

The objective in (5) is to ensure that the possibility that real labels and predictable labels are same is maximized.

We adopt the sub-gradient descent algorithm to solve the optimization problem in (5) to update the parameters of our model:

$$w^l = w^l - \frac{\eta}{m} \sum_x \delta^{x,l} (\alpha^{x,l-1})^T. \tag{6}$$

$$b^l = b^l - \frac{\eta}{m} \sum_x \delta^{x,l}. \tag{7}$$

where $\frac{\eta}{m}$ is the learning rate, $2 \leq l \leq L - 1$.

These are the basic ideas of BP Neural Network, we can find that although the methods indeed optimize the parameters in some degrees, the model isn't robust and would result in a decline in the classification due to the fluctuation of parameters. Based on these, we have improved the BP Neural Network algorithm by combining theory of Differential-weighted Global Optimum. We store the parameters as follows:

$$w_{best} = w_{best} - \alpha(w_{best} - w^l). \tag{8}$$

$$b_{best} = b_{best} - \alpha(b_{best} - b^l). \tag{9}$$

The potential global optimal values will be initiated and updated in (8) and (9) after each iteration. The specific details are that if parameters trained are better than the potential global optima, the optima will be updated, on the contrary, parameters which are used to train will be replaced by the potential global optima. This will ensure the stability of our method and improve the results of classification.

3.2 Classification

Given a testing image set $X^q = X_1^q, X_2^q, \dots, X_{N_q}^q$, where x_j^q is the j th image ($1 \leq j \leq N_q$) in this set and N_q is the number of images in this set. Now we put the image feature x_j^q into the model we trained, the output layer represents the possibilities that the image belongs to each class. Then we will choose the class which has the highest possibility as the predictable label.

4 Experimental Results

The effectiveness of our method is evaluated in the situations of semi-supervised image classification, where the amount of labeled data is sparse relative to the total amount of data. Eight standard datasets are used for the evaluation: Texture-25 [19], Caltech-101

[20], STL-10 [21], Scene-15 [22], Indoor-67 [23], Event-8 [24], Building-25 [25], LandUse-21 [26]. Texture-25 has 25 texture classes, with 40 samples per class, Caltech-101 has 101 object classes, with 31 to 800 images per class, and 8677 images in total, STL-10 has 10 object classes including airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck, with 500 training images per class. Scene-15 has 15 scene classes with both indoor and outdoor environments, 4485 images in total. Each class has 200 to 400 images. Indoor-67 has 67 indoor classes such as shoe shop, mall and garage, with a total of 15620 images and at least 100 images per class. Event-8 has 8 sports event classes including rowing, badminton, polo, bocce, snowboarding, croquet, sailing, and rock climbing, with a total of 1574 images. Building-25 has 25 architectural styles such as American craftsman, Baroque, and Gothic, with 4794 images in total. LandUse-21 has 21 classes of satellite images in terms of land usage, such as agricultural, airplane, forest, and 2100 images in total, with 100 images per class. For semi-supervised classification, five classifiers were adopted to evaluate the method which are SVMs with RBF kernels, Logistic Regression (LR), k-NN, BP Neural Network, Deep NNs and our method (DBP).

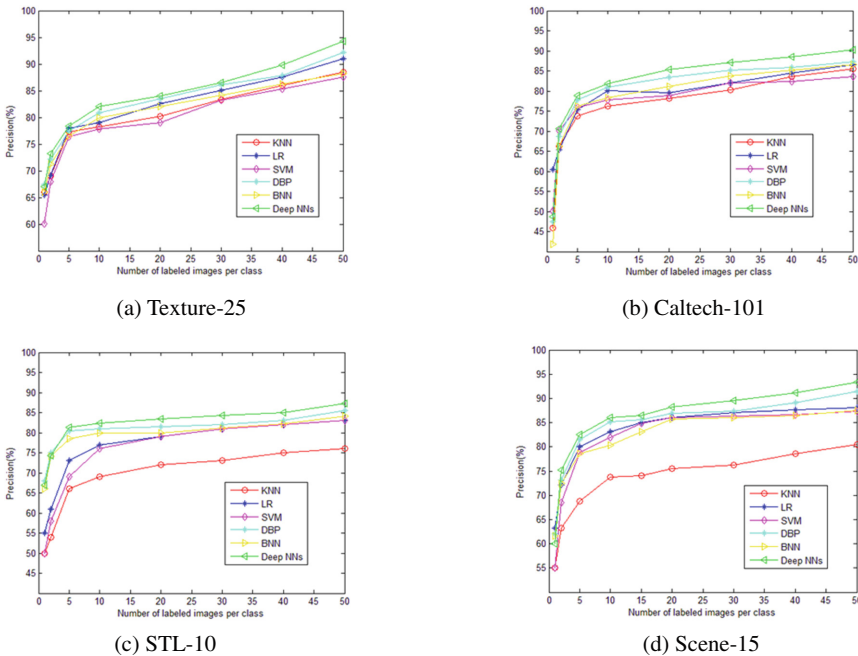


Fig. 2. Classification results of DBP and the other five classifiers on the Texture-25, Caltech-101, STL-10, Scene-15 datasets, where five classifiers are used: k-NN, Logistic Regression, SVMs with RBF kernels, BP Neural Network and Deep NNs. All methods were tested on the EP features.

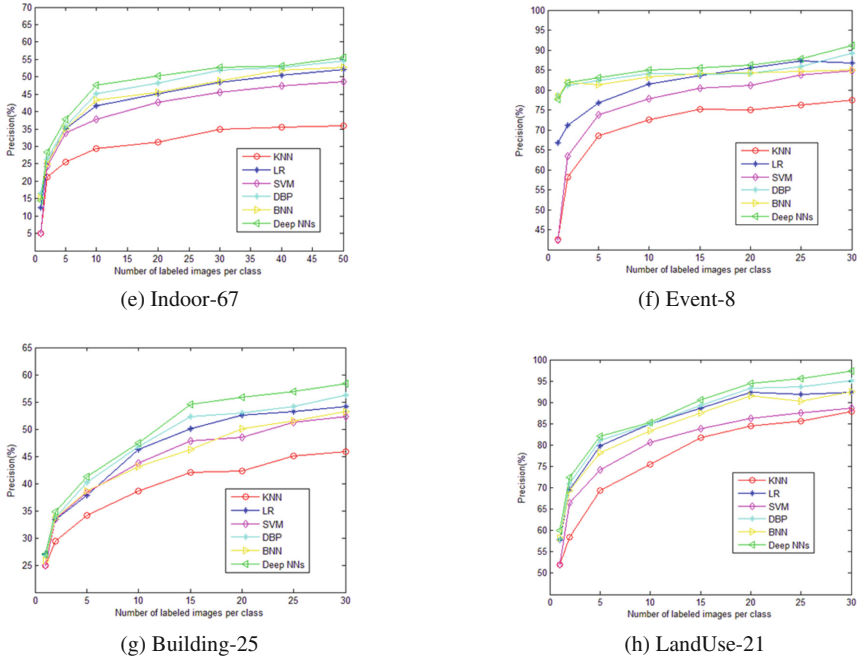


Fig. 3. Classification results of DBP and the other five classifiers on the other four datasets, where five classifiers are used: k-NN, Logistic Regression, SVMs with RBF kernels, BP Neural Network and Deep NNs. All methods were tested on the EP features.

As to the parameters of our method, a wide variety of values for them were tested in experimental setup and we fixed them to the following values: $T = 100$, $r = 30$, $n = 6$, $L = 20$ and $m = 50$, which to a feature vector of 3000 dimensions.

4.1 Comparison to Other SSL Method

In the section, we compare our method (DBP) with the four SSL methods k-NN, LR, SVM, BNN, Deep NNs. Figures 2 and 3 show the results of the six classifiers based on eight datasets and Table 1 lists the precision of the methods when 10 labeled training examples per class are used. From Figs. 2 and 3, we can clearly see that our method is superior to the other four methods on the experimental results. For instance, if 10 labeled training examples per class are used, our method (DBP+EP) improve over the best competing method LR which is proposed in the paper [18] by 3.6% in terms of precision on Event-8. Apart from testing the methods in the paper [18], we also compare our method with the BP Neural Network and find that our approach increases the accuracy and stability on the basis of the BP Neural Network by 6.7% on LandUse-21. As we can see in the Tables 1 and 2, our method (DBP) can distinguish objects with significant objectivity in a single context well, while in complex background such as Indoor-67, our method (DBP) may not be effective. However, the combination of the optimized method and EP image feature is still superior to other methods [18].

Table 1. Precision (%) of image classification on the eight datasets, with 10 labeled training examples per class. The first row is our method (DBP).

Methods	Texture-25	Caltech-101	STL-10	Scene-15	Indoor-67	Event-8	Building-25	LandUse-21
DBP	93.2	87.3	80.6	79.3	50.3	84.5	48.9	91.2
k-NN	89.2	85.2	68.4	60.2	29.3	72.2	38.5	78.1
LR	92.0	86.3	77.2	76.1	41.6	81.6	46.2	86.2
SVM	89.1	84.1	75.1	74.3	38.2	78.4	43.1	82.4
BNN	91.5	86.1	78.0	77.0	47.3	83.1	45.2	85.9
DNNs	94.3	88.3	81.1	80.6	50.8	85.6	50.3	92.6

4.2 Robustness to Features

In this section, we examine the robustness of our method against different image features with 10 labeled training examples per class. Different types of image features such as EP, CNN and SIFT features were tested for the DBP classifier across the eight datasets. It could be found from Table 2 that even though our method was used on different image features, the mean average precision is still higher than other classifiers. This suggests that our approach is robust to different image features.

Table 2. The mean average precision (%) of DBP on the eight datasets, with EP, CNN and SIFT features.

Features	Texture-25	Caltech-101	STL-10	Scene-15	Indoor-67	Event-8	Building-25	LandUse-21
EP	93.2	87.3	80.6	79.3	50.3	84.5	48.9	91.2
CNN	95.3	88.5	82.3	82.1	49.6	72.2	50.6	90.1
SIFT	90.5	85.6	78.4	72.3	41.6	75.5	41.9	85.7

5 Conclusion and Future Work

In order to avoid the BP Neural Network into the local optimal solution, in this paper, we have proposed DBP, a semi-supervised image classification strategy which optimizes the BP Neural Network. We help the model jump out of the local optimization trap through storing potential global optimum and substituting the parameters when the result of penalty function declines. Experiments on eight datasets and three different image features show the superior performance of our method, especially for DBP+EP, for image classification and different features. In addition, the method has been, in some degrees, improved with respect to the BP Neural Network, either in accuracy or stability.

For future work, the number of the hidden layer in the neural network would waste the computing time if we blindly pursue accuracy. We are interested in achieving how to reduce the number of the hidden layer as well as keeping the higher accuracy or carrying out some research in the deep-learning.

Acknowledgement. The research was partly supported by the program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, USST incubation project (15HJPY-MS02), National Natural Science Foundation of China (No. U1304616, No. 61502220). We would like to appreciate Zhong hui for modifying English spelling during the whole work.

References

1. Coates, A., Ng, A.Y., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. *J. Mach. Learn. Res.* **15**, 215–223 (2011)
2. Dosovitskiy, A., Springenberg, J., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: *Neural Information Processing System*, Montreal, Canada, pp. 766–774, December 2014
3. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: *International Conference on Computer Vision*, Santiago, Chile, pp. 1422–1430, December 2015
4. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: *Computer Science (2015)*
5. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: *Neural Information Processing System*, Vancouver, Canada, pp. 522–530, December 2009
6. Guillaumin, M., Verbeek, J.J., Schmid, C.: Multi-model semi-supervised learning for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, America, pp. 902–909, June 2010
7. Dai, D., Van Gool, L.: Ensemble projection for semi-supervised image classification. In: *International Conference on Computer Vision*, Sydney, Australia, pp. 2072–2079, December 2013
8. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A.: Discovering objects and their location in images. In: *International Conference on Computer Vision*, Beijing, China, pp. 370–377, October 2005
9. Dai, D., Wu, T., Zhu, S.C.: Discovering scene categories by information projection and cluster sampling. In: *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, America, pp. 483–497, June 2010
10. Chapelle, O., Schölkopf, B., Zien, A., et al.: *Semi-Supervised Learning*. MIT Press, Cambridge (2006)
11. Elgammal, A., Lee, C.S.: Separating style and content on a nonlinear manifold. In: *Computer Vision and Pattern Recognition*, Washington, DC, America, pp. 478–485, June 2004
12. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)
13. Zhu, P., Zhang, L., Zuo, W., Zhang, D.: From point to set: extend the learning of distance metrics. In: *International Conference on Computer Vision*, Sydney, Australia, pp. 2664–2671, December 2013
14. Huang, G.B., Lee, H., Learned-Miller, E.: Learning hierarchical representations for face verification with convolutional deep belief networks. In: *Computer Vision Pattern Recognition*, Providence, Rhode Island, pp. 2518–2525, June 2012
15. Dai, D., Gool, L.V.: Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering. In: *Computer Vision & Pattern Recognition*, Las Vegas, America, pp. 254–260, June 2016

16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**(2), 1097–1105 (2012)
17. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition. *Comput. Sci.* **50**(1), 815–830 (2014)
18. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, America, pp. 580–587, June 2014
19. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. *IEEE Pattern Anal. Mach. Intell.* **27**(8), 1265–1278 (2005)
20. Li, F.F., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Washington DC, USA, pp. 59–70, June 2004
21. Coates, A., Ng, A.Y., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. *Int. Conf. Artif. Intell. Stat.* **15**, 215–223 (2011)
22. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, pp. 17–22, June 2006
23. Quattoni, A., Torralba, A.: Recognizing indoor scenes, In: *IEEE Conference on Computer Vision and Pattern Recognition*, Florida, USA, pp. 413–420, June 2009
24. Li, L.J., Li, F.F.: What, where, and who? Classifying event by scene and object recognition. In: *International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp. 1–8, October 2007
25. Xu, Z., Tao, D., Zhang, Y., Wu, J., Tsoi, A.: Architectural style classification using multinomial latent logistic regression, *European Conference on Computer Vision*, Zurich, Swiss, pp. 600–615, September 2014
26. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: *International Conference on Advances in Geographic Information Systems*, pp. 270–279. ACM, San Jose, November 2010

Classification Model for Skin Lesion Image

Nontachai Danpakdee and Wararat Songpan^(✉)

Department of Computer Science, Faculty of Science,
Khon Kaen University, Naimuang, Muang, Khon Kaen 40002, Thailand
nontachai@kkumail.com, wararat@kku.ac.th

Abstract. The problem of image classification mostly focuses on feature extraction which depends on image data. However, the famous feature is extracted from images using Gray Level Co-occurrence matrix (GLCM) in order to cover all feature, which are range differently. The problem is less accuracy when it is inputted into classification model. This idea of paper is proposed for feature normalization 2 types called local-normalization and global-normalization from all feature extraction using GLCM in preprocessing of classification method. These feature values extracted from GLCM are transformed to proper normalization and given input classification model as Back Propagation in Multi-layer perceptron and Multi-Class support vector machine methods (polynomial and RBF) to compare these classification models. The skin disease image classification which occurs from skin lesions has divided into four classes: Tinea Corporis, Pityriasis Versicolor, Molluscum Contagiosum and Herpes Zoster. The experimental results are shown comparison between non-normalization and normalization within the same class called local-normalization and all classes called global-normalization. The accuracy of MLP with normalization by min-max normalization with local-normalization is highest to 92%. The methods of polynomial-SVM and RBF-SVM are given accuracy as 85% and 81% respectively. Whereas, the accuracy of classification model with non-normalization, and global-normalization are given average of accuracy as only 35% approximately.

Keywords: Grey Level Co-occurrence Matrices (GLCM) · Multilayer perceptron · Multi-class support vector machine · Skin lesion

1 Introduction

Diagnosis errors may cause to patients who is harm from interpreter of images or lab testing. These are depended on experience and training of medical expertise to accurate learning which knowledge can be diagnosed as any kinds of disease. Lesion's dermatitis is a type of disease of skin that occurs frequently in diary life. This disease can found in childhood until old age. Some diseases require medical attention. Some diseases can be diagnosed by themselves. If the patient disinterested to treat or diagnostic error is dangerous, may cause spreading throughout the body. The image processing is very important in healthcare research. However, the problem is various features that can be extracted from an image. There are many researches referred to feature extraction of skin image which has the objective is to feature extraction from lesion images in [1–3]

and skin cancer in [4–6]. In addition, these papers have different method for feature extraction from image such as feature extraction is performed using Grey Level Co-occurrence Matrix (GLCM) in [1, 4–6] which are famous method, using color, texture and RGB histogram to extract the features in [2] and using Otsu thresholding and morphological reconstruction algorithms that was proposed in [3].

Our method integrates feature extraction based on texture feature using only GLCM features. The feature normalization in each type of dermatitis are using min-max normalization, which the value range [0, 1] input to classifier models which also is compared with all types of dermatitis and non-normalization. The results are presented to the performance between non-normalized feature and normalized feature with SVM with polynomial kernel and Back Propagation Neural Network classifier (Fig. 1).



Fig. 1. (a) Tinea Corporis; (b) Pityriasis Versicolor; (c) Molluscum Contagiosum; (d) Herpes Zoster [7–10]

The paper is arranged as follows, Sect. 2 presents the related work. Section 3 presents proposed a step of feature normalization to classification for skin disease of 4 types. The experimental results are presented in Sect. 4. Finally, the conclusion and future work are presented in Sect. 5.

2 Related Works

In the present, the skin disease is interesting. Many researches are applied to image processing and classifier models. The features are many ways to extracted from skin image. Therefore, the prime objective in [1] is to feature extraction from lesion images. Feature extraction is performed using Grey Level Co-occurrence Matrix (GLCM). GLCM is used to calculate the different combinations of gray levels in four different directions ($\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$). These features were used to build a model for classification with the artificial neural network classifier. In addition, segmentation and classification of skin lesions for disease diagnosis presented an approach for classification of skin lesions. The proposed was using color, texture and RGB histogram to extract the features. Also, the classification results using SVM and k-NN with 46.71% and 34% of F-measure respectively. However, the accuracy is still inadequate. In [3], the paper tried to a high performance algorithm to diagnosis of skin lesions deterioration in dermatoscopic images using new feature extraction. An algorithm was proposed four major steps for classification of pigmented skin lesions. Otsu thresholding and morphological reconstruction algorithms used to segmentation. The ABCD rule used to extract features and normalized features between $[-1, 1]$. The SVM classifier with RBF kernel classified

begin and malignant lesions. The experimental results show specificity 90.03%, sensitivity 79.89% and accuracy 84.09%.

For other researches are concerned with skin disease. In [4] feature extraction and classification is hybrid genetic algorithm artificial neural network classifier for skin cancer detection. Classification of non-cancerous and cancerous or malignant are from lesion images using GLCM feature and RGB color feature extraction. The paper focused on classifier models so the genetic algorithm was applied for artificial neural network. The overall accuracy of the system is 88%. However, the feature extraction is important. In [5], texture and Color feature based WLS Framework aided Skin Cancer Classification Using MSVM and ELM. The proposed skin cancer classification was using GLCM and histogram of oriented gradients (HOG) to extract texture feature and color histogram to extract color feature. In [6] GLCM and multi-class Support Vector Machine was based automated skin cancer classification. The paper proposed an approach for classification of four different types of skin cancers. The features are extracted by GLCM based texture feature and classification with multi-class Support vector machine. The accuracy of proposed method is 81.43% for 75 training images.

3 Proposed Methodology

In this paper is proposed feature normalization within the same class and all class to test with four types of lesion image from dermatitis as Tinea Corporis, Pityriasis Versicolor, Molluscum Contagiosum and Herpes Zoster. The proposed method includes four steps: gray scale transformation, feature extraction, feature normalization and classification for skin disease. The experimental result show that the performance in feature normalization and classification properties with Support vector machine and Back propagation neural network.

The skin lesion images is used 100 images from online sources, including 25 lesion image from Tinea Corporis disease, 25 lesion image from Pityriasis Versicolor disease, 25 lesion image from Herpes Zoster disease and 25 lesion image from Molluscum Contagiosum disease. These images are format in .jpg extension (Fig. 2).

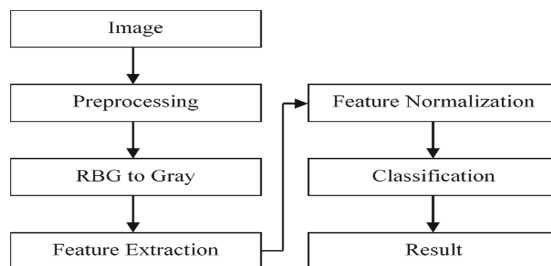


Fig. 2. Overview of proposed methodology for classification of skin lesion images

3.1 Preprocessing and RGB to Gray

The images are cropped and resized lesion images to 200×200 pixels. After that, the images are converted from RGB lesion images to grayscale images [11] are shown in Fig. 3. The images are transformed to grayscale, which the Luminance value (Y) by forming a weighted sum of Red (R), Green (G) and Blue (B) components using the Eq. (1) as follows,

$$Y = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \quad (1)$$

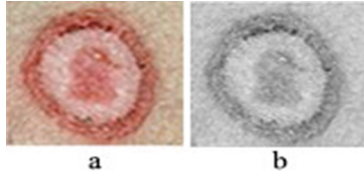


Fig. 3. (a) Input image; (b) Grayscale image

Where,

- R is weighted summation of red pixels
- G is weighted summation of green pixels
- B is weighted summation of blue pixels.

3.2 Feature Extraction

Feature extraction using Gray Level Co-occurrence matrix (GLCM) based on statistical method to extract texture. GLCM are extracted from grayscale lesion image. There are 24 features for each lesion image including: Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity, Energy matlab, Entropy, Homogeneity matlab, Homogeneity, Maximum probability, Sum of squares variance, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, Information measure of correlation1, Information measure of correlation2, Inverse difference normalized (INN), Inverse difference moment normalized, Mean, Variance and Angular second moment. [12–14], which are cover feature in image processing.

3.3 Feature Normalization

Feature Normalization used to transform very different value ranges. For example, some values are greater than one hundred and some values are less than zeros. Therefore, we proposed the attribute are normalized between zeros to one in each the class, which the class is a type of disease called local-normalization as follows:

$$V'_A{}^C = \frac{V_A^C - Min_A^C}{Max_A^C - Min_A^C} \quad (2)$$

Where,

V'_A^C is the normalized value for each class (c) with that feature attribute (a).

Min_A^C is the minimum value of an attribute within the same class.

Max_A^C is the maximum value of an attribute within the same class.

V_A^C is a feature value, which will be transformed in the class (Figs. 4, 5 and 6).

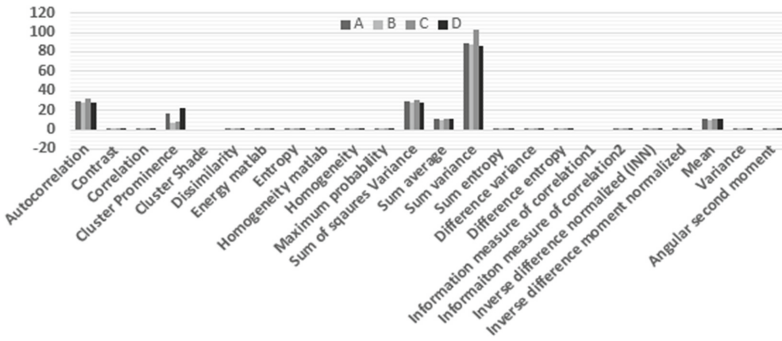


Fig. 4. Average value of each feature with non-normalization: (A) Tinea Corporis; (B) Pityriasis Versicolor; (C) Molluscum contagiosum; (D) Herpes Zoster

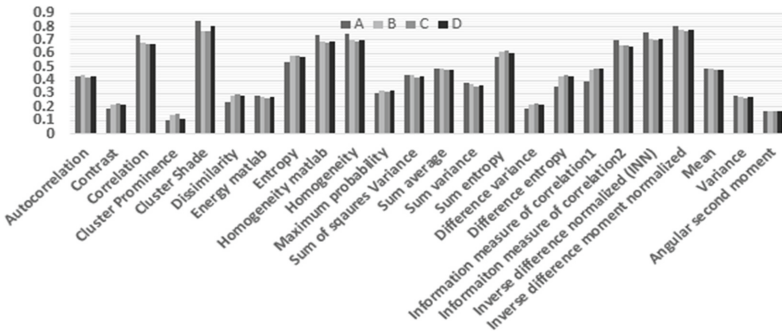


Fig. 5. Average value of all feature normalization within all class

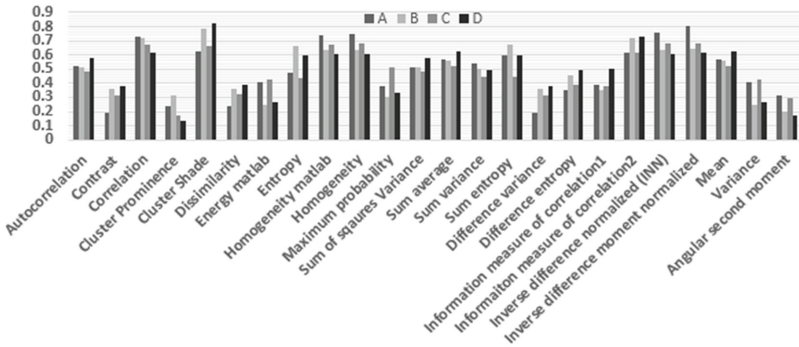


Fig. 6. Average value of all feature normalization within the same class

3.4 Classification

Normalized features are inputted to classifier model as multi-class SVM classifier with Polynomial, RBF kernel and multi-layer perceptron with Back propagation algorithm.

4 Experimental Results

Image preprocessing (Grayscale) and GLCM feature extraction were done in MATLAB R2008b. In Table 1 present the percentage of accuracy value with 100 images of training for 10 simulations and 4 classes obtained from different classifier as polynomial-SVM, RBF-SVM and BP-MLP. The percentage of accuracy for polynomial-SVM is 85.21%, RBF-SVM, the percentage of accuracy is 81% and finally, the percentage of accuracy for Multi-layer perceptron using Back Propagation is highest to 92% applied feature normalization within the same class (local-normalization). Whereas, non-normalization given the highest accuracy only 36% and global-normalization given the percentage of accuracy as 35% respectively.

Table 1. Performance of feature normalization

Method of normalization	Accuracy (in %) obtained using			Average of F-measure
	SVM		MLP	
	Polynomial	RBF	Back propagation	
Non-normalization	36	34	35	0.350
Global-normalization	37	33	35	0.371
Local-normalization	85	81	92	0.824

The performances are also evaluated in terms of F-measure in Table 1. The classification models of four difference classes with Back Propagation Neural Network used to feature attribute normalization. The average of F-measure is 0.824 within the same class to

normalization so the propose method is better results that non-normalized features and using all classes to feature normalization. However, the consuming time of back propagation is longer than other model (Table 2).

Table 2. Consuming time to build model of Classification

Method of normalization	Build model (in seconds)		
	<i>SVM</i>		<i>MLP</i>
	<i>Polynomial</i>	<i>RBF</i>	<i>Back propagation</i>
Non-normalization	0.01	0.01	0.36
Global-normalization	0.01	0.01	0.29
Local-normalization	0.09	0.03	5.78

Table 3. F-measure (10 simulations) of classification

Skin lesion classes	F-measure obtained from		
	<i>SVM</i>		<i>MLP</i>
	<i>Polynomial</i>	<i>RBF</i>	<i>Back propagation</i>
Tinea Corporis	0.898	0.846	0.920
Pityriasis Versicolor	0.824	0.776	0.816
Molluscum Contagiosum	0.868	0.784	0.941
Herpes Zoster	0.809	0.875	0.920

Table 4. True positive rate and false positive rate (10 simulations) of classification

Skin lesion classes	Performance obtained using					
	<i>SVM</i>				<i>MLP</i>	
	<i>Polynomial</i>		<i>RBF</i>		<i>Back propagation</i>	
	<i>TP</i>	<i>FP</i>	<i>TP</i>	<i>FP</i>	<i>TP</i>	<i>FP</i>
Tinea Corporis	0.88	0.027	0.84	0.040	0.92	0.040
Pityriasis Versicolor	0.84	0.067	0.76	0.093	0.84	0.027
Molluscum Contagiosum	0.92	0.067	0.80	0.067	0.96	0.027
Herpes Zoster	0.76	0.040	0.84	0.053	0.96	0.013
Average	0.85	0.050	0.81	0.063	0.92	0.027

Therefore, the method of separated classes features normalization (within the same classed) is used to extract features and present in Table 3. Multi-layer perceptron with Back propagation and SVM classifier with polynomial and RBF kernels are experimented for classification. Back propagation classifier is given the best performance to 0.899 of average F-measure where SVM classifier using polynomial and RBF kernel given 0.849 and 0.820 of average f-measure respectively for each of the four classes.

The Back Propagation Neural Network is used all 24 features of GLCM and feature normalization is within the same class (local-normalization) before input all feature into input layer. There are two hidden layers each hidden layer has eight nodes and four nodes

of output. The learning rate is 0.01, momentum is 0.25 and training time is 10000 which given the best results. In Table 4, Back Propagation classifier has achieved as 0.92 of average true positive (TP) rate where SVM classifier using polynomial and RBF kernel has given 0.85 and 0.81 of average true positive rate respectively for each of the four classes.

5 Conclusion

Skin disease images are very important interpreter-image for treat skin. Our proposed model used to extract texture feature with GLCM. These features need to be normalized within the same class as local min-max value within a feature attribute which is extracted from image. Using local min-max value is better than global or all min-max of feature attributes. Therefore, the classification is performed with artificial neural network classifier as Multi-Layer perceptron using Back Propagation and compare with other methods as SVM using Polynomial and RBF kernel to classify four types of skin disease from Skin Lesion Image. The experimental results found that the Back propagation neural network has overall the higher performance than the SVM classification. In the future, the false positive is still low value. Therefore, the proposed method will be improved this performance value to be more effective to classify the skin lesion image.

References

1. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **3**(6), 610–621 (1973)
2. Sumithra, R., Suhil, M., Guru, D.S.: Segmentation and classification of skin lesions for disease diagnosis. *Procedia Comput. Sci.* **45**, 76–85 (2015)
3. Razazzadeh, N., Khalili, M.: A high performance algorithm to diagnosis of skin lesions deterioration in dermatoscopic images using new feature extraction, pp. 1207–1212 (2015)
4. Aswin, R.B., Jaleel, J.A., Salim, S.: Hybrid genetic algorithm - artificial neural network classifier for skin cancer detection. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 1304–1309 (2014)
5. Choudhury, D., Naug, A., Ghosh, S.: Texture and color feature based WLS framework aided skin cancer classification using MSVM and ELM. In: 2015 Annual IEEE India Conference (INDICON), pp. 1–6 (2015)
6. Maurya, R., Singh, S.K., Maurya, A.K., Kumar, A.: GLCM and multi class support vector machine based automated skin cancer classification, pp. 444–447 (2014)
7. “Skin Problems and Treatments”, WebMD. <http://www.webmd.com/skin-problems-and-treatments/default.htm>. Accessed 25 Aug 2016
8. “Medical Press|Noticias de salud y avances médicos”, Medical Press. <http://www.medicalpress.es/>. Accessed 25 Aug 2016
9. “Siamhealth.net”. <http://www.siamhealth.net/>. Accessed 25 Aug 2016
10. “What causes Herpes Zoster, Infection, Treatment Plans”. <https://blog.eduzones.com/nanasara/133203>. Accessed 25 Aug 2016
11. “Convert RGB image or colormap to grayscale - MATLAB rgb2gray”. <http://www.mathworks.com/help/matlab/ref/rgb2gray.html>. Accessed 25 Aug 2016

12. Mohanaiah, P., Sathyanarayana, P., Kumar, L.G.: Image texture feature extraction using GLCM approach. *Int. J. Sci. Res. Publ.* **3**(5), 1–5 (2013)
13. Soh, L.-K., Tsatsoulis, C.: Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Trans. Geosci. Remote Sens.* **37**(2), 780–795 (1999)
14. Clausi, D.A.: An analysis of co-occurrence texture statistics as a function of grey level quantization. *Can. J. Remote Sens.* **28**(1), 45–62 (2002)

Software Engineering

Generation of Use Cases for Requirements Elicitation by Stakeholders

Junko Shirogane^(✉)

Department of Communication, Tokyo Woman's Christian University,
Suginami, Japan
junko@lab.twcu.ac.jp

Abstract. Use case diagrams and scenarios are often used in the requirements elicitation phase in software development. It is difficult for developers to create them based on appropriate stakeholders' requirements. Meanwhile, stakeholders can survey existing applications to find functions and interactions that are similar to their requirements. This paper proposes a method to generate the bases of use case diagrams from the operation histories of existing applications. Operation histories are divided into operations of individual windows, and the entire window-switching sequence in an existing application is represented as a directed graph. Then the directed graphs are analyzed to extract the window-switching sequences that correspond to use cases. Finally, use case diagrams are generated.

1 Introduction

Requirements analysis of software development consists of requirements elicitation, analysis, specification, and validation [1]. Based on these phases, various requirements engineering processes have been proposed [2, 3], and these subphases are iteratively performed. Especially, more detailed processes for requirements elicitation phase have been provided [4, 5], however, few processes have been provided for other phases. In addition, many methods to elicit stakeholders' requirements have also been proposed [6]. Thus, because stakeholders often do not understand their requirements, and developers cannot extract their requirements enough [7], researches for requirements elicitation are more active than other phases, and appropriate elicitation of stakeholders' requirements by developers considered to be quite difficult.

In the requirements elicitation phase, use case diagrams and scenarios are often used to describe users' requirements and to communicate between stakeholders and developers. However, developers often have difficulty understanding the stakeholders' intent as use case diagrams and scenarios do not always represent the stakeholders' requirements appropriately. To describe accurately the stakeholders' requirements as use case diagrams and scenarios, it is desirable for the stakeholders themselves to describe the requirements. However, stakeholders without knowledge of software development often have difficulty describing use case diagrams and scenarios.

Meanwhile, in many cases, various applications already exist in the domain that stakeholders' require. Surveying existing applications is one important method to elicit

requirements [8, 9]. Although it is difficult for stakeholders to describe use case diagrams and scenarios, they can use existing applications as references to realize use case diagrams and scenarios in new applications.

This paper proposes a method to generate the bases of use case diagrams from the operation histories of existing applications operated by the stakeholders. Concretely, the operation histories are divided into individual windows, and window-switching sequences are represented. Next the main window, which is the window where all functions originate, is identified. Then the entire window sequence in the application is represented. Finally, use cases are extracted from the entire window sequence, and use case diagrams are generated. Scenarios are generated from the events of the window-switching sequences in the use cases, and scenario generation method was proposed previously [10]. This paper describes details of use case diagram generation.

The rest of the paper is organized as follows. Section 2 compares this work to other research. Section 3 describes the basic concepts of this method, while Sect. 4 describes the details of the use case diagram generation. Section 1 evaluates this method. Finally, Sect. 6 concludes this paper.

2 Related Works

Several works support creating use case diagrams. They can be classified into three types; languages to describe use cases and requirements, generation of diagrams, and use case patterns.

For languages to describe use cases and requirements, Silva et al. proposed a pattern language to describe use cases [11]. The authors defined patterns to specify the use case type. Next they determined patterns to correspond between the use case and the main entity of a domain model. They also defined patterns to classify four types of interaction blocks, which represented whether scenario events were actor behaviors or system processes. Finally, they described patterns to classify the interaction blocks as detailed behaviors to apply Model Driven Development approaches. For another research, Aspect-oriented User Requirements Notation (AoURN) is a requirements engineering language combining goal-oriented, scenario based, and aspect-oriented [12]. User Requirements Notation (URN) [13] is extended. URN is also a requirements engineering language and consists of Goal-oriented Requirements Language (GRN) [14] for goal modeling and Use Case Maps (UCM) [15] for scenario modeling. By these researches, requirements and specific types of use cases can be described formally. However, it is difficult for stakeholders to use the notations. The method of this paper does not require any software development knowledge to stakeholders.

For generation of diagrams, A method to generate use case diagrams based on the results of analyzing problems by problem frames [16] was also proposed [17]. First, problems were analyzed by the problem frame approach, and subproblems were identified. Next, two types of formal context were created based on the Formal Concept Analysis (FCA) approach [18]. Finally, FCA concept lattices corresponding to relationships between use cases, such as include and extend, and relationships between use cases and actors were created. However, it is difficult for stakeholders to analyze by

problem frame approach. In the method of this paper, stakeholders can easily prepare operation histories of existing applications.

For use case patterns, Cruz focused on data oriented systems and proposed use case patterns [19]. In this method, use cases were classified into two types. One was “Independent use cases” that had interactions with actors and were not include and extend use cases of other use cases. The other was “Dependent use cases” that did not have interactions with actors and were include or extend use cases of other use cases. Then, five use case patterns were defined based on the typical operations for data management. Also, Ko et al. proposed a method to extract use case patterns for supporting completeness of software requirements [20]. First, this method extracts agents (actors) and verbs from event sentences of scenarios, and they are classified into high appearance group and low appearance group. Next, semantic distances between verbs are calculated, verbs in the low appearance group are linked to verbs in the high appearance group. The linked verbs become clusters. Then, use case flow graphs are generated. Finally, strength of the relationships between clusters are calculated. Use case flow patterns are extracted applying thresholds to the calculated values. Although these methods provide or extract use case patterns, support to create use cases are beyond.

3 Basic Concepts

To extract use cases, herein the operation histories that stakeholders try to use in existing applications are analyzed. The operation histories record sequences of users’ events. Events indicate operations of widgets in windows and window switching.

3.1 Event Sequence of a Function

In most applications, functions begin from the main window or top page. For a desktop application, when users select a function in the main window, the process of the function corresponding to the menu item begins, and windows related to the function are displayed. Once the function process is complete, the related windows disappear, and only the main window is displayed.

For a web application, the top page has many links and buttons. To initiate the process of the corresponding link or button, users select a link or button and web pages are switched. Once the function process is finished, the top page is displayed again. Hereafter, both main windows and top pages are referred to as “main windows”, while all other windows and web pages are called “windows”. This method targets applications that functions start from the main window.

3.2 Role of Windows

To extract use cases, the whole window sequence is constructed from all the operation histories. Concretely, event sequences in the operation histories are divided into windows. Next identical windows are merged, and the entire window sequence of an application is represented. Use cases are extracted based on the identified main window.

Thus, identical windows must be identified. Window title indicates the contents assigned to a window [21]. Consequently, window titles can be used to identify windows. Windows with the same title are identical.

In this research, the locations of titles in 15 desktop applications and 15 web applications were surveyed. For desktop applications, titles of 12 applications were located in the title bar, and an application was the top label. For web applications, titles of 10 applications were located in the title bar, and three applications were the first h1 tag. If titles were displayed in both the title bar and the h1 tag, only the title bar was counted.

Most of the titles were displayed in the title bar for desktop applications, whereas the titles were displayed in both the title bar and the h1 tags for web applications. Thus, the character strings in the title bar are identified as the window title in desktop applications. For web applications, if the character strings in the title bars differ by window, the character strings are identified as the title. On the other hand, if the character strings in the title bars of all windows are the same, the character strings in the first h1 tags of the windows are identified as the title. The titles are used as the window identifiers.

4 Use Case Diagram Generation

Use case diagrams are generated from the operation histories by five steps:

1. Divide the event sequences by windows
2. Merge windows
3. Identify the main window
4. Extract use cases
5. Generate scenarios

4.1 Divide the Event Sequences by Windows

The input and selection of a window consist of some events in a scenario, and a scenario can be represented by a sequence of windows. The operation histories record the events in the executing scenarios. In this research, the operation histories are assumed to record event sequences and window switching. Thus, the events in the operation histories are divided into windows using the entries of window switching. In this manner, the operation histories are represented by the window sequences (Fig. 1).

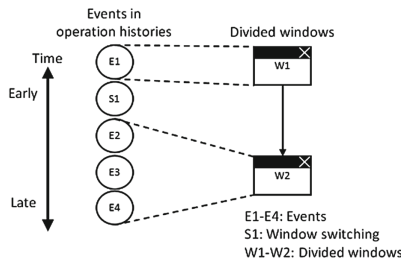


Fig. 1. Schematic of operation histories

4.2 Merge Windows

The operation histories record many scenario executions. Certain scenarios may be executed multiple times. That is, after dividing the operation histories into windows in Sect. 4.1, appearances of identical windows may exist. As described in Sect. 3.2, window titles are used as identifiers. Thus, to specify a certain window sequence of a scenario, windows with the same titles are merged in this step. Window sequences are represented as a directed graph (Window Switching Graph, WSG). In a WSG, windows and window switches are represented as nodes and edges, respectively. The number of window switches from each window is represented by an edge.

Figure 2 depicts the merging windows. The left side shows the divided windows in Sect. 4.1, while the right side is the merged window sequences. The number “2” indicates that window “W1” switches to window “W2” twice in the operation histories.

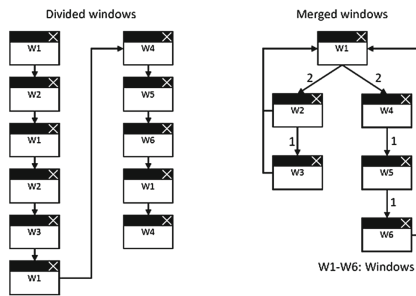


Fig. 2. Depiction of merging windows

4.3 Identify the Main Window

Because a use case consists of several scenarios, execution of a use case also starts from the main window. That is, it is necessary to identify the main window in the WSG. Window switching from the main window to another window occurs using any function, whereas window switching to other windows only occurs when using a specific function. That is, more window switches occur from the main window. Thus, the number of window switches from each individual window is calculated, and the window with the most switches is identified as the main window.

Table 1 shows the number of window switches from individual windows shown in Fig. 2. Because window “W1” has is most switches, it is identified as the main window.

Table 1. Identification of the main window

	W1	W2	W3	W4	W5	W6
Number of window switches	4	2	1	1	1	2

4.4 Extract Use Cases

After constructing a WSG, the use cases are extracted. As described in Sect. 3.1, because all functions start from the main window in this research, individual window sequences from the main window are extracted as use cases. Use cases often have precondition and include relationships [22]. Thus, three patterns are defined to extract use cases from a WSG; basic pattern, pre-condition pattern, and inclusion pattern.

Use case diagrams represent the interactions between actors, including users, and use cases. In this research, the operation histories record the interactions between applications and stakeholders. Thus, the name of the actor is “user”, and the use case diagrams are generated as the interactions between the users and the extracted use cases.

4.4.1 Basic Pattern

This is the base for all patterns to extract use cases. For each use case, the main window is the starting point, and each branch of a window sequence from the main window is extracted as a use case. Because a use case often consists of several scenarios (e.g., main scenarios, alternative scenarios, and exceptional scenarios [23]), branches of window sequences from windows except the main window are recognized as different scenario executions in a use case.

Figure 3 shows an example of a WSG to extract use cases based on basic pattern where the two branches of window sequences (“M” -> “W1” and “M” -> “W6”) are from the main window (“M”). Two use cases are extracted, such as U-B1 and U-B2. Figure 4 shows the generated use case diagram from the WSG in Fig. 3.

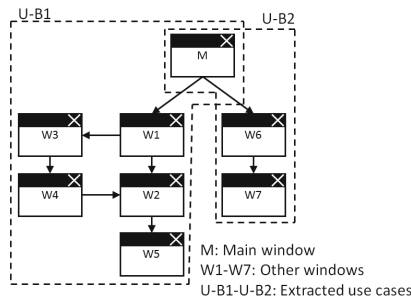


Fig. 3. Example of basic pattern

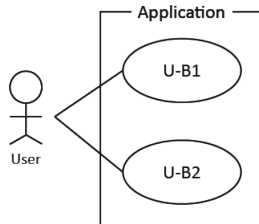


Fig. 4. Generated use case diagram from basic pattern

4.4.2 Pre-condition Pattern

Some interactions are performed before the main window is displayed. For example, in an internet banking application, users login to the application before using functions of the application. In this case, the login is always performed before the main window. Hence, completing the login is a pre-condition for functions in the internet banking application.

In this pattern, window sequences from the main window are extracted as use cases (post-use cases) based on basic pattern in Sect. 4.4.1. Window sequences before the main window are also extracted as use cases (pre-use cases). Then event sequences of the operation histories in the pre and post-use cases are analyzed. If all event sequences of the pre-use cases are executed before the post-use cases, the pre-use case is identified as a pre-condition of the post-use case.

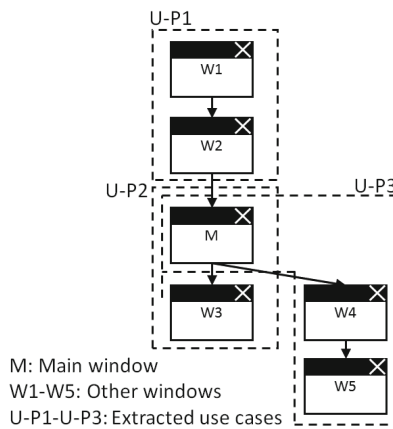


Fig. 5. Example of pre-condition pattern

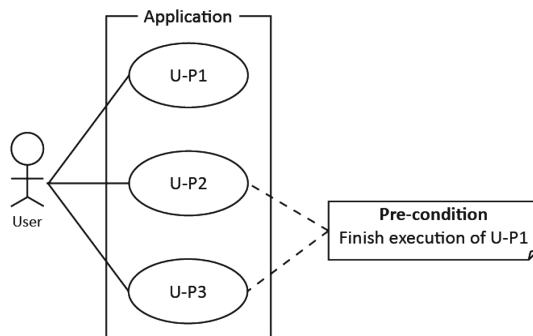


Fig. 6. Generated use case diagram from pre-condition pattern

Figure 5 shows an example of a WSG to extract the use case of a pre-condition and the extracted use cases. Because the window sequence (“W1” -> “W2”) is always displayed before the main window, it is extracted as pre-use case “U-P1”. Window

sequences after the main window “M” are extracted as post-use cases “U-P2” (“M” -> “W3”) and “U-P3” (“M” -> “W4” -> “W5”) based on the basic pattern. Then the pre-use case “U-P1” is identified as the pre-condition for post-use cases “U-P2” and “U-P3”. Figure 6 shows the generated use case diagram from the WSG in Fig. 5.

4.4.3 Inclusion Pattern

Different use cases sometimes have common event sequences, and the event sequences are often represented by dependent use cases. That is, different use cases can employ the same event sequences as dependent use cases.

Figure 7 shows an example where two use cases can be extracted based on the basic pattern in Sect. 4.4.1, such as the window sequences “M” -> “W1”-> “W2” -> “W3” -> “W4” (use case (a)) and “M” -> “W5” -> “W6” -> “W2” -> “W3” -> “W7” (use case (b)). Both use cases (a) and (b) include the common window sequence “W2” -> “W3”.

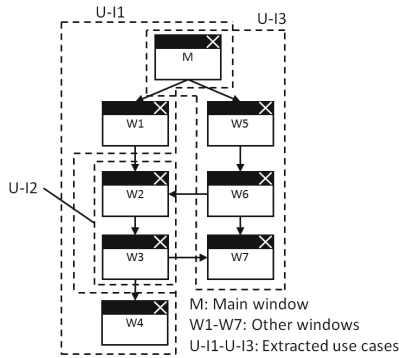


Fig. 7. Example of inclusion pattern

Thus, this common window sequence is separated from use cases (a) and (b) and extracted as a dependent use case. That is, U-I1, U-I2, and U-I3 are extracted as use cases. Use case U-I2 is included by use cases U-I1 and U-I3. Figure 8 shows the generated use case diagram from the WSG in Fig. 7.

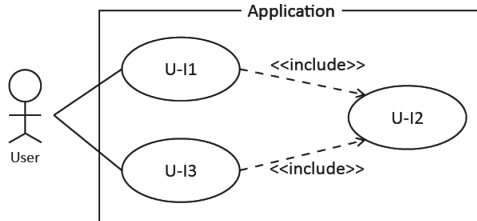


Fig. 8. Generated use case diagram from inclusion pattern

5 Evaluation

The appropriateness of the use case diagrams generation was evaluated using five applications; internet banking (App A), food delivery (App B), ticket reservation (App C), travel (App D), and e-Learning (App E). Table 2 shows the results of use case generation, where “Actual (Basic)”, “Actual (Pre)”, and “Actual (Include)” indicate number of use cases based on the extraction patterns of the basic, pre-condition, and inclusion patterns in Sect. 4.4, respectively by manual. Hereafter, the strategy specifying use cases by the author is called “actual”. “Extract (Basic)”, “Extract (Pre)”, and “Extract (Include)” indicate the number of use cases extracted by this method.

Table 2. Results of use case generation

	Actual (Basic)	Actual (Pre)	Actual (Include)	Extract (Basic)	Extract (Pre)	Extract (Include)
App A	2	1	2	2	1	2
App B	1	0	1	1	0	1
App C	3	0	0	3	0	0
App D	1	0	0	2	0	2
App E	6	1	0	6	1	0

The number of use cases between the actual and this method differed in App D. For App D, the actual use case was only one, but this method extracted two use cases for the basic pattern for hotel selection. The window titles included the area names of hotels, preventing the windows from being merged in the step of Sect. 4.2. Thus, window sequences of the hotel selection from the main window were separated by area, and they were extracted as different use cases.

In addition, after the hotel selection, login and inputting customer information were required. This sequence consisted of two windows (e.g., login and inputting customer information), and the actual use case indicated that they were part of one window sequence in the hotel selection. However, the inclusion pattern in this method extracted two use cases for this sequence. Because this sequence was executed after the two use cases, they were extracted by the inclusion pattern. In addition, once login was completed for the first hotel selection, login for a second hotel selection was not required. To select a second hotel, only the customer information must be inputted. Because both extracted use cases included window sequences that executed/did not execute login, login and inputting customer information were extracted as different use cases.

For “App E”, although the number of use cases between the actual and this method were the same, there were window sequences that this method did not use for use case extraction. In “App E”, the main window was identified as the window of functions to manage a lecture. However, the window title included the lecture name. Thus, the title of the window differed by lectures and were not merged in the step of Sect. 4.2. Consequently, window sequences of one lecture but not the other are used in the use case extraction. In the actual case, the windows were merged because they were recognized as the same window.

These results reveal some problems using window titles. Parsing a natural language is required to merge windows. If the area names of “App D” and the lecture names of “App E” can be deleted by parsing a natural language, windows that should be merged in this method can be merged. However, words should not always be deleted. Thus, strategies to delete words must be carefully considered.

Although problems merging windows due to the window titles were elucidated, other problems were not identified. Thus, except the problems, this method could be confirmed that use cases could be extracted appropriately.

6 Conclusion

Because developers difficultly define stakeholders’ requirements appropriately, it is desirable for stakeholders themselves to use case diagrams and scenarios that reflect their requirements directly. To facilitate the description of use case diagrams and scenarios, this paper proposes a generation method that uses the operation histories of existing applications.

To generate use case diagrams in our method, first, the event sequences in the operation histories are divided into windows. Then windows with the same titles are merged. Next the window with the most window switches is identified as the main window. Then the use cases are extracted based on the main window by three patterns: basic, pre-condition, and inclusion patterns. To confirm the appropriateness of the generation, five existing applications with different domains were operated. Then the use case diagrams and scenarios were generated from the operation histories. Although there were some problems, the use cases and scenarios were almost appropriately extracted and generated.

In the future strategies to merge windows to extract use cases must be considered. Because window titles included proper nouns, some windows were not merged, but should have been in the evaluations. Appropriately merging windows can realize more suitable use case extraction. To realize this, parsing window titles as a natural language could be considered.

Next, the types of users must be identified. Currently, the actors in the use case diagrams are only “users”. However, use cases that users can execute often differ by the types of user, and the use case diagrams must represent each type of user. To realize this, recording users operating the applications and then analyzing the use cases that each user operates can be considered. Additionally, naming the extracted use cases should be considered. Although the names of buttons or links for window switching from the main window to the next window can often be used as the name of the use case, this is inappropriate in some cases.

References

1. Bourque, P., Fairley, R.E.: SWEBOK V3.0 Guide to the Software Engineering Body of Knowledge. IEEE (2014)
2. Lamsweerde, A.: Requirements Engineering: From System Goals to UML Models to Software Specifications. Wiley, Hoboken (2009)
3. Wiegers, K.: Software Requirements, 3rd edn. Microsoft Press, Redmond (2013)
4. Potts, C., Takahashi, K., Anton, A.I.: Inquiry-based requirements analysis. *IEEE Softw.* **11**(2), 21–32 (1994)
5. Christel, M.G., Kang, K.C.: Issues in requirements elicitation, Technical report CMU/SEI-92-TR-12 (1992)
6. Carrizo, D., Dieste, O., Juristo, N.: Systematizing Requirements Elicitation Technique Selection. *Inf. Softw. Technol.* **56**(6), 644–669 (2014)
7. Alexander, I.F., Beus-Dukic, L.: Discovering Requirements How to Specify Products and Services. Wiley, Hoboken (2009)
8. Laplante, P.A.: Requirements Engineering for Software and Systems, 2nd edn. Auerbach Publications, Boca Raton (2013)
9. Kotonya, G., Sommerville, I.: Requirements Engineering: Processes and Techniques. Wiley, Hoboken (1998)
10. Shirogane, J.: Scenario description method based on existing software operation history. In: Proceedings of 9th International Conference on Software Technologies (ICSOFT2014) (2014)
11. Silva, A.R., Savi, D., Vlaji, S., Antovi, I., Lazarevi, S., Stanojevi, V., Mili, M.: A pattern language for use cases specification. In: Proceedings of the 20th European Conference on Pattern Languages of Programs (2015)
12. Mussbacher, G., Amyot, D., Whittle, J.: Composing goal and scenario models with the aspect-oriented user requirements notation based on syntax and semantics. Moreira, A., Chitchyan, R., Araújo, J., Rashid, A. (eds.) *Aspect-Oriented Requirements Engineering Part II*, Springer, Heidelberg (2013)
13. Z.151: User Requirements Notation (URN) - Language definition. <http://www.itu.int/rec/T-REC-Z.151/en>. Accessed 14 Oct 2016
14. GRL. <http://www.cs.toronto.edu/km/GRL/>. Accessed 14 Oct 2016
15. Buhr, R.J.A.: Use case maps as architectural entities for complex systems. *IEEE Trans. Softw. Eng.* **24**(2), 1131–1155 (1998)
16. Jackson, M.: Problem Frames Analyzing and Structuring Software Development Problem. Addison Wesley, Boston (2000)
17. Imam, A.A., Hamza, H.S., Moneim, R.A.: Automated generation of use case diagrams from problem frames using formal concept analysis. In: Proceedings of 10th International Conference on Information Technology: New Generations (2013)
18. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) *Ordered Sets*, vol. 83, pp. 445–470. Springer, Netherlands (1982)
19. Cruz, A.M.R.: A pattern language for use case modeling. In: Proceedings of the 2nd International Conference on Model-Driven Engineering and Software Development (Modelsward 2014) (2014)
20. Ko, D., Park, S., Kim, S., Hwang, M.: Automatic use case flow pattern generation using verb clustering. *Int. J. Softw. Eng. Appl.* **9**(7), 201–212 (2015)

21. Guidelines. [https://msdn.microsoft.com/ja-jp/library/windows/desktop/dn688964\(v=vs.85\).aspx](https://msdn.microsoft.com/ja-jp/library/windows/desktop/dn688964(v=vs.85).aspx). Accessed 14 Oct 2016
22. Object Management Group, OMG Unified Modeling Language TM (OMG UML) Version 2.5 (2013)
23. Schneider, G., Winters, J.P.: Applying Use Cases: A Practical Guide. Addison-Wesley, Boston (1998)

Smart Learner-Centric Learning Systems

Naseem Ibrahim^{1(✉)} and Ismail I.K. Al Ani²

¹ School of Engineering, Penn State Behrend, Erie, PA, USA
ni11@psu.edu

² The Emirates College for Management and Information Technology, Dubai, UAE
iialani@ecmit.ac.ae

Abstract. This paper is concerned with knowledge delivery in learning systems. A learning system is a system through which learners can obtain knowledge. Providers deliver the knowledge in the way they decide is most appropriate. With the wide popularity of e-learning, learners can obtain knowledge from any source in any location in the world. On the other hand, each learner has his/her own learning style(s). But current learning systems are provider oriented. We believe that this is not sufficient. Hence, this paper introduces a smart learner-centric architecture. Smart in the sense that it allows the learner to decide the source of the knowledge he/she is requiring depending on his/her preferred learning style(s). Learner-centric is in the sense that knowledge providers publish their knowledge in a rich definition that specifies the used learning style(s). The architecture allows knowledge requesters to control the source of the knowledge and the learning style used to deliver the knowledge. The proposed architecture is an extension of traditional service-oriented architectures. It extends the definition of traditional service by adding context.

Keywords: Learning systems · Smart learners · Learning styles · Context · Service-oriented architecture

1 Introduction

It is essential to distinguish between learning and teaching. Teaching is usually associated with students in schools, colleges or universities. On the other hand, learning does not have to be associated with a learning institution. A learner does not have to be student; every single individual can be a learner. A driver who is looking for direction is a learner. A parent who is looking for information on available schools is a learner. A tourist who is looking for information on attractions in a city is a learner. A student who is doing a research on “Search Algorithms” is also a learner. Each of these learners has different needs and abilities.

Different learning styles exist [2]. Each learner prefers a single or combination of learning styles to comprehend information. Each learning style also has a set of techniques that improves the speed and quality of the learning process.

There are unlimited methods for obtaining knowledge by a learner, where each method represents a learning style. Current learning systems can provide a subset of these methods. We define a learning system to be a system through which a learner can

obtain knowledge. For example Web sites such as Triptadvisor provide a list of attractions with a simple description and a gallery of pictures. The available systems are provider oriented, in the sense that they do not adapt to the different needs of learners, specifically their preferred learning style. They can provide a specific amount of knowledge and they make this knowledge available regardless of the requirements of the learners.

We believe that such systems are not sufficient. To be competitive in the current market where a learner can obtain knowledge from any location over the world, the learning system should be able to provide specific knowledge that is both adaptive and specific to the learner requirements.

On the other hand, learners in this current age are much smarter. They know the knowledge they are looking for and the learning style(s) that is most appropriate for them. Current learning systems are provider oriented, in the sense that they do not consider the smartness of the knowledge requester. They make assumptions on behalf of the learner.

To obtain knowledge, the learner searches for available providers and then a communication happens between the learner and the provider. This interaction closely resembles the interactions in Service-oriented architectures (SOA). The learner represents the service requester, the knowledge provider represents the service provider, and the place where the user searches for knowledge represents the service registry. Hence, in this paper we propose an SOA based architecture for learner oriented learning.

Traditional SOA based architecture [5, 6] focuses on functionality. Functionality drives service publication and discovery. Knowledge can be represented as functionality. But to support learning oriented learning, a much richer SOA architecture is necessary. Hence, in this paper we propose extending traditional service-oriented architectures by including context as a first-class element. Context will be used in the publication and discovery of services. Context can include any information including the learning style and associated techniques.

The rest of this paper is structured as follows. Section 2 briefly introduces learning styles. Section 3 discusses related work. Section 4 introduces the smart learner-centric architecture. Section 5 presents an example. Finally, Sect. 6 presents some concluding remarks and discussion of future work.

2 Learning Styles

Learning styles includes [2] but are not limited to: Visual (spatial), Aural (auditory-musical), Verbal (linguistic), Physical (kinesthetic), Logical (mathematical), Social (interpersonal), and Solitary (intrapersonal). For each learning style the learner prefers a set of specific method to understand information. In Visual, the learner prefers using pictures, images and spatial understanding. In Aural, the learner prefers using sound and music. In Verbal, the learner prefers using words both in speech and writing. In Physical, the learner prefers using body, hands and sense of touch. In Logical, the learner prefers using logic, reasoning and systems. In Social, the learner prefers to learn in groups or with other people. And in Solitary, the learner prefers to work alone and use self-study.

To illustrate the different needs and abilities of learners let's take as an example a tourist looking for points of interest in New York.

- The tourist might prefer a map with pinpoints of all attractions in downtown New York; “Visual”
- The tourist might prefer a list of all family friendly attractions in New York; “Verbal”
- The tourist might prefer a gallery of pictures for a specific attraction; “Visual”
- The tourist might prefer a short movie of one or multiple attractions that are family friendly and in downtown New York.; “Aural”
- The tourist might prefer to participate in activities with others to understand New York culture; “Social”
- The tourist might prefer to discover downtown New York by him/her-self; “Solitary” and “Physical”
- The tourist might prefer using a tourist-agency to arrange different tours to historical places; “Social” and “logical”

The information required in the above example can be elucidated in single/multiple styles of learning. To be more specific each of the requirements is associated with one or more of the seven learning styles proposed above. It is very clear that the same learner may use different styles of learning to get the desired information.

Current learning systems can provide a subset of these methods. We believe that such systems are not sufficient. Learning systems should enable knowledge requesters to select their preferred learning style(s).

3 Related Work

Research in learning and knowledge delivery goes back hundreds of years. But recently, with the significant advancement in technology, research interest has dramatically increased in learning and especially online learning [8–12]. The literature is rich with approaches for knowledge delivery to learners; be it a consumer, a marketer, a student or any sort of learner. These approaches can be categorized in the way they consider context into: (1) Location based; where they consider the geographical position of the learner in a variety of contexts such as health, work, personal life or any other sort of context, (2) Consumer centric; where they are concerned with taking consumer insights and ensuring that brands can act on them, deliver their communications, their products and services to focus on those insights, which means it is ideally adapting to the lifestyle, attitudinal and behavioral patterns of the target consumer in all contexts, (3) User generated context and marketing; where users are participating, developing, writing, and publishing the context. Web 2.0 is a good example that demonstrates the use of a browser and an internet connection to publish any content, i.e. the user act as a provider not learner, (4) Another approach that uses Cloud Computing as a supporting environment in sustainable development in higher education has also emerged [1].

All of these approaches are provider centric, in the sense that they concentrate on (1) provider oriented content and (2) provider context delivery. Different supporting

software is available today to help in creating a learning service by taking screenshots, record screen-casts and collaborating. Semantic web languages are used to express a crude meaning of the page and then use it to organize and filter data to meet the user needs [7] such as XML [3], RDF [17], OWL [18], and XSL [19].

To remedy the shortcomings of available approaches we suggest a new model that starts with the use of service-oriented architecture (SOA). We first extend the current SOA model to take into consideration the different requirements and abilities of the knowledge requesters. Then we use the extended SOA model, to publish, discover and provide knowledge that is learner-centric.

4 Smart Learner-Centric Architecture

This section introduces a Smart Learner-centric Architecture. Smart is in the sense that learners are able to make the smart selection of the preferred learning style(s). Learner-centric is in the sense that knowledge is represented to support multiple learning styles. Hence, the three main goals of this architecture is to: (1) enable knowledge learners to select the way knowledge is delivered to them, (2) enable knowledge providers to publish knowledge in different formats that support different learning styles, and (3) enable the matching between knowledge requesters needs and knowledge providers’ knowledge.

To support the above goals this paper proposes a SOA-based architecture. This is an extension of the work presented in [13–16]. In traditional SOA, a service provider publishes a service definition in the service registry. The service requester searches and selects from the services published in the service registry. After selecting a service, the service requester interacts with the service provider by sending requests and receiving responses.

Knowledge can be represented as a service. But traditional definition of services depends on functionality. Service providers publish service functionalities and service requesters search for service functionalities. This is not enough for this architecture. Hence, this section introduces an extended SOA model that supports smart learner-centric learning.

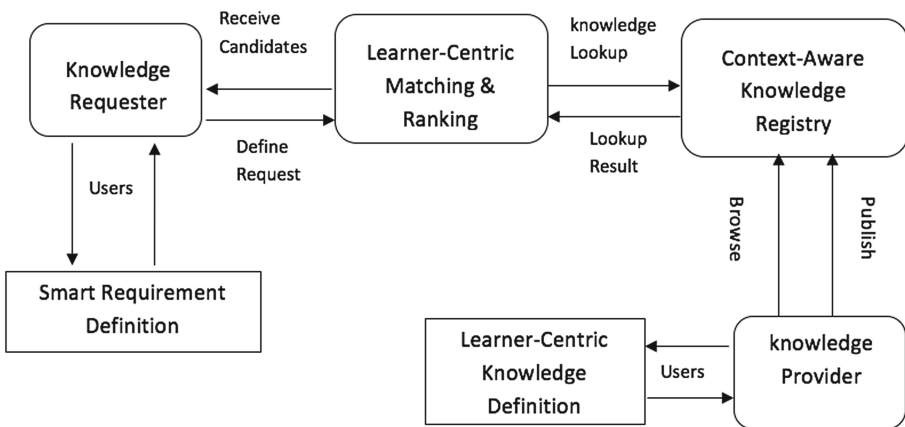


Fig. 1. Smart learning-centric architecture

The extended SOA is illustration in Fig. 1. The architecture introduces the following elements:

- **Knowledge Requester:** It is the learner that is requiring knowledge. It represents the service requester. It is the client side of the interaction. It can also be an application or another service.
- **Knowledge Provider:** It is the entity that provides knowledge. It is responsible for representing knowledge in different learning styles. It represents service providers.
- **Smart Requirement Definition:** In this architecture, a knowledge requester is able to specifically define and list all the requirements and abilities in a rich definition. These requirements and abilities are defined and stated using smart requirements definition.
- **Learner-Centric Knowledge Definition:** Knowledge providers define and publish their knowledge using this entity. It resembles service definitions. Traditional SOA service definitions such as WSDL [4] definitions focuses on service functionalities and some nonfunctional properties. Our definition of it is much richer. The service definition contains enough information to facilitate a better matching between knowledge requester requirements and knowledge provider services. This is achieved by introducing context as a first class element in a knowledge definition. This context information is learner-centric. It lists the different learner styles that it can meet.
- **Context-aware Knowledge Registry:** It represents the service registry. Traditional service registries publish services focusing on service functionalities. In this architecture, the published definition is much richer.
- **Learner-Centric Matching and Ranking:** This unit is responsible of two main roles. First, it matches the requirements of the knowledge requester to the available knowledge definition in the the context-aware knowledge registry. Second, it ranks the candidates according the requester requirements.

The interactions in the newly introduced architecture can be summarized as follows:

- **Defining knowledge:** Knowledge providers use the learner-centric knowledge definition to specify information about the knowledge. This information will be made public.
- **Publish knowledge:** Knowledge providers publish the knowledge definition through the context-aware knowledge registry.
- **Defining requirements:** The knowledge learner defines his requirements using the smart requirements definition. The definition of the requirements will include the required learning styles. A priority will be included in the definition to guide the matching and ranking.
- **Request knowledge:** The learner will pass the requirements definition to the learner-centric matching & ranking unit.
- **Matching and ranking:** The learner-centric matching & ranking unit will match the knowledge request with available knowledge definition in the context-aware knowledge registry. It will rank the candidates while considering the required learning styles and the assigned priorities.

5 Example

The architecture proposed is independent of implementation. It can be used globally for any type of knowledge in any domain or in a specific domain. An example would be the domain of higher education. The implementation of the architecture would provide a platform where Universities can publish knowledge. That knowledge can be complete courses or a specific topic that belongs to a course. The published courses or topics will be published with associated learning style. The learner (student), will be able to use this platform to find the course or topic that is delivered in his/her preferred learning style(s).

To clarify the use of the architecture, let us consider as an example a subject called “Data Structure and Algorithm Design” from the field of Computer Science. Let us assume that the learner would like to cover the topic “Learn the primitive term ‘type of graphs’ and the terminology for representing a graph”. The learner can use the Learner-centric Requirement Definition to define his/her request. Then the Learner-centric Knowledge Matching & Ranking Unit searches the Context-aware Knowledge Registry for a match to that request. Obviously the requested knowledge may have documented in different styles. Knowledge providers have already published specifications of their knowledge with associated learning styles in the Registry. The matching and ranking unit will match the available registry entries with the student requirements to get the best match. In many cases there might not be a specific match, so a ranking will be necessary.

The knowledge might take different forms that belong to different learning styles.

The primitives may take a type of definitions, explanations, theories or/and analysis. Definitions may take a form of text, audio or/and video. Explanation may take a form of figures, drawing, diagrams or/and discussion through round table or peer- to-peer. Theory may take a form of simulation, step-by-step solution or algorithm, video or/and drawing. Analysis may take a form of statistical charts, data, information or/and graphs.

Accordingly, an interaction process wherein the learner may require different representation styles of knowledge which may be transferred to the learner as the appropriate knowledge.

6 Conclusion and Future Work

The extended SOA architecture represents a new approach to structure the relationship between knowledge requesters and knowledge providers. This approach is smart and learner-centric. Smart is by allowing knowledge requesters to have an input in their preferred learning style(s) and delivery method. Learner-centric in that it forces knowledge providers to consider learning styles and delivery methods in the way they define and deliver knowledge. It is also noticeable that the model has given the freedom to the provider to include any extra component demanded by the requester without affecting the structure of the model or even its main concept.

The educational example has demonstrated the concept of learner-smartness in the learner context request. It is very clear that the request of the smart learner has been

reflected in the learner context of the selection process. It also shows the usefulness of both knowledge requisition and knowledge acquisition to both of the knowledge provider and the knowledge requester.

Moreover, the architecture has demonstrated its power of granting learners with the privilege of selecting one or more alternatives styles of investigation and additionally allowing the learner to follow the one that seems suitable to his/her context.

Our future work includes extending the architecture to support: learner motivation and learning strategies. We are also working on a complete implementation of the architecture.

References

1. Idowu, S.A., Osofisan, A.O.: Cloud computing and sustainable development in higher education. *J. Emerging Trends Comput. Inform. Sci.* **3**(11), 1466–1471 (2012)
2. Pashler, H., McDaniel, M., Rohrer, D., Bojork, R.: Learning styles: concepts and evidence. *Psychol. Sci. Public Interest* **9**(3), 105–119 (2008)
3. Tidwell, D.: *Mastering XML Transformation XSLT*. O'Reilly, Sebastopol (2001)
4. WSDL, Web services description language 1.1, W3C Note. March 2001. <http://www.e3.org/TR/wsdl> (2001)
5. Dan, X., Shi, Y., Tao, Z., Xiang-Yang, J., Zao-Oing, L., Jun-Feng, Y.: An approach for describing soa. In: *International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM* (2006)
6. ter Beek, M.H., Gnesi, A.S., Koch, N., Mazzanti, F.: Formal verification of an automotive scenario in service-oriented computing. In: *Proceedings of the 30th international conference on software engineering, ser. ICSE 2008*. ACM, New York (2008)
7. Ogbu, U.: *The Language of the Semantic Web*. New Architect (2002)
8. Muna, S., Hatem, H., Ramadan, A., Neagu, D.C.: e-Learning based on context oriented semantic web. *J. Comput. Sci.* **1**(4), 500–504 (2005)
9. Bonk, C.J., Khoo, E.: *Adding some TEC-Variety: 100 + activities for motivating and retaining learners online* (2014)
10. Gagne, R.M., Briggs, L.J., Wager, W.F.: *Principles of Instructional Design*. Wadsworth, Belmont (1985)
11. Keller, J.M., Suzuki, K.: Use of the ARCS motivation model in courseware design. In: Jonassen, D.H. (ed.) *Institutional designs for microcomputer courseware*. Lawrence Erlbaum, Hillsdale, NJ (1988)
12. Brinthaup, T.M., Fisher, L.S., Gardner, J.G., Raffo, D.M., Woodard, J.B.: What the best online teachers should do. *MERLOT J. Online Learn. Teach.* **7**(4) (2011)
13. Ibrahim, N., Mohammad, M., Alagar, V.: Publishing and discovering context-dependent services. *Human-Centric Comput. Inform. Sci.* **3**(1), 1–22 (2013). Springer
14. Ibrahim, N., Al Ani, I.I.K.: An architecture for providing and defining student-oriented services. In: *The 3rd IEEE International Conference on Current Trends in Information Technology* (2013)
15. Ibrahim, N., Al Ani, I.I.K.: Domain ontology for trustworthy context-dependent services. In: *The 5th IEEE International Conference on Computer Science and Information Technology (CSIT 2013)*. IEEE Xplore (2013)
16. Ibrahim, N., Al Ani, I.I.K.: Languages for the publication and discovery of context-dependent services. In: *The International Conference on Computing, Networking and Digital Technologies (ICCNDDT 12)* (2012)

17. Powers, S.: Practical RDF. O'Reilly & Associates Inc., Sebastopol (2013)
18. Antoniou, G., van Harmelen, F.: Web Ontology Language OWL. *International Handbooks on Information Systems*, pp. 67–92. Springer, Heidelberg (2004)
19. Fitzgerald, M.: Learning XSLT. O'Reilly & Associates Inc., Sebastopol (2003)

Prioritized Process Test: More Efficiency in Testing of Business Processes and Workflows

Miroslav Bures^{1(✉)}, Tomas Cerny^{2(✉)}, and Matej Klima¹

¹ Department of Computer Science, FEE, Czech Technical University in Prague,
Karlovo Namesti 13, 121 35 Prague, Czech Republic

miroslav.bures@fel.cvut.cz

² Department of Computer Science, Baylor University, Waco, TX 76798, USA
tomas_cerny@baylor.edu

Abstract. Testing business processes and workflows in information systems, while aiming to cover all possible paths, requires high efforts demanding considerable costs. In this paper, we propose an algorithm generating a path-based test cases from the system model, based on weighted directed graph. The approach brings an alternative to the currently established test requirements concept. The algorithm reflects various levels of priorities of particular functions in the tested system, previously defined by the test designer. When compared to simulated naive approaches based on reverse reduction of test set, our proposed algorithm produces more efficient test cases in terms of number of the total test steps, whilst keeping the same level of test coverage of the priority functions of the tested system.

1 Introduction

Testing business processes and workflows of information systems represents significant portion of the overall test efforts. The efficiency of testing strongly depends on creation of consistent and efficient test cases. This implies the needs for an underlying model of the System Under Test (SUT). For the majority of information system processes documented by UML Activity Diagrams, a variant of directed graph is usually used as such an underlying model. Multiple algorithms for generation of paths-based test cases have been proposed and compared [1–6].

The particular intensity of testing leads into a sequence of steps that need to be executed as well as to efforts related to the tests execution. Intensive techniques based on prime paths are excellent measure to test critical systems with an adequate test coverage. Despite the previous efforts to minimize produced test cases, whilst keeping defined test coverage [2–4] for non-critical systems, these techniques lead into a set of test cases, which would be overly extensive to execute in the economic scope of a given project. Naive reduction of such test set can lead to uncontrolled reduction of the coverage and to overall inefficiency of the process. Systematic prioritization and reduction of the test set is desired. Recently, several strategies to prioritize the path-based test cases have been explored [4–6]. Nevertheless, due to frequent demands of the current ICT industry, we believe that this area deserves to be explored further.

In this paper, we present the Prioritized Process Test (PPT), addressing the described problem. The technique works with explicit prioritization of the individual parts of the tested workflow and generates test cases with less steps, while keeping high test coverage of workflow priority parts.

For the purpose of this paper, we define the SUT model as the directed graph $G = (D, A)$, where D is a set of decision points (graph nodes), $D \neq \emptyset$, and A is a set of actions in the SUT (graph edges). A is a subset of $D \times D$. In the model, one start decision point (initial node) $d_s \in D$ is defined. $D_e \subseteq D$ is a set of end decision points (final nodes of the graph), $D_e \neq \emptyset$.

The action a is an abstraction for either

- (1) one physical step in the SUT, or
- (2) a linear sequence of more physical steps in the SUT without a possibility to select an alternative way (no decision point is implemented in this sequence of physical steps).

The test case t is a sequence of decision points (nodes) d_1, d_2, \dots, d_n with a sequence of actions (edges) a_1, a_2, \dots, a_{n-1} , where $a_i = (d_i, d_{i+1})$, $a_i \in A$, $d_i \in D$. Further on, $d_1 = d_s$, and $d_n \in D_e$. When these conditions are satisfied, we can further denote the test case as a sequence of actions a_1, \dots, a_{n-1} . The test set T is a set of the test cases.

This paper is organized as follows. In Sect. 2 we summarize the related work. The Prioritized Process Test technique is introduced in Sect. 3. The Sect. 4 presents and discusses results of the experiments. In the last section we conclude the paper.

2 Related Work

In the current research on prime paths testing techniques, the common goal is to find algorithms that minimize the produced test cases in terms of different criteria as number of nodes, edges or paths [2, 3]. Test requirements concept is used in the majority of published work [1]. Nevertheless, from our observations, definition of test requirements, which must be covered in the result test cases could not satisfy practical demands for a prioritization of particular workflow parts. Usually, more levels are used in practical prioritization [7] - in contrast to the test requirement concept, by which practically only one level of priority can be set.

A number of alternative strategies to prioritize the path-based test cases have been addressed. An example approaches base on neural network clustering [4], fuzzy clustering [5] or Firefly optimization algorithm [6]. In these prioritization techniques, information about internal structure of the system under test is used. Still, the problem of production of efficient prioritized paths-based test cases (being one of the major and the mostly used testing techniques) deserves further attention and exploration of new alternative approaches.

In our definition of Prioritized Process Test (PPT), we use the Test Depth Level (TDL) criterion concept of the Process Cycle Test (PCT) technique, as defined in TMap *Next* [8]. The PCT uses G as SUT model and produces test cases t (as defined in the Introduction). In our previous work, we implemented this algorithm in the PCTgen platform [9].

In PCT, the TDL criterion is defined as:

- (1) $TDL = 1$ if $\forall a \in A$ the action a occurs at least once in at least one test case $t \in T$.
- (2) $TDL = n$ if the following conditions are satisfied: For each decision point $d \in D$, the S_d is a set of all possible paths in G starting with an edge (action) incoming to the decision point d , followed by a sequence of $(n-1)$ edges outgoing from the decision point d .

Then, $\forall d \in D$, the test cases of the test set T contain all paths from S_d .

Nevertheless, importance or priorities of individual functionalities of the SUT are not reflected in this technique. By TDL, we can set only overall test coverage level for particular G .

3 Prioritized Process Test

Prioritized Process Test (PPT) is a technique that generates test cases focused to cover the priority parts of the workflows with high intensity and deliberately cover the non-priority parts with lower intensity. First, we extend the SUT model to include prioritization. Next, we define coverage criteria for the PPT technique. Finally, we present the PPT algorithm, which generates the test cases.

The SUT model extension includes the priorities, determining the intensity of produced test cases. We extend G to weighted graph: for each action $a \in A$, a priority p is defined. In our model, $p \in \{high, medium, low\}$. When priority is not defined, it is considered as low. Then, A_h is a set of high priority actions, A_m is a set of medium priority actions and A_l is a set of low priority actions, $A_h \cup A_m \cup A_l = A$, $A_h \cap A_m = \emptyset$, $A_m \cap A_l = \emptyset$, $A_h \cap A_l = \emptyset$.

When modelling the SUT, the priorities are determined by test analysts. Various techniques and approaches can be used, such as Product Risk Analysis (PRA) presented in the BDTM approach [10] or others [7].

To determine the intensity of testing in individual parts of the SUT, the Prioritized Process Test uses two concurrent coverage criteria: (1) the Prioritized Test Level (PTL), which we define further on, and (2) Test Depth Level (TDL) of Process Cycle Test, which is used in a modified way in comparison to the original technique [8].

Prioritized Test Level (PTL) can be set to values $\{high, medium\}$ and is defined as:

- (1) $PTL = high$ if $\forall a \in A_h$ the action a occurs at least once in at least one test case $t \in T$.
- (2) $PTL = medium$ if $\forall a \in A_h \cup A_m$ the action a occurs at least once in at least one test case $t \in T$.

Table 1. Specification of TDL for Prioritized Process Test.

Coverage Criteria	PTL = <i>high</i>	PTL = <i>medium</i>
TDL = 1	$P = A_h$	$P = A_h \cup A_m$
TDL = n , $n > 1$	$P =$ set of all paths identified in G by TDL = n criterion of PCT, which start with any of $a \in A_h$	$P =$ set of all paths identified in G by TDL = n criterion of PCT, which start with any of $a \in A_h \cup A_m$

Next, to determine the test coverage, the Test Depth Level (TDL) criterion is used as specified in Table 1. Let P is a set of paths in G , which must exist in the test cases of the test set T to satisfy the coverage criteria. These paths have length 1 for TDL = 1.

The PPT test cases are generated by the Algorithm 1. Model of the SUT G and selected TDL and PTL values are input to the algorithm. Test set T is the output.

Algorithm 1. GenerateTestCases (G, TDL, PTL)	Output: Test set T
$T := \emptyset; P := \emptyset; ALLTDL := \emptyset$; Initialize new stack S ; depth := TDL	
For (each $d \in D$) do	
$ALLTDL := ALLTDL \cup \text{GetAllTDLPathsForNode}(d, \text{depth}, ALLTDL, S)$	
End for	
$P := \text{SelectRelevantTDLPaths}(D, ALLTDL, TDL, PTL)$	
$ALLE2E := \{ z \mid z \text{ is path in } G \text{ starting with decision point } d_e \in D \text{ and ending with any decision point } d_e \in D_e \text{ and there exist a path } p \in P \text{ such that } p \text{ is subpath of } z \}$	
For (each $p \in P$) do	
add p to indexed table PTAB, p is indexed by the second decision point from p	
End for	
$T := \text{CreateTestCases}(PTAB, ALLE2E)$	
GetAllTDLPathsForNode (d, depth, ALL, S)	Output: $ALLTDL$
depth := depth-1	
If depth < 0 then Create new path from edges in S and add it to ALL ; End if	
$O :=$ outgoing edges from d	
For (each $o \in O$) do	
Push o to stack S ; $d_o :=$ decision point at the end of edge o	
$\text{GetAllTDLPathsForNode}(d, \text{depth}, ALLTDL, S)$; Remove o from stack S	
End for	
If stack S is empty then return $ALLTDL$; End if	
SelectRelevantTDLPaths ($D, ALLTDL, TDL, PTL$)	Output: P
For (each $c \in ALLTDL$) do	
$e :=$ the first edge of c	
If ((PTL=high and priority of e is high) or (PTL=medium and priority of e is high or medium)) then $P := P \cup \{c\}$	
End for	
If (TDL > 1) then	
For (each $a \in A$) do	
If (a is not contained in any path of P) then	
If ((PTL=high and priority of a is high) or (PTL=medium and priority of a is high or medium)) then $P := P \cup \{c\}$	
End if	
End if	
End for	
End if	
CreateTestCases (PTAB, ALLE2E)	Output: T

```

T := ∅
While (PTAB contain any elements) do
    b := SelectBestE2EPath(PTAB, ALLE2E)
    ALLE2E := RemoveUnnecessaryE2EPaths(PTAB, ALLE2E)
    PTAB := RemoveUsedTDLPaths(PTAB, b)
    T := T ∪ b
End While

```

SelectBestE2EPath(PTAB, ALLE2E) **Output:** bestE2EPath

```

bestE2EPath := ∅; bestScore := 0
For (each x ∈ ALLE2E) do
    score := 0
    For (each key k from PTAB) do
        Pk := set of all paths for key k
        For (each p ∈ Pk) do
            If (p is subpath of x) then score := score + 1; End if
        End for
    If (score > bestScore) then bestScore := score; bestE2EPath := { p } End if
End for

```

RemoveUnnecessaryE2EPaths(PTAB, ALLE2E) **Output:** ALLE2E

```

For (each x ∈ ALLE2E) do
    score := 0
    For (each key k from PTAB) do
        Pk := set of all paths for key k
        For (each p ∈ Pk) do
            If (p is subpath of x) then score := score + 1; End if
        End for
    If (score = 0) then ALLE2E := ALLE2E - { x } End if
End for

```

RemoveUsedTDLPaths(PTAB, b) **Output:** PTAB

```

For (each key k from PTAB) do
    Pk := set of all paths for key k
    For (each p ∈ Pk) do
        If (p is subpath of b) then Pk = Pk - { p }; Remove p from PTAB; End if
    End for
    If (Pk = ∅) then Remove key k from PTAB; End if
End for

```

4 Experiments

To test functionality and efficiency of proposed PPT algorithm, we implement the algorithm and create a set of 50 testing models of SUT (G weighted by defined priorities of the actions). Using these testing models, we compare:

- (1) Test cases produced by preliminary implementation of PCT [9],
- (2) Set of PCT test cases, which has been reduced by removing of all test cases not containing:

- a. any action with priority *high* for $PTL = high$ (further denoted as **DCT(h)**),
 - b. any action with priority *high* and *medium* for $PTL = medium$ (further denoted as **DCT(m)**), and
- (3) Test cases produced by the PPT, which is defined above (further denoted as **PPT(h)** for $PTL = high$ and **PPT(m)** for $PTL = medium$).

In these experiments, we compare all relevant combinations of the coverage levels $TDL = \{1, 2, 3\}$ and $PTL = \{high, medium\}$.

For the algorithm comparison we use specially modified version of the PCTgen framework [9]. Results for 11 randomly selected SUT models and $TDL = 2$ are presented in Table 2. The complete result set is beyond the scope of this paper and can be provided on demand.

In Table 2, $|D|$, $|A|$, $|A_h|$, $|A_m|$ and $|A_l|$ describes properties of particular SUT model G (refer to the definitions above), *cycles* denote the number of cycles in G ,

$|T|$ denotes number of produced test cases,

α denotes total number of actions (graph edges) in test set T ,

β denotes number of unique actions which are contained in the test set T ,

$ec = (\beta / |A|)$ is ratio of unique actions contained in the test set T in percentage.

For PCT, $ec = 100\%$ by principle of the technique.

$\Delta\alpha = (\alpha \text{ for particular technique} / \alpha \text{ PCT})$ in percentage,

α_h denotes total number of actions of priority *high* in test set T ,

α_m denotes total number of actions of priority *high* and *medium* in test set T ,

$\lambda_h = (\alpha_h/\alpha)$ is “priority path testing efficiency” metric for *high* priority actions for particular testing technique. By analogy, $\lambda_m = (\alpha_m/\alpha)$ for *high* and *medium* priority actions.

When test cases were created, we verified their consistency: all of the produced test cases, all *high* priority actions have been covered by **DCT(h)** and **PPT(h)** and all *high* and *medium* priority actions have been covered by **DCT(m)** and **PPT(m)** for all SUT models. Further on, all of the test cases were starting in d_s and ending in a $d_n \in D_e$. All of the test cases represented a valid path in G .

The values β for individual techniques presented in Table 2 indicate, that PPT has a weaker overall coverage than DCT and PCT in terms of unique actions, which are contained in the test set T . Here, please note that the coverage was reduced only for non-priority actions. All priority actions were covered by the produced PPT test cases: high priority actions were covered for $PTL = high$ and high and medium priority actions were covered for $PTL = medium$. Nevertheless, as documented by $\Delta\alpha$ and λ_h (or λ_m) values, mentioned decrease in coverage of unique low-priority actions was amply compensated by significant decrease in number of total steps of test cases produced by PPT. This would lead to significantly less resource demands during practical application of the test cases. For PPT test cases, TDL test coverage was kept in priority parts of G : TDL criterion was kept for the paths from P containing high (or high and medium for $PTL = medium$) priority actions.

The results show that PPT produces consistent test cases suitable for lower intensity tests, regression testing and smoke tests, directly focusing on priority parts of business workflows in information systems. TDL test coverage is kept in these priority parts and

Table 2. Properties of selected experimental models of SUT and comparison of test cases produced by PCT, reduced PCT and PPT techniques

Value	1	2	3	3	4	5	7	8	9	10	11
D	25	11	10	11	22	26	30	35	40	45	50
A	27	16	13	15	27	38	45	48	54	61	74
A _h	3	1	1	2	5	6	8	5	7	10	13
A _m	5	2	4	2	7	3	4	9	6	9	6
A _l	19	13	8	11	15	29	33	34	41	42	55
<i>cycles</i>	0	0	0	4	0	3	5	5	0	4	7
T PCT	17	8	9	6	12	14	11	25	23	17	19
α PCT	69	53	38	52	97	94	200	171	197	352	292
T DCT(h)	14	3	2	6	11	11	11	14	13	16	19
α DCT(h)	61	21	7	52	92	83	200	121	132	350	292
β DCT(h)	22	13	6	15	25	37	45	37	42	61	74
<i>ec</i> DCT(h) (%)	81,5	81,3	46,2	100	92,6	97,4	100	77,1	77,8	100	100
Δα DCT(h) (%)	88,4	39,6	18,4	100	94,8	88,3	100	70,8	67	99	100
α _h DCT(h)	19	3	2	10	21	14	26	19	8	71	49
λ _h DCT(h)	0,31	0,14	0,29	0,19	0,23	0,17	0,13	0,16	0,06	0,20	0,17
T PPT(h)	4	1	1	3	4	7	4	5	6	2	6
α PPT(h)	19	6	3	18	36	39	55	40	33	58	112
β PPT(h)	13	6	3	9	17	26	29	25	24	44	61
<i>ec</i> PPT(h) (%)	48,1	37,5	23,1	60	63	68,4	64,4	52,1	44,4	72,1	82,4
Δα PPT(h) (%)	27,5	11,3	7,9	34,6	37,1	41,5	27,5	23,4	16,8	16,5	38,4
α _h PPT(h)	8	1	1	5	10	9	16	8	9	13	24
λ _h PPT(h)	0,42	0,17	0,33	0,28	0,28	0,23	0,29	0,20	0,27	0,22	0,21
T DCT(m)	11	8	9	6	11	13	11	23	20	16	19
α DCT(m)	69	53	38	52	97	91	200	163	184	350	292
β DCT(m)	27	16	13	15	27	38	45	47	50	61	74
<i>ec</i> DCT(m) (%)	100	100	100	100	100	100	100	97,9	92,6	100	100
Δα DCT(m) (%)	100	100	100	100	100	96,8	100	95,3	93,4	99,4	100
α _m DCT(m)	27	11	18	16	32	23	36	56	42	124	72
λ _m DCT(m)	0,39	0,21	0,47	0,31	0,33	0,25	0,18	0,34	0,23	0,35	0,25
T PPT(m)	8	2	7	3	10	8	5	13	10	6	9
α PPT(m)	32	11	27	29	78	46	69	81	75	140	135
β PPT(m)	19	8	12	15	25	32	35	37	40	55	63
<i>ec</i> PPT(m) (%)	70,4	50%	92,3	100	92,6	84,2	77,8	77,1	74,1	90,2	85,1
Δα PPT(m) (%)	46,4	20,8	71,1	55,8	80,4	48,9	34,5	47,4	38,1	39,8	46,2
α _m PPT(m)	17	3	15	9	27	16	25	32	25	43	40
λ _m PPT(m)	0,53	0,27	0,56	0,31	0,35	0,35	0,36	0,40	0,33	0,31	0,30

total number of test steps is significantly reduced compared to PCT or alternatively a naive test set reduction approach simulated by DCT.

5 Conclusion

In the paper we proposed the Prioritized Process Test algorithm generating the paths-based test cases from SUT model abstracted as weighted directed graph, where more levels of priority can be defined for particular SUT actions. This approach represents an alternative to currently established test requirements concept, as in this concept practically only one level of priority can be set.

Compared to naive approach (which we simulated as DCT), proposed PPT produces consistent, but much more economic test cases exercising priority actions of SUT. The PPT(h) has reduced the number of test steps by 53,1% averaged for all presented SUT instances. The price for this optimization was a decrease of number of unique low-priority actions which were contained in the test set. For the PPT(h), average of this decrease for all presented SUT instances was 30,7%. Introduction of more priority levels gives the proposed technique more flexibility in scaling of testing intensity than standard concept of test requirements. In the future work we will elaborate the PPT algorithm to work more intensely with this prioritization leveling and, thus, to present more significant alternative to currently published approaches.

This research is conducted as a part of the project TACR TH02010296 Quality Assurance System for Internet of Things Technology.

References

1. Yoo, S., Harman, M.: Regression testing minimization, selection and prioritization: a survey. *Softw. Test. Verification Reliab.* **22**(2), 67–120 (2010). Willey
2. Dwarakanath, A., Jankiti, A.: Minimum number of test paths for prime path and other structural coverage criteria. In: Merayo, M.G., Oca, E.M. (eds.) *ICTSS 2014*. LNCS, vol. 8763, pp. 63–79. Springer, Heidelberg (2014). doi:[10.1007/978-3-662-44857-1_5](https://doi.org/10.1007/978-3-662-44857-1_5)
3. Nan, L., Fei, L., Offutt, J.: Better algorithms to minimize the cost of test paths. In: *IEEE 5th International Conference on Software Testing, Verification and Validation*, pp. 280–289. IEEE (2012)
4. Gökçe, N., Eminov, M., Belli, F.: Coverage-based, prioritized testing using neural network clustering. In: Levi, A., Savaş, E., Yenigün, H., Balcısoy, S., Saygın, Y. (eds.) *ISCIS 2006*. LNCS, vol. 4263, pp. 1060–1071. Springer, Heidelberg (2006). doi:[10.1007/11902140_110](https://doi.org/10.1007/11902140_110)
5. Belli, F., Eminov, M., Gökçe, N.: Coverage-oriented, prioritized testing – a fuzzy clustering approach and case study. In: Bondavalli, A., Brasileiro, F., Rajsbaum, S. (eds.) *LADC 2007*. LNCS, vol. 4746, pp. 95–110. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-75294-3_8](https://doi.org/10.1007/978-3-540-75294-3_8)
6. Panthi, V., Mohapatra, D.P.: Generating prioritized test sequences using firefly optimization technique. In: Jain, L.C., Behera, H.S., Mandal, J.K., Mohapatra, D.P. (eds.) *SIST*, vol. 32, pp. 627–635. Springer, Heidelberg (2015). doi:[10.1007/978-81-322-2208-8_57](https://doi.org/10.1007/978-81-322-2208-8_57)
7. Achimugu, P., et al.: A systematic literature review of software requirements prioritization research. *Inf. Softw. Technol.* **56**(6), 568–585 (2014)

8. Koomen, T., Broekman, B., van der Aalst, L., Vroon, M.: TMap Next: for Result-Driven Testing. UTN Publishers, pp. 598–602 (2013)
9. Bures, M.: PCTgen: automated generation of test cases for application workflows. In: Rocha, A., Correia, A.M., Costanzo, S., Reis, L.P. (eds.) AISC, vol. 353, pp. 789–794. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-16486-1_78](https://doi.org/10.1007/978-3-319-16486-1_78)
10. van der Aalst, L., Roodenrijs, E., Vink, J., Baarda, R.: TMap Next: Business Driven Test Management, pp. 93–113. UTN Publishers (2013)

Static Testing Using Different Types of CRUD Matrices

Miroslav Bures^{1(✉)} and Tomas Cerny^{2(✉)}

¹ Department of Computer Science, FEE, Czech Technical University in Prague,
Karlovo Namesti 13, 121 35 Prague, Czech Republic
miroslav.bures@fel.cvut.cz

² Department of Computer Science, Baylor University, Waco, TX 76798, USA
Tomas_Cerny@baylor.edu

Abstract. Static testing leads to early detection of defects throughout a project software development. This results in reduced costs and risks in the development process. Various types of static tests can be performed. In this paper, we propose extensions to contemporary static testing techniques based on CRUD matrices. In particular, we consider cross-verification between various types of CRUD matrices made by different parties at different stages of the project. This leads into extended the verification consistency of a CRUD matrix. In our evaluation, proposed techniques lead to significantly more consistent test Data Cycle Test cases, when involving our static testing techniques. Moreover, our results indicate positive impact on lowering the number of defects that usually remain undetected under the system test.

1 Introduction

Static testing is an efficient method detecting software defects in a phase, where the defect fixing is rather inexpensive, when compared to the later project phases. Various concepts and methods exist in this area. In this paper, we focus on static testing related to consistency of business data objects in the Enterprise Information Systems (EIS). Usually data-flow based techniques apply to data consistency in EIS. On the conceptual level, the Data Cycle Test (DCyT) [1] is considered as a template for data consistency tests. The DCyT bases on a CRUD matrix, introduced by Martin [2]. The DCyT proposes basic methods of static testing using CRUD matrices. We propose extensions to the static testing methods.

To define the terms, a *data entity* is an object consisting of data that are stored in the database of the System Under Test (SUT). For test design purposes, the data entities are commonly identified on a conceptual level of the design. Typically, we are interested in capturing principal business data entities that correspond to a reality that is modeled by the SUT.

The *function* is a SUT feature, that performs any Create, Update, Read and Delete (C, R, U, D) *operation* on a data entity. Further on, $F = \{f_1, \dots, f_n\}$ is a set of all the SUT functions, and $E = \{e_1, \dots, e_p\}$ is a set of all the data entities taken into account for the test design. Then, the *CRUD matrix* is defined as $\mathbf{M} = (m_{i,j})_{n,p}$, $n = |F|$, $p = |E|$,

$m_{i,j} = \{o|o\{C, R, U, D\} \Leftrightarrow \text{function } f_i \in F \text{ performs the respective Create, Read, Update or Delete operations on the data entity } e_j \in E\}$.

This paper is organized as follows. In Sect. 2 we discuss the related work. In Sect. 3 we present proposed approach for more extensive static testing using various types of CRUD matrices. The Sect. 4 presents and discusses results of the experiments. In the last section, we conclude the paper.

2 Related Work

The principle of static testing based on the CRUD matrix is introduced, for example, in DCyT technique, previously published as a part of several testing methodologies or textbooks [1, 3, 4]. During the research of static testing techniques, more extensive methods have been elaborated. Previous work exploring the data-flow analysis principle considers detection of design errors in workflow design [5–8]. In these proposals, the main use case is validation of the process design and notations different to CRUD matrix are used for SUT modeling.

For instance, Sundari et al. uses UML activity diagrams [6]; Ryndina et al. uses UML state chart diagrams [7]; proposals by Trecka et al. [5] and Awad et al. [8] use Petri's nets as a data-flow modeling structure. In practice, this approach is suitable for static testing, where Business Process Model Notation (BPMN) diagrams or UML state chart diagrams are available as the test basis. Similarly, data-flow analysis is sometimes used for verification of web services models in WS-BPEL notation [9]. In such case, data dependencies are identified and reflected in the verification process.

An alternative approach to static testing of a database design stands on the Formal Concept Analysis [10]. Conceptually, this approach is similar to a CRUD-matrix technique, however the proposal focuses only on verification of the SUT design. Sun et al. proposes an alternative approach to data-flow verification in SUT processes [11]. It uses Data-Flow Matrices, which similarly to CRUD matrix contain read and write operations performed by workflow actions on particular data objects. The difference with the CRUD Matrix is that the Data-Flow Matrix uses only read and write operations.

In the TMap *Next* description of the Data Cycle Test [1], static testing uses a standard CRUD Matrix and bases on verification of the completeness of the C, R, U, D operations for each entity $e \in E$. This approach is valid; nevertheless, it can be extended by some other techniques, introduced in this paper.

3 Static Testing Using CRUD Matrices

The common presentation of Data Cycle Test technique [1, 3, 4] does not discuss the methods of CRUD Matrix preparation. The DCyT implicitly works with CRUD Matrix created by analyst or architect designing the SUT, or CRUD Matrix created by test analyst from available test basis of other type. Nevertheless, there are other possible methods, how to create a CRUD Matrix on a software development project. Having

more versions of CRUD Matrices gives us more possibilities of static testing. Generally, there are four ways for CRUD Matrices to be created in a software development process:

Type 1: A CRUD Matrix is constructed directly from already implemented SUT code by a manual analysis or a semi-automated way.

Type 2: A CRUD Matrix is created by technical designer or developer during the system design phase of the project.

Type 3: A CRUD Matrix is assembled by test designer from the business or technical specification of SUT behavior. In this specification, data entities and SUT functions using these data entities were identified. Respective C, R, U, D operations performed on the data entities by these functions were then added to the matrix.

Type 4: A CRUD Matrix is designed by test designer in a way different to Type 2. The test designer summarizes only a list of data entities and SUT functions. Then, the designer independently proposes corresponding C, R, U, D operations by his/her domain knowledge. In this process, the designer uses the basic facts from the test basis only - he/she tries to create the CRUD Matrix in a most independent way, separately from the detail of the test basis. To get more information for this process, we can consult with the potential business users.

From our observations of industrial projects, if any type is created, the CRUD Matrix Type 2 is the most common type. By introducing Types 1, 3 and 4, we provide new opportunities for static testing: a cross-verification of CRUD Matrices.

3.1 Cross-Verification of the CRUD Matrices

To extend the opportunities for static testing based on CRUD Matrices, we propose the following cross-verification method:

1. Prepare the test basis: 2 or more independently prepared types of the CRUD Matrices $\mathbf{M}_1, \dots, \mathbf{M}_n$, $n = 2..4$ for the SUT.
2. For the two selected CRUD Matrices \mathbf{M}_1 and \mathbf{M}_2 , E_1 is a set of entities in the matrix \mathbf{M}_1 ; E_2 is a set of entities in the matrix \mathbf{M}_2 ; F_1 is a set of functions in the matrix \mathbf{M}_1 , and F_2 is a set of functions in the matrix \mathbf{M}_2 .
3. Organize the matrices \mathbf{M}_1 and \mathbf{M}_2 to list the functions F_1, F_2 and entities E_1, E_2 in the same order by using the same criteria (e.g., alphabetical sort).
4. If $E_1 \neq E_2$:
 - a. Analyze to determine whether some of the entities from E_1 and E_2 appearing as different entities are actually the same entity. If yes, unify identification of the entities;
 - b. If $E_1 \neq E_2$ is still not valid, report the difference $E_1 - E_2$ to the backlog of issues that must be clarified.
5. If $F_1 \neq F_2$ is not valid:
 - a. Analyze and determine if some of the functions from F_1 and F_2 , which appear as different functions, are actually the same functions. If yes, unify identification of the functions;

- b. If $F_1 \neq F_2$ is still not valid, report the difference $F_1 - F_2$ to the backlog of issues that must be clarified.
6. For each of the cells of the compared matrices that correspond to e and f , $e \in E_1$, $e \in E_2$, $f \in F_1$, $f \in F_2$, if the cell content differs in the C, R, U, D operations, report the difference to the backlog of issues that must be clarified.
7. When the issues in the backlog are clarified, merge matrices \mathbf{M}_1 and \mathbf{M}_2 to a final CRUD Matrix \mathbf{M} , which will represent a corrected version of the expected behavior for the SUT.

A difference reported to the backlog can denote either incomplete information in one of the CRUD Matrices \mathbf{M}_1 and \mathbf{M}_1 or a potential defect, which is the subject of our investigation. The next set of static tests can be defined for one CRUD Matrix.

3.2 Extended Consistency Verification of a CRUD Matrix

For static testing performed on a single CRUD Matrix, as described in TMap *Next* [1], we propose the following extension:

1. For each of the cells in the matrix \mathbf{M} , E is a set of entities in the matrix \mathbf{M} , and F is a set of functions in the matrix \mathbf{M} ;
2. If entities $e_1 \in E$ and $e_2 \in E$ have the same set of C, R, U, D operations in their respective columns in the CRUD Matrix: analyze the situation to determine if the entities are separated for a certain reason or if this situation indicates unnecessary duplicity in the code;
3. If functions $f_1 \in F$ and $f_2 \in F$ have the same set of C, R, U, D operations in their respective lines in the CRUD Matrix: analyze to determine whether the functions are separated for a certain reason or if this situation indicates an unnecessary duplicity in the code;
4. For each entity $e \in E$, verify how the deletion operation is specified in the source documentation and how it is reflected in the CRUD Matrix:
 - a. Entity e has to be deleted, which should be captured by a D operation;
 - b. Entity e has to be archived, instead of deleted, which should be captured by a U operation for the entity e , or by a D operation for the entity e and a C operation for an archive entity e' , which is copied from e .
5. If there are requirements to maintain a history of changes for data entity $e \in E$ after an update of this entity by function $f \in F$, this fact may be explored in detail. In addition to the respective U operations in the CRUD Matrix in the cells corresponding to e and f , the situation may be captured by other U or C operations in the matrix line corresponding to the function f . These additional U or C operations would be performed on the entity that maintains a record of the changes in entity e . The particular situation depends on the technical details of the implementation process.

The situations described in steps 2 and 3 are interesting from the general redundancy point-of-view in a SUT, which is a frequent source of defects. In this analysis, possible planned future extensions of the system should be considered.

Step 4 aims to detect another possible type of defect that is related to the proper deletion or archival of the data entities. From our experience, this is an area where design mistakes can occur, these are expensive to correct after the implementation phase. The same analysis applies for step 5, which focuses on the requirements for maintaining a history of the changes in the data entities that are processed by a SUT.

4 Experiments

To verify the proposed static testing approach, we simulated a situation where an incomplete and inconsistent test basis has been used as an input to creation of DCyT test cases. Using the defined structure as the artificial SUT, we provide a group of test designers with a test basis and certain number of inserted inconsistencies.

The experimental group has 12 test designers, who vary by their previous praxis in software testing area from 1 year to 8 years. First, we measure the efficiency of the prepared test cases by TMap Next DCyT [1] without any static testing. Next, the test designers apply the proposed static testing techniques, and again, we evaluate the efficiency of the test cases. In the experiment, the aim was to answer the following questions: When test basis is inconsistent, **(1)** how many inconsistent test case steps are produced by DCyT [1] with static testing compared to situation, when no static testing is used? Moreover, **(2)** how many data consistency defects remains undetected by DCyT with static testing compared to situation, when no static testing is used? Finally, **(3)** how resource consuming is the proposed static testing method; moreover, how is this method efficient?

For the experiment, we create two instances of artificial SUT. Also, the number of artificial defects are defined in particular artificial SUT instances (simulating presence of defects in the SUT). The created instances are retained as a baseline to determine the actual state of the SUT, which is be tested (we use a term *baseline SUT* for these instances further on). The baseline SUT is not known to the experiment participants.

The instance of an **artificial SUT** A is a six-tuple (F, E, S, D, W, L) ; where F is a set of SUT functions, E is a set of data entities used by the functions, S is a set of possible states of the SUT. The SUT changes its state when a function $f \in F$ is executed.

D is a set of inserted data consistency defects. An inserted data consistency defect is a quaternion $d = (e, f_c, o_c, F_d)$; where $e \in E$ is a data entity, which is in an inconsistent state that causes a defect, $f_c \in F$ is the function that causes the data entity e to be inconsistent, $o_c \in \{C, U\}$ is the particular create or update operation that causes the data entity e to be inconsistent when accessed by the function f_c , F_d is a set of pairs (f_d, o_d) , where $f_d \in F$ is a function that activates a defect in the SUT as a result of the inconsistency of the data entity e . This is caused by function f_c , the $o_d \in \{C, R, U, D\}$ is the particular operation that activates the defect in the SUT when accessed by function f_d .

W is a set of workflows implemented in the SUT. The workflows describe the possible sequences of functions $f \in F$ in the SUT. The workflow is a directed graph (S_w, F_w) , whose nodes $S_w \in S$ are the states of the SUT, and its edges $F_w \in F$ are the functions of the SUT. L is a set of data entity lifecycles in the SUT. For each data entity $e \in E$, the data lifecycle $l_e \in L$ is defined. The data entity lifecycle is a directed graph (S_e, F_e) ,

whose nodes $S_e \in S$ are states of the SUT, and whose edges $F_e \in F$ are functions of the SUT. For each of these edges, the C, R, U, D operations that are performed on an entity e are defined.

Test case c for data entity e is a sequence of steps $\{s_1, \dots, s_n\}$, where relevant of these steps are pair (f_s, o_s) , and $f_s \in F, o_s \in \{C, R, U, D\}$ is an operation that is performed on the data entity e by the function f_s (see an example in Fig. 1). Possible sequences of the SUT functions in the test cases are determined by W and L .

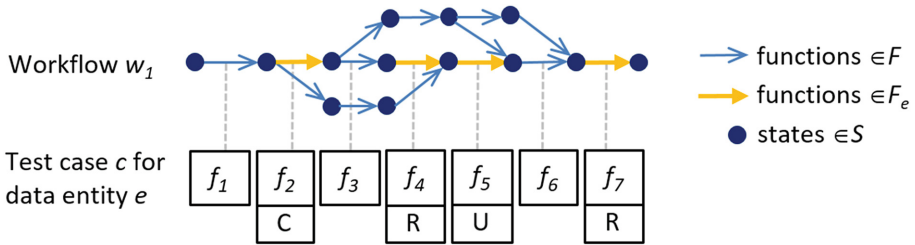


Fig. 1. Example of artificial SUT workflow and test case.

For the experiment, two instances of artificial SUT are created (the details are in Table 1). Then, before providing the instances to the test designers, we change the test basis in several places, making it inconsistent with the baseline SUT. In particular, we induce several changes to the structured map of the artificial SUT, consisting of F, E, W, L , and I , as defined above. We also change a few randomly selected edges of selected workflows from W and the data entity lifecycles from L . In the future, we will use the term *inconsistent test basis* to describe this structure.

Table 1. Artificial SUT instances created for the experiments

Instance ID	$ E $	$ F $	$ W $	$ I $	$ D $	$C1$	$C2$
1	12	57	5	10	22	10	14
2	14	82	6	12	35	16	20

The artificial SUT instances are randomly distributed to the test designers. The testing group is provided with a structured map of the artificial SUT that consists of F, E, W and L , as defined above. This artifact simulates a test basis.

In this experiment, all the test designers create a Type 2 CRUD Matrix from this inconsistent test basis (denoted as *CRUD Matrix A* further on). Then two groups of testers (each of the groups consists of six testers) are testing the following situations for the both artificial SUT instances (see Table 1):

Group 1: The test designers create the DCyT test cases using an inconsistent test basis (*CRUD Matrix A*) without any static testing. By this, we simulate a situation when an inaccurate test basis is used in DCyT.

Group 2: In addition to the inconsistent test basis, the test designers are given another CRUD Matrix (in our taxonomy, corresponding to Type 1), denoted as *CRUD Matrix B* further on. Matrix B is created directly from the baseline SUT, but we randomly

make changes to the C, R, U, and D operations in the matrix cells. This simulates an inaccurate or out-of-date CRUD Matrix, which may be provided to the test designers in an actual project. *The testers in this group are performing cross-verification between CRUD Matrices A and B and extended consistency verification of a CRUD Matrix proposed in this paper.* In this group, the designers are not told which version of the test basis was correct (if the *CRUD Matrix A* or *B*). Actually, both of the matrices contained inconsistencies. When asking for clarification, the designers are told the correct information by us (this information is taken directly from the baseline SUT). Then, the test designers create DCyT test cases.

In Table 1, column *C1* specifies how many changes in edges of graphs in sets *W* and *L* we did in the test basis of baseline SUT to create the inconsistent test basis (used by the both groups of test designers to create *CRUD Matrix A*).

The column *C2* specifies, how many C, R, U, D operations are changed in the CRUD Matrix. It is created directly from the baseline SUT to produce the *CRUD Matrix B* (given to the Group 2). The Group 2 then performs static testing as described above. In this part, we measure the duration of the static testing. Then participants from the both groups create test cases by the DCyT and finally, the test cases are simulated against the artificial SUT. Table 2 shows the results.

Table 2. Experiment results

Instance ID	Group	<i>TS</i>	<i>TIS</i>	ΔIS	<i>TD</i>	ΔD	<i>TIME</i>
1	1	90.7	28.3	31.2%	8.1	36.8%	–
1	2	91.2	6.2	6.8%	3.0	13.6%	3.7
2	1	134.4	36.8	27.4%	11.7	33.4%	–
2	2	133.9	13.8	10.3%	5.2	14.9%	5.6

In Table 2, column *TS* contains the total number of steps of produced test cases, averaged for all test sets produced by testers in the experimental group. Column *TIS* contains total number of inconsistent steps in produced test cases, again averaged for all produced test sets. Column ΔIS contains relative difference between *TS* and *TIS*.

The test step *s* is inconsistent when two subsequent C, R, U, D, B operations performed by the functions $f_1 \in F$ and $f_2 \in F$ in the test case cannot be performed in the SUT. This occurs because we cannot reach a proper state in the SUT to execute the function f_2 from the SUT state reached by function f_1 .

The column *TD* contains averaged total number of undetected defects for all produced test sets. Column ΔD contains relative difference between $|D|$ (refer to Table 1) and *TD*. In this experiment, we considered a data consistency defect $d \in D$ undetected when, for all $f_d \in F_d \in d$, the function $f_c \in d$ is either not present, or it is not followed by an f_d in any of the test cases that were created for a particular instance of artificial SUT. Finally, the column *TIME* contains average time in hours spent by testers in the experimental group performing the proposed static tests.

For the total number of steps of produced test cases, no significant change has been observed using the static testing technique (column *TS*). Nevertheless, test cases produced by DCyT with static testing have significantly less inconsistent steps by

approximately 25% (question 1, columns *TIS* and ΔIS). Moreover, the number of defects that remained undetected after the simulation of test cases in the baseline SUT was lower by approximately 20% when applying the proposed static testing technique (question 2, columns *TD* and ΔD). Next, as the results show, an initial time investment in static testing (column *TIME*) is most likely to be amply returned, especially when we consider the possible overhead caused by inconsistent test cases and potential defects, which remain undetected in the SUT (question 3).

5 Conclusion

In this paper, we proposed an extension to the common approach of static testing based on CRUD Matrices [1, 3, 4]. In particular, we propose (a) cross-verification between various types of CRUD Matrices created by different parties at different stage of the project, and (b) extension of consistency verification of the CRUD Matrix.

In the conducted experiment, was simulated a case, when the test basis is inconsistent with the SUT. When the test basis differs from the SUT, the situation leads to inconsistent test cases and a higher ratio of undetected defects. In our experiment, the proposed static testing (even when using a CRUD Matrix with defects) led to significantly more consistent DCyT test cases when using a test basis corrected after these tests. The result led to a lower number of defects, which remained undetected in the SUT. Moreover, initial time investment in performed static testing (in our experiments 5 h on the average) can be considered much lower than potential overhead arising from further detection and fixes related to defects caused by a wrong or inconsistent test basis.

This research is conducted as a part of the project TACR TH02010296 Quality Assurance System for Internet of Things Technology.

References

1. Koomen, T., van der Aalst, B., Brokeman, M., Vroon, M.: TMap Next, for Result-Driven Testing, pp. 625–627. UTN Publishers, Den Bosch (2013)
2. Martin, J.: Information Engineering. Prentice Hall, Englewood Cliffs (1990)
3. De Groot, D.: TestGoal: Result-Driven Testing, pp. 208–210. Springer, Heidelberg (2008)
4. Van Veenendall, E.L.: The Testing Practitioner, pp. 241–243. UTN Publishers, Den Bosch (2002)
5. Trčka, N., Aalst, Wil, M.,P., Sidorova, N.: Data-flow anti-patterns: discovering data-flow errors in workflows. In: Eck, P., Gordijn, J., Wieringa, R. (eds.) CAiSE 2009. LNCS, vol. 5565, pp. 425–439. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-02144-2_34](https://doi.org/10.1007/978-3-642-02144-2_34)
6. Sundari, M.H., et al.: Detecting data flow errors in workflows: a systematic graph traversal approach. In: 17th Workshop on Information Technology and Systems (2007)
7. Ryndina, K., Küster, Jochen, M., Gall, H.: Consistency of business process models and object life cycles. In: Kühne, T. (ed.) MODELS 2006. LNCS, vol. 4364, pp. 80–90. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-69489-2_11](https://doi.org/10.1007/978-3-540-69489-2_11)
8. Awad, A., Decker, G., Lohmann, N.: Diagnosing and repairing data anomalies in process models. In: Rinderle-Ma, S., Sadiq, S., Leymann, F. (eds.) BPM 2009. LNBIP, vol. 43, pp. 5–16. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-12186-9_2](https://doi.org/10.1007/978-3-642-12186-9_2)

9. Moser, S., Martens, A., Gorlach, K., Amme, W., Godlinski, A.: Advanced verification of distributed WS-BPEL business processes incorporating CSSA-based data flow analysis. In: International Conference on Services Computing, pp. 98–105. IEEE (2007)
10. Poelmans, J., Dedene, G., Snoeck, M., Viaene, S.: Using formal concept analysis for the verification of process-data matrices in conceptual domain models. In: Proceedings of the IASTED International Conference on Software Engineering, pp. 79–86. Acta Press (2010)
11. Sun, S.X., et al.: Formulating the data flow perspective for business process management. *Inf. Syst. Res.* **17**(4), 374–391 (2006)

Extracting Test Cases with Message-Sequence Diagram for Validating the Photovoltaic Energy Integrated Monitoring System

Woo Sung Jang¹, Bo Kyung Park¹, Hyun Seung Son¹, Byung Kook Jeon²,
and R. Young Chul Kim^{1(✉)}

¹ SELab, Department of CIC (Computer and Information Communication),
Hongik University, 2639, Sejong-ro, Jochiwon-eup, Sejong special self-governing city,
30016, Republic of Korea

{jang, park, son, bob}@selab.hongik.ac.kr

² Department of Software, Gangneung-Wonju National University,
Wonju, Gangwon Province 26403, Korea
jeonbk@gwnu.ac.kr

Abstract. Recently, in the photovoltaic energy integrated monitoring software system, it has more complex, and accordingly may be possible to occur more errors. In this industrial services, a small error can lead to a huge accident to make the power failures. To completely build this system, it should verify whether it is or not stability of software through measuring the full coverage with generating test cases in detail level based on a message sequence model. In this paper, we apply to verify a system stability of this monitoring system with our previous research such as the automatic test case generation based on UML 2.4.1 message-sequence diagram via cause-effect diagram. With this, we extract automatically test cases on coverage.

Keywords: Automatic test case generation · Message-sequence diagram · Renewable energy · Integrated monitoring system

1 Introduction

Approximately it occurs sixty percent of the software errors in the pre-design stages as well as the design stage, while only 40% of them in the post-design stages [1]. Furthermore, the requirements from the pre-design stages are likely to have uncertain and incomplete defects, so they are not easily detected. Also, when the requirement is misinterpreted, it would cause another issue that new software needs to be redeveloped. That is, one of the main reasons that causes software errors is a test case based on incomplete requirements [2].

Model-based testing tools generally decrease the number of incomplete test cases caused by requirements. If requirement-based test cases are generated in the design stage and test cases are implemented for a system development, the modules being developed based on misinterpreted requirements would be identified more quickly.

The existing methods generate model-based test case with Use Case Diagram [3]. In this paper, we use Message-Sequence Diagram method [4, 5]. Using the Sequence Diagram, we are to generate test cases satisfied with 100% coverage of software with “Metamodel

oriented Test Case Generation Method Based on transforming UML 2.4.1 Message-Sequence Diagram via Cause-Effect Diagram” [6].

In this paper, a test case would be extracted by applying a Message-Sequence Diagram drawn in the design stage to the previous approach [6] in order for the stability of the solar energy total monitoring system to be verified.

The outline of this paper is as introduced below. Section 2 describes related works including a method of metamodel oriented test case generation on the new & renewable energy total monitoring systems. In Sect. 3, Message-Sequence Diagram of solar energy total monitoring system is to be designed. And Sect. 4 describes the test case extraction of the solar energy total monitoring system, followed by conclusion and further studies.

2 Related Works

2.1 The Integrated Monitoring System for New and Renewable Energy

On the integrated monitoring system for new & renewable energy, we need to have a standard interface that interprets different types of data to be delivered to various kinds of energy plants. Because the standard interface is designed based on metamodel, a new data type is easily added into it. This is, plug and play on heterogeneous solar devices. Therefore, it provides total monitoring services based on web server, so each customer can easily track the current power via web browser. It also may provide prediction of the power using statistical methods for big data [7]. Figure 1 shows its total structure of this system.

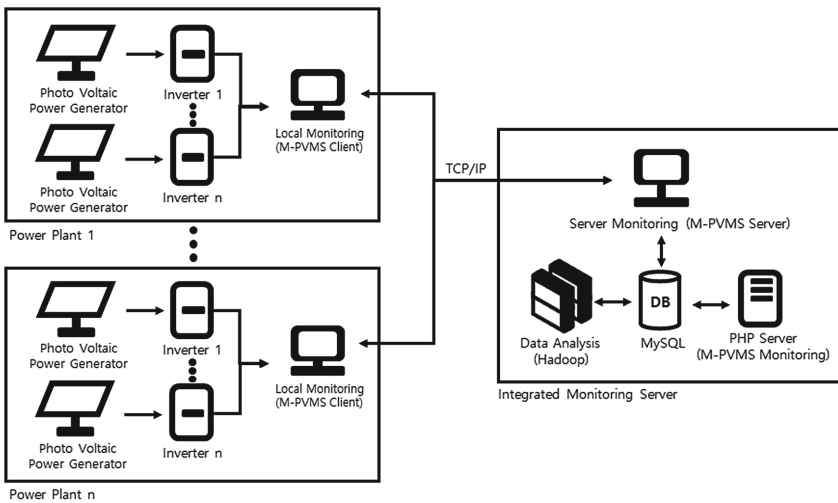


Fig. 1. The total structure of integrated monitoring system

2.2 Metamodel Oriented Test Case Generation

We use Message-Sequence Diagram to be transformed to Cause-Effect Diagram, and then generate test cases based on Cause-Effect Diagram with 100% of functional requirement coverage, which would be fulfilled by using minimal test cases only.

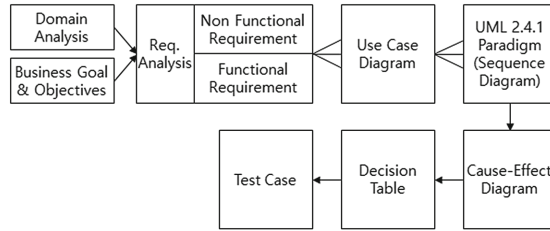


Fig. 2. A mechanism of test case generation in detail level from requirements

Figure 2 shows the detail of the process of test case generation that we analyze and design Use Case Diagram; Message-Sequence Diagram is designed by using each Use Case in the Use Case Diagram; Message-Sequence Diagram is transformed to Cause-Effect Diagram; Cause-Effect Diagram is transformed to decision table; and decision table turns into a test case.

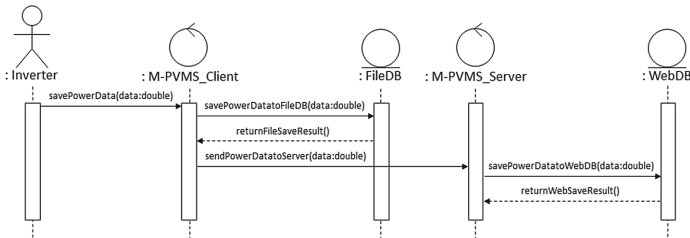


Fig. 3. The flow of saving the power data at most cases

3 A Message-Sequence Diagram in the Integrated Monitoring System

In this paper, with requirements, we draw the interaction with the system, that is, Message-Sequence Diagram in order to increase stability of the integrated monitoring system for new & renewable energy. As an example, the saving function of the power data in the system is drawn with Message-Sequence Diagram. From the diagram, test cases are to be extracted.

The saving of the power data into the server is a function to store data delivered from inverters. The flow of saving the power data at the most cases is seen in Fig. 3. The M-PVMS Client delivers data of inverter to M-PVMS server, and stores it into FileDB. M-PVMS server then stores the delivered data into WebDB.

Table 1. Message-sequence diagram codes for power data saving function in most cases

```

<?xml version="1.0" encoding="UTF-8"?>
<sed:SEDModel xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:sed="http://sed/1.0" name="Model">
  <mMessage name="savePowerData" startMessage="//@mLimeline.0"
endMessage="//@mLimeline.1"/>
  <mMessage name="savePowerDatatoFileDB" startMessage="//@mLimeline.1"
endMessage="//@mLimeline.2"/>
  <mMessage name="returnFileSaveResult" startMessage="//@mLimeline.2"
endMessage="//@mLimeline.1"/>
  <mMessage name="sendPowerDatatoServer" startMessage="//@mLimeline.1"
endMessage="//@mLimeline.3"/>
  <mMessage name="savePowerDatatoWebDB" startMessage="//@mLimeline.3"
endMessage="//@mLimeline.4"/>
  <mMessage name="returnWebSaveResult" startMessage="//@mLimeline.4"
endMessage="//@mLimeline.3"/>
  <mLimeline name="Inverter" type="Actor" ownedMessage="//@mMessage.0">
    <mObkind xsi:type="sed:Actor"/>
  </mLimeline>
  <mLimeline name="M_PVMS_Client" type="Control"
ownedMessage="//@mMessage.0 //@mMessage.1 //@mMessage.2
//@mMessage.3">
    <mObkind xsi:type="sed:Control"/>
    <ecaRule mMessage="//@mMessage.0 //@mMessage.1"/>
    <ecaRule mMessage="//@mMessage.2 //@mMessage.3"/>
  </mLimeline>
  <mLimeline name="FileDB" type="Service" ownedMessage="//@mMessage.1
//@mMessage.2">
    <mObkind xsi:type="sed:Service"/>
    <ecaRule mMessage="//@mMessage.1 //@mMessage.2"/>
  </mLimeline>
  <mLimeline name="M_PVMS_Server" type="Control"
ownedMessage="//@mMessage.3 //@mMessage.4 //@mMessage.5">
    <mObkind xsi:type="sed:Control"/>
    <ecaRule mMessage="//@mMessage.3 //@mMessage.4"/>
  </mLimeline>
  <mLimeline name="WebDB" type="Service" ownedMessage="//@mMessage.4
//@mMessage.5">
    <mObkind xsi:type="sed:Service"/>
    <ecaRule mMessage="//@mMessage.4 //@mMessage.5"/>
  </mLimeline>
</sed:SEDModel>

```

The designed Message-Sequence Diagram needs to turn into XMI codes. In the power data saving function, Message-Sequence Diagram codes of the most cases are as seen in Table 1.

4 Test Case Extraction for New and Renewable Energy Monitoring System

The process to extract test case for solar energy total monitoring system is as described below. The XML codes for Message-Sequence Diagram are to be inputted to the automation tool, and then test case is to be extracted. The automation tool is applied by the model transformation rules based on ATL [6].

Generated XMI code for test case can be showed a chart format in Microsoft Excel. The Fig. 4 shows a test case extraction for power data saving function.

	A	B	C	D	E
1	ns2:version	no	testpre	testcon	testresult
2	2	TC1	savePowerData=F,savePowerDatatoFileDB=F	NORMAL	savePowerDatatoFileDB=F
3	2	TC2	savePowerData=T,savePowerDatatoFileDB=T	NORMAL	savePowerDatatoFileDB=T
4	2	TC3	returnFileSaveResult=F	NORMAL	returnFileSaveResult=F
5	2	TC4	returnFileSaveResult=T	NORMAL	returnFileSaveResult=T
6	2	TC5	sendPowerDatatoServer=F	NORMAL	sendPowerDatatoServer=F
7	2	TC6	sendPowerDatatoServer=T	NORMAL	sendPowerDatatoServer=T
8	2	TC7	savePowerDatatoWebDB=F	NORMAL	savePowerDatatoWebDB=F
9	2	TC8	savePowerDatatoWebDB=T	NORMAL	savePowerDatatoWebDB=T
10	2	TC9	savePowerDatatoWebDB=F	NORMAL	returnWebSaveResult=F
11	2	TC10	savePowerDatatoWebDB=F	NORMAL	returnWebSaveResult=T

Fig. 4. Test cases for the power data saving function

In model based testing based on requirements, we generate test cases from message sequence diagram via the cause effect diagram, which covers in detail level of the system. To validate the right requirements, we should generate all possible test cases which are satisfied by the requirements.

5 Conclusion

The volume of software is positively related to the number of errors in software. Because there are possibilities of life damage in the industrial field, an emphasis on development should have one of methods to increase reliability of software to industrial system.

By using “Metamodel oriented Test Case Generation Method Based on Transforming UML 2.4.1 Message-Sequence Diagram via Cause-Effect Diagram” we automatically generate test cases with any, and accordingly would be reduced the probability of inaccurate implementation in software.

In this study, test case is extracted by applying “Metamodel oriented Automatic Test Case Generation Method based on Transforming UML 2.4.1 Message-Sequence Diagram via Cause-Effect Diagram” in order to verify reliability of solar energy total monitoring system. Also, it is ascertained when the system functions are consistent with the demands.

In the future, we would be researched in the further study on methods of automated transformation of Message-Sequence Diagram into XMI codes.

Acknowledgments. This work was supported by the Human Resource Training Program for Regional Innovation and Creativity through the Ministry of Education and National Research Foundation of Korea (NRF-2015H1C1A1035548).

By Prof. Jeon, This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2014R1A1A2058667).

References

1. NIPA, 2014 Software Engineering White Book, National IT Industry Promotion Agency (2014)
2. Chae, H.S.: Model-based test-concepts and issues. In: Korean Institute of Information Scientists and Engineers, vol. 32, no. 4, pp. 59–71 (2014)
3. Kim, R.Y.C., Joo, B.-G., Kim, K.-C., Joen, B.-K.: Scenario based testing & test plan metrics based on a user case approach for real time UPS (Uninterruptible Power System). In: Kumar, V., Gavrilova, M.L., Tan, C.J.K., L'Ecuyer, P. (eds.) ICCSA 2003. LNCS, vol. 2668, pp. 646–655. Springer, Heidelberg (2003). doi:[10.1007/3-540-44843-8_71](https://doi.org/10.1007/3-540-44843-8_71)
4. Kim, Y., Son, H.S., Kim, W., Seo, J., Kim, D., Seo, Y., Ryu, D., Kim, R.Y.C.: A study on modeling the obstacle avoidance UGV using extended executable UML. In: Joint Workshop on Software Engineering Technology 2007 (KSEJW-2007), pp. 108–116 (2007)
5. Woo, S., Son, H.S., Kim, R.Y.C.: A study on extending message-sequence diagram for mapping cause-effect diagram. In: Korea Information Processing Society, vol. 19, no. 1, pp. 1251–1254 (2012)
6. Woo, S.: Metamodel oriented automatic test case generation based on transforming UML 2.4.1 message-sequence diagram via cause-effect diagram. Hongik University (2012)
7. Jang, W.S., Son, H.S., Park, B.K., Kim, R.Y.C.: Implementation of effective web integrated monitoring system for small renewable energy business industries. In: Korean Institute of Smart Media, vol. 5, no. 1, pp. 303–305 (2016)

Automatic Test Case Generation with State Diagram for Validating the Solar Integrated System

Bo Kyung Park, Woo Sung Jang, Hyun Seung Son,
Keunsang Yi, and R. Young Chul Kim (✉)

SELab, Department of CIC (Computer and Information Communication), Hongik University,
Sejong Campus, Seoul 339-701, Korea
{park, jang, son, Keunsang, bob}@selab.hongik.ac.kr

Abstract. For safe software development on the solar integrated monitoring system, it is very important how to identify safe behaviors of the system behaviors. Therefore, it needs to test the system behaviors after the software development. To solve this problem, the existing studies have proposed the use case based test coverage analysis at all software development stages [1]. With this method, we identify the test cases based on priority of the system behaviors. In this paper, we propose automatic test case extraction method based on state diagram among the use case-based test coverage extraction methods. That is, we can use state diagram for a system behaviors with which generates test cases to validate the system. We show an applicative case on the system behaviors of a solar integrated system with this approach.

Keywords: Automatic test case generation · State diagram · Use case based test coverage · A renewable energy integrated monitoring system

1 Introduction

With the recent development of convergence software, the relative importance of software is increasing in the automobile, aviation, and railroad industries. Also, software is applied in various fields and thus requires complex functions. Therefore, it is the importance of quality in increasing software, for example, safety, reliability, etc. As suggested in 2010 Toyota recalls, software defects cause personnel and material loss. For this reason, safe software development is an important issue.

To development a safer software, software test is very important. Software defects are discovered after carrying out a test. If software test is executed earlier for that error can be discovered quickly, we can reduce period of development and cost. That is, the actual cost for correcting deficiency becomes cheaper. But the actual test of the system behaviors is executed after the software is implemented. The existing studies have focused on a use case-based testing combined with software development stage [1]. This method can discover and modify the problem that can occur in early development stage in advance as it conducts a testing from the requirement stage. The current method can test all possible input values. Also, the more the generated test cases are existed, the more the time and cost are consumed for testing. This method extracts all possible test

cases and focuses on effective testing. The finally extracted levelled test case identifies order of priority. The identified order or priority executes maximum coverage testing through minimum test case. This paper proposes an automatic test case establishment method through State Diagram. This method is included in Use case-based test coverage extraction method. The extraction process is generating State Diagram, State Table, and State Transition Tree and extracting test case. This method can detect error quickly and reduce period of development and cost because it can conduct a test in design stage. HIMEM, developed in this research laboratory as case study was applied to photovoltaic monitoring system [2].

This paper is as introduced below. Section 2 describes our test coverage on use case approach. Section 3 describes automatic test case generation with state machine based on the system behaviors. And Sect. 4 describes a case study on apply the solar energy total monitoring system with this approach, followed by conclusion and further studies.

2 Our Test Coverage on Use Case Approach

Figure 1 shows the use case oriented testing procedure [1]. For use case oriented testing, users should analyze domain and set business goals. In other words, users should understand the entire system clearly through domain analysis and set business goals. This is why the testing should also appreciate for business organization to be a great expense.

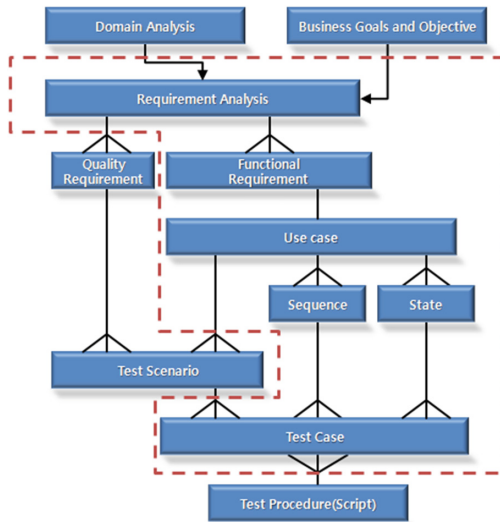


Fig. 1. A procedure of use case oriented testing

If such test is conducted despite unclear goal, it may lead to indiscriminate tests. This may result in a waste of time and cost, which has also the most critical impact on the entire cycle of development. Most entire software errors occur in requirement analysis

stage. If such requirement is analyzed through exact identification, it can reduce the time, cost, and personnel for software test. Requirement analysis identifies functional and non-functional requirements. The identified requirements can have a 1:1 or 1: n relationship through the existing requirement analysis. Use case is extracted from the functional requirements. Sequence and State Diagram are generated from Use case. Test scenario is extracted from non-functional requirements. Test scenario has a 1:1 or 1: n relationship. Finally, all possible test cases is extracted from use case, sequence, and state diagram. Test case has a 1:1 or 1: n relationship. Through this process, a test is conducted based on requirements. Table 1 shows the abstract levels of a test case coverage matrix.

Table 1. The abstract levels of a test case coverage metrics

Use Case Test Case Coverage	Message Sequence			State Test Case Coverage	Object Test Case Coverage	Method Test Case Coverage	
	Dialogue Test Case Coverage	MLU Test Case Coverage	Reusable Pattern Test Case Coverage				
UC 1	D 1	MLU 1	P1	s0, s1	Object 1	m1	
					Object 2	m2	
					Object 3	m4	
		MLU 2	P2		s1, s2	Object 3	m3
						Object 5	m4
						Object 1	m5
	MLU 4	P6	s5, s0	Object 3	m10		
				Object 5	m11		
				Object 6	m12		
				Object 3	m6		
UC 2	D 2	MLU 3	P3,P4,P5	s2,s3,s4	Object 3	m8	
					Object 4	m9	
					Object 3	m7	
					Object 4	m8	
					Object 3	m3	
UC 3	D 3	MLU 2	P2	s1,s2	Object 3	m4	
					Object 5	m3	
					Object 5	m5	
	D 4	MLU 4	P3,P4,P5		s2,s3,s4	Object 3	m8
						Object 5	m9
						Object 5	m7
					Object 5	m8	

3 Automatic Test Case Extraction Based on State Diagram

State transition testing is a model based technique. This technique generates test case based on the system behaviors of the solar integrated monitoring system. We implement a test case extraction tool based on state diagram.

Figure 2 shows a test case generation procedure [3]. First, we model the state diagram of a target system. The state model is converted to state table. The state table is composed of all events of states and events. The state can express each one of all situations. The event is a factor to cause the transition of state. The state table displays each state in the top side of table, and also represents the event in the left side of table. And the rest displays a movable state when a state meets an event. N/A represents a case that can't be movable. After then, we use the information of state table, and generate state transition tree. The state transition tree includes all information that a next state is transferred by the event in state table. Therefore, to look for the transition route of state and generate test case, the state table is converted to the state transition tree. This transition tree generation method displays all states in the upper side of table. The next stage represents all accessible states. Through this process, the test case is made by considering all accessible cases. The generated test case is a scenario executed by each state in state transition tree.

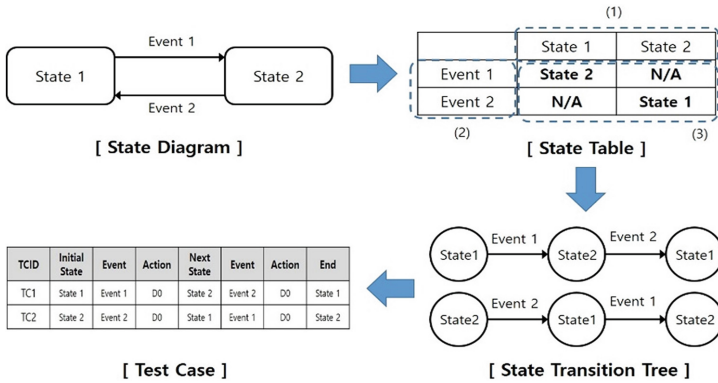


Fig. 2. State transition testing method

4 A Case Study

This study is an automatic test case generation based on a state diagram, that is, the system behaviors of our integrated renewable energy monitoring system. For this, we use our tool, that is, Hongik MDA based Embedded Software Development Methodology (HIMEM v1.0), developed in our research laboratory of this study. HIMEM v1.0 is an automatic test case generation tool.

This System monitors such information of electric energy and temperature collected from the solar cell. Figure 3 shows the communication architecture of a photovoltaic monitoring system. This system stores the data transmitted from many power plants into our web-server. And the data sent from each power plant are integrated through meta-model based standard interface. This meta-model is used because the inverter installed to each power plant uses different communication packet. Meta model converts different packets to the same types of packet.

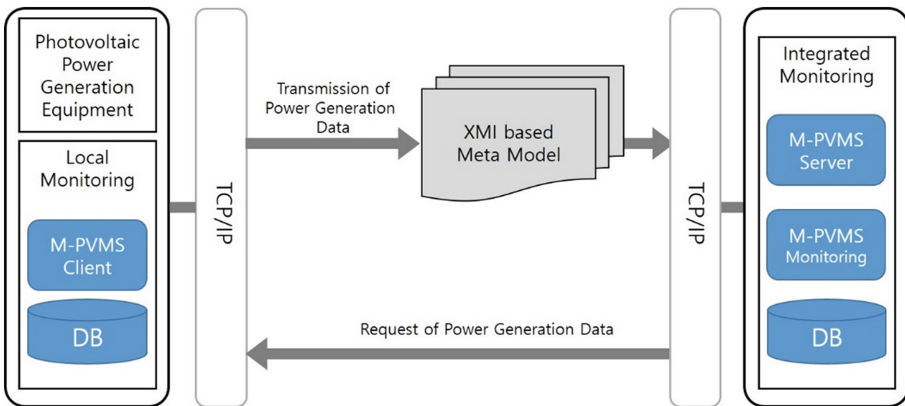


Fig. 3. The communication architecture of a photovoltaic monitoring system

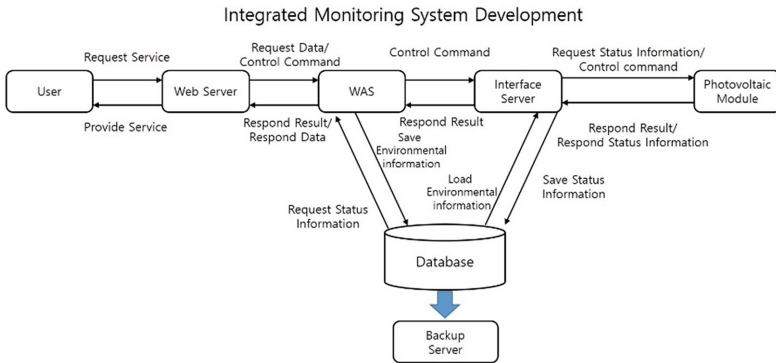


Fig. 4. The whole system structure of integrated monitoring system

Figure 4 shows the whole system structure of Integrated monitoring system. In this current system, a central server does manage information data occurred from the existing monitoring system. An individual user can possibly monitor the generated energy on the web. Therefore, we can diagnose and resolve problems through monitoring at the central server. Figure 5 shows a test case extraction method based on state diagram for validating the system behaviors. We use our HIMEM v1.0 tool to draw the state diagram to represent the system behaviors of our monitoring system. In the state diagram of our integrated photovoltaic monitoring system in Fig. 5(A), the 'Idle' state is the initial state of the integrated monitoring system. The 'Ready' state represents the beginning state and resetting state of monitoring system. The 'Run' state represents the execution state of system, and processes the data, graph, and connection information of electric energy. 'Run' state is divided into two sub states like data collection (collectData) and data analysis (analyzeData). In a case of data collection, the data of electric energy and environmental sensor are collected through the inverter in a power plant. After then, the collected data are transmitted to monitoring system. If it does occur defects of the equipment, Alarm sends a message to administrator. In other case of data analysis, the collected data are analyzed. This data are printed out in graph after daily/monthly/yearly electric energy analysis. If disorderData occurs, the system is converted to watch out state. At this time, the system checks MPVMSClient, MPVMSServer, or MPVMSMonitor and sends a warning message. Figure 5(B) shows a state table of integrated monitoring system. The state table represents a state and event when a particular event is generated in an applicant state from the state diagram. The upper side of table represents state, and the left side of table represents an event. The intersection represents an accessible state, that is, to the next stage. Figure 5(C) shows the state transition tree of integrated monitoring system. The State Transition Tree is generated based on state table. The state transition tree expresses all possible states in the form of tree. The test case is generated based on state transition tree. Figure 5(D) shows the finally extracted test case. The extracted test case shows an event and action occurred in the current state. Also, the next state can be identified in the current state. The total of 107 test cases were extracted.

Comparison of Software Complexity Metrics in Measuring the Complexity of Event Sequences

Johanna Ahmad^(✉) and Salmi Baharom

Software Engineering and Information System Department, UPM,
Serdang, Malaysia
gs43485@student.upm.edu.my, salmi@upm.edu.my

Abstract. One of the main challenges in software development is the complex structure of a system. The software development for event sequences is complex. It is a challenge to define a complexity metric for event sequences application. Lack of knowledge in complexity metric can lead to issues such as rises in software cost and delays in project timing. Numerous complexity metrics have been proposed and published, such as information flow complexity, lines of code, function points, and unique complexity metric. However, in the context of the event sequences, most of the research focuses on measuring web graphs, measuring the web traffic and how the complexity of the web impacts the customer. In this paper, the researchers studied and compared five different software complexity metrics. This paper describes the on-going research that addresses the issue to produce a unique weight to prioritise event sequences test cases.

Keywords: Software complexity metric · Event sequences · Unique complexity metric

1 Introduction

The software process model to develop a system comprises a sequence of steps. A system that has high performance in terms of the reliability, testability, safety, availability, maintainability and security can be categorised as a high-quality system [1]. The measurement is needed to make sure that the system fulfils all the aspects mentioned above. The measurement results allow us to conclude whether the structure of the system is easy or difficult to understand and whether the relationship between modules needs to be reengineered or not [2]. The software complexity measurement is needed to understand the boundaries and requirements of the system, behaviour of the system and connections between components.

Understanding the complexity of the system is the first step before customising it to achieve the desired usability, reliability, availability, maintainability and performance [3]. More recently, literature has emerged that few factors such as the number of lines, the total occurrence number of operators and number of control structure in the program and function affect the complexity of the program [4]. There is a relationship between complexity and number of detected faults [5]. If the complexity measurement of a system is a complex system, most probably the number of detected faults is high. The

relationship between the complexity results and number of detected faults can be used as a surrogate for fault measurement in the subsequent testing phase. Understanding the complexity of the event sequences in a test case is the first step towards the solutions to cater for a large number of test cases that may exist for the event sequences application. Although a number of software complexity metrics have been proposed and published for the research on the event sequences application, they do not analyse the complexity of events.

For this reason, in this paper, the researchers studied existing complexity metrics and compare five commonly used complexity metrics: Lines of Code (LOC), McCabe Cyclomatic Complexity, Information Flow Complexity, Function Points (FP) and Unique Complexity Metric (UCM). The main addressed issue was to find the complexity value for each of the event sequences that probably occurred in a test case. The rest of the paper is organised as follows. In Sect. 2, the researchers summarise numerous existing software metrics, while in Sect. 3; the researchers present the results of applying the five complexities metric on two functions from a case study and two functions from the industry. The last Sect. 4 includes the conclusions drawn.

2 Related Work

2.1 Lines of Code

Line of code (LOC) is the most common measure of source codes program size. There are four types of LOC; blank lines, comment lines, data declarations and lines that contain several separated instructions. Each type needs to be clearly understood before applying it to avoid confusion resigns. However, there is a possibility for the function to have closer behaviour and characteristic. It is an easy metric whereby the LOC is a physical entity, and normally people will do it manually. In reviewing the literature, some researchers stated the disadvantages of LOC. First is a lack of cohesion with functionality. This will happen when expert programmers may develop the same function with the intermediate programmer with far less code [6]. Second is the difference in languages. The amount of effort would be different writing in COBOL and C++ [6].

2.2 McCabe Cyclomatic Complexity

Thomas McCabe defined McCabe Cyclomatic Complexity in 1976. It will calculate based on the graph-theoretic concept where the number of linearly independent paths in a program will be counted. McCabe claimed that a greater value of v means that the modules are likely to be fault-prone and hard to maintain. The threshold for the value of v is 10. If the value is greater than 10, McCabe classified the program as possibly problematic. One of the most significant current issues by McCabe is, does the counting number of nodes give the true measure of complexity [7], while some researchers believed that relying on the cyclomatic complexity can be misleading. The cyclomatic complexity is calculated using the formula shown below:

$$v(F) = e - n + 2n \quad (1)$$

The formula can only be applied for a program with flow graph only, where e is the number of edges and n is the number of nodes available in the program. It is not satisfying if it is applied to the structural program. Besides that, the cyclomatic complexity fails to differentiate the difference between simple cases. When one program with single conditions and another program opposed the multiple conditions in conditional statements [8], both programs will present the same flow graph, but they have different complexities as the logical expression is different.

The following two segment codes illustrate the above scenario:

```
IF (a % 2 == 0) THEN x=1
    ELSE x=2
IF (a % 2 == 0) && (b % 2 == 0) THEN x=1
    ELSE x=2
```

2.3 Information Flow Complexity

The information flow complexity is suitable to measure a large scale system because it reflects the system structure [1]. The information flow complexity was validated in UNIX and proposed in 1981 [9]. It measures the total information flow between modules by considering the data structure of the modules. Three steps and two definitions have been defined by [1] to make it easy to understand the theories mentioned by [9]. The fan_in means information flowing into the function while fan_out means information flowing from the function.

Definition 1. The fan_in belongs to a number of parameters from the outside function and number of global variables read by the function.

Definition 2. The fan_out is the number of the return value and global variables written by the function.

The fan_in and fan_out represent the total possible number of input and output combinations. The given weight for the fan_in and fan_out is based on the assumption that the complexity is more than linear in terms of the connections between function. After listing the numbers of fan_in and fan_out, the information flow for each of the function will be calculated as follows:

$$IFC = length * ((fan_in) * (fan_out))^2 \quad (2)$$

This metric is good for data-driven programs. However, some researchers stated that for a function that has no external interactions with other function, the complexity value is 0.

2.4 Function Point

The Albrecht's function points (FPs) is one of the approaches to measuring the complexity of the software. FPs is used to measure how well the functionality of a system

achieved the requirements. FPs involves three steps. First, it needs to compute an unadjusted function point count (UFC). To calculate the UFC, the item in the software first needs to be categorised first. There are four categories, and each of the categories represents different functionality. They are the external inputs, external outputs, external inquiries, external and internal files. The external inputs involve inputs provided by the users of the system excluding the inquiries. Inquiries are under the external inquiries. External outputs are those items produced by the system to the users such as reports and messages. While the external files are those items that are machine readable from other systems. Lastly, the internal files are logical master files stored in the system. To get the final value for the UFC, the weight needs to be assigned to the item. The weight value is based on three ordinal scales, simple, average and complex. The UFC formula is as shown below:

$$UFC = \sum_{i=1}^n (No\ of\ items\ of\ variety_i)X(weight_i) \tag{3}$$

Second, the technical complexity factor (TCF) needs to be computed. There are 14 contributing factors already defined as depicts in Table 1. The following formula combines the 14 technical complexity factors:

$$TCF = 0.65 + 0.01 \sum_{i=1}^{14} F_i \tag{4}$$

Table 1. Components of the technical complexity factor

Factor name	Description
F1	Reliable backup and recovery
F2	Data communications
F3	Distributed functions
F4	Performance
F5	Heavily used configuration
F6	Online data entry
F7	Operational ease
F8	Online update
F9	Complex interface
F10	Complex processing
F11	Reusability
F12	Installation ease
F13	Multiple sites
F14	Facilitate change

If each F_i is set to 0, the factor varies value will start from 0.65, while if each F_i is set to 5 then the factor varies from 1.35. By multiplying the UFC and TFC, the final value of FPs will be produced.

2.5 Unique Complexity Metric

The unique complexity metric (UCM) includes all major factors that consider the effects of the complexity of programs. In the UCM, it focuses on the internal attribute and most of the researcher believed that it is important to focus on the internal attributes responsible for complexity [4]. Examples of internal attributes are complexity, bugs, testability and size of the systems. Three factors have been identified as the major contributing factors for the complexity of a system in UCM. First is the size of the code. Second is the total occurrence of operands and operators in the program. The last factor is the researchers believed that the complexity depends directly on the cognitive weights of Basic Control Structure (BSC). [4] reviewed a few factors that may influence the complexity of a system and came up with three opinions as follows:

1. The size of the code as the first factor with the assumption complexity for any single line of code is considered as 1.
2. [4] also suggested to measure the total of operators and operands and the complexity due to i^{th} line of code that can be calculated as

$$SOO_i = N_{i1} + N_{i2} \quad (5)$$

N_{i1} : Total number of operators at line i

N_{i2} : Total number of operands at line i

3. The complexity of the system should be directly correlative with the cognitive weights and basic control structures (BSC).

Cognitive weight is the degree of difficulty or effort taken to understand the number of BSC in a program [4]. Most of the existing software metrics only cover the internal structures of the program which differs from this cognitive complexity where it also covers the input-output as its processes. BSC is a basic flow control mechanism in any software system. The sequence, iteration, function call, recursion, parallel, interrupt, and branch are the common BCSs [10]. The cognitive weight is known as the maximum time and effort needed to understand the number of BCS in a software system [10]. Table 2 depicts the weights for BCSs which is defined by [4]. Those weights are assigned based on the classification of cognitive phenomenon discussed by [10]. In the effort of calculating the complexity of event sequences, the researches applied three factors mentioned above and calculated the UCM for the two simple programs using the cognitive weights as depicted in Table 2. The UCM is calculated as follows:

$$UCM = \sum_{i=1}^n (1 + SOO_i * CW_i) \quad (6)$$

[10] reviewed the BCSs details to determine the complexity and component functionality which are based on the theory of cognitive informatics. The cognitive informatics refer to the functional complexity found in the software and it depends on input, output and internal processing [10]. The UCM followed the discipline, and some of the basic measurement requirements are stated in the Measurement Theory (MT) [11].

Table 2. Assigning weights for the BCSs

Category	Basic control structures	Cognitive weight
Sequence	Sequence	1
Branch	If-Then-Else	2
	Case	3
Iteration	For-do	3
	Repeat-until	3
	While-do	3
Embedded component	Function Call	2
	Recursion	3
Concurrency	Parallel	4
	Interrupt	4

3 Case Study

The complexity of the software depends on two factors [9]. First is the complexity of the function and second is the complexity in terms of connections with other functions. In this research, the researchers look into the complexity of events in each of the test cases. Two small functions written in Java taken from the circular queue program will be used for the analysis of the case study. The segment codes are given in Table 3. In order to hold an enduring fascination with the industry, the researchers extracted two functions from the industry that consist of more line of codes and the operations are more complex compared with the small functions shown in Table 3. The programs were written in VB.Net. Table 4 shows segment codes for *Move Change Language in Textbox Fields* Function and *Move Textbox Fields to DB* Function. Those functions are possible events that may occur in a test case.

Table 3. Segment codes function add and function remove.

Function Add	Function Remove
<pre> /** This is add function */ private void add(int inVal) { if (len < QSIZE) { rear = (rear + 1) % QSIZE; dataQ[rear] = inVal; dataQori[rear] = inVal; len = len + 1; } else if (len == QSIZE) { } } /**end of if statement */ } </pre>	<pre> private void remove() { if (len > 0) { dataQ[front] = null; front = (front + 1) % QSIZE; len = len - 1; } else if (len == 0) { } } </pre>

Table 4. Segment codes function move change language in textbox fields and function move textbox fields to DB

Function Move Change Language in Textbox Fields	Function Move Textbox Fields to DB
<pre> Private Sub sbMoveChangeLanguageInTextBoxFields () If ddlQUES_LANG.SelectedValue = "BS" Then txtQUES_QUESTION.Content= bQUES_QUESTION_BS.Text txtQUES_ANS1.Text = txtQUES_ANS1_BS.Text txtQUES_ANS2.Text = txtQUES_ANS2_BS.Text txtQUES_ANS3.Text = txtQUES_ANS3_BS.Text txtQUES_ANS4.Text = txtQUES_ANS4_BS.Text ElseIf ddlQUES_LANG.SelectedValue = "EN" Then txtQUES_QUESTION.Content= lbQUES_QUESTION_EN.Text txtQUES_ANS1.Text = txtQUES_ANS1_EN.Text txtQUES_ANS2.Text = txtQUES_ANS2_EN.Text txtQUES_ANS3.Text = txtQUES_ANS3_EN.Text txtQUES_ANS4.Text = txtQUES_ANS4_EN.Text ElseIf ddlQUES_LANG.SelectedValue = "CN" Then txtQUES_QUESTION.Content= lbQUES_QUESTION_CN.Text txtQUES_ANS1.Text = txtQUES_ANS1_CN.Text txtQUES_ANS2.Text = txtQUES_ANS2_CN.Text txtQUES_ANS3.Text = txtQUES_ANS3_CN.Text txtQUES_ANS4.Text = txtQUES_ANS4_CN.Text End If End Sub </pre>	<pre> Private Sub sbMoveTextBoxFieldsToDBFields(ByVa l strMode As String)With objBE .QUES_CERTTYPE = hdnCER_TTYPE.Value .QUES_UID = SysUtility.clsEmbeddedQuote (txtQUES_UID.Text.Trim) If RDBTN_Single.Checked = True Then .QUES_TYPE_MOD = "S" ElseIf RDBTN_Scenario.Checked = True Then .QUES_TYPE_MOD = "SC" End If .CR_LVL_STRING_MOD= hdnCR_LVL_STRING.Value .QUES_QUESTION_MOD= SysUtility.clsEmbeddedQuote (txtQUES_QUESTION.Content.Trim) .QUES_ANS1_MOD= SysUtility.clsEmbeddedQuote(txtQUE S_ANS1.Text.Trim) .QUES_ANS2_MOD= SysUtility.clsEmbeddedQuote(txtQUE S_ANS2.Text.Trim) .QUES_ANS3_MOD= SysUtility.clsEmbeddedQuote(txtQUE S_ANS3.Text.Trim) .QUES_ANS4_MOD= SysUtility.clsEmbeddedQuote(txtQUE S_ANS4.Text.Trim) .QUES_CORRECTANS_MOD= ddlQUES_CORRECTANS.SelectedValue .QUES_RMK_MOD= SysUtility.clsEmbeddedQuote (txtQUES_RMK.Text.Trim) .QUES_VERIFY_STATUS = 1 .QUES_STATUS = 0 If rdbSTATUS10.Checked = True Then QUES_AVAILABLE_MOD = 0 End Sub </pre>

Complexity values for each of the functions above are calculated and results are shown in Table 5. However, the details of the calculations are not included in this paper because of lack of space. Based on the LOC values in Table 5, it show that *add* function and *move textbox fields to DB* function have high values since both have more lines of codes. Line of code is based on counting the lines of the codes. However, the values become not realistic if one of the functions has more comments since LOC will calculate comments as one line. Because of that, some researchers agree that one of the lines of code drawback is in terms of the language syntax and style.

Table 5. Complexity values calculations

Complexity metric	Simple case study (Java language)		Industry case study (VB.Net language)	
	Function add	Function remove	Function move change language in textbox fields	Function move textbox fields to DB
Line of code	13	8	49	53
McCabe cyclomatic complexity	3	3	7	10
Information flow complexity	16	4	121	256
Function points	28.08	20.28	38.0	46.06
Unique complexity metric	37	29	180	205

For example, practically programmers will leave blank spaces in their codes so that it will be easy for others to read. However, to estimate the programming effort, in reality, the blank space does not contribute anything compared to the lines that involved calculations or algorithms. Since line of code is sensitive with the issue of language syntax and style, this metric will not be selected for future research. McCabe cyclomatic complexity is based on measuring the linearly independent path. Basically, there are four categories assigned by McCabe. If the complexity value is below 10, the code is simple program and without much risk. If the complexity value are ranges between 11 and 20, the code is more complex and under moderate risk. For the complexity value ranges from 21 to 50, the code is complex and high risk. While for the complexity value above than 50, the code is under category not testable and very high risk. Based on values stated in Table 5, it shows that all the functions are simple program since their complexity value is under 10. Some researchers stated that the codes are structured and easy to maintain if the complexity value is fewer than 10.

However, the researchers conclude that if the structures of functions are almost same like *add* function and *remove* function, there is a possibility to have similar complexity value even tough number of nodes and edges are different. McCabe cyclomatic complexity is a quality metric as it can produces relative complexity value for various designs but it only focus on the flow of the program. Some researchers agree that the

weighing scheme in McCabe cyclomatic complexity is too simple. With that reasons, the researchers will not select McCabe cyclomatic complexity for future research. For the information flow complexity, *add* function and *move textbox fields to DB* function still have high value compared to other functions as well. Based on the concept of the information flow complexity, the calculation is based on the flow of information through available input parameters, output parameters and global data structures. Many researchers agree that it is a good complexity metric as it can be derived during the design phase but it will produce complexity value 0 if there are no external interactions in the functions. There is a possibility for the event sequences test case to have simple functions whereby no external interactions. For that reason, researchers believed that information flow complexity is not suitable for their research. Function points are known as a metric that focus on to measure the amount of functionality in a system. However, because of the evaluation of FP from a specification of the system, it cannot be done automatically. It will be evaluate manually and different group and expertise will evaluate each of the specifications differently. FPs is not suitable to measure the complexity event sequences in terms of the differentiating specified items. There are similarities in opinions between previous researchers regarding limitations of the FPs.

Some of the limitations discussed are problems with the subjectivity in the TCF since the range is between 0.65 and 1.35 and problems of accuracy, changing requirements and subjective weighting. This metric will not be selected by the researchers for the future research even though the differences complexity value for the function points metric is realistic. In Table 5, under the UCM metric, *add* function and *Move Change Language in Textbox Fields* function are consider complex since both functions have high values. The difference of UCM values are realistic even the number of lines codes not too different. In UCM, function that has highest UCM value means the function is complex compared with other functions. One of the reasons the *Move Change Language in Textbox Fields* function has high UCM value because the segment code consists many function calls. In the classification of cognitive phenomenon, function call is categorised under embedded component which is under category medium complex.

Furthermore, different languages will not affect the UCM values since UCM is language independent. However, UCM does not assign the upper and lower bound complexity values. Further analysis should be done for this restriction to make sure the complexity values more comprehensive to be used. The determination of the complexity of event in this paper is used to assign appropriate weight before calculating the priority value of event sequences test cases. The event needs to be ranked to identify the most important events. The researcher will choose UCM as a complexity metric to measure the complexity of event sequences because of the reasons mentioned above. Besides that, the weight is given based on classification of the BCSs. Like previous work done on GUI test cases, the assigning weight is based on the sensitivity for each of the events that may occur in a test case.

4 Conclusions

In this paper, the researchers studied and compared five commonly software complexity metrics. Two functions taken from a simple case study and two functions taken from the industry has been chosen to evaluate the effectiveness of applied each of the software complexity metrics. This is part of on-going research which aims to find a technique that can produce a unique priority value while prioritize event sequences test cases. The researchers believed that the complexity value gained from this analysis can help future work to create new technique to produce a unique priority value for each of the test cases. UCM is selected as a metric to calculate the complexity of event sequences based on discussions on the importance, strengths, and weaknesses during the analysis. UCM is a simple metric but can fulfill the requirement of a good metric and it will aid developers to evaluate the software complexity as an analyser for the software engineering measurement.

References

1. Azim, A., Ghani, A., Tieng, K., Geoffrey, W., Muketha, M., Wen, W.P.: Complexity metrics for measuring the understandability and maintainability of business process models using goal-question-metric (GQM). *J. Comput. Sci.* **8**(5), 219–225 (2008)
2. Bhatt, K., Tarey, V., Patel, P.: Analysis of source lines of code (SLOC) Metric, **2**(5), 3–7 (2012)
3. Briand, L.C., Morasca, S., Basili, V.R.: Property-based software engineering measurement. *IEEE Trans. Softw. Eng.* **22**(1), 68–86 (1996). doi:[10.1109/32.481535](https://doi.org/10.1109/32.481535)
4. Butkiewicz, M., Madhyastha, H. V., Sekar, V.: Understanding website complexity: measurements, metrics, and implications. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, IMC 2011, pp. 313–328, November 2011
5. Henry, S., Kafura, D.: Software structure metrics based on information flow. *IEEE Trans. Softw. Eng.* **SE-7**(5), 510–518 (1981). <http://doi.org/10.1109/TSE.1981.231113>
6. Kaner, C.: Rethinking software metrics. *Softw. Test. Qual. Eng.* **2**(2), 50–57 (2000)
7. Madi, A., Zein, O.K., Kadry, S.: On the improvement of cyclomatic complexity metric. *Int. J. Softw. Eng. Appl.* **7**(2), 67–82 (2013)
8. Misra, S., Akman, I.: A unique complexity metric. In: Gervasi, O., Murgante, B., Laganà, A., Taniar, D., Mun, Y., Gavrilova, Marina, L. (eds.) ICCSA 2008. LNCS, vol. 5073, pp. 641–651. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-69848-7_52](https://doi.org/10.1007/978-3-540-69848-7_52)
9. Munson, J.C.: Software measurement: problems and practice. *Ann. Softw. Eng.* **1**(1), 255–285 (1995). doi:[10.1007/BF02249053](https://doi.org/10.1007/BF02249053)
10. Shi, Y., Xu, S.: A new method for measurement and reduction of software complexity. *Tsinghua Sci. Technol.* **12**(S1), 212–216 (2007)
11. Shao, J., Wang, Y.: A new measure of software complexity based on cognitive weights. *Can. J. Electr. Comput. Eng.* **28**(2), 69–74 (2003). doi:[10.1109/CJECE.2003.1532511](https://doi.org/10.1109/CJECE.2003.1532511)

Implementation of Ceph Storage with Big Data for Performance Comparison

Chao-Tung Yang^(✉), Cai-Jin Chen, and Tzu-Yang Chen

Department of Computer Science, Tunghai University, Taichung City 40704, Taiwan
ctyang@thu.edu.tw, amranchen@yahoo.com, applepaoo@gmail.com

Abstract. High Available share storage becomes one of the important resource information to expand our system especially for Big Data implementation system. To consider the world demand of reduce high risk data corrupt and improve the reading and writing storage performance, through our research we mainly apply Ceph storage with Big Data Performance testing in order to solve the best reading and write speed performance and data backup. This system is started from Hadoop operations. The data is stored in the Hadoop Distributed File System (HDFS) and copied to Alluxio MEM space. The data through Map Reduce processing (Mapping – Sorting – Filtering – Reducing) got the result and the output will be stored in to Alluxio MEM space. For the first experimental, we use S3 API and Rados Gateway of Ceph components as a bridge between Alluxio and Object Storage Daemon (OSDs). The second experimental is the same like first environment, but the output of Map Reduce will be directly connect to Object Storage Daemon using Ceph File System (CephFS). The data is more safety in the Ceph than in the Alluxio MEM only, because OSDs can back up the data with object storage levels. We also can use S3 browser (GUI) to maintenance the OSD's data, e.g.: grant access, keep folder, create user account, move data location etc. The last one, we use Inkscope to monitor all system, if there is any problem the system will respond the error or giving warning alerts to the user.

Keywords: High available · Big data · Share storage · Recovery · Monitor

1 Introduction

In recent years, the importance of continuous delivery tooling continues grown up and needed, the requirements for availability and scalability also increase. Tools must be High Available so engineers can always deliver new software. Many companies need to have automatically expand share storage and their system service to remain available even when a component of that service fails. A component could for example be some kind of process. To make that more reliable you can run multiple instances of this process and load balance between them. A component can also be a collection of data. Data can be stored in a database or on a file system, or whatever, but it will end up on some kind of storage device, like an SSD, a Hard Drive or even a Tape.

Many things can go wrong with storage, it need High Available share storage to reduce high risk of device breaks, the file system get corrupt, the system that is attached

to the device can break and so on. To avoid outages, it would like to have multiple instances of all components. If something wrong happened, the other components could automatically take over and both of replicated and distributed all data. Users do not worry any outages happened.

In this paper, we use Ceph Storage, Alluxio and Apache Hadoop integrated technology to collect a fully working system and using Hi-Benchmark performance testing tool to measure sequential and partial read/write speeds then found increase the performance solution that looked for many developers and enterprises nowadays.

2 Background and Related Work

2.1 Background – Ceph Storage

Ceph is an object storage based free software storage platform that stores data on a single distributed computer cluster, and provides interfaces for object-, block- and file-level storage. Ceph storage clusters are designed to run on commodity hardware, using an algorithm called CRUSH (Controlled Replication under Scalable Hashing) to ensure data is evenly distributed across the cluster and that all cluster nodes can retrieve data quickly without any centralized bottlenecks (Fig. 1).

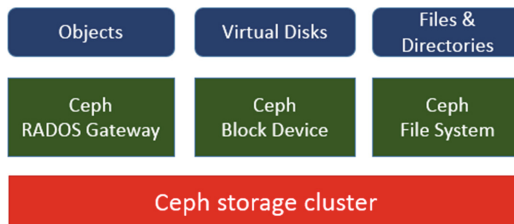


Fig. 1. General Ceph architecture

Ceph consists of three component services, that are RADOS Gateway (Ceph Object Gateway Daemon), Block Device and CephFS.

- RADOS Gateway is a bucket –based REST gateway, compatible with S3 and Swift.
- RBD is a reliable and fully-distributed block device, with a Linux kernel client and QEMU/KVM driver
- CephFS is a POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE.

2.2 Background – Alluxio

Alluxio (former known as Tachyon) is an open source memory speed virtual distributed storage system. It unifies data access and bridges computation frameworks and underlying storage system. Applications only need to connect with Alluxio to access data stored in any underlying storage systems. The following is Alluxio Architecture layout (Fig. 2).

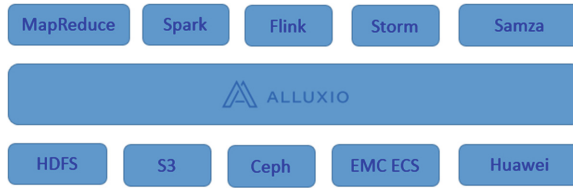


Fig. 2. Alluxio architecture

2.3 Background – Apache Hadoop

Apache Hadoop is now the most popular solutions of big data processing, which is Apache Software Foundation open source framework. The Apache Hadoop framework is built on top of the Hadoop Distributed File System (HDFS), which supports a stable and automatic distributed processing system. Hadoop implements Map Reduce programming framework, which composed by the map and reduce and the input divided into the same size, it allow data fragment executed on any node in the cluster (Fig. 3).

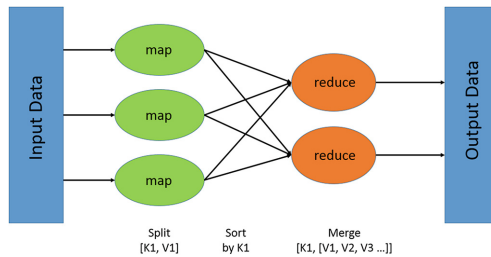


Fig. 3. Map reduce architecture

3 System Design and Implementation

3.1 Experimental Environment

In order implement Ceph integration system well, we prepare 9 CPUs as servers of 12 CPUs available as an experimental environment. We use internal physical IP Address on whole servers and putty.exe as client remote tool (Fig. 4).



Fig. 4. Experiment servers

3.2 System Architecture

Figure 5 overall system architecture that we deployed on our system. In this figure shows you rough sketch system that consists of Ceph, Alluxio, Hadoop and Inkscope.

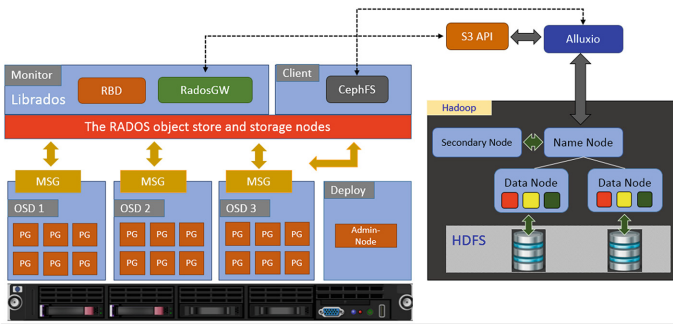


Fig. 5. Experiment architecture

Above experimental environment is described as below:

First experimental, this system uses word count data loads to Map Reduce environment. This data can be adjusted based on user requirement. In here, we set three kinds of sample data size: 5 GB, 10 GB and 15 GB. Through Map <key, value> , sort by key and Merge [key, [value-1, value-2, value-n]] algorithm, the data is sent in to Alluxio memory-speed virtual storage system. We also activate S3 API (one of Ceph components) in the Alluxio file configuration, through S3 and RADOS Gateway the data is also stored in to the Object Storage Daemon (OSDs). For the second experimental, this system is the same data load like the first one system, but in this system there is not through S3 API and RADOS Gateway to store the data in to Object Storage Daemon (OSDs), OSDs directly connects to the Alluxio plugin. The second experiment reduces S3 API and RADOS Gateway levels. These environments have monitored by Inkscope monitoring system. Inkscope monitor all system in Ceph. If there is any outstages, the Inkscope will appear alerts for user.

4 Experimental Environmental and Results

4.1 Experimental Environment

For the hardware specification of the computer that we use 6 servers for Ceph, 1 server for Hadoop and Alluxio and 1 server for Inkscope. These servers are the physical machine. Ubuntu 14.04 with 64 bit is adopted as our operating system.

4.2 Experimental Results

4.2.1 MapReduce Input and Output

First, we show the real-time Map Reduce in Hadoop environment. The source data is stored to HDFS and then copy to Alluxio memory space environment. The data will be stored in to alluxio/wordcount/myfile (Fig. 6).

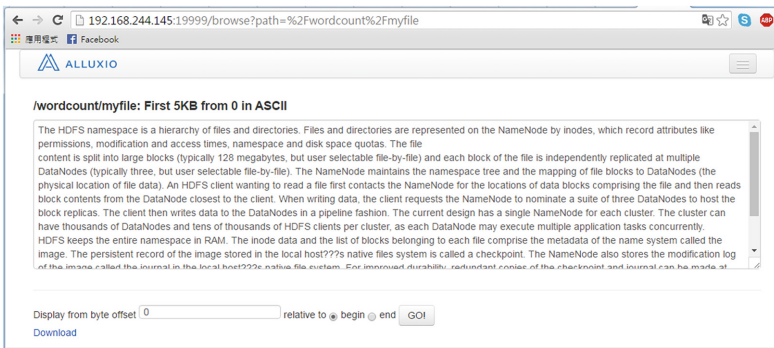


Fig. 6. Data source in Alluxio

The output result is stored in the Alluxio memory space like below (Figs. 7 and 8):

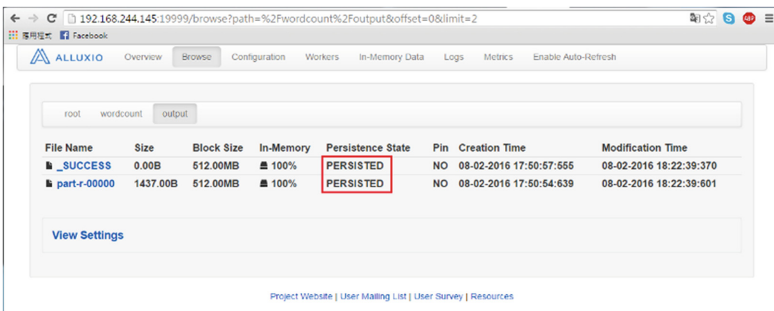
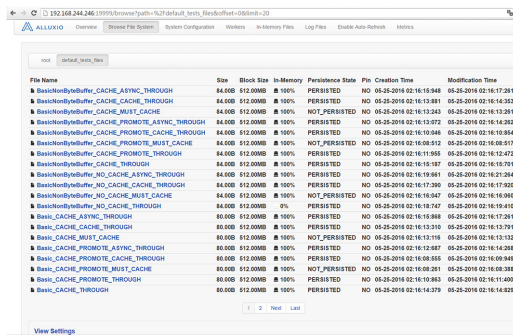


Fig. 7. Map reduce output 1

```
userceph@hadoop:~/alluxio-1.2.0$ bin/alluxio fs cat /wordcount/output/part-r-0
9000
(the 1
(typically 2
128 1
An 1
DataNode 2
DataNodes 5
During 1
Files 1
For 1
HDFS 4
NameNode 7
RAM 1
The 10
When 1
a 6
access 1
also 1
and 12
application 1
ake 2
as 1
at 2
attributes 1
be 1
```

Fig. 8. Map reduce output 2

Besides that, all the MapReduce data have been store in to the Ceph OSDs that used S3 API and Rados Gateway as the first way and directly CephFS for the second way (Fig. 9).



File Name	Size	Block Size	In-Memory	Persistence State	Prio	Creation Time	Modification Time
■ Basic OSD by Buffer_CACHE_ASYNC_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:15.948	05-25-2016 02:16:17.281
■ Basic OSD by Buffer_CACHE_CACHE_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:13.881	05-25-2016 02:16:14.253
■ Basic OSD by Buffer_CACHE_MUST_CACHE	84.00B	512.00MB	100%	NOT_PERSISTED	NO	05-25-2016 02:16:13.243	05-25-2016 02:16:13.281
■ Basic OSD by Buffer_CACHE_PROMOTE_ASYNC_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:13.072	05-25-2016 02:16:14.262
■ Basic OSD by Buffer_CACHE_PROMOTE_CACHE_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:10.046	05-25-2016 02:16:19.854
■ Basic OSD by Buffer_CACHE_PROMOTE_MUST_CACHE	84.00B	512.00MB	100%	NOT_PERSISTED	NO	05-25-2016 02:16:05.512	05-25-2016 02:16:05.517
■ Basic OSD by Buffer_CACHE_PROMOTE_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:11.955	05-25-2016 02:16:12.472
■ Basic OSD by Buffer_CACHE_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:15.187	05-25-2016 02:16:19.791
■ Basic OSD by Buffer_NO_CACHE_ASYNC_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:15.915	05-25-2016 02:16:21.264
■ Basic OSD by Buffer_NO_CACHE_CACHE_THROUGH	84.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:17.390	05-25-2016 02:16:17.200
■ Basic OSD by Buffer_NO_CACHE_MUST_CACHE	84.00B	512.00MB	100%	NOT_PERSISTED	NO	05-25-2016 02:16:15.047	05-25-2016 02:16:16.800
■ Basic OSD by Buffer_NO_CACHE_THROUGH	84.00B	512.00MB	0%	PERSISTED	NO	05-25-2016 02:16:15.747	05-25-2016 02:16:19.619
■ Basic_CACHE_ASYNC_THROUGH	80.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:15.888	05-25-2016 02:16:17.281
■ Basic_CACHE_CACHE_THROUGH	80.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:13.310	05-25-2016 02:16:13.791
■ Basic_CACHE_MUST_CACHE	80.00B	512.00MB	100%	NOT_PERSISTED	NO	05-25-2016 02:16:15.118	05-25-2016 02:16:13.132
■ Basic_CACHE_PROMOTE_ASYNC_THROUGH	80.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:12.887	05-25-2016 02:16:14.268
■ Basic_CACHE_PROMOTE_CACHE_THROUGH	80.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:09.555	05-25-2016 02:16:09.840
■ Basic_CACHE_PROMOTE_MUST_CACHE	80.00B	512.00MB	100%	NOT_PERSISTED	NO	05-25-2016 02:16:09.241	05-25-2016 02:16:08.288
■ Basic_CACHE_PROMOTE_THROUGH	80.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:10.883	05-25-2016 02:16:11.400
■ Basic_CACHE_THROUGH	80.00B	512.00MB	100%	PERSISTED	NO	05-25-2016 02:16:15.279	05-25-2016 02:16:14.829

Fig. 9. Ceph OSD output

4.2.2 Inkscope Monitoring System

The following is our monitoring system for Ceph environment by Inkscope. Inkscope monitors all server hardware, network, pools, and services that show as below (Fig. 10): If your OSD servers in Ceph have any problem, this system will give you alert as the following layout (Fig. 11).

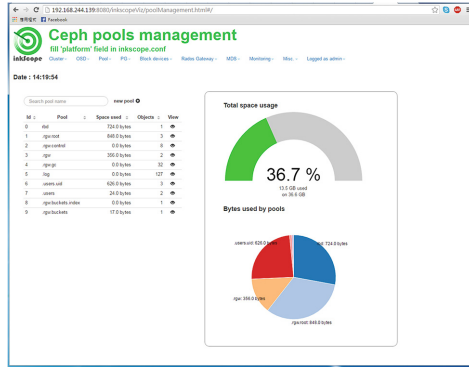


Fig. 10. Ceph pools management

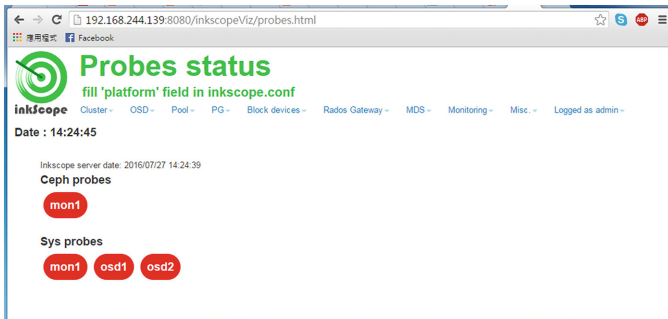


Fig. 11. Inkscope Ceph pools management

4.2.3 The Performance Between RadosGW and CephFS Comparison

After this system has been built, we are also using FIO tool to measure the performance using Rados Gateway and CephFS. The following is our experimental result (Table 1).

Table 1. Speed Performance Test with Read, Write, Randomize for each OSDs

	osd1			osd2			osd3			CephFS		
	1G	3G	5G	1G	3G	5G	1G	3G	5G	1G	3G	5G
S.Read	2.4	6.1	5.8	2	4.9	2.1	1.9	2.1	2.1	4	3.8	7.3
S.Write	10.3	17.6	19.1	11.9	16.3	18.5	18.3	17.7	17.7	33.5	32.7	32.1
Rand.Read	0.9	1.6	1.6	0.7	1.2	1.1	0.8	1.3	2.4	0.9	0.8	0.8
Rand.Write	1.5	1.7	0.9	1.4	1.5	1.2	1.4	1.7	0.9	2.6	0.5	0.5
S.Read(30%)	1.9	3.3	3.4	1.7	4.1	4	2	4.2	3.8	4.8	5	5
S.Write(30%)	0.8	1.4	1.5	0.7	1.8	1.7	0.8	1.8	1.6	2	2.1	2.1
Rand.Read(30%)	0.6	0.8	0.9	0.6	0.5	0.8	0.6	0.8	0.5	0.6	0.5	0.4
Rand.Write(30%)	0.2	0.3	0.4	0.2	0.2	0.3	0.2	0.3	0.3	0.2	0.2	0.1

The following is each OSDs speed based on above table.

Table 2. IOPS each OSDs environment

	osd1			osd2			osd3			CephFS		
	1G	3G	5G	1G	3G	5G	1G	3G	5G	1G	3G	5G
S.Read	156	383	362	131	306	134	123	133	134	252	240	461
S.Write	645	1103	1198	749	1022	1159	1143	1109	1109	2099	2044	2009
Rand.Read	56	104	104	45	80	71	49	82	152	54	51	53
Rand.Write	98	107	55	88	94	81	92	111	58	37	34	31
S.Read(30%)	121	208	218	110	259	254	126	263	241	304	314	317
S.Write(30%)	50	90	94	46	112	110	53	113	103	130	134	135
Rand.Read(30%)	37	53	60	37	34	52	37	54	54	37	34	24
Rand.Write(30%)	15	21	24	15	14	20	15	21	21	15	14	10

Based on above table, three OSDs (osd1, osd2 and osd3) is using Rados Gateway and S3 API way to store the data in to OSDs. We measured each OSD speed and got above result. The second way is through CephFS stored the data to OSDs. According to both of experiment, we can conclude that CephFS performance is better than through Rados Gateway & S3 API. Based on IOPS measuring, it is the same result. The higher the value, the better the read/write performance (Fig. 12 and Table 2).

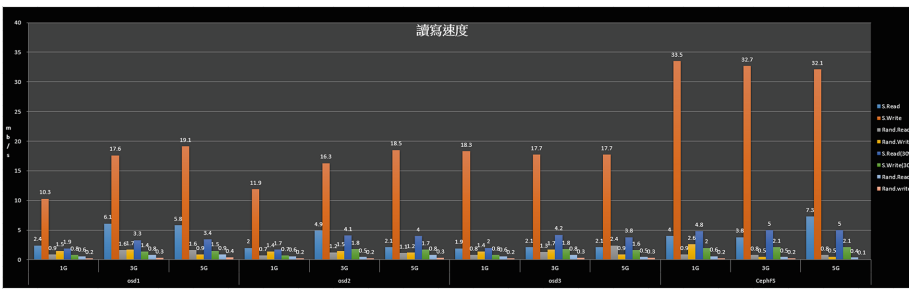


Fig. 12. Speed performance test with read, write and randomize for each OSD

IOPS (Input/Output Operations per second) is a performance measurement used for Ceph Storage. This is used to give the response time on workload per seconds, the higher value the better performance (Fig. 13).

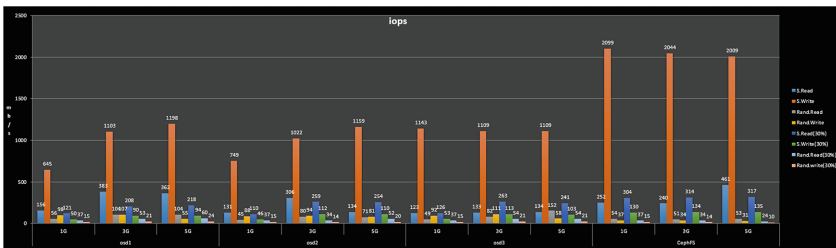


Fig. 13. IOPS performance test with read, write and randomize for each OSD

5 Conclusion

This experimental uses open-source software components that are Ceph, Alluxio and Hadoop environment. The benefit of its operating performance as below:

1. High speed read/write storage and flexibility to expand storage size with OSD additional and NFS formatted.
2. Ceph data is more secure because it consists of RADOS Gateway as a middle level security.
3. CephFS way is better performance than Rados Gateway performance, because the data is directly connected from HDFS to OSDs.

Acknowledgements. This work is supported in part by the Ministry of Science and Technology, Taiwan ROC, under grants number MOST 104-2221-E-029-010-MY3, 105-2622-E-029-003-CC3 and MOST 105-2634-E-029-001-.

References

1. Azzedin, F.: Towards a scalable HDFS architecture. In: Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013, pp. 155–161 (2013)
2. Peng, C., Jiang, Z.: Building a cloud storage service system. *Procedia Environ. Sci.* **10**, 691–696 (2011)
3. Zheng, Q., Chen, H., Wang, Y., Zhang, J., Duan, J.: Cosbench: cloud object storage benchmark. In: 4th ACM/SPEC International Conference on Performance Engineering (ICPE 2013). ACM (2013)
4. Zhan, L., Fang, X., Li, D.: Source of the Document Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2016 (2016)

Web Technology

Predicting Engaging Content for Increasing Organic Reach on Facebook

Natthaphong Phuntusil and Yachai Limpiyakorn^(✉)

Department of Computer Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Natthaphong.Ph@student.chula.ac.th, Yachai.L@chula.ac.th

Abstract. Over the past few years, many people have been concerned about declines in organic reach for their Facebook Pages. This has been a pain for many businesses, especially those small businesses and startup. Organic reach refers to how many people you can reach for free on Facebook by posting to your page. The declined organic reach results from some key changes to improve how News Feed chooses content. News Feed is aimed at becoming more engaging, even as the amount of content being shared on Facebook continues to grow. This paper presents a technique to increase Facebook organic reach. The method investigates some promising factors to predict the engaging content posting on business Pages, so that the post would gain exposure in News Feed of the liking users on Facebook. The proposed approach provides the alternative for businesses to increase the organic reach without more expense on advertising posted on Facebook Pages.

Keywords: Social commerce · Decision support · Engaging content · Organic reach · Social network

1 Introduction

Facebook users have probably felt that more and more content is being created and shared on social media every day. Thanks to devices like smartphones, many people can share important moments and experiences, photos and videos, or articles with just a few swipes of the finger or taps on a button. In addition to the growth in content, people are also liking more Pages. As a result, competition in Facebook News Feed is increasing, resulting in harder for any story to gain exposure in News Feed.

News Feed is the place on Facebook where people view content from their family and friends, as well as businesses [1]. It is designed to show each person on Facebook the content that is most relevant to them. Some key changes to improve how News Feed chooses content have resulted in declined organic reach. Organic reach refers to how many people you can reach for free on Facebook by posting to your page [2]. Declines in Pages' organic reach have been a pain for social commerce. Many businesses, especially small businesses and startup, are concerned about the rising expense of advertising their business Pages. This research has thus investigated some factors potentially predict the engaging content posting on business Pages, so that the post would gain exposure in News Feed of the liking users on Facebook. Finally, this would result in the increase of organic reach on Facebook Pages, and less advertising cost for businesses.

2 Facebook

2.1 Facebook Pages

Pages are for brands, businesses, organizations and public figures to create a presence on Facebook, whereas profiles represent individual people. Anyone with an account can create a Page or help manage one, if they have been given a role on the Page like admin or editor. People who like a Page and their friends can get updates in News. There are 6 primary categories [3–5] to choose from, including: Local Business or Place; Company, Organization or Institution; Brand or Product; Band or Public Figure; Entertainment; Cause or Community.

2.2 Facebook Graph API

The Graph API is the primary way to get data in and out of Facebook's platform. It is a low-level HTTP-based API that can be used to query data, post new stories, manage ads, upload photos and a variety of other tasks that an application might need to do [6].

3 Similarity Measurement

Term Frequency–Inverse Document Frequency, or TF-IDF, is a numerical statistic used as a weighting factor in information retrieval and text mining [7]. The intent is to reflect how important a word is to a document in a collection or corpus. Zhang and Pennacchiotti [8] studied the correlation between Facebook categories and eBay meta-categories. The result suggested that the set of Facebook categories may be predictive of purchase behaviors. The list of Facebook categories contains 214 features. For each user u and Facebook category f the feature value is computed using TF-IDF to reflect the user interest associated with a particular category.

4 Research Methodology

Figure 1 illustrates steps of the proposed method. In brevity, to predict the engaging content, the process starts with creating the selected business Page, then comparing the similarity of user interests between people found in the business Page and content page sources. Cosine similarity and Pearson correlation are applied for the similarity analysis in this work. It is observed that many null values exist in the dataset. However, it does not mean there is none of interests on null category. It might be pages not found in null category. Therefore, Pearson correlation is used to test the result by minus with the average for shifting the value to centre.

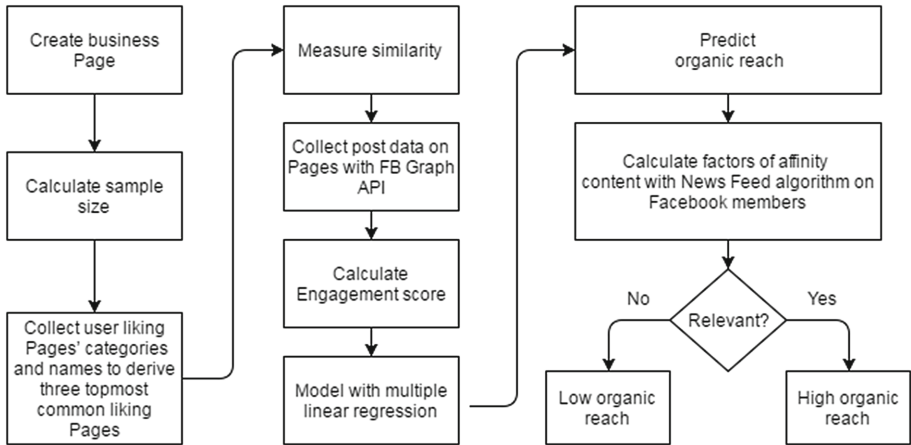


Fig. 1. Process of identifying engaging contents to increase Facebook organic reach

4.1 Create Business Page

For the experiment, a new business Page was created. The audience was built by promoting the Page with Facebook ads to the target goal of 1,000 fans (Page likers). Meanwhile, we had been posting to sell products on the business Page timeline. Figure 2 shows the findings reported by Facebook Insight. We found that 96% is Women and 4% is Men. The information suggested that the target group would be women. The post plan will then focus on the engaging content for women.

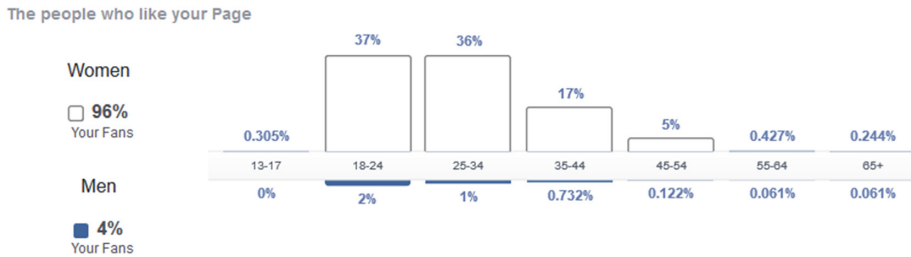


Fig. 2. Findings of created business page reported by Facebook insight

4.2 Calculate Sample Size

Based on Yamane [9] formula, the size of dataset containing 400 users is used as the sample size n in this work, while the population size $N = 1.18$ Billion daily active users. The result provides 95% confidence level, and level of precision $e = 0.05$.

4.3 Collect User Liking Pages' Categories and Names

A set of 400 users liking the created business Page was randomly selected. The library of Python 3, Selenium Webdriver [10], is used to control Firefox browser to scrape the information of all 400 users liking FB pages' categories and names. This is the challenging step to collect the important data of the counts and categories of other FB liking Pages from the 400 users who likes the created business Page. Table 1 summarized the details of the top three liking FB Pages of the selected 400 users.

Table 1. Top three of liking FB pages collected from 400 users liking created business Page

FB page	Content type	Fans' likes	$Like(u, f)^a$
1. Spice	Beauty, Healthy	686,563	197
2. Vonvon.me	Entertainment, Game	28,922,326	149
3. Jatiewpainai	Travel	788,509	145

^a $Like(u, f)$ is the number of page likes by user u in page f

Figure 3 shows the Facebook users' interests based on Pages they like. Although there are many reasons for users to like their Pages, we assume that people want to get more useful information from Facebook Pages. In this work, the amount of liking page

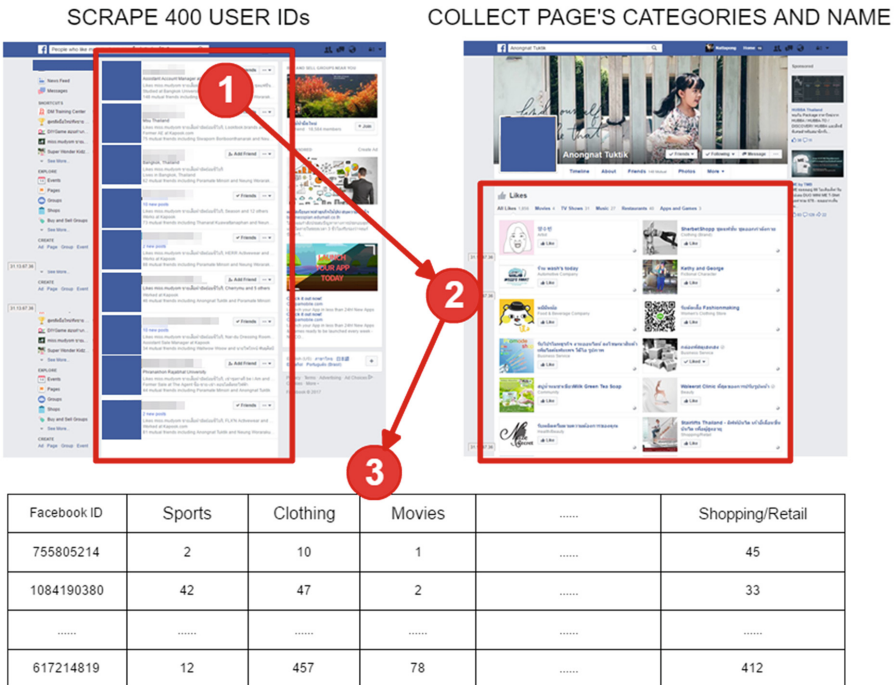


Fig. 3. Collected data of liking user Pages' categories

categories is used to measure similarity of user interests between the business Page and content page sources.

4.4 Measure Similarity

Currently, the list of Facebook categories contains 156 features. User page’s categories in terms of frequency is used to calculate the similarity of people [11]. The sampling data of 400 users are used for the computation of cosine similarity between the business Page and sources of content page. If pairwise vectors are same orientation, then cosine similarity equals 1, whereas 0 denotes both vectors are not similar. The similarity threshold is set to 0.5 or 50% in this work. Example similarity results are:

Spice (41%), Jatiewpainai (32%), Starvingtime (43%), WomenBeautyCommunity (40%), SistaCafePage (38%), Jeban(30.5%).

4.5 Collect Post Data on Pages with Facebook Graph API

The post data on Page were collected from Facebook timeline to calculate engagement scores and post scores. The collected fields consist of: “*status_id*”, “*status_message*”, “*link_name*”, “*status_type*”, “*status_link*”, “*status_published*”, “*num_reaction*” are summation of all user reactions such as “*num_comments*”, “*num_loves*”, “*num_wows*”, “*num_hahas*”, “*num_sads*”, “*num_angrys*”, “*num_shares*”, “*num_likes*”.

4.6 Calculate Engagement Score

There are many Facebook engagement score formulas. Equation 1 shows the engagement score formula from the website unmetric.com. The formula is derived from user research and observations on features and functionalities of different social media platform [12, 13]. Table 2 describes the variables used in Eq. 1. The default unmetric values of α , β are used in the formula.

$$E_{score} = \frac{(N_{like} + \alpha N_{comment} + \beta N_{share})}{N_{audience}} \times 10^4 \tag{1}$$

Table 2. Variable description.

Symbols	Meaning
E_{score}	Engagement rate
N_{like}	Number of likes on for each posts
$N_{comment}$	Number of comments for each posts
N_{share}	Number of shares for each posts
$N_{audience}$	Audience reception rate = (Number of brand fans) ^{0.8}
α	5
β	10

In this work, Share is considered as the factor with more influence on the audience reach, compared to Like and Comment. We focus on the highest shares of content page post and then calculate the engagement score for prediction. The result of extrapolation is not sensible scoring because the similarity of each page does not equal. The Similarity score is thus applied as the weight to adjust the engagement score in order to obtain the Post score as shown in Eq. 2.

$$Post_{score} = E_{score} \times Sim_{score} \tag{2}$$

The Post score will then be used to pick the content for posting on the business Page. Next, the evaluation of the effect results of engagement scores between two pages is carried out. The content with the highest score from Facebook source pages was posted as well as a variety of contents posted in random time within the same day.

Figure 4 illustrates how to derive the values of Sim_{score} , E_{score} , $Post_{score}$ in sequence.

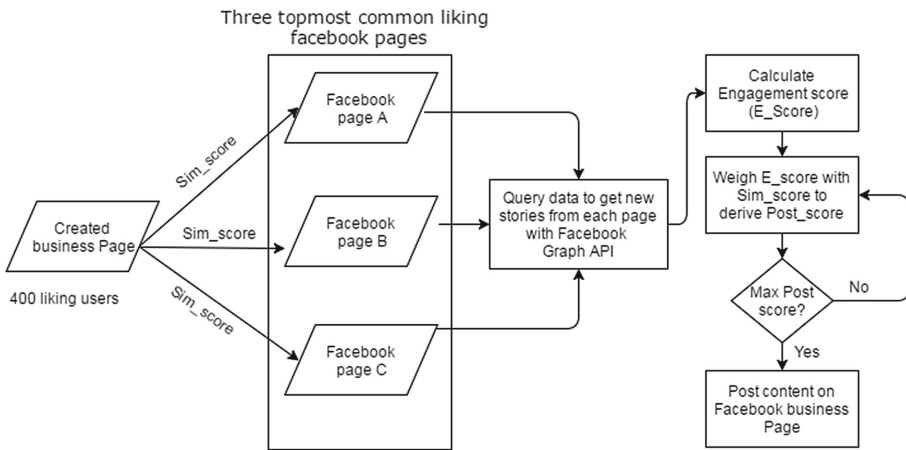


Fig. 4. Steps to derive values of similarity score, engagement score, and post score

4.7 Multiple Linear Regression

Multiple Linear Regression as Eq. 3 is used to predict \hat{Y} on the basis of p predictors (X_1, X_1, \dots, X_p) .

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p \tag{3}$$

The initial set of 12 predictors consisting of:

E_{score} , Sim_{score} , $Post_{score}$, *Reactions*, *Comments*, *Shares*, *Likes*, *Love*, *Wows*, *Hahas*, *Sads*, *Angrys* was used as independent variables to predict organic reach using multiple linear regression. Throughout several iterations of adjustment, we obtain the best regression model with three independent variables: (1) Similarity scores, (2) Shares, and (3) Wows, used as the factors to predict the value of dependent variable, organic reach, in this work.

5 Results

The highest twenty E-score engaging contents were posted on the created business Page in each period of content page sources. It is observed that some contents can penetrate the reach baseline. This may assume that the posts could be considered as affinity contents. The multiple correlation between reach and the three predictors = 0.9246 and $R^2 = 0.8549$ or 85.49%. However, we are more interested in Adjusted $R^2 = 0.779$, that is 77.9% of the variation in organic reach can be predicted on the basis of the three predictors. The value of standard error is 130.874.

The hypothesis was established as below:

H_0 : If there is no relationship between organic reach and selected factors.

H_1 : If there is relationship between organic reach and selected factors.

According to Table 3, where $\alpha = 0.05$; then $p - value = 4.05E - 07 < 0.05$,

Table 3. ANOVA Summary

Source	df	SS	MS	F	Significance F
Regression	3	1715860.74	571953.58	33.392	4.05E-07
Residual	17	291176.25	17128.01		
Total	20	2007037			

H_0 is rejected because *significance F* is less than the predetermined value 0.05.

We found that all the p-values of each independent variable were significant: SimScore = 0.000941, Shares = 0.00001713, Wows = 0.0001152. The regression coefficients can be used for weighing three predictors, that is, let b_0 represent the intercept = 0, $b_1 = 387.771$, $b_2 = 0.3759$, $b_3 = -7.818$.

Equation 4 shows the model derived for supporting the prediction of organic reach on Facebook with the use of three factors: Similarity score, Shares, and Wows.

$$\hat{Y} = 387.771 \text{ SimScore} + 0.3759 \text{ Shares} - 7.818 \text{ Wows} \tag{4}$$

6 Conclusion and Future Work

The News Feed ranking system offers people a better, more engaging experience on Facebook. However, the policy has affected social commerce since it causes Pages' organic reach to decrease. This paper presents an approach to supporting the decision for post plan. Rather than spending on more advertising, small businesses and startup could selectively post contents on their business Pages to increase Facebook organic reach, so that businesses can succeed on Facebook. We present a method to predict the engaging content for increasing organic reach on Facebook. A set of 12 factors was investigated whether they are potential predictors for engaging content. With multiple linear regression, the preliminary results on the clothing category showed that there is relationship between Facebook organic reach and the selected factors, which are Similarity score, Share, and Wow. The findings support the assumption that posting the viral affinity content could raise organic reach on Facebook. This would alleviate pain on

social commerce due to the policy of Facebook News Feed. However, further investigation of the relationship type, the potential predictors associated with individual coefficient need be carried out so that the output model could be extrapolated to other similar domain business Pages.

References

1. Facebook for Business. <https://www.facebook.com/business/news/Organic-Reach-on-Facebook>. Accessed 11 Sep 2016
2. An Update to Facebook News Feed: What it Means for Business Pages Facebook for Business. <https://www.facebook.com/business/news/update-to-facebook-news-feed>. Accessed 01 Dec 2016
3. Facebook Help Center. <https://www.facebook.com/help/282489752085908>. Accessed 11 Sep 2016
4. List of Facebook Fan Page Categories. <http://www.birdsonganalytics.com/list-of-facebook-fan-page-categories/>. Accessed 26 Nov 2016
5. Create a Page. <https://www.facebook.com/pages/create/>. Accessed 27 Nov 2016
6. Overview - Graph API - Documentation - Facebook for Developers. Facebook Developers. <https://developers.facebook.com/docs/graph-api/overview>. Accessed 12 Sep 2016
7. Karen Spärck Jones – Wikipedia. https://en.wikipedia.org/wiki/Karen_Sp%C3%A4rck_Jones. Accessed 20 Nov 2016
8. Zhang, Y., Pennacchiotti, M.: Predicting purchase behaviors from social media, pp. 1521–1532 (2013)
9. Yamane, T.: Statistics: An Introductory Analysis. Harper and Row, New York (1967)
10. Introduction — Selenium Documentation. http://www.seleniumhq.org/docs/01_introducing_selenium.jsp#introducing-selenium. Accessed 24 Oct 2016
11. Beginners Guide to Learn About Content Based Recommender Engines, Analytics Vidhya. 11 Aug 2015
12. Engagement - Unmetric. <https://unmetric.com/engagement/>. Accessed 07 Nov 2016
13. Bernstein, M.S., Bakshy, E., Burke, M., Karrer, B.: Quantifying the invisible audience in social networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 21–30 (2013)

Learning Performance Evaluation in eLearning with the Web-Based Assessment

Cheng-Ying Yang¹, Tsai-Yuan Chung², Min-Shiang Hwang^{3,5(✉)},
Cheng-Yi Li⁴, and Jenq-Foung JF Yao⁶

¹ Department of Computer Science, University of Taipei, Taipei 100, Taiwan

² Center for Teacher Education and Career Development, University of Taipei,
Taipei 100, Taiwan

³ Department of Computer Science and Information Engineering, Asia University,
Taichung, Taiwan
mshwang@asia.edu.tw

⁴ Department of Management Information Systems, National Chung Hsing University,
Taichung, Taiwan

⁵ Department of Medical Research, China Medical University Hospital,
China Medical University,
Taichung, Taiwan

⁶ Department of Computer Science, Georgia College and State University, Milledgeville,
GA 31061, USA

Abstract. With the prevalence of Internet, eLearning provides a platform for education that enables students to take classes online. While eLearning provides a flexible learning environment, it also has drawbacks. This research investigated the potential benefit of the proposed method in an informal formative web-based assessment. The data were collected from college students in three separated groups. The statistical analyses showed mixed results. Some possible reasons were discussed along with other methods that could be further explored in the future.

1 Introduction

Nowadays, many people have experiences with Internet. In Taiwan, the estimated Internet users are 18.83 million out of the total population, 23 million. There are more than 83.7% of the people who can access broadband network. Most users access Internet at home with a high frequency of 91.2% [1]. These bring about new alternatives in pursuing higher education. ELearning provides opportunities that utilize Internet technologies to access educational curriculum outside of a traditional classroom.

Among Internet applications, eLearning offers a flexible learning environment, an easy access to online learning resources, and a feasible solution to self-learning [2]. Since eLearning provides a flexible learning environment, one could learn without the restriction of time and space [3, 4]. However, there are some drawbacks in eLearning, such as, the basic computer manipulation skills are necessary, and it may be inconvenient to the individuals who seldom use a computer. Hence, it minimizes the learning results and learners'

satisfaction [5]. In addition, eLearning environment usually lacks interaction [6, 7]. The learners may misunderstand the learning materials. The learners in the eLearning platform may perform worse than the students in the traditional classes. As to the drop-out rate, in the traditional course is around 10%–20%. And, it rises to 25%–40% in eLearning [8].

In the traditional face-to-face learning environment, students have to attend the classes and arrive on time. They have to schedule study time regulated by the instructor. In contrast, in the eLearning environment, students could study whenever they have free time without the regulations from their instructor. Although eLearning has the advantage of flexibility, it lacks face-to-face regulations and enforcement. If students are short on motivation, they could become procrastinating or even stop using the eLearning system altogether [9].

Comparing with the traditional classrooms, eLearning may have no real-time interactions, learners could study passively, unable to ask the instructor questions and to discuss study materials with classmates. In the meantime, they could have difficulty navigate the eLearning system, resulting in low satisfaction, thus discourage students in the eLearning environment. Positive correlations were found between user satisfaction and system interactivity investigated by a previous research [10]. In addition, the formative assessment is often used as a separated part embedded in the eLearning system. Although the feedbacks from students who have taken the formative assessment are mostly positive [11], this assessment needs to be taken by students voluntarily in the eLearning environment. If the students do not self-regulate to use it, the effect of the formative assessment may be lower than that it should be. Thus, it brings about embedding an assessment in the eLearning system. This setting forces students to study the learning material and to take the assessment in the same eLearning session. It could be helpful in increasing the successful rate of the eLearning system. Hence, this paper proposes a scheme to prompt students' interest in using eLearning and to increase students' usage of the system. The scheme contains a formative assessment that embedded in a lecture video. It is designed to use the informal formative assessment to prompt the student's interest in the eLearning.

The following section introduces the formative assessment in the eLearning, and the evidence that the assessment supports students' learning by keeping their interest in the eLearning. The third section describes this research's research method. The fourth section gives the discussions about research hypothesis and the statistical analysis of the experiment results. Finally, the conclusion and research challenges are addressed in the last section.

2 The Formative Assessment

There are two kinds of assessment in general: the summative assessment and the formative assessment. The summative assessment is used for the purpose of grading. It is used to judge the degree of understanding. An exam is a good example of summative assessment. On the other hand, the formative assessment is used to help students' learning. It provides the instructor insight in terms of what students have understood and confused. Similar to the summative

assessment, it could be an exam and its result could be considered part of grading. However, the grade for formative assessment is not as important as that for summative assessment. The initiation–response–evaluation or initiation–response–feedback sequence is the best to describe formative assessment [12]. The instructor initiates the process and asks a question. S/he retrieves responses from the students and provides a feedback to the students. This way, instructors can gather all information from the students and know what the students don't understand. Then, they adjust teaching style if it is necessary. With a different teaching style, students could get a deeper understanding from the feedback. Hence, the formative assessment is helpful to improve learning.

There are two different types of formative assessment [13]. One is the formal formative assessment. Instructors pre-schedule the assessment and they can manage the content and get feedbacks from the students. The other one is the informal formative assessment. As its name suggests, the assessment can be held during the interactions between teachers and students. Both of these two formative assessments have a similar process including initiation–response–evaluation and collecting student's learning progress. Because the formal formative assessment is planned ahead, the instructors can predict what types of information will be collected and how to evaluate such information. Informal formative assessment is only implemented at times during the course of the term. Both formal and informal formative assessments are helpful to student learning.

Since the purpose of the formative assessment is used to help students to learn, it could be significantly related to the positive learning performance. Most researchers use the formal formative assessment for the experiments [14]. They built the assessment separated from the teaching material. This paper proposes an informal formative assessment method in the eLearning environment, which is similar to the traditional classroom. The purpose is to explore whether the informal formative assessment still has the same effectiveness in the eLearning environment.

3 Research Questions and Research Method

As mentioned in section two, the formative assessment can be used to improve students' learning, and the feedbacks from the students are also positive. In that thought, the proposed eLearning system has an informal formative assessment embedded. This research is designed to investigate whether the proposed embedded assessment could prompt students' interest and increase their eLearning usage.

There are four questions that are related to our experiment:

Question 1: Does the formative assessment prompt students' eLearning usage?

Question 2: Does the informal formative assessment improve students' learning performance?

Question 3: Is there a significant relationship between the usage frequency and the learning performance?

Question 4: Does the roll call increase students' eLearning usage?

Due to the limitation of resources, this research has held an experiment in an undergraduate course, Introduction to Statistic. This course is one-year long spanning in two

semesters. There were 55 students registered for this course in the first semester. In the subsequent semester, there were 68 students including the 51 students from the previous semester. Only the 51 original students taking the class from the previous semester were chosen for the experiment for the obvious reason. According to the performance on the mid-term exam before the experiment, students are randomly clustered to 3 groups named with Question-Embedded eLearning System group (QEES), Roll-call-Embedded eLearning System group (REES) and Control group. Each group consists of 17 students. The composition of students is shown in Table 1.

Table 1. Summary of group composition

		QEES	REES	Control
Gender	Male	6	6	10
	Female	11	11	7
Age	20	9	15	13
	21	7	2	4
	25	1	0	0

*Notice: Every participant is a sophomore majored in MIS.

The eLearning system used in this research is implemented with PHP and MySQL. There are two parts in the system. One contains the information related to class and the other contains the teaching material including PowerPoint slides and the teaching videos. Depending on the experimental groups, there are 3 different appearances to play the videos as follows.

1. QEES group: The students will have a screen with two areas. One area is used for playing a teaching video and the other area is used for showing an assessment question. The video playing area attaches with a functional control panel bar including Pause, Play, Forward or Backward and Seek. Beside to the video playing area is the question showing area. When the instruction video reaches a specific time, the video player will pause and display an assessment question in the question showing area.
2. REES group: Without the meaningful assessment questions, the system contains an area to play a teaching video, similar to the one for QEES group. There is no area to show the question. Instead, there is a text entry to record that the students are using the system.
3. Control group: This system only contains an area for playing a teaching video. Without any inspection, it recognizes the students are studying with the eLearning after logging in the system. Comparing to the system for REES group, there is no inspection to assure if the students are using the eLearning. However, the mechanism is still existing for checking the usage frequency.

4 Hypothesis and Experimental Results

In order to answer the research questions, we have derived the hypotheses and variables from the proposed questions.

Hypothesis 1: students in the QEES group will use the eLearning system more frequently than those in the control group.

Hypothesis 2: students in the REES group will use the eLearning systems more frequently than those in the control group do.

Hypothesis 3: students in the QEES group will have a better learning performance than those in the control group.

Hypothesis 4: students' learning performance is positively correlated with the eLearning usage frequency.

After the experiment, we analyzed the collected data and obtained the following results. Because Hypothesis 1 and Hypothesis 2 discuss the eLearning usage frequency, their results are presented together. We investigated the usage frequency for each chapter and determined any difference. The results of the ANOVA analysis with respect to the Average Usage Frequency are shown in Table 2. The significance value is 0.004 which indicates a significant difference among the groups. Then, LSD test is used to compare these groups and the result is shown in Table 3. The control group is significantly different from the other two groups because of the p-values. Furthermore, the mean differences between Control group and the other two groups are positive. It shows the Control group has the highest usage frequency. Hence, the students in Control group use the eLearning system more frequently than those in the other groups do. Since the usage frequency in the eLearning system is related to the students' intention of using a specific technology, this research adopts Technology Acceptance Mode (TAM) [15] for the experiment.

Table 2. ANOVA analysis to average usage frequency

	Sum of squares	df	Mean square	F	Sig.
Between groups	222.275	2	111.137	6.352	.004
Within groups	839.765	48	17.495		
Total	1062.039	50			

Table 3. Pairwise comparison of average usage frequency

Mean difference						
(I) Group	(J) Group	(I-J)	Std. error	Sig.	95% Confidence interval	
					Lower bound	Upper bound
QEES	QEES					
	REES	1.88235	1.43466	.196	-1.0022	4.7669
	Control	-3.17647(*)	1.43466	.032	-6.0610	-.2919
REES	QEES	-1.88235	1.43466	.196	-4.7669	1.0022
	REES					
	Control	-5.05882(*)	1.43466	.001	-7.9434	-2.1742
Control	QEES	3.17647(*)	1.43466	.032	.2919	6.0610
	REES	5.05882(*)	1.43466	.001	2.1742	7.9434
	Control					

*Notice: The mean difference is significant at the .05 level.

The results of ANOVA analysis with respect to the mean grade are given in Table 4. According to the results, there are no significant differences among these groups because of the p-values. The students in the different group have the identical performance without the influence of the assessments. From the previous related works, the formative assessment is able to increase students’ learning performance. However, the analysis result is not able to demonstrate such effect in this research. The reason could be that informal formative assessment does not have a similar effect to that the formal formative assessment does in the eLearning environment.

Table 4. ANOVA analysis of mean grade

	Sum of squares	df	Mean square	F	Sig.
Between groups	16.193	2	8.096	.007	.993
Within groups	55835.216	48	1163.234		
Total	55851.408	50			

With the regression analysis, the usage frequency is the independent variable and the score of the exam is the other dependent variable, the results are given in Table 5. According to the standardize coefficient, Beta, it is significantly positive. It shows the learning performance is positively related to the usage frequency. Comparing with the pure eLearning courses, there still exists some external influential factors. For example, the registered students in this course are familiar with each other. They could get the knowledge of the course by using the eLearning system, or by studying and discussing the material with other students. It is difficult to determine if the eLearning prompt students’ learning.

Table 5. Regression analysis

	Beta	t	Sig.
Frequency of use	.309	2.276	.027

*Notice: Dependent variable is Grade.

5 Conclusion

This research attempts to apply an informal formative assessment in the eLearning environment. Due to the research limitation of time and available research subjects, it adopts a hybrid approach from a traditional course in Statistics. TAM model is employed to obtain the results. The students in Control group can use the system easily without any interruption. It is reasonable to explain they have the highest usage frequency in the eLearning.

Referring to the hypothesis of this work, the proposed method should have added the interactivity to the eLearning system to increase learners’ interest and willingness to use the system. However, the results have shown otherwise. Other methods, such as applying Web 2.0 technologies, may increase the desired interactivity. Web 2.0 is a Web service in which the users could also be the contributors. Such service often requires the

high involvement of users and possesses high interactivity. Hence, Web 2.0 is a good candidate with the potential to solve the concerned problem.

Acknowledgements. This work was supported by Ministry of Science and Technology Grant (MOST 105-2410-H-468-009).

References

1. TWNIC, A Survey on Broadband Internet Usage in Taiwan (2015). <http://www.twnic.net.tw/download/200307/20150901d.pdf>
2. Rapp, C., Gülbahar, Y., Adnan, M.: e-Tutor: a multilingual open educational resource for faculty development to teach online. In: *The International Review of Research in Open and Distributed Learning*, vol. 17, no. 5 (2016)
3. Brand-Gruwel, S.: Learning ability development in flexible learning environments. In: Michael Spector, J., David Merrill, M., Elen, J., Bishop, M.J. (eds.) *Handbook of Research on Educational Communications and Technology*, pp. 363–372. Springer, New York (2014)
4. Gooley, A., Lockwood, F.: *Innovation in Open and Distance Learning: Successful Development of Online and Web-based Learning*. Routledge (2012)
5. Bouhnik, D., Marcus, T.: Interaction in distance learning courses. *J. Am. Soc. Inform. Sci. Technol.* **57**(3), 299–305 (2006)
6. Abrami, P.C., Bernard, R.M., Bures, E.M., Borokhovski, E., Tamim, R.M.: Interaction in distance education and online learning: using evidence and theory to improve practice. In: Moller, L., Huett, J.B. (eds.) *The Next Generation of Distance Education*, pp. 49–69. Springer, USA (2012)
7. Rennie, F., Morrison, T.: *E-learning and Social Networking Handbook: Resources for Higher Education*. Routledge (2013)
8. Xenos, M.: Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Comput. Educ.* **43**(4), 345–359 (2004)
9. Tuckman, B.W.: The effect of motivational scaffolding on procrastinators distance learning outcomes. *Comput. Educ.* **49**, 414–422 (2007)
10. Fredericksen, E., Picket, A., Shea, P., Pelz, W., Swan, K.: Student satisfaction and perceived learning with on-line courses: principles and examples from the SUNY learning network. *J. Asynchronous Learn. Netw.* **4**(2), 7–41 (2000)
11. Burrow, M., Evdorides, H., Hallam, B., Freer-Hewish, R.: Developing formative assessments for postgraduate students in engineering. *Eur. J. Eng. Educ.* **30**, 255–263 (2005)
12. Fisher, D., Frey, N.: *Checking for Understanding: Formative Assessment Techniques for your Classroom*. Association for Supervision and Curriculum Development: Alexandria, VA (2007)
13. Ruiz-Primo, M.A., Furtak, E.M.: Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *J. Res. Sci. Teach.* **44**(1), 57–84 (2007)
14. Ricketts, C., Wilks, S., Freeman, A.: How can CAA support the acquisition of medical knowledge in an integrated medical curriculum? In: *Proceedings of the 8th International Computer Assisted Assessment Conference*, Loughborough (2004)
15. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **13**(3), 319–340 (1989)

Improving Teaching and Learning in Southeast Asian Secondary Schools with the Use of Culturally Motivated Web and Mobile Technology

Sithira Vadivel^(✉), Insu Song, and Abhishek Singh Bhati

School of Business/IT, James Cook University Australia,
Singapore Campus, Singapore, Singapore
{sithira.vadivel, insu.song, abhishek.bhati}@jcu.edu.au

Abstract. Improving and stimulating teaching and learning are an interesting topic among educational researchers. As technology advances and with mobile technology and the Internet being used widely, it has become a vital tool for knowledge gathering and information sharing. It can foster new directives for teachers and stimulate the minds of learners, improving learning outcomes. However, the process of this triangulation of interaction has been overlooked in the Southeastern Asian region and requires an in-depth study into its culturally diverse background to identify its core problems and benefits. We propose Student Motivated Integrated Learning & Education with culture (SMILE c) model in order to integrate education with web and mobile technology with an emphasis on Asian learning culture to promote active learning reduce overall costs and improves student learning outcome. We illustrate how this model can be implemented in Southeast Asian schools to improve teaching to suit students' learning style during lessons through an alert system and motivates student to participate in discussions which can be used by the institution to identify student's skill set early in the learning process.

Keywords: Mobile education · Education with social media · Technology and education · Biometric and education · Technological and pedagogical · Smart learning · ICT in southeast Asian schools · Culture and pedagogical approaches

1 Introduction

For more than a decade, technology has been a tool for success in every industry. The diversity of technology advancement has projected a new era of science and technology development in which technology developers are inspired by innovating and improvising technology features. Integration of the Internet, social media, web 2.0 and mobile devices are the current focus for industry players in this modern digital age, thus harnessing the power of social media and web technologies to advance their knowledge and businesses to the next level [1].

Proper technology integration guides students toward a better understanding of all concepts taught in class. It boosts student capacity, motivates students, increases student performance, amplifies and guides the cognitive process of learners [2].

However, technology integration process in the education industry is challenging due to the constantly changing nature of technology. Ravenscroft argues that “we cannot truly transform educational practice for the better through using new technologies unless we examine the roles the computer can play in truly stimulating, supporting and favoring innovative learning interactions that are linked to conceptual development and improvements in understanding” [3].

Measuring students’ progress and the learning outcome requires a systematic and gradual update for the educators in identifying issues with delivery in a progressive manner. Social media can facilitate the creation of Personal Learning Environment that help learners to share results of learning achievement and participate in collective knowledge generation [4].

The underpinning factor for student excellence is motivation. Motivated students perform better in class and ultimately improve learning outcomes. Adopting technology in education and streamlining knowledge attainment are inimical for schools. Modifying the education curriculum with the essence of technology requires intense study, which necessitates the optimization of features for an optimal learning outcome as well as identifying cultural resistance.

Our proposed method includes a background survey on literature on the applications of ICT in education. The result identifies the efficiency and deficiency in technology application and approaches in the Southeast Asian education industry. The significance of the result proves the differences between Western and Asian culture in teaching and learning and Asian teachers who do not understand the potential of technology. These deficiencies provide the foundation on a proposal for an integrated learning model for teachers who can identify students learning style during lesson and motivation on student participation due to cultural barriers.

The rest of the paper is organized as follows. Section 2 includes literatures which are categorized in ICT applications; Learning style and culture and Approaches. Section 3 provides discussion on the analysis and findings and the result is illustrated in a diagram. Section 4 identifies the proposed model and implementation approach and Sect. 5 concludes the paper.

2 Background and Survey on ICT in Education

In order to address the problems faced by teachers in using the right technology and student motivation during lessons with cultural issues that we stated, we have conducted a systematic survey of eighty (80) recent research articles. We briefly summarized the main themes and trends in this section, and the meta-analysis results of the survey are discussed in Sect. 3.

2.1 ICT Applications

Wastiau’s study attempts to determine the optimal usage of mobile devices under certain technical and organizational conditions. The findings show that teachers who are confident in their digital skills and positive about ICT’s impact on learning organize more

ICT-based activities with their students and confident and supportive teachers are needed to use the ICT infrastructure effectively to understand its potential [5].

Clark's study was designed to explore learners' perceptions and their experience of technology-mediated activity in school. A mapping study was used to explore these technologies and their use in greater detail with relation to technologies, practices, and context. There was evidence in Clark's study that few learners use these technologies with a high level of sophistication and institutions and teachers do not fully understand their potential benefits [6].

Spire's study was focused on students' perspectives of school and the use of technologies during school activities. The results show that students want the school to be aesthetically a pleasing environment that inspires and motivates them to learn and achieve with creative and ubiquitous use of technology. They find that learning is more fun when they get to use technology [7].

2.2 Learning Style and Culture

Tweed assessed whether Western-influenced students approach to learning more by questioning, evaluating and generating ideas compared to Chinese-influenced students. The findings from Tweed's study suggest that Western and Chinese-Canadian students have a more Socratic approach to learning compared to Asian-born Chinese students. The Western students were more likely question their instructors publicly. In contrast, Chinese-influenced students were more passive learners [8].

Liu's study was to find "a negative attitude toward participation" among students in Hong Kong Polytechnic University. The survey results show that 43% of students feel uncomfortable speaking in English as they lack practice. Students' had inadequate opportunities to practice spoken English and they adopted passive roles in the classroom [9].

Hofstede's study identified four dimensions: power distance, uncertainty avoidance, individualism versus collectivism and masculinity versus femininity. Hofstede's study on dimension shows that the largest universal shift is individualism and showed divergence among the countries rather than convergence [10].

2.3 Approaches

Cain used audience response systems (ARS) as a tool to aid faculty members in engaging and interacting with students. This tool allows instructors to pose questions to students and receive immediate feedback. Students' immediate feedback via ARS assist instructor to allocate more time on topics of which students lack an understanding. The results from Cain's study show that almost every student ($n = 110$) responded that ARS usage helped them maintain attention, and 98% ($n = 109$) felt that discussions stemming from the ARS were beneficial [11].

Joyce used remote-sensing computer-aided learning (RSCAL) to facilitate students' active engagement with foundational knowledge & skills, which was responsive to newer pedagogical perspectives and emerging learner needs. RSCAL incorporates video, animation, narrations that align with lectures, interactive play and quizzes that

appeal to different styles of learners, such as auditory- and visual-style learners. The system facilitated students’ learning and engagement with materials extremely well [12].

3 Analysis and Finding

We did the survey in the previous section in order to find the efficiency and deficiency of technology (mobile technology, social media, cloud services & E-learning application) use among students, teachers, schools and government. Figure 1 summarizes the result in a wheel. Technology usage and acceptance level in secondary education were categorized using the four main factors: students, teachers, schools and government. The outer layer specifies the deficiency in each section, and the second outer layer specifies the efficiency of using technology in the classroom. The cluttered area of efficiency in students’ use of technology indicates the promising features students benefit from in classroom learning. It shows the positive change and influence of using technology in the learning domain. However, the cluttered area in the teachers and schools sections is a deficiency and indicates the lack of the use of technology in the teaching domain and practices and policies supported in the schools that will fail to facilitate the positive integration of technology into the teaching and learning domain.

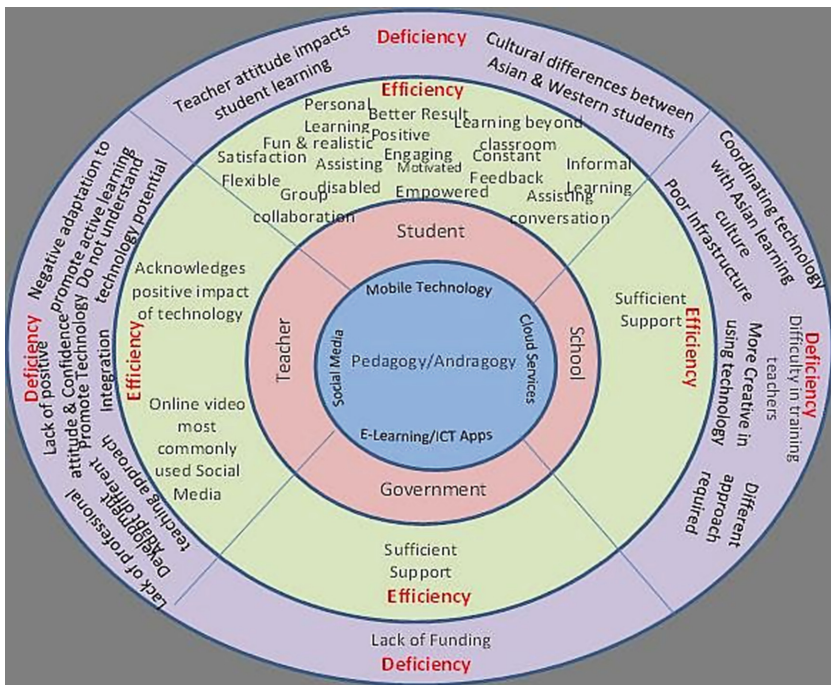


Fig. 1. Result summary

The presence of insufficient technology support in the teachers and schools sphere confirms students' inability to use high standards of technological tools and services and hence affects successful learning outcomes to achieve better results. The crux of the issue lies in the staff and organization's inability to understand the perfect match between teaching and technology, particularly by exploiting the anonymity that such technology entities could provide in the learning realm. One vital deficiency found in the schools domain, "D1-Coordinating technology with Asian learning culture," can be closely tied to the deficiency found in the teachers domain, "D2-Negative adaptation to promote active learning," "D3-Do not understand technology potential," and "D4-Adapt different teaching approach;" hence; these factors affects students' learning processes, as there are "D5-cultural learning differences presence between Asian and Western students" in the students domain that require profound attention and transformation of the teaching and learning approaches.

The current teenage generation (Gen Y) is more familiar with new technologies, and hence, using technology in the classroom seems to be a crucial factor in achieving effective results in teaching and learning. The identified barriers can be manifested into a positive approach if the right perspective is executed.

The five main deficiencies, D1–D5 were chosen as the base factors in developing the new model. These barriers were considered influencing factors for an effective learning outcome for the Asian region, and it is prudent to reflect upon these factors in future research. Institutions and teachers play a dynamic role in their manifestation, and a higher degree of study and analyses is required to provide a concrete foundation in the best interest of all parties and to move to the next phase of secondary education in the technological era. Keeping up with advancements is pivotal to improving performance and learning. The inseparability of modern technology and modern teenagers can be used to our advantage to motivate their educational journey; the education industry should consider the SMILE c: model an acronym for "Student Motivated Integrated Learning & Education with culture" to educate and motivate students using modern technology with appropriate cultural values for an effective teaching and learning approach: "giving them the tools to achieve what they want to achieve in an effective way".

The Table 1 below summarizes the factors considered for the model construction. The table shows the five deficiencies (D1–D5) identified as the main factors for deficiency and the second column shows the method to overcome the deficiency (OD1–OD5) and the third column highlights the features (F1–F5) that will be considered in the proposed model which tackles the deficiency.

4 Proposed Model and Approach

Figure 2 illustrates SMiLE C: model which is designed to motivate students for active learning during the lesson and for an effective learning outcome. Teachers can post questions during lessons from the smart library repository (F3), which consists of a pool of discussion questions and quizzes to promote interactive and active learning spontaneously (F2). Students post their answers using smartphones; this process attempts to

Table 1. Deficiency to features matrix

Deficiency	Overcome deficiency	Features
D1-Coordinating technology with Asian learning culture	OD1-Integrating Asian learning behavior	F1-features to identify weak or passive learners (Asian students)
D2-Negative adaptation to promote active learning	OD2-Promote active learning during lesson	F2-Post discussion questions to motivate active learning
D3-Do does not understand technology potential	OD3-Provide a smart learning model	F3-Easy to use model for teachers with readily available repository of questions/quiz
D4-Adapt different teaching approach	OD4-Changing teaching methods quick to adapt weak students	F4-Alert notification on the lack of students participation & teachers can change mode of teaching instantly
D5-Cultural learning differences presence between Asian and Western students	OD5-New model required for teaching Asian students	F5-Feature to identify and categorize students skills (passive & active) early for future planning

resolve issues for non-participative students (F1) who are not willing to voice their answers or opinions verbally during lesson. This process also draws students' attention and maintains student's attention span during lesson for an effective learning outcome. In this process, the teacher becomes more of a facilitator or motivator during the lesson. Students' answers are added to the smart repository, which can categorize students' special qualities and skills (F5) as artistic, creative, innovative and intellectual early in the learning process. Teachers and the school can use these reports for future planning and activities. The poll system detects the number of responses from students in real time, and the teacher receives an alert (F1) if the number of responses falls below the threshold level (number of students) with a graphical representation. The teacher can adjust the mode of teaching based on the number of responses received. If the response is low, teachers can switch to more interactive game-based learning or a short video on the topic to accommodate students' negative learning curves (F4). This system also delivers paperless technology to the education industry and hence improves the environment and helps cut back on costs.

A few secondary schools in Southeast Asia will be selected to test the SMiLE c: model. Teachers and school administrators will be briefed on the model, and an empirical test will be conducted during lessons. Teachers and students will be surveyed before and after the usage of the SMiLE c: model to discover the model's ease of use and its efficiency in producing an effective learning outcome. This model can be implemented using web and mobile technologies.

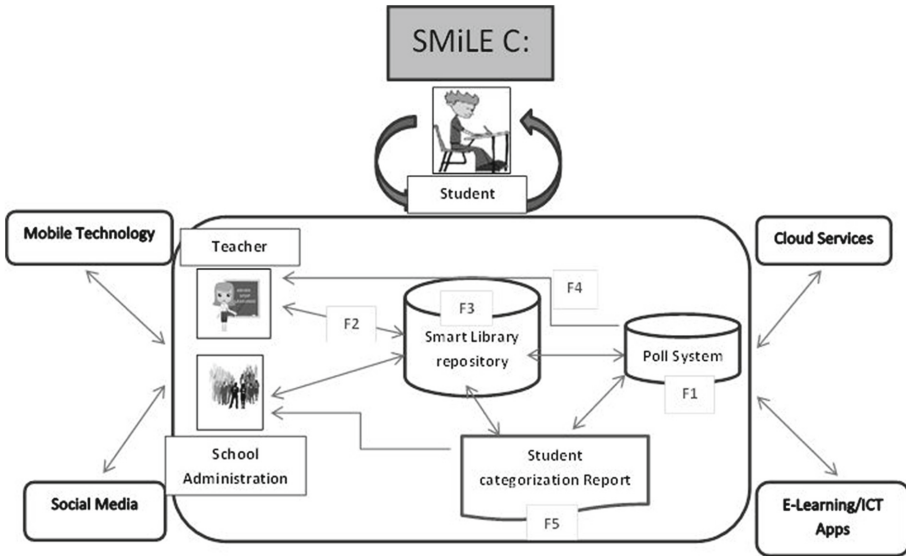


Fig. 2. SMiLE C:

5 Conclusion

Traditional classroom teaching requires a boost in energy and a vibrant atmosphere for active learning and to accommodate different learning cultures. As technology becomes a vital tool for knowledge gathering and information sharing, it can foster new directives for teachers and stimulate the minds of learners during the lesson, improving the learning outcome. The analysis of SMiLE c model shows that an integrated system is required for teachers to simplify their teaching and learning domain which can identify the lack of student participation and to accommodate student learning style during lesson for an effective learning outcome. This can be an integral part of student motivation and effective learning in classroom education. The features of SMiLE c: are efficient for immediate feedback in real time for teachers to understand students’ learning progress; the model identifies types of students early and categorizes students’ expertise level for future planning and activities. It promotes active learning, and most importantly, it reduces overall costs and improves student learning outcomes.

References

1. Mtega, W.P., Benard, R., Dettu, M.: The prospects of Web 2.0 technologies in teaching and learning in higher learning institutes: the case study of the Sokoine University of Agriculture in Tanzania. *Knowl. Manag. E-Learn. Int. J. (KM&EL)* **5**(4), 404–418 (2014)
2. Darling-Hammond, L.: Teacher quality and student achievement. *Educ. Policy Anal. Arch.* **8**, 1 (2000)

3. Ravenscroft, A.: Designing e-learning interactions in the 21st century: revisiting and rethinking the role of theory. *Eur. J. Educ.* **36**(2), 133–156 (2001)
4. Dabbagh, N., Kitsantas, A.: Personal learning environments, social media, and self-regulated learning: a natural formula for connecting formal and informal learning. *Internet High. Educ.* **15**(1), 3–8 (2012)
5. Wastiau, P., Blamire, R., Kearney, C., Quittre, V., Van de Gaer, E., Monseur, C.: The use of ICT in education: a survey of schools in Europe. *Eur. J. Educ.* **48**(1), 11–27 (2013)
6. Clark, W., Logan, K., Luckin, R., Mee, A., Oliver, M.: Beyond web 2.0: mapping the technology landscapes of young learners. *J. Comput. Assist. Learn.* **25**(1), 56–69 (2009)
7. Spires, H.A., Lee, J.K., Turner, K.A., Johnson, J.: Having our say: middle grade student perspectives on school, technologies, and academic engagement. *J. Res. Technol. Educ.* **40**(4), 497–515 (2008)
8. Tweed, R.G., Lehman, D.R.: Learning considered within a cultural context: confucian and Socratic approaches. *Am. Psychol.* **57**(2), 89 (2002)
9. Liu, N.-F., Littlewood, W.: Why do many students appear reluctant to participate in classroom learning discourse? *System* **25**(3), 371–384 (1997)
10. Hofstede, G.: National cultures in four dimensions: a research-based theory of cultural differences among nations. *Int. Stud. Manag. Organ.* **13**(1/2), 46–74 (1983)
11. Cain, J., Black, E.P., Rohr, J.: An audience response system strategy to improve student motivation, attention, and feedback. *Am. J. Pharm. Educ.* **73**(2), 21 (2009)
12. Joyce, K.E., Boitshwarelo, B., Phinn, S.R., Hill, G.J., Kelly, G.D.: Interactive online tools for enhancing student learning experiences in remote sensing. *J. Geogr. High. Educ.* **38**(3), 431–439 (2014)

Game-Based Learning to Teach Assertive Communication

ClickTalk for Enhancing Team Play

Bah Tee Eng ^(✉)

School of Business/IT, James Cook University Australia,
Singapore Campus, Singapore, Singapore
bahtee.eng@my.jcu.edu.au

Abstract. The rise of the computer as an “entertainment medium” has been achieved today only through computer games. But computer or video games have the potential and capability to function as “mediums of education” too. Can game-based learning provide learning experience and yet there is fun in changing behavior (assertive communication) for the individual? Game-based learning has been used to teach various skills to people with quite encouraging results. In this paper, a study was carried out to confirm the hypothesis that game-based learning can be a good platform to teach assertive communication delivering learning and fun because of its benefits and encouraging results from other research. A high-fidelity game-based learning prototype, ClickTalk was created for this purpose and it was evaluated with some interesting results.

Keywords: Game-based learning to teach assertive communication

1 Introduction

The first computer game was played on a PDP-11 computer (Bellis 2016) and over the years, the rise of the computer as an “entertainment medium” has been achieved only through computer games (Jayakanthan 2002). But computer or video games have the potential and capability to function as “mediums of education” too (Jayakanthan 2002).

Games-based learning (GBL), is also known as “Serious Games” (Corti 2006). Kevin (2006) mentioned that GBL has the potential to greatly improve training activities and initiatives in the organisation.

Game-based learning has the motivational virtues of video games and allows a simulated environment, learning by experience and make the experience compelling so that the learners can remember what and why something happened (Corti 2006).

Thus, game-based learning has been employed in various contexts such as teaching social skills to children (Thomas and DeRosier 2010), database analysis and design to IT graduates and post-graduates (Connolly et al. 2006), and historical knowledge to secondary schools students with quite encouraging results (Huizenga et al. 2009).

Incidentally, the global games market has now reached \$99.6 billion this year (NewZoo 2016) and is growing rapidly. Is game-based learning a good platform to teach assertive communication delivering learning and fun because of the benefits and encouraging results from other research mentioned above?

As such, a study was carried out to confirm the above hypothesis that game-based learning can be a good platform to teach assertive communication delivering learning and fun because of the above benefits and encouraging results from other research.

In order to carry out this study, a high-fidelity game-based learning prototype was created for this purpose and it was evaluated with some interesting results. (Note: a low-fidelity game-based learning prototype by the author in a different paper has proven the hypothesis that game-based learning is more effective than traditional classroom teaching or presentation).

What is novel or different about this paper's approach as compared to previous research is that users were recruited to develop game-based learning prototypes besides the one created by the author. Further research will develop on the one or two good prototypes. In other research papers, a development team designed only a single game prototype for evaluation and further research.

Section 2 of this report then looks at the background research on game-based learning and provides a summary of applications of game-based learning and their respective previous approaches.

Section 3 presents the details about ClickTalk, the high-fidelity game-based learning prototype that has been used to teach assertive communication in this study.

The methodology adopted for this paper is covered in Sect. 4. Section 5 looks at the results. Section 6 discusses the results and Sect. 7 presents the conclusions. The references for this paper are in Sect. 8.

2 Background

Here, we cover the background research related to game-based learning and also take a look at overall summary of previous game-based learning applications and the approaches used.

It has been mentioned that the trend is now going towards serious games - interesting games that inject learning for the individual (Squire 2003). Kurt (2003) mentioned that game-based learning is basically using computer technology to delight and "engage" players for the purpose of developing their new knowledge and skills.

It has been said that the younger generation of learners who have grown up in the midst of mobile phones, IPADs, graphic-rich movies, Xboxes and so forth cannot continue to be taught using traditional classroom teaching methods (Prensky 2001).

Kevin (2006) also mentioned that it is the motivational virtues of a game that "initially prompted training and development professionals" to look to games-based approaches. Games-based learning has more things to offer than just using fun to engage learners (Corti 2006).

In the past, "game-based learning environments were rather expensive" for many organisations (Pho and Dinscore 2015). However in recent years, health care organisations and medical schools have started to rely on games and simulations, and using such tools for practice is now the norm or even encouraged. In the US for instance, medical staff need to undergo virtual reality training for the placement of some stents and also many medical schools have come up with centres dedicated to simulation training.

Table in sub-Sect. 2.1 below shows some game-based learning employed over the years.

2.1 Summary of Applications and Previous Approaches used

Main category	Application areas	Title of approaches	Methods	Result 1: Significant 2: Mild 3: No improvement
Soft skill training	Collaboration, communication	Controlled study	Computer game	1 (Thomas and DeRosier 2010)
Hard skill training	Computer science	Experimental approach	Computer game	1 (Connolly et al. 2006)
	Science	Experimental approach	Mobile game	1 (Sung and Hwang 2013)
	History	Experimental approach	Mobile game	1 (Huizenga et al. 2009)

3 ClickTalk

ClickTalk is a high-fidelity game-based learning prototype created using HTML5, Javascript and the popular Phaser game engine for the purpose of teaching assertive communication to young adults and above. Its creation was inspired in fact after the focus group of users has created six game-based learning prototypes.

ClickTalk initially starts with a basic assertive quiz of five questions and if these questions are successfully answered, the learner can proceed to three stages.

Stage 1 – In this stage (Fig. 1), the learner is introduced to team communication. By pressing relevant keys on keyboard(one player use A,S,D,F keys and the other player, the arrow keys), two players can communicate assertively and help “rescue” koala bears to safety in a van from an impending earthquake. The notion here is that team communication is crucial as koala bears’ lives are at stake in such a dangerous situation. With a score above 20, the player can proceed to next stage of the game. A successful rescue of big koala bear fetches 40 points while that of a small koala bear fetches only 20 points.



Fig. 1. Stage 1

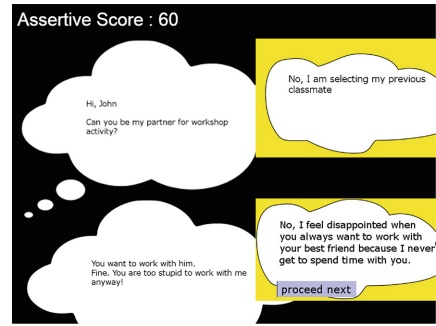


Fig. 2. Correct dialogue selected and a proceed next choice to go to last stage

Stage 2 – In this stage, the learner will learn the use of I-message in assertive conversation and its format. In Fig. 2 above, there is an initial request given (“Hi John, can you be my partner for workshop activity?”). The learner has two choices as shown by the two yellow highlighted dialogues. The correct choice selected will gain for the learner some points and with that, the learner can proceed to the last stage. If wrong choice is selected, then points will be deducted.

Stage 3 – this is the only fun part which is a shooting game to “shoot down” arguments using “fog” fired from a machine gun turret.

4 Methodology

Currently, there has been little published research that get participants to actually “design” a game-based learning prototype. Much research has been on many participants using an already-designed game-based learning prototype and how they evaluated the results of using the game.

As we know, nowadays people from all walks of life play games. Go around the city and you will notice many people play mobile games. People have ideas from the many games they played but their ideas are not being tapped. In general, people(users) around us may offer fresh perspectives or creative ideas on a game-based learning prototype to teach assertive communication.

To capitalise on such a great source of ideas, this paper therefore decided to employ a user-centred design approach to develop a game-based learning prototype to teach assertive communication.

With this in mind, six participants were recruited on the James Cook University Singapore campus to form a focus group. Informed consent was first obtained from the participants before their participation in the study commenced.

The six participants attended a workshop session conducted on the topic of basic assertive communication for half an hour. They were then briefed on what is needed in game-based learning - such as assessment, scoring, some fun element and learning. With that knowledge in hand and some basic requirements and guidelines, they went on to

design their own game-based learning prototypes. Their prototypes are mentioned in the Results section.

From here, we also created ClickTalk mentioned earlier in Sect. 3 above (a high-fidelity prototype using HTML5/Javascript and Phaser game engine) and the focus group then helped to evaluate this ClickTalk prototype. Their evaluation is discussed in the Results section. The entire data collected was hence analysed and evaluated and alternatives were considered in the design. The response format being used in the questionnaires was the 5-point Likert scale.

5 Results

A total of six students of James Cook University formed the focus group. The following diagrams showed their demographics (Figs. 3, 4 and 5):-

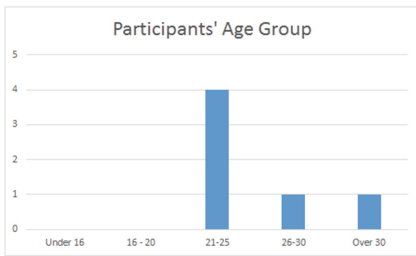


Fig. 3. Age group of participants

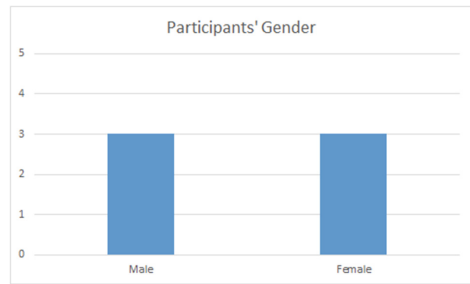


Fig. 4. Gender of participants

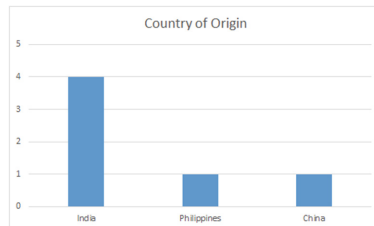


Fig. 5. Country of origin of participants

On evaluating the game-based learning HTML app prototype, ClickTalk, the participants views are summarised as follows:-

- All agree that ClickTalk makes the participants more effective, gives them a good overview of team communication and it is a good way to learn assertive communication.
- All also agree that ClickTalk is simple, user-friendly, pleasant, easy to learn and convenient to use.

- All agree that the app helps them to understand the different behavior types (passive, assertive and aggressive) and it will be popular in future classrooms and it is a good supplement to traditional classroom teaching.
- Majority feels ClickTalk is refreshing, fun to use, need to have and can be used successfully every time. Majority also finds it is productive to use ClickTalk and it gives them an idea of what to say during conflict and control over activities in a team.
- However, half of the participants prefers a neutral stand when it comes to whether there is any inconsistencies in using ClickTalk or whether they have seen anything like ClickTalk before.

The six user-designed prototypes are “Assertive Car Race”, “Multi-Scenarios”, “Assertive Trust”, “Shoot the Alien”, “Learn and Play” and “Just Coin It”.

6 Discussion (Results Compared to Previous Works)

The six prototypes created by the focus group of participants are rather promising and the ideas are quite fresh. They are also of different game genres. From the evaluation of ClickTalk prototype given by the six participants in the previous section, the hypothesis about whether game-based learning is a good platform to teach assertive communication with learning and fun has therefore been proven right - participants agreed there is fun as they learn assertive communication concepts at the same time and they recommend the game-based learning app to anyone who wants to know more about assertive communication.

What then does this paper offer as compared to previous research?

Firstly, this paper employs user-centred design and the prototypes created are the “solutions” of users to the problem of a game-based learning prototype that can teach and yet fun to play.

Secondly, such an approach is more efficient and effective as compared to someone creating only one prototype and that may not be the best prototype. In this case, we can easily select one or two best prototypes out of the seven prototypes created (six prototypes including ClickTalk, the one created by author).

Also, the adage “users know what they want” has never been more true than in this instance. Users will always create something they like to see.

Overall, the results obtained are very encouraging as the case with other previous research on game-based learning.

ClickTalk can be further improved with more features to cover a basic assertive communication curriculum employed in classroom or business setting and possibly with more clearer learning outcomes. In further research, a bigger group of participants will be recruited and ClickTalk game can also be enhanced with the use of more sophisticated game engine and more random questions to test learner. The current prototype is a showcase of game-based learning.

Further research will also involve selecting one or two prototypes from the seven prototypes and develop them into interesting games for teaching assertive communication.

7 Conclusions

This paper uses a high-fidelity game-based learning (HTML5/Javascript/Phaser) prototype to conduct its experiment.

From the evaluation of the game-based learning prototype, it is worth to note that all participants agree that ClickTalk is simple, easy to use and remember, user-friendly, easy to learn, convenient and pleasant to use, refreshing and fun – all the adjectives of good product usability.

It is also important to note that the majority of participants want more fun than learning or equal amounts of fun and learning in a game-based learning app if it is to be implemented.

The above study shows how user-centered design involving users can help provide quite refreshing insights and solutions.

References

- Bellis, M.: The History of computer and video games. In: *Timelines of Inventions and Technology* (2016)
- Connolly, T., Stansfield, M., McLellan, E.: Using an online games-based learning approach to teach database design concepts. *Electron. J. e-Learn.* **4**(1), 103–110 (2006)
- Corti, K.: Games-based learning; a serious business application. *Informe de PixelLearning* **34**(6), 1–20 (2006)
- Garris, R., Ahlers, R., Driskell, J.E.: Games, motivation, and learning: a research and practice model. *Simul. Gaming* **33**(4), 441–467 (2002)
- Huizenga, J., Admiraal, W., Akkerman, S., Dam, G.T.: Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *J. Comput. Assist. Learn.* **25**(4), 332–344 (2009)
- Jayakanthan, R.: Application of computer games in the field of education. *Electron. Libr.* **20**(2), 98–102 (2002)
- Malone, T.W.: Toward a theory of intrinsically motivating instruction. *Cogn. Sci.* **5**(4), 333–369 (1981)
- Moreno-Ger, P., Burgos, D., Martínez-Ortiz, I., Sierra, J.L., Fernández-Manjón, B.: Educational game design for online education. *Comput. Hum. Behav.* **24**(6), 2530–2540 (2008)
- Muller, M.J., Wildman, D.M., White, E.A.: “Equal opportunity” PD using PICTIVE. *Commun. ACM* **36**(6), 64 (1993)
- NewZoo: The Global Games Market Reaches \$99.6 Billion in 2016, Mobile Generating 37% (2016). 1
- Papastergiou, M.: Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Comput. Educ.* **52**(1), 1–12 (2009)
- Pho, A., Dinscore, A.: *Game-Based Learning* (2015)
- Prensky, M.: *Digital game-based learning* (2001)
- Squire, K.: Video games in education. *Int. J. Intell. Games Simul.* **2**(1), 49–62 (2003)
- Sung, H.-Y., Hwang, G.-J.: A collaborative game-based learning approach to improving students’ learning performance in science courses. *Comput. Educ.* **63**, 43–51 (2013)
- Teoh, C.: User-centred design (UCD) - 6 methods (2006). 1

- Thomas, J.M., DeRosier, M.E.: Toward effective game-based social skills tutoring for children: an evaluation of a social adventure game. In: Paper Presented at the Proceedings of the Fifth International Conference on the Foundations of Digital Games (2010)
- Yusoff, A., Crowder, R., Gilbert, L., Wills, G.: A conceptual framework for serious games. In: Paper Presented at the 2009 Ninth IEEE International Conference on Advanced Learning Technologies (2009)

Cloud Storage Federation as a Service Reference Architecture

Rene Ivan Heinsen, Cindy Pamela Lopez, Tri D.T. Nguyen,
and Eui-Nam Huh^(✉)

Department of Computer Engineering, Kyung Hee University, 1732,
Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Republic of Korea
{reneheinsen,cindylopez,tringuyendt,johnhuh}@khu.ac.kr

Abstract. Cloud storage is one of the leading technologies to address today's data storage demand. However, facing Big Data storage challenges relying on a single cloud storage provider is almost impossible. Cloud storage federation model provides the integration of multiple cloud storage providers into a single virtual storage pool, eliminating the dependency on a single provider and decreasing vendor lock-in problem. Moreover, federating multiple cloud storage providers improve data availability, storage scalability and data processing performance. In this paper, we propose a reference architecture for Cloud Storage Federation Service implementation. Moreover, a demo cloud storage federation service implementation is presented.

Keywords: Cloud storage federation · Reference architecture · Cloud storage service

1 Introduction

As a consequence of the constantly increasing data volumes and the heterogeneous data sources and formats, IT industry is facing enormous challenges regarding data storage and management. For instance, 2013 world's generated data was more than duplicated by 2015, going from four to nine zettabytes, and is expected to reach 40 zettabytes by 2020 [1]. Today's data growth and speed are being driven for the enormous amount of devices connected and transmitting data in real time through the Internet (IoT) [2]. One of the leading approaches addressing this reality is Cloud Storage. However, Big Data storage continues being a big challenge for any cloud storage provider.

Cloud storage federation brings to the table a suitable and scalable way to improve cloud storage technologies. It integrates multiple autonomous cloud storage providers into a common management service that takes care about how data is distributed, managed, and migrated through the participant providers. Furthermore, the federation model is useful for improving data availability, reliability, durability and storage scalability, as well as preventing vendor lock-in issue. General cloud federation model allows different cloud providers to work

in collaboration, meaning that some kind of interoperability is present between providers, more specific, one provider can ask for resources to other provider in order to fulfill its customer needs. Nevertheless, Vellante [3] defined federated storage as “the collection of autonomous storage resources governed by a common management system that provides rules about how data is stored, managed, and migrated throughout the storage network”. Therefore, the federation model we are proposing does not required interoperability between providers since it is managed by an external entity.

In this paper we aim to step toward into a Cloud Storage Federation as a Service, providing a reference architecture for future federation services implementation. Furthermore, an example cloud storage federation service is presented. The rest of this paper is organized as follows; Sect. 2 describes our proposed Cloud Storage Federation service reference architecture. Section 3 introduces a demo Cloud Storage Federation service. Section 4 discusses related work. Section 5 summarizes our contributions and recognizes future work.

2 Cloud Storage Federation as a Service Reference Architecture

This section describes our proposed reference architecture for implementing a Cloud Storage Federation Service. A Cloud Storage Federation Service allows clients applications to manage multiple cloud storage accounts as a single and unified one. It provides a single access and control point for vast heterogeneous cloud storage providers APIs. Moreover, it must implement a policy mechanism that allows clients to define their own data processing rules. Detailed Cloud Storage Federation service architecture is shown in Fig. 1. The architecture shows seven main components:

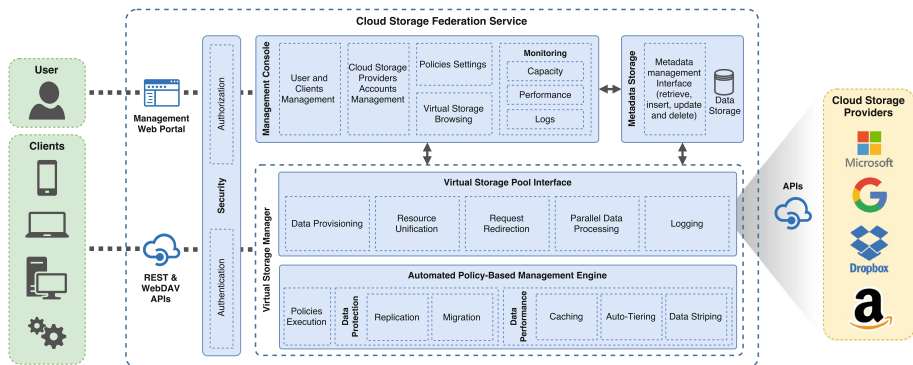


Fig. 1. Cloud storage federation service reference architecture

2.1 Management Web Portal

This component provides users a web UI for accessing the Management Console functionalities. As we are proposing a Cloud service it should have a self-service interface for user registration and resource management.

2.2 APIs

The APIs provide resources access for third-party application. Application must be registered by a user through the management web portal in order to grant the access to the APIs. Two technologies are shown in the architecture but it does not mean that it must be limited to them. REST APIs are suitable for resources management and application development, while WebDAV allows to access the service as a network drive on a PC.

2.3 Security

Obviously a security must be present on any service design. This module provides authentication and authorization for the Management Portal and the APIs. A token-based authentication should be used, like OAuth 2.0.

2.4 Management Console

This component exposes an interface for managing all the service resources and more. Its functionalities can be summarized as: User registration and account information management; Client applications management, to grant third-party applications credentials the access to the APIs endpoints or revoke them; Add or remove Cloud Storage accounts in order to scale or reduce the federated storage size, cloud Storage Providers that provides OAuth 2.0 protocol are recommended since it eliminates the need of storing user credentials; Setting policies, users must be able to create their own storage policies in order to automatically manage how data is processed by the federation service; Allows users to browse the virtual storage pool and manage the data in it, through the Management Portal; Monitoring the virtual storage properties, like performance, capacity and logs.

2.5 Metadata Storage

Provides a persistent storage for all service's metadata. Moreover, it exposes an interface that allows other components to easily access the metadata.

2.6 Automated Policy-Based Management Engine

This component provides automated and optimal decisions based on predefined policies by the users. It is in charge of executing users policies and based on them take care of how data is distributed across all cloud storage providers or how it is retrieved from them. Moreover, it is responsible for data protection

and performance decision making. In case of data protection, it should be able to replicate and migrate data in order to improve availability and durability. On the other hand, to improve data performance, techniques like caching, auto-tiering and data striping can be implemented.

2.7 Virtual Storage Pool Interface

Component responsible to provide access to the virtual storage as a seamless unified storage pool based on users configured cloud storage accounts. The component distributes and redirects the access request to the corresponding cloud storage providers, utilizing parallel processing technique in order to improve access performance, finally it collect provider’s responses and unify them into a single response to the client. Furthermore, requests and responses logging is done for further analysis and monitoring.

3 Personal Cloud Storage Federation Service

Personal cloud storage is a widely used technology today and there are plenty of providers to choose from. Often people end using more than one account for different reasons, e.g. free space offers or inclusion of account with other services as Google and Microsoft do with their emails accounts. Managing multiple cloud storage accounts can be very difficult and most of the time there are not fully exploited. As a proof of concept for our proposed model, we developed a personal cloud storage federation service, which integrates multiple cloud storage accounts into a single one. This section presents the results achieved and implementation details.

Our service allows users to self-register, add or remove their cloud storage accounts, create their own policies and browse the unified virtual storage. For data protection and performance, replication and data striping techniques are used. Other functionalities like client applications management and resource monitoring are not yet implemented.

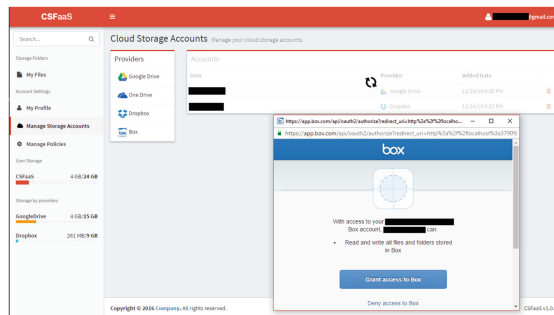


Fig. 2. Cloud storage accounts management

3.1 Cloud Storage Account Management

After users registration and account verification they are allow to register their cloud storage accounts through our service portal. Storage accounts from four different providers are supported; Google Drive, Microsoft OneDrive, Dropbox and Box. Moreover, a user can register multiples accounts from the same provider. Thanks to OAuth 2.0 protocol our service can access user accounts without storing their credentials. In first place, user selects the provider he wants to aggregate. Then, he is redirected to the provider's authentication page, where the user has to login and grant full access to our service. Finally, our service receives an authentication token, a refresh token, and an expiration date. These parameters are stored and used for transparently accessing user data. Figure 2 shows the cloud storage account management interface.

3.2 Policy Management

Our service offers two kind of policies, static and dynamic policies. There are three static policies that manage the default behavior of our service. The first, *global replication factor*, specifies the number of replicas must be stored for any uploaded file through our service. The second, *restrict to service's folder*, it allows our service to control all the files on the user accounts or be restricted only to its own folder. Finally, *account selection factor*, tells our service how to select where to save files in case of no dynamic policies were created. Policy management interface is shown in Fig. 3.

On the other hand, users can create as many dynamic policies as they want. Dynamic policies are composed by three parameters. The first one is the *trigger action*, e.g. uploading or deleting a file. The second one is the *trigger condition*, it specifies with files are going to be affected for the policy, e.g. file of a specific type, files with a size between a range or files uploaded to a specific folder. The third and last parameter is the *policy action*, it specifies what should our service do with the files, e.g. change the replication factor, store in a specific folder, store in a specific account or change the account selection factor. Figure 4 shows dynamic policy creation interface.

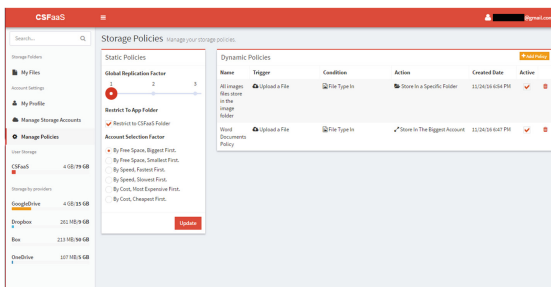


Fig. 3. Policy management interface

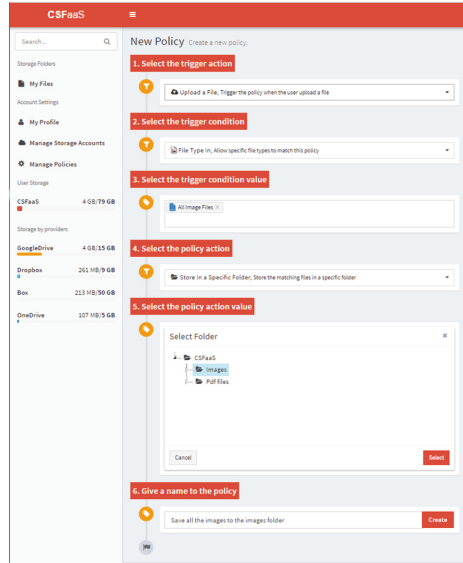


Fig. 4. Dynamic policy creation

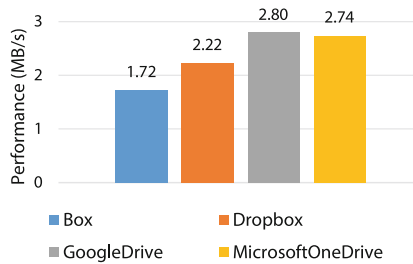


Fig. 5. Cloud storage provider performance evaluation

As can be noted in our static policies, Fig. 3, our decision model relies on three factors; cost, available space, and performance. The cost was measured according to 1 TB of space price. The available space depends on each account. For performance a previous evaluation of each provider was done, see results on Fig. 5.

3.3 Browsing Virtual Storage

Our service provides a unified storage view for the users. User are allowed to browse the virtual storage the same way as any other cloud storage provider interface. Moreover, file management operations are also provided; upload, download, delete, move, share and search. Figure 6 shows the storage browsing interface. Column “Location” was added for demonstration purpose.

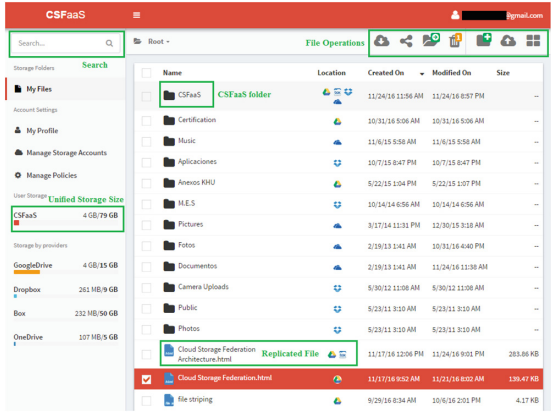


Fig. 6. Virtual storage browsing interface

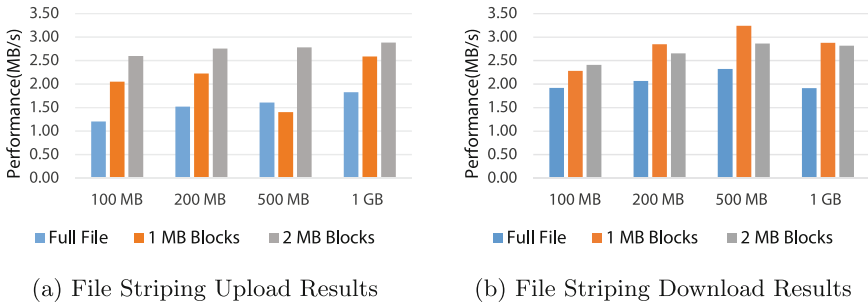


Fig. 7. File striping performance analysis results

3.4 File Striping

We conducted experiments with big files and distinct block sizes in order to evaluate the feasibility of using striping technique in our service. Results show that data performance and throughput can be improved with the combination of parallel processing and file striping. Figures 7(a) and (b) show achieved results for upload and download performance respectively.

4 Related Work

Several works regarding cloud storage federation and multi-cloud storage systems were studied. Yang and Ren [4], provide a framework for federating open cloud storage providers. Vernik et al. [5], present an on-boarding federation mechanism for adding a special layer on cloud storage providers allowing them to import data from other providers. Janviriyaya et al. [6], approach consist in a Multiple Cloud Storage Integration Systems based on RAID 0 striping technology. Zhao et al. [7], propose a middleware that enable any end-user application to automatically

and securely store files in multiples cloud storage accounts. However, none of the reviewed approaches provides a policy-based decision model for data distribution and retrieval, their decision models are fixed. Moreover, the focus point of these studies is end-user, while we are focusing on heterogeneous clients.

5 Conclusion and Future Work

We have introduced our reference architecture for the implementation of new storage federation services. Our approach presents a federation model where no interoperability between providers is required since it is managed for an external service. Moreover, we have described our personal cloud storage federation service as a proof of concept. Finally, an evaluation of the performance that can be achieved with the combination of file striping and parallel processing was shown.

Our future work is driven to improve the proposed decision model, allowing our service to make more complex selection of providers as well as using more aspects of storage (e.g. availability, workload and QoS). In addition, we aim to continue extending our service with the research and implementation of the concepts introduced in our reference architecture.

Acknowledgements. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-(H8501-16-1015) supervised by the IITP (Institute for Information & communications Technology Promotion)).

References

1. Turner, V., Gantz, J.F., Reinsel, D., Minton, S.: The digital universe of opportunities: rich data and the increasing value of the internet of things. IDC Analyze the Future, April 2014
2. Heller, P., Piziak, D., Stackowiak, R., Licht, A., Luckenbach, T., Cauthen, B., Misra, A., Wyant, J., Knudsen, J.: An enterprise architects guide to big data. Oracle Corporation, March 2016
3. Vellante, D.: What is federated storage. Wikibon, May 2010. http://wikibon.org/wiki/v/What_is_Federated_Storage%3F
4. Yang, D., Ren, C.: VCSS: an integration framework for open cloud storage services. In: IEEE 10th World Congress on Services (SERVICES 2014), Alaska, USA, pp. 155–160, June 2014
5. Vernik, G., Shulman-Peleg, A., Dippl, S., Formisano, C., Jaeger, M.C., Kolodner, E.K., Villari, M.: Data on-boarding in federated storage clouds. In: IEEE Sixth International Conference on Cloud Computing (CLOUD 2013), Santa Clara CA, USA, pp. 244–251, June 2013
6. Janviriyaya, P., Ongarjithichai, T., Numruktrakul, P., Ragkhitwetsagul, C.: Cloudy-Days: cloud storage integration system. In: Third ICT International Student Project Conference (ICT-ISPC 2014), Nakhon Pathom, Thailand, pp. 125–128, March 2014
7. Zhao, R., Yue, C., Tak, B., Tang, C.: SafeSky: a secure cloud storage middleware for end-user applications. In: IEEE 34th Symposium on Reliable Distributed Systems (SRDS 2015), Montreal, Canada, pp. 21–30, September 2015

Internet of Things

A Study on the IoT Framework Design for Ginseng Cultivation

Kyung-Gyun Lim¹ and Chang-Geun Kim^{2(✉)}

¹ Department of IT Convergence Engineering,
GyeongNam National University of Science and Technology, Jinju 660-758, Korea
i7027@naver.com

² Department of Computer Science and Engineering,
GyeongNam National University of Science and Technology, Jinju 660-758, Korea
cgkim@gntech.ac.kr

Abstract. The Republic of Korea's ginsengs are a high-priced special produce cultivated since the Koryo Dynasty (AD 918–1392) in the Middle Ages. Their efficacy has been studied for a long time and published in both domestic and foreign papers of internal medicines. Several Korean pharmaceutical companies are producing the immune enhancers with them and some of them are waiting for the FDA's approval while conducting clinical trials, during which their excellent efficacy has been proven. Having the cultivation period of 3 to 6 years, ginseng farmers can draw a high income if they keep an adequate growing condition for these expensive produce favored by many Koreans and foreigners. Thus, by grafting the IoT technology onto the ginseng growing conditions, the farmers will be able to increase their outputs and incomes, as well as increasing the competitiveness of the Korean ginseng. Such a method can also contribute to the reduction of labor force which is one of the serious problems in the agricultural sector where the population is continuously declining. While this study focuses on the designing of an IoT framework considering the characteristics in ginseng cultivation conditions, the results can be standardized and used for the other special produce that require a long-term cultivation period.

Keywords: IoT · Framework · Ginseng cultivation · Smart farm

1 Introduction

In recent years, the Smart Agriculture or Smart Farm has become an issue in the Republic of Korea (ROK) along with the IoT technology as the FTA agreements have been completed between US, China and Vietnam. Accordingly, the ROK government is implementing a subsidy package totaling up to 5.4 billion dollar in financial and taxation support to compensate the expected damages to the Korean farmers and improve their competitiveness by transforming the agricultural sector as an export-oriented industry. The administrative supports include various strategic export support in production and promotion of fresh and processed agricultural products. The government is also planning to lead uncompetitive crop or fish farmers to find other more lucrative businesses.

Due to the development of IT-related technologies in the ROK, more and more unimaginable things in the past are now being actualized. The smart agriculture is one of them. The existing labor-intensive crop cultivation methods are currently in the process of automation which is becoming much more effective, efficient and sophisticated as the result of integration of IT technology with farming technology. For example, the irrigation, light adjustments and crop-dusting process have been automated in several major farmlands where a variety of advanced technologies and statistical data are applied. It has become quite clear that only those who are able to analyze and use the data wisely will remain competitive in a fierce global competition. For this reason, development of ICT and its equipments has become essential.

Meanwhile, the ‘framework design’ in this study refers to a designing method that provides a consistency and user convenience by adopting a model where all the components used for the separate technical developments are being integrated.

Understanding the information related to the crop or fish cultivation is possible owing to the development of various sensors such as temperature, illuminance, and steering sensors, and the data obtained through these sensors must be processed in the form of signal that can be recognized by the users who will be receiving them with their communication equipments or devices. Also, it is important to miniaturize the sensors for portability, usability or power saving purpose. In the end, farmers can receive or find adequate information related to the current cultivation technologies and environments through their smartphones just by entering basic present data of their crops or fishes. However, these technologies still have shortcomings to overcome: inaccurate statistical data, unclear signals, equipment/device malfunctions, power consumption rate, or etc.

2 Related Research

Korean crop farmers are facing difficulties from FTAs, weather anomalies, aging workforce, and etc., especially when cultivating special produces that require a long growing period and more careful attention to maintain their quality. Accordingly, the researches for the ICT-integrated smart farms are being actively carried out to deal with these problems [1]. The Free Trade Agreement (FTA) refers to an agreement that lowers or eliminates all the trade barriers to ensure free movement of goods between nations. The trend in the world’s agriculture industry is to promote the smart farming technology that is expected to provide safe foods and solve the problem of price inflation or decrease in yield caused by typhoons, cold weathers or other natural disasters. Additionally, this technology can be used to produce crops in hostile environments like deserts or wild lands. Thus, a platform building plan based on the Internet of Things (IoT) technology is introduced in this study to promote smart farming [2–8].

Ginsengs are species of Araliaceae which are usually cultivated in the shaded soils where water drains well and propagated with seeds. However, it is not easy to grow them as their seeds are hard to obtain and require much effort to sprout from them so that, as a herbaceous perennial plant, it is practical to transplant one-year-old ginseng in shaded areas where their roots will become thicker over the years [8–17].

Most of the shade net-covered farms around the country are ginseng-growing farms. The shade nets are used to create shadows artificially but otherwise, farmers have to choose the naturally shaded areas. The water in the soil must be drained well all the time but the soil still has to maintain adequate humidity. Choosing the one-year-old tiny plants with much dirt covering their roots is advised but it is better to buy the nursery plants exclusively grown for a seeding purpose even if they are more expensive. Each plant should be planted in a 20 cm x 20 cm area to let the roots spread sufficiently. As it takes at least 4 years to be marketable, ginseng farms should be located at the quite desolated shaded areas where shades are so dense that other foreign seeds or grasses cannot grow easily. The ginsengs artificially cultivated at the farm are usually harvested after 3 years but recently, 6-year-old ginsengs are more welcomed at the market and receive much better prices. If composts or exclusive fertilizers are not added to the soil, it is estimated that the farmers may have to wait around 10 years before harvesting. The suitable harvesting period is from September to November when the leaves fall. In the end, water-draining, adequate level of humidity, shade and grass controls are essential for ginseng cultivation but they will have keep in mind that the ginseng population in each growing ground will be decreased over the years. As mentioned earlier, ginsengs become marketable after 3 to 6 years but the 6-year-old ones account 1/10 of entire production as they often become vulnerable to the gray mold and rot when they approach 6th years, reducing the number of harvestable 6-year ginsengs. It is advised to set up the farm where its temperature is cool, well-ventilated, and has a reddish soil. The disinfection works should be conducted 8 (min.) to 15 (max.) a year.

3 The IoT Framework Design for Ginseng Cultivation

It is quite difficult and unproductive to cultivate ginsengs by just observing them with farmer’s eyes or conducting pest control, fertilization and management based on the immediate forecasting. Also, the knowledges obtained through relevant websites or other information sources may not be useful as the farm environments are not the same everywhere. Therefore, utilization of sensors that measure the condition of the farm soil (e.g., temperatures, humidity and etc.) along with the big data is proposed to carry out precise management of cultivation process (Fig. 1).

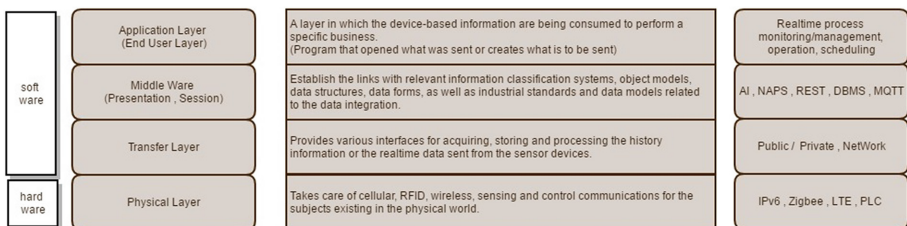
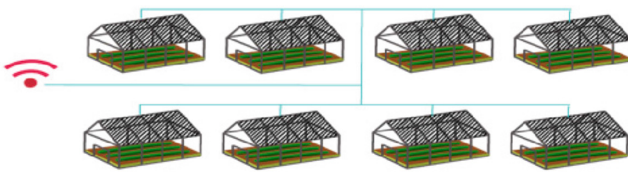
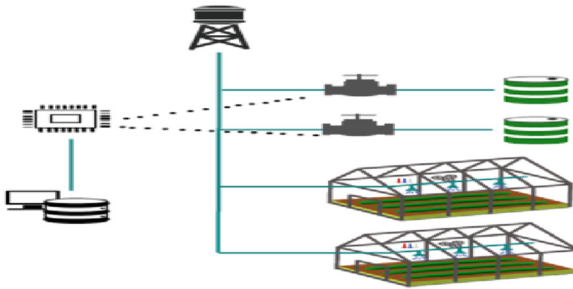


Fig. 1. Farm IoT base solution layer (Natures of OSI Layers).

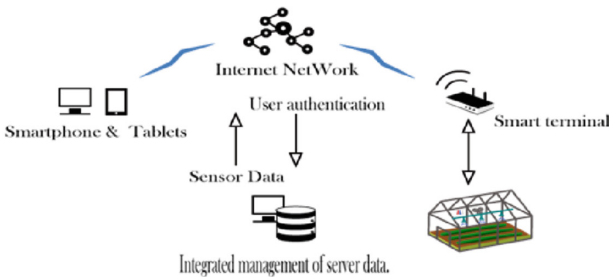
Through 7 OSI layers, check the basic network resources need to construct a ginseng cultivation environment. In a ginseng cultivation IoT environment, the most important part of its network should be designed based on the standardized design. Above table presents a re-interpreted design model of an IoT cultivation system based on the 7 OSI layers. This aims to establish a systematic network through these OSI layers whose purpose is to process every aspect associated with the electrical specification, physical connection, physical address, transfer time and reliability when one device in the network support layer attempts to transfer data to another device through the Physical Layer, Datalink Layer and Network Layer. In the ginseng cultivation environment, the construction work of a stable network and wireless control have been designed based on the Physical layer. This is to design various standard interfaces through the Transport



(a) Configuration of a multi-farm house network in the farm



(b) Simulation of a pest control spray method based on sensors



(c) An overall communication flow for the cultivation

Fig. 2. Configuration of an IoT-based ginseng cultivation environment.

Layer and to achieve realtime data transfer. Also, a Middle Ware environment has been provided between the transfer Layer and the Application Layer to store and integrate various data. The Application Layer (a user support layer) was designed focusing on the user interface in a realtime environment and the factors involved in data management/operation by providing a form that enables interactions between separate and unrelated softwares.

Figure 2 shows the IoT network design for the ginseng growing houses. The houses with a basic cultivation environment can communicate with other houses but they are bound in a single network group. Similar to other plants, more profits can be guaranteed for the ginseng farmers if they grow more such that this study aims to establish an integrated management system that can be used for the multiple number of houses. The sensors installed in each house performs Bluetooth communications and the data are delivered through the central communication system. Individual houses can control temperatures, humidity, water levels and sunlight amount through sensors and the collected data are delivered and stored at the integrated server through the gateway. The involved functions are: control of DC motor and fans based on the measured temperature and humidity, operation of ventilation fans and windows by the temperature and humidity levels, plant LED light control after sensing the level of illumination, and realtime monitoring of ginseng's growth in the house through CCTVs. The IoT-based ginseng cultivation management system is comprised of the Gateway system and the mobile program. The former operates multiple actuators (e.g., water pump, fans, DC motor and plant LEDs) based on the stored sensor data collected by the temperature, humidity and illuminance sensors. At the server, collected data are analyzed to set an adequate growing environment.

4 Implementation of the IoT Framework for Ginseng Cultivation

If the collected data from the database is for the germination phase and the sensor temperature obtained through Arduino is not in a temperature range between 10°C to 15°C, system fans will be operated until the temperature fits in the range. The same process will be carried out for the blooming phase but the temperature range will be 21°C to 25°C instead. The temperatures will be checked repeatedly in realtime. Figure 3 shows temperature sensor data processing in each growing phase and execution algorithm.

Figure 4 shows flow chart of storage process of temperature sensor data. Temperature is a key factor that affects the yields of ginseng products and their remedial effects. The temperature control sensor in a growing house sends the data to the web server through its network in realtime [18]. Then, the web server delivers the realtime data of house's environment and temperatures to the DB server where the Ginseng Cultivation Table is located. All the data containing house information, dates, and sensor temperatures are stored and managed here.

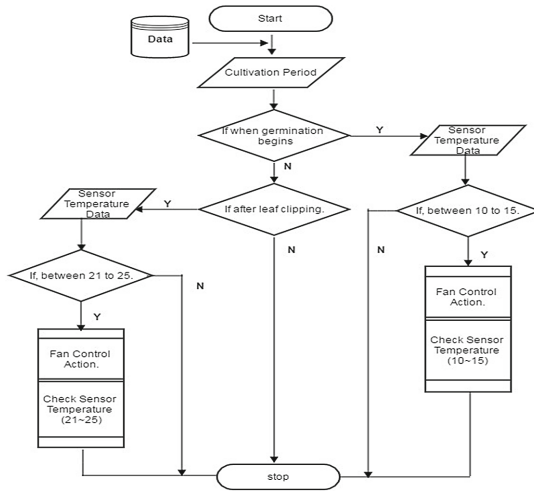


Fig. 3. Temperature sensor data processing in each growing phase and execution algorithm.

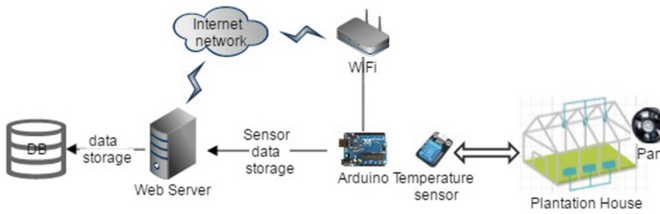


Fig. 4. Flow chart of storage process of temperature sensor data.

Figure 5 shows flow chart of sensor data processing using the scheduling service. The realtime sensor controls based on the data analysis service at the DB server will be performed at the growing house(s) through network communications.

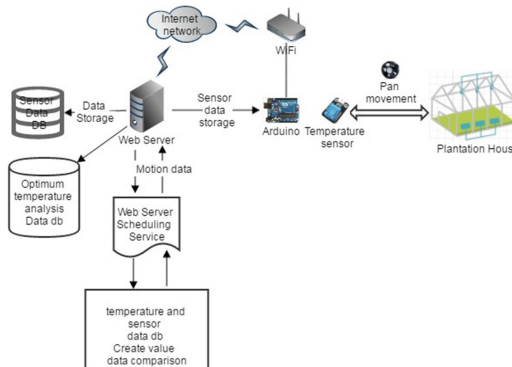


Fig. 5. Flow chart of sensor data processing using the scheduling service.

The web server will then use the scheduling service (realtime notifications) to compare the DB-stored sensor temperature data with the optimal temperature database and calculate the optimal temperature which will be then used to execute house control commands automatically. This process will reduce much operation time and prevent operational mistakes.

The signal control process between a client and a web server has been put to the test through simulation using socket programming. First connection was made between the client and web server, followed by later connection between the web server and Raspberry Pi that acted as a gateway. The signal was then transmitted to Arduino (serial communication) where the sensor values were collected. This process was reversed and the values were delivered to the client. Additionally, the streaming-based tests (Picture-image data transmissions) through the camera device have been conducted. The results of from these tests were satisfactory. Figure 6 shows Simulated implementation test And Arduino UNO R3 (+Duemilanoves).

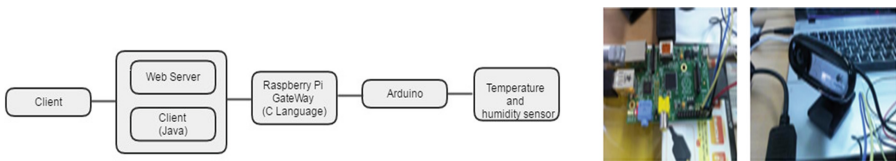


Fig. 6. Simulated implementation test and arduino UNO R3 (+Duemilanoves)

5 Conclusion and Future Work

In this study, a platform for the IoT-based ginseng cultivation system was designed instead of other popular produces such as mushrooms or strawberries, which have been considered in other preceding researches, as ginsengs are an expensive produce and difficult to cultivate due to their long-term growing process and sensitive environment. By constructing the platform, farmers will be able to reduce the damages by diseases and insects or rapid weather changes and increase their output, in addition to reducing the labor cost and manpower problems, both of which are becoming serious problems in rural areas.

Our future extended study includes implementation of the Ginseng History Information System which allows identification of fake Ginsengs and provision of safe and reliable products through Origin Checking system. This study proposes a basic platform design for the IoT-based ginseng cultivation system and the system should be discussed further for its implementation.

Even if an appropriate IoT technology has been applied to the ginseng farms, the priority is to solve the problems associated with diseases and insects to maintain ginseng’s growth and remedial effect. In the future system implementation, an integrated communication network will be constructed between 2 or 3 small-scale model houses. The data from the earth moisture, temperature and illuminance sensors will be processed through Arduino and delivered to the central system server with Raspberry Pi. Also, by

constructing a database concerning the cultivation environment, data-based sensor operation will be implemented for the test.

References

1. Kim, K.-A., Jeong, Y.-M., Park, D.-Y.: The implementation of farm management system based on IoT. In: 2016 KICS of Winter Conference, pp. 366–367 (2016). (In Korean)
2. Huh, J.-H., Je, S.-M., Seo, K.: Design and configuration of avoidance technique for worst situation in zigbee communications using OPNET. In: Kim, K., Joukov, N. (eds.) Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 331–336. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_33](https://doi.org/10.1007/978-981-10-0557-2_33)
3. Zhu, J., Zeng, K., Kim, K.H., Mohapatra, P.: Improving crowd-sourced wi-fi localization systems using bluetooth beacons. In: Sensor 2012 9th Annual IEEE Communications Society Conference on Mesh and Ad Hoc Communications and Networks (SECON), pp. 290–298. IEEE (2012)
4. Huh, J.H.: Design and android application for monitoring system using PLC for ICT-Integrated Fish Farm. In: Park, J., Jin, H., Jeong, Y.S., Khan, M. (eds.) MUE 2016. LNEE, vol. 393, pp. 617–625. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-1536-6_80](https://doi.org/10.1007/978-981-10-1536-6_80)
5. Kajioaka, S., Mori, T., Uchiya, T., Takumi, I., Matsuo, H.: Experiment of indoor position presumption based on RSSI of Bluetooth LE beacon. In: 2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE), pp. 337–339. IEEE (2014)
6. Huh, J.H., Yohan, B., Seo, K.: Bluetooth-tracing RSSI sampling method as basic technology of indoor localization for smart homes. *Int. J. Smart Home* **10**(10), 9–22 (2016). SERSC, Australia
7. Huang, H., Gartner, G., Schmidt, M., Li, Y.: Smart environment for ubiquitous indoor navigation. In: New Trends in Information and Service Science, NISS 2009, pp. 176–180. IEEE (2009)
8. Huh, J.-H., Je, S.-M., Seo, K.: Communications-based technology for smart grid test bed using OPNET simulations. In: Kim, K., Joukov, N. (eds.) Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 227–233. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_23](https://doi.org/10.1007/978-981-10-0557-2_23)
9. Chawathe, S.S.: Low-latency indoor localization using bluetooth beacons. In: 12th International IEEE Conference on Intelligent Transportation Systems, pp. 1–7. IEEE (2009)
10. Herrera, M.M., Bonastre, A., Capella, J.V.: Performance study of non-beaconed and beacon-enabled modes in IEEE 802.15.4 under bluetooth interference. In: Mobile Ubiquitous Computing, Systems, The Second International Conference on Services and Technologies, UBICOMM 2008, pp. 144–149. IEEE (2008)
11. Huh, J.H., Seo, K.: RUDP design and implementation using OPNET simulation. In: Park, J., Stojmenovic, I., Jeong, H., Yi, G. (eds.) Computer science and its applications, CUTE 2014. LNEE, vol. 330, pp. 913–919. Springer, Berlin Heidelberg (2015). doi:[10.1007/978-3-662-45402-2_129](https://doi.org/10.1007/978-3-662-45402-2_129)
12. Huh, J.H., Lee, D.G., Seo, K.: Design and implementation of the basic technology for realtime smart metering system using power line communication for smart grid. In: Park, D.S., Chao, H.C., Jeong, Y.S., Park, J. (eds.) Advances in computer science and ubiquitous computing, CUTE 2015. LNEE, vol. 373, pp. 663–669. Springer, Singapore (2015). doi:[10.1007/978-981-10-0281-6_94](https://doi.org/10.1007/978-981-10-0281-6_94)

13. Otgonchimeg, J.H., Huh, S., Seo, K.: Advanced metering infrastructure design and test bed experiment using intelligent agents: focusing on the PLC network base technology for Smart Grid system. *J. Supercomputing* **72**(5), 1862–1877 (2016). Springer
14. Park, J.H., et al.: Design of the Real-Time Mobile Push System for Implementation of the Shipboard Smart Working. In: Park, D.S., Chao, H.C., Jeong, Y.S., Park, J. (eds.) *Advances in Computer Science and Ubiquitous Computing, CUTE 2015*. LNEE, vol. 373, pp. 541–548. Springer, Singapore (2015). doi:[10.1007/978-981-10-0281-6_78](https://doi.org/10.1007/978-981-10-0281-6_78)
15. Inoue, Y., Sashima, A., Ikeda, T., Kurumatani, K.: Indoor emergency evacuation service on autonomous navigation system using mobile phone. In: *ISUC 2008*, pp. 79–85. IEEE (2008)
16. Huh, J.-H., Seo, K.: Design and test bed experiments of server operation system using virtualization technology. *Hum.-centric Comput. Inf. Sci.* **6**(1), 1–21 (2016). Springer, Berlin Heidelberg
17. Huh, J.-H., Seo, K.: Smart grid framework test bed using OPNET and power line communication. In: *Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems*, pp. 736–742. IEEE (2016)
18. Lee, J., et al.: A study on the necessity and construction plan of the internet of things platform for smart agriculture. *J. Korea Multimedia Soc.* **17**(11), 1313–1324 (2014). (In Korean)

An IPS Evaluation Framework for Measuring the Effectiveness and Efficiency of Indoor Positioning Solutions

Jacqueline Lee Fang Ang¹(✉), Wai Kong Lee¹, Boon Yaik Ooi¹,
and Thomas Wei Min Ooi²

¹ Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia
jacquelineang@lutar.my, {wklee, ooiyby}@utar.edu.my

² Intel Penang, FIZ 3, 11900 Bayan Lepas, Pulau Pinang, Malaysia
thomas.wei.min.ooi@intel.com

Abstract. The indoor positioning system (IPS) has been attracting great attention from researchers, thanks to the rapid adoption of smartphone technologies. Although there are many IPS proposed in the past decade that claimed to have good performance, all of them use their own method to evaluate and compare the accuracy of the proposed solution. During the evaluation phase, the method of gathering ground truth data (original position) is often not well described. As such, it is very difficult for other researchers to reproduce the work and improve on the existing methods. In this paper, we proposed a simple to implement framework to facilitate the process of evaluating IPS accuracy. Under this framework, the IPS position coordinates and ground truth are sent to the server using REST protocol when the phone reads an event triggered from tags scan placed on a fix position. We evaluated an existing well-known IPS technique, the Pedestrian Dead Reckoning (PDR) technique using our IPS evaluation framework. From our experiments, we showed that in addition to measuring the accuracy of IPS, the proposed solution can also measure the IPS accuracy deviation over time. Instead of relying on precision and recall, the framework also includes visualization tool for researchers to observe the overall accuracy of an IPS.

1 Introduction

An indoor positioning system (IPS) is a system that tracks and locates objects within indoor spaces (e.g. inside a building, underground and etc.). The need for IPS has been attracting a lot of attention in recent years because the Global Navigation Satellite System (GNSS) does not work well in indoor environment even though it enjoys great success in outdoor environment. The inaccuracy of GNSS in indoor is caused by multiple interferences from the indoor environment such as pillars and ceilings which reflected the transmitted satellite signals. On the other hand, modern smartphones are widely used to implement many important indoor positioning techniques such as Wi-Fi fingerprinting [3–5, 10], magnetic fingerprinting, Pedestrian Dead Reckoning (PDR) [2, 6, 9] and many others.

Proper visualization and interpretation of gathered data from the smartphones sensors and the ground truth are crucial to verify the accuracy of proposed IPS. A survey conducted by Adler et al. [7] concluded that many authors did not report the detailed process of gathering ground truth data. As such, it is assumed that “manual measurements using rulers and distance meters were used for ground truth positions” [7]. Furthermore, calculation of accuracy percentage in the IPS is not well described and solely depends on the manual measurements from the ground truth. This may create confusions among the readers and prevent a fair comparison for various IPS proposed in the literature.

In this paper, we introduce an easy to implement framework to show the IPS accuracy deviation over time instead of portraying only the final positioning error by evaluating a well-known IPS using the PDR technique. Besides that, every developed IPS has more than one methods to be implemented and compared. For example, there are more than one way to obtain the step length estimation of the PDR technique (e.g. Weinberg approach, fixed step length, etc.). This makes the comparison among them cannot be fair when every sampling for each method is collected at a different point of time as they are affected by several factors like time drift and noise and also the human walking attitude. With this framework, every position coordinates from all of the methods of the developed IPS can be sent to the server simultaneously which can minimize the errors mentioned.

Furthermore, besides measuring the mapping accuracy, this framework also enables visualization comparison between various IPS proposed in the literature with the ground truth in a 2D plane via a web browser. The system can also show more than one positioning in the same 2D plane. The proposed framework can reduce the time and effort spent in the positioning error measurement process, which in turn helps to create a fair platform to compare existing proposed solutions. To ease the deployment of gathering data from the smartphones, position coordinates will be sent to the server in real time.

The rest of the paper proceeds in the following order. In Sect. 2, we present the related existing IPS solutions and in Sect. 3, we describe the method of implementation of our proposed solution, followed by evaluation and discussions in Sect. 4. Then, directions of future work and conclusion will be in Sect. 5.

2 Related Works

According to a survey done by authors in Adler et al. [7], they concluded that a high percentage of publications describe their methods to gather ground truth data poorly. Many authors did not present the detail process and it is assumed that manual measurements using rulers and distance meters were used for ground truth positions. Without full disclosure of the experiments, it is hard to validate and reproduce the results independently.

2.1 Methods Employed in the Evaluation of Positioning Accuracy from Existing IPS that Uses Dead Reckoning Technique

Another popular study on indoor positioning solution is the PDR technique. This technique uses smartphone sensors like accelerometer, magnetometer and gyroscope to observe pedestrian movement [6]. User's stride length and heading orientation were derived from the raw data of these sensors. [9] The basic idea of dead reckoning is to derive the user's current location based on the previous location by extracting the current stride length and the respective heading orientation.

An experiment done by Radu et al. evaluate their accuracy by calculating the localization error using Euclidean distance between the known position of these reference points and their proposed IPS location estimated at the time of encounter. The experiments were conducted by selecting a corridor track of 100 m and along the corridor track, 20 reference points representing entrances to offices is set.

Kang et al. evaluate the accuracy of proposed solution by representing them on a two-dimensional building floor plan. Before starting their experiments, a walking path is set on the floor plan with total length of 165.55 m and the walking path is tracked with a step counter multiple times to get the approximate number of steps walked under normal walking speed. Then, the trajectory of proposed solution is mapped on to the same building plan and compared with the pre-set walking path. Localization error is calculated based on the walking distance. For example, for every 20 m distance walked, the localization error was computed by subtracting 20 m with the distance walked produced by the proposed solution.

Chen et al. conducted their experiments in an office zone. To obtain the ground truth data, a camera is used to record the whole walking process and each step walked is manually marked. Then, to evaluate the performance of proposed solution, the criterion of Root Mean Square Error (RMSE) calculation is applied at the end of the experiment.

2.2 Summary

Based on these reviews, we can conclude that many existing works presented its own method to gather ground truth data and evaluate the performance of their proposed solution in many different ways. In addition, the same conclusion is also found in [7], whereby the authors concluded that there is no standard benchmarking framework to measure the accuracy of IPS solutions. Therefore, we are motivated to design a framework to measure the effectiveness and efficiency of IPS solutions, which is also simple to be implemented.

In this paper, the proposed framework describes clearly the process to gather ground truth data and allows visual comparison on multiple IPS. With this framework, benchmarking can be performed between existing IPS and newly developed IPS easily.

3 Methodology

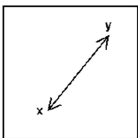
In this section, the concept of our framework will be discussed. First of all, for every developed IPS solution, the overall percentage of positioning accuracy and their positioning errors are often obtained based on the ground truth. As mentioned in the literature review, many existing works did not describe well on how they gather the ground truth for their IPS.

In this paper, ground truth coordinates will be set permanently with every coordinate representing a checkpoint. Then, in every checkpoint, tags will be set up as a means for triggering an event depending on the required precision. Tags like QR or barcode scan or even NFC scan can be used for triggering the event. With this triggered event in every checkpoint, coordinates of the ground truth and coordinates of developed IPS will be sent to the server simultaneously. This automation is proposed to reduce the error due to time drift, noise and the human walking attitude. The communication between the smartphone with the framework is basically carried out via REST protocol. This makes the process of collecting ground truth to be in a more systematic manner and by sending the data to the server in real time instead of manually transferring the data from the smartphone to the PC, it speeds up the process in performing the experiments.

3.1 Calculation of Distance Error Between the Ground Truth and the IPS

Calculation of positioning error is performed for every developed IPS at every checkpoint by using two methods as shown below.

1. Euclidean distance function.

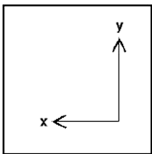


Euclidean

The Euclidean distance function can be defined as a straight line between two points. If $x = (x_1, y_1)$ and $y = (x_2, y_2)$, then the distance is given by

$$d(x,y) = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)} \tag{1}$$

2. Manhattan distance function.



Manhattan

The Manhattan distance function can be defined as the distance travelled to get from one data point to the other if a grid-like path is followed. If $x = (x_1, y_2)$ and $y = (x_2, y_2)$, then the distance is given by

$$d(x,y) = |x_1 - x_2| + |y_1 - y_2| \tag{2}$$

3.2 Evaluation of Overall Positioning Accuracy

The overall percentage of positioning accuracy can be determined by using a method called precision and recall which is usually used in evaluating search strategies as shown in Fig. 1.

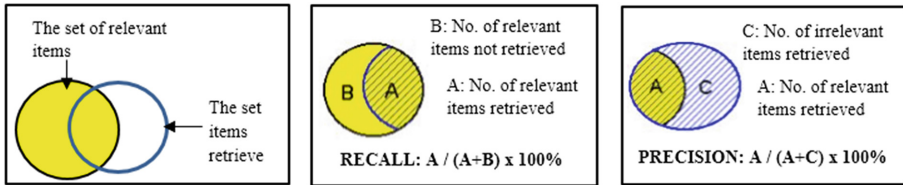


Fig. 1. Simple diagrams on precision and recall

In every search topic, there is a set of items which is relevant and items are assumed to be either relevant or irrelevant. The actual retrieval set may not perfectly match the set of relevant items.

Precision is the ratio of the number of relevant items retrieved to the total number of irrelevant and relevant items retrieved. On the other hand, recall is the ratio of the number of relevant items retrieved to the total number of relevant items. These ratios are usually expressed as a percentage.

With this, the overall percentage of the positioning accuracy of the IPS can be calculated.

Besides precision and recall, this framework also allows visualization perspectives in a 2D plane to show the comparison of coordinates between the ground truth and the developed IPS in every checkpoint. The main features of the visualization components are listed as follows:

- Allows visualized comparison of multiple IPS solutions with the ground truth in a single graph.
- Trend lines of accuracy deviation over time shown from the first point to the last point for each IPS.

In the web page, checkboxes with all the files from experiment will be prompted. User can choose to view the positioning graph of each developed IPS separately or choose to view the comparison between ground truth data with the developed IPS on the same graph in a separate browser. With this feature, positioning error can be seen clearly through the graph when several of them were being compared. In addition, this graphical view of the positioning coordinates can support unlimited number of files for comparison in a single graph.

4 Evaluation and Discussions

In this section, we present the experimental setup, software features and discuss our findings. The experiment on this framework is done in a lab environment with layout illustrated in Fig. 2.

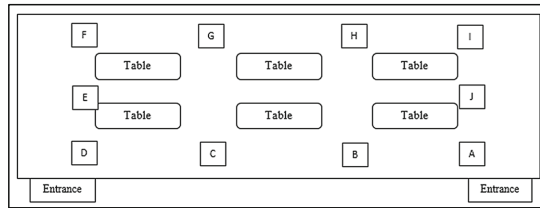


Fig. 2. Checkpoints set for experiment in a lab environment

Firstly, the distance between each checkpoint from checkpoint A to J is set to be 4 meters apart each other. The ground truth coordinates is set as shown in Table 1 and a simple example of visualization of the ground truth coordinates is shown in Fig. 3. Every coordinate generated from developed IPS will be compared with this ground truth to determine its positioning accuracy. Then, in every checkpoint, barcode scan were used for the event triggering.

Table 1. Checkpoints set with its own coordinates

Checkpoint	A	B	C	D	E	F	G	H	I	J
Coordinate (in cm)	0,0	0,400	0,800	0,1200	400,1200	800,1200	800,800	800,400	800,0	400,0

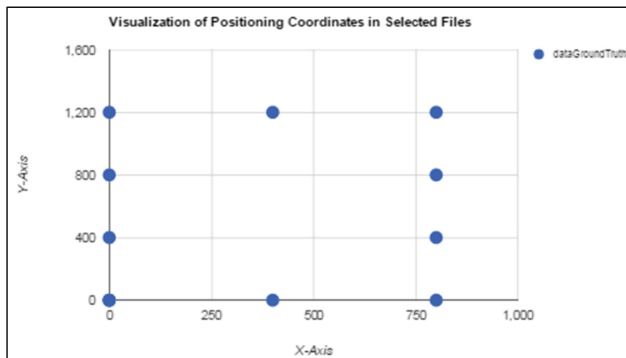


Fig. 3. Graph of ground truth visualized in web page

Then, an IPS application is installed in the smartphone. When user reaches a checkpoint and scans the barcode, selected data like position coordinates will be sent to the server and saved into specific text files accordingly. Example file types are ground truth files and developed IPS (e.g. dataPDR) files. These files will be automatically generated every time the application is made to run with unique numbering so that all the files generated in each run will not be erased when new experiment takes place.

Positioning error in each checkpoint will be done by using the calculation formula from the Euclidean and Manhattan distance function in this experiment. Then, trend lines of accuracy deviation over time from point A to point J will be shown with accumulated Euclidean and Manhattan distance functions respectively.

4.1 Developed IPS Using the PDR Technique

To evaluate this framework, an IPS solution using the PDR technique is implemented. As mentioned, PDR technique greatly relies on not only the heading orientation but also the step length estimation. Several methods to obtain the step length were reviewed as follows

- i. Fixed step length.
- ii. Weinberg Approach.

$$\begin{aligned} \text{step length} &= k * \sqrt[4]{a_{max} - a_{min}}, k \\ &= 0.41, a_{max} \text{ and } a_{min} \text{ are accelerometer peaks value} \end{aligned}$$

- iii. Preconfigured height of a person.

$$\text{step length} = (\text{height} * k), k = 0.413(\text{for women}) \text{ and } k = 0.415(\text{for men})$$

When a barcode scan is performed, the position coordinates derived from every method is sent to the server simultaneously.

Based on Fig. 4, the web page is firstly equipped with checkboxes for user to choose which files to be compared. After checking the desired files, click the submit button, then, the comparison of the coordinates can be visualized on a separate browser as shown in Fig. 5.

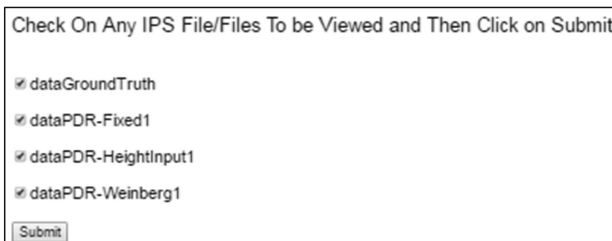


Fig. 4. Checkboxes on web page to select files for visualization

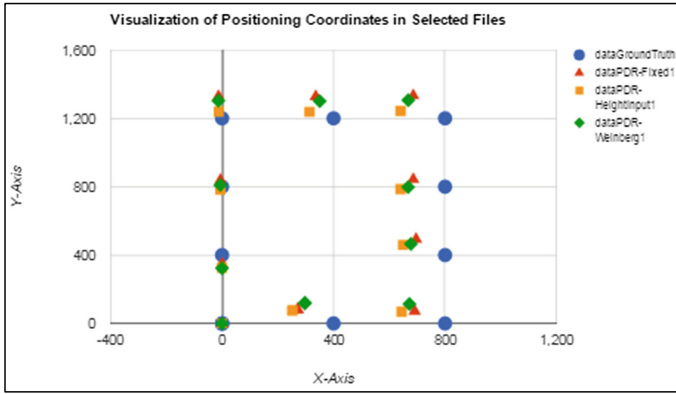


Fig. 5. Comparison of ground truth and developed IPS solution

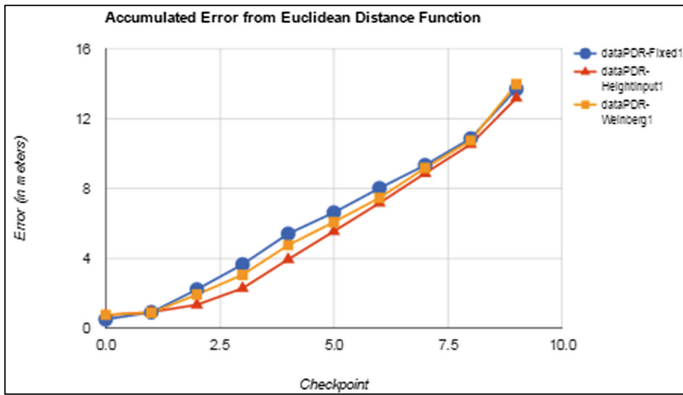


Fig. 6. Accumulated error from Euclidean distance function

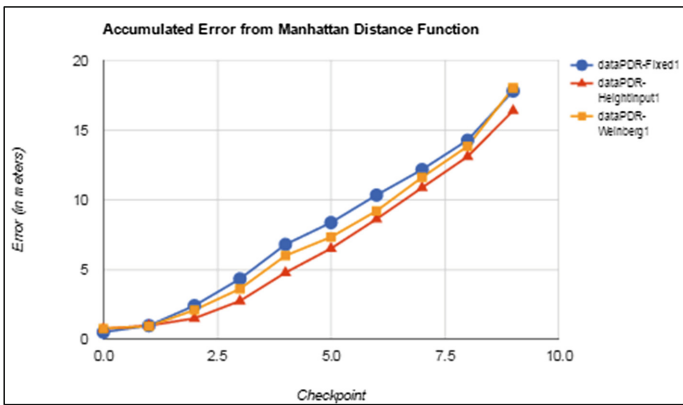


Fig. 7. Accumulated error from Manhattan distance function

The trend lines of accumulated distance error from the first point to the last point calculated using Euclidean distance function and Manhattan distance function are shown in Figs. 6 and 7 respectively.

With this framework, the positioning accuracy between ground truth and developed IPS can be evaluate in a systematic manner and wireless transfer of data to the server also helps ease and speed up experiments process whereby the outcome can be viewed instantly. Furthermore, by sending the position coordinates of several different IPS at the same time, the time drift is reduced. In future, this framework will be expanded to evaluate IPS using the Wi-Fi Fingerprinting technique and newly developed IPS can also adopt our proposed framework to perform fair benchmarking against other existing solutions.

5 Conclusion and Future Work

This work presents a visualization framework for IPS solutions that allows visual comparison between the ground truth data and multiple IPS solutions on the same graph. Instead of manually transferring the data and put them into excel files for interpretation, real-time data collection and automation of transferring data to the server eases the development during experiments. Besides, the collection of position coordinates from every developed IPS at the same time reduces the time drift as well. Although this work is in the preliminary stages which only applied for the PDR technique, but in future, this framework can be adopted in the development of new IPS solutions (e.g. by using the Wi-Fi Fingerprinting method) whereby the new position coordinates can be compared with existing IPS directly. With this, it allows fair comparison between the newly developed IPS with the state-of-the-art.

References

1. Niu, L., Saiki, S., Masumoto, S., Nakamura, M.: Implementation and evaluation of cloud-based integration framework for indoor location. In: iiWAS 2015 Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, p. 72 (2015)
2. Radu, V., Marina, M.K.: HiMLoc: indoor smartphone localization via activity aware pedestrian dead reckoning with selective crowdsourced WiFi fingerprinting. In: 2013 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–10 (2013)
3. Schussel, M., Pregizer, F.: Coverage gaps in fingerprinting based indoor positioning: the use of hybrid gaussian processes. In: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–9 (2015)
4. Jedari, E., Wu, Z., Rashidzadeh, R., Saif, M.: Wi-Fi based indoor location positioning employing random forest classifier. In: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–5 (2015)

5. Boonsriwai, S., Apavatjrut, A.: Indoor WIFI localization on mobile devices. In: 2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1–5 (2013)
6. Kang, W., Han, Y.: SmartPDR: smartphone-based pedestrian dead reckoning for indoor localization. *IEEE Sens. J.* **15**(5), 2906–2916 (2015)
7. Adler, S., Schmitt, S., Wolter, K., Kyas, M.: A survey of experimental evaluation in indoor localization research: a look back on IPIN conferences 2010, 2011, 2012, 2013 and 2014. In: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–10 (2015)
8. Tao, P., Wang, X., Wang, C., Shi, D.: Hybrid wireless indoor positioning with iBeacon and Wi-Fi. In: 11th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2015), pp. 1–5 (2015)
9. Chen, Z., Zhu, Q., Jiang, H., Soh, Y.C.: Indoor localization using smartphone sensors and iBeacons. In: Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference, pp. 1723–1728 (2015)
10. Wang, B., Zhou, S., Liu, W., Mo, Y.: Indoor localization based on curve fitting and location search using received signal strength. *IEEE Trans. Ind. Electron.* **62**(1), 572–582 (2015)

An IoT-Based Virtual Addressing Framework for Intelligent Delivery Logistics

Omar Hiari¹(✉), Dhiah el Diehn I. Abou-Tair¹, and Ismail Abushaikha²

¹ School of Electrical Engineering and Information Technology,
German Jordanian University, Amman, Jordan
{omar.hiari,dhiah.aboutair}@gju.edu.jo

² School of Management and Logistics Sciences,
German Jordanian University, Amman, Jordan
ismail.abushaikha@gju.edu.jo

Abstract. The Internet of Things (IoT) has been one of the influential paradigms in the development of logistics transport functions. The introduction of IoT in logistics has impacted application areas such as capacity sensing, planning, route optimization, and energy management. However, most works presented so far assume the existence of physical addresses for all applications. This paper, as a result, deals with the logistics delivery inefficiency represented by the lack of physical addresses that is common developing countries. We adopt an IoT approach to propose a virtual addressing framework for tracking, monitoring, and managing package deliveries efficiently. The framework consists of a node network that provides address information virtually to enable better deliveries.

Keywords: IoT · Internet of Things · Logistics · Intelligent transportation systems

1 Introduction

The IoT paradigm provides the promise of connected devices to all objects that surround humans. This enables a multitude of solutions for problems that existed traditionally in various areas. Most importantly, it could provide a positive impact on the environment, economy, and society. Moreover, developing countries in particular have unique problems that can be addressed by the promise of IoT. An example of such an area with unique problems is logistics. Logistics has long been considered by academics and practitioners as a dynamic function of business that would provide substantial savings for organizations. One problem particular to developing countries is the lack of a physical address for delivery targets in both rural and urban areas. In a recent market research by the united postal union (UPU), structural constraints has been cited as one of the common trends preventing the growth of postal systems in some developing regions [9]. Needless to say, this results in a lot of inefficiencies when attempting to manage a logistical operation.

The logistics industry itself has witnessed in the past few years a rise in the use of van freight for secondary transport within cities and residential areas [1, 2]. This has resulted in a rising challenge of arranging efficient and responsive deliveries to end-users. From a logistics and operations management perspective, this falls under what has been termed in recent literature as city logistics [5]. The concept was first coined in the logistics and operations management literature in order to optimize the last-mile delivery of goods. Despite the large body of research on logistics and transport systems in the field of supply chain management, recent literature suggests that the last-mile delivery logistics still holds a substantial research opportunity [1, 5, 7].

The purpose of the framework presented in this paper is to develop and exploit a virtual address framework for the efficient delivery of mail and parcels. The virtual addressing framework proposed, as a result, would facilitate the effective and efficient management of transport and delivery of packages. Although virtual addressing has been suggested by recent literature [8], the question that remains is how to achieve highly efficient delivery systems in the absence of physical addresses to delivery companies. The context of developing countries is particularly relevant to this framework due to the limited accessibility to physical addresses in most cities. Hence, the proposed framework in this paper tries to fill this gap for the last-mile challenge being faced by express delivery companies in the context of insufficiency of the physical address information.

2 Background and Related Work

Much work has been done trying to improve the logistics framework based on existing physical addresses. In addition, the use of ICT technologies in the transport and logistics industry is mostly limited to track-and-trace applications [3, 4, 6]. Harris et al. [3] review 33 EU framework program projects in freight transport to examine the major efforts in ICT developments in this field. Perego et al. [6] found that the most common applications for wireless technology in the logistic industry focus on order tracking and vehicle location monitoring. Verdouw et al. [11] proposed an IoT-based logistic information systems in agri-food supply chains. The framework introduced in [11] focused on preservation of perishable products. The cited works, however, do not consider solving the barrier of physical addressing in developing countries.

The United States Postal Service introduced the concept of virtual addressing in [8]. However, the idea was to transform existing physical addresses to a virtual addressing framework to provide more or better services. [8] also addresses the challenge of transforming the existing information technology systems to adopt the virtual addressing system. The framework introduced in [8], nevertheless is still dependent on the existence of a physical home address.

The framework proposed in this paper addresses the physical addressing problem by introducing virtual addresses for areas that lack physical addresses. The virtual addresses, through an IoT framework, would map to a physical location that is determined by a static node. Nevertheless, the framework does have

challenges that have to be addressed adequately. Going forward, this type of framework, enables also global logistics in such a manner that everybody would have one unique virtual address that could map to any physical location.

3 Case Study

The motivation for conducting this research stemmed from a logistical inefficiency facing the package delivery companies in the developing countries that lack physical addresses [10]. In order to gain a greater understanding of the challenges faced by delivery companies, this paper adopts a case study approach. We studied the day-to-day operations of a local logistics service provider. Moreover, the logistics provider studied is considered one of the largest express operators in terms of market share and reach. The case study is used in this paper to provide an understanding of delivery operations in developing regions, but also to provide a context-based case for implementing the framework proposed in this paper. Also one of the authors of this paper, based on his background in logistics, provided insight on the daily operations and the challenges faced by package delivery companies. This allowed us to understand the phenomenological context in which the operations occur and identify challenges and gaps.

In addition to the traditional delivery challenges facing most delivery companies worldwide in terms of dealing with incorrect addresses, the unavailability of physical addresses in developing countries has been a big challenge in the logistics business. Package delivery companies in developing regions depend largely on the knowledge and experience of their ground couriers in a particular area in order to allocate the accurate consignee's address. Nevertheless, operations and ground couriers still face challenges in identifying the exact home address, although the street address might be correct. As a result, most deliveries either fail or are retried after multiple delivery attempts due to poor addressing.

Other challenges faced by package delivery companies are that customers do not provide accurate address because of the oral culture. Therefore, customers tend to provide the name, city, town, and phone number of the consignee. The delivery function becomes even more challenging when the consignee is located in urban areas or a small village. Most villages in developing countries still lack a street name, house, or even block number. Moreover, some areas like refugee camps do not have street names or lot numbers.

This case study background suggests that the lack of a proper address system has resulted in logistical challenges for delivery companies. Hence, the purpose of this research is to explore and develop a virtual addressing system to address the problem at hand. Needless to say, such a framework would benefit more than just delivery companies.

4 IoT Virtual Addressing Framework

Virtual addressing as a concept would provide an address space larger than what can be physically addressed. As a result, and in the context of logistical

operations, utilizing IoT to enable virtual addressing would provide tremendous assistance to package delivery services among others. The following sections describe in detail the main aspects of the proposed framework.

4.1 Architecture

The physical architecture of the virtual addressing framework is based on a centralized model as depicted in Fig. 1. The framework includes the following subcategories:

Static Node Network. The static node network includes multiple components including the static nodes that would be attached to a physical location or a building. The nodes would store location information and collect geographical data. Each node would have a unique ID and is fixed in a certain location. Additionally, the nodes can collect information about existence of individuals that can receive a package at a particular location. This would save the courier the trip if nobody is available for pick up.

The local networks connected to the individual nodes are tasked with the collection of the geographical data and node information from all the nodes. The local networks could be either a cellphone network or a nearby wifi network. The local networks are also tasked with sending the collected information over the internet to a central database.

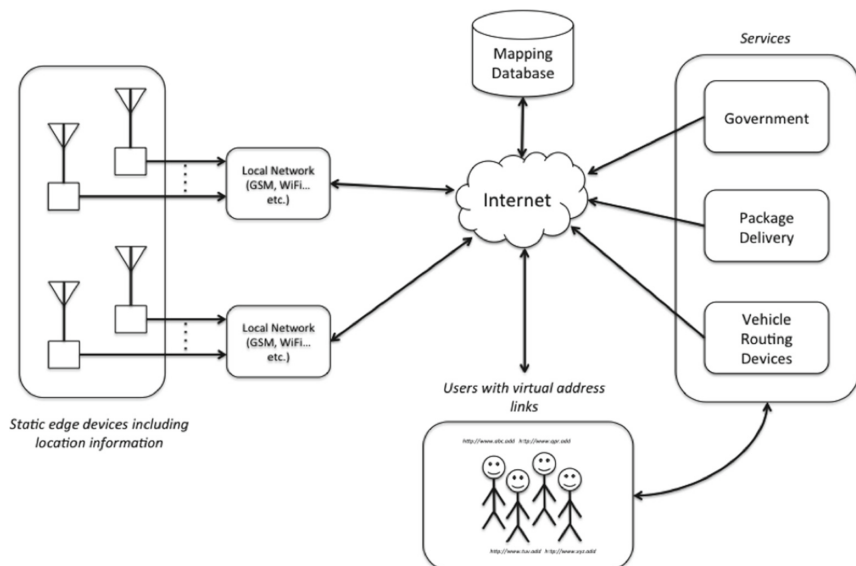


Fig. 1. Simulation results for the network.

Users. The users include the individuals each using a virtual link to map to a physical address. Although each user may bear a unique virtual address, multiple virtual addresses can be mapped to a single physical address. For example, family members living in the same home or employees of a single company.

The idea here also is that every user or individual will always have the same virtual address even if their physical addresses changes. This enhances the user experience such that the user would update only one central location when their physical address changes. This also prevents the user from having to update their address for all services individually.

Services. The services combine all services that could benefit from access to a virtual addressing framework. This includes, but is not limited to: emergency services, governmental services like delivery of court order notifications or driving violation notices, banks, package delivery services, and vehicle routing devices.

Mapping Database. The mapping database maps the users unique virtual links to the unique IDs of static devices. The unique IDs in turn maps to the physical information of the node ex. a GPS coordinate or a drop pin in a map.

4.2 Package Delivery Algorithms and Data Analysis

Figure 2 shows the various data processing modules for a package delivery use case described in this paper. The data flow involves mapping package data to address data and then providing updates to operation managers in addition to couriers picking up or making deliveries. The data flow consists of the following pillars:

Geo Data Collection and Address Verification. This module represents the collection of geographic data from the different static and dynamic nodes. Static nodes are the physical nodes that exist at static addresses representing a delivery target. The static nodes are expected to be also equipped with a short range wireless communication scheme (ex. Bluetooth) to enable short range or indoor guidance. Delivery couriers often can reach to a general area but have trouble finding the address when on foot. The short range communication scheme aids in that regard.

The dynamic nodes represent nodes with location data that changes frequently. The dynamic nodes could be either part of a fleet management system or nodes that are integrated into courier end devices (ex. PDA with GPS).

This module also has an address verification part for security purposes. In the framework proposed in this paper, the location data of the static nodes is stored along with a unique ID in a central data store. The address verification part is expected to verify every time the address is accessed for mapping in the data store that the static node has not been moved from its original position. In order to do address verification, a process needs to be introduced for first time installation that guarantees the authenticity of the location data.

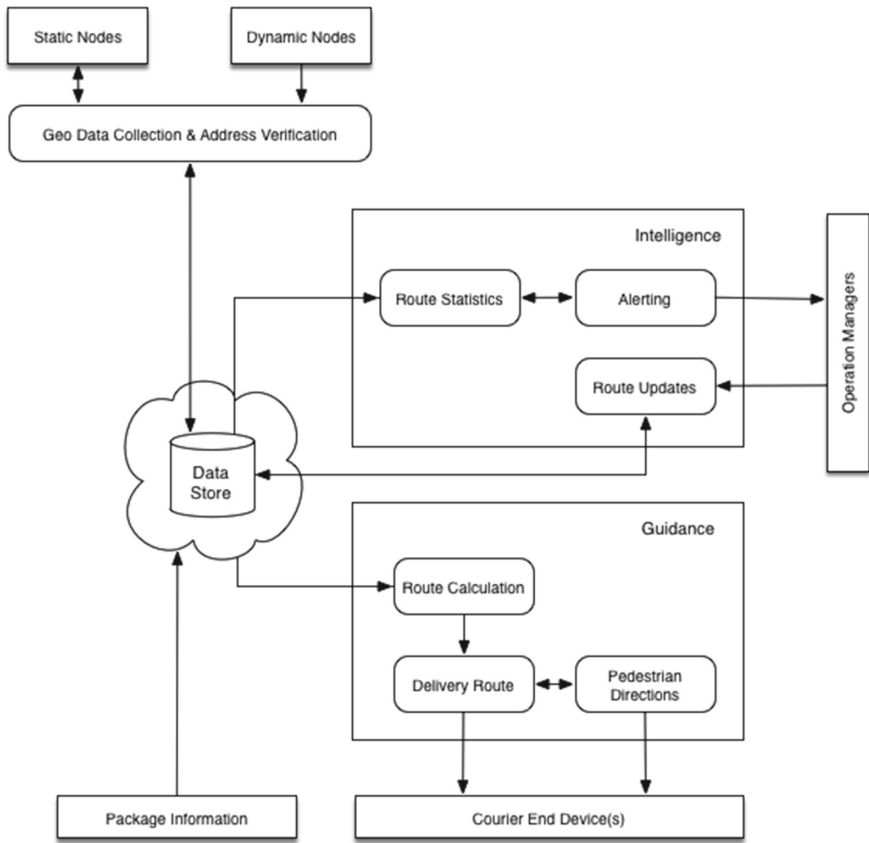


Fig. 2. Data flow

Intelligence. The intelligence module delivers intelligent analysis of the data to operation managers. This includes real time and projected package delivery data. The real time delivery data is compiled from the current positions of all couriers in a certain region and the package information. The package delivery data can be used by operation managers to proactively take actions and monitor couriers. On the other hand, the projected delivery data of packages are used to forecast package delivery statistics for various geographical locations. This is then used to derive possible future failed or late deliveries. The outcomes of this module can be viewed graphically on the client application. In addition, the alerting module creates and broadcasts alerts to operation managers. The alerts include current courier deviation from route that exceeds a predefined distance threshold, and possible future failed deliveries that exceeds the predefined limit. The module also allows the operation managers to provide route updates in case a pick up order comes in or a change in delivery priority occurs. The module would update the route data and provide new statistics to the operation managers.

Guidance. The guidance module delivers the route that needs to be taken by the courier. It includes real-time optimized vehicle routing data in addition to on-foot directions when the courier is in range of the address. The vehicle routing data would be compiled from the package data, operation manager updates, and route data essential for efficient route calculation (ex. traffic, road closures...etc.). The outcomes of this module can be viewed on courier end devices such as PDAs or smart phones.

5 Challenges

The proposed framework faces multiple challenges that need to be individually addressed. Some of the challenges, however, already have proposed solutions through other IoT works though some others do not. The challenges can be divided into the following two main categories:

5.1 Security Challenges

As in all of IoT applications, security is of paramount importance. In the proposed framework, security challenges exist at multiple levels. Security challenges include ensuring that the node remains static, detection of device tampering, information privacy, and user authentication.

5.2 Infrastructure Challenges

As evident through the proposed framework, the amount of data that has to be collected and managed will be significant. This results in challenges for managing the framework, being able to log all the change history, and ensuring that the infrastructure is reliable. Moreover, a question arises about the entity responsible for managing and paying for the framework infrastructure. Processes have to also be identified for items such as change of address, new node installation, and node removal among others.

6 Conclusions

In this paper, an IoT based framework for tracking, monitoring, and managing package deliveries efficiently has been presented. The solution framework architecture and the data analysis methodologies that have been developed have been described in detail. The value of this paper is that it proposes a novel framework for solving this delivery logistics problem. IoT has been considered by much of recent literature as a major contributor to the development of the field of transport logistics [7]. This work contributes to the extant literature on the role IoT in supporting different disciplines, in particular, the logistics function. Thus, this work has both theoretical and practical implications.

For future work, a full evaluation of the proposed framework in partnership with a local delivery courier is planned.

References

1. Bonilla, D.: Urban vans, e-commerce and road freight transport. *Prod. Plann. Control* **27**(6), 433–442 (2016). <http://dx.doi.org/10.1080/09537287.2016.1147093>
2. Boyer, K.K., Prud'homme, A.M., Chung, W.: The last mile challenge: evaluating the effects of customer density and delivery window patterns. *J. Bus. Logistics* **30**(1), 185–201 (2009). <http://dx.doi.org/10.1002/j.2158-1592.2009.tb00104.x>
3. Harris, I., Wang, Y., Wang, H.: ICT in multimodal transport and technological trends: unleashing potential for the future. *Int. J. Prod. Econ.* **159**, 88–103 (2015). <http://www.sciencedirect.com/science/article/pii/S0925527314002837>
4. Lacey, M., Lisachuk, H., Ogura, A., Giannopoulos, A.: Shipping smarter IoT opportunities in transport and logistics. An article in Deloitte's series examining the nature and impact of the Internet of Things (2015)
5. Montoya-Torres, J.R., Muoz-Villamizar, A., Vega-Meja, C.A.: On the impact of collaborative strategies for goods delivery in city logistics. *Prod. Plann. Control* **27**(6), 443–455 (2016). <http://dx.doi.org/10.1080/09537287.2016.1147092>
6. Perego, A., Perotti, S., Mangiaracina, R.: ICT for logistics and freight transportation: a literature review and research agenda. *Int. J. Phys. Distrib. Logistics Manage.* **41**(5), 457–483 (2011)
7. Shakshuki, E.M., Karakostas, B.: The 4th international conference on ambient systems, networks and technologies (ANT 2013), the 3rd international conference on sustainable energy information technology (SEIT-2013) a DNS architecture for the Internet of Things: a case study in transport logistics. *Procedia Comput. Sci.* **19**, 594–601 (2013). <http://www.sciencedirect.com/science/article/pii/S187705091300687X>
8. United States Postal Service: Virtual post office boxes. Technical report, Office of Inspector General United States Postal Service (2013)
9. Universal Postal Union (UPU): Market research on international letters and light-weight parcels and express mail service items. Technical report, Universal Postal Union (UPU) Berne, Switzerland (2010)
10. Universal Postal Union (UPU): Measuring postal e-services development. a global perspective. Technical report, Universal Postal Union (UPU) Berne, Switzerland (2016)
11. Verdouw, C., Robbmond, R., Verwaart, T., Wolfert, J., Beulens, A.: A reference architecture for IoT-based logistic information systems in agri-food supply chains. *Enterp. Inf. Syst.* 1–25. <http://dx.doi.org/10.1080/17517575.2015.1072643>

Context-Aware Security Using Internet of Things Devices

Michal Trnka¹(✉), Martin Tomasek¹, and Tomas Cerny²

¹ Computer Science, FEE, Czech Technical University, Technicka 2, Prague, Czech Republic
{trnkami1, tomasma5}@fel.cvut.cz

² Computer Science, Baylor University, One Bear Place #97356, Waco, TX 76798-7356, USA
Tomas_Cerny@baylor.edu

Abstract. Current trends aim to extend software applications with context-awareness. Nowadays, there are already various approaches enabling security based on context, unfortunately there have limitations. However, the challenging topic is how to obtain as much context information about user as possible. Current progress in Internet of Things domain could be leveraged to obtain more context data. We propose a method to formalize context based on Internet of Things devices and use it for application context-aware security. Our approach is based on composition of a tree topology correlating to the user's devices for recurring situations. Based on changes in the tree we determine unusual behavior, trigger events or invoke specific actions.

1 Introduction

The emerging amount of mobile technologies [4], as well as the growing users' demands for personalized applications provide a base for current trends moving software applications towards context-awareness (CA) [1, 6]. Applications provide personalized content based on user's context or the application's context [5]. This brings novel experience to the applications users. However, securing applications is usually done the traditional way, assigning users various application roles, permissions for resources or security rules independent to the context. There are only few applications having the security based on context information. Nevertheless, we can expect that users and application owners would take the advantage of application security that uses context to provide specific resource control.

Applications using Context-Aware Security [11] (CAS) can be much less obtrusive for users. They can be asked for different authentication methods based on context. The result of the authorization for specific resource may vary depending on their context. For example, access from City A can have different access rights then access from City B. They can even sometimes omit authentication because their context is trustworthy by itself (e.g. access from inner company network). The context even could be created based on devices that a user uses. Each device has unique ID and it communicates with another devices that also communicate with another devices, therefore their interaction and position could be used to create user context. Similar to users, also application operators can profit from the context-based authentication. Different application might define stricter security rules for suspicious users' behavior (e.g. Internet access to

system's confidential resources at night). The usage of context allows system administrators to manage more fine-grained security rules, which would otherwise tangle through multiple rules and make them unsustainable for maintenance. Another advantage is that system may automatically flag suspicious users and prevent them from doing certain actions.

However, the problem is how to obtain context from the user. Some information is obvious for the system (e.g. time, frequency of log-ins, history of application-user communication), other can be guessed but not guaranteed (e.g. geographical location determined from the IP address) while a lot of information are difficult to obtain (e.g. biometric information about the user). All of those mentioned information about the user's context may significantly increase security of the system, while significantly improving the application's user experience. In the following pages we will describe our approach to the issue by involving Internet of Things(IoT) devices to obtain user's context.

This paper is organized as follows: The following section describes related work, followed by our promising solution. The solution is demonstrated in case study section and the paper ends with conclusion remarks.

2 Related Work

Kranz et al. [8] describes the general interaction of the IoT devices with people. It focuses on few use cases with various augmented objects to verify that those areas are suitable for the concept of IoT interaction and that there are benefits. The results of this work indicate that certain areas of the IoT interactions are repeating in all scenarios, while some are unique. Nevertheless, there is no conclusion (or even framework/method proposal) and the paper just states that IoT is promising solution for many areas of human activities.

Petriu et al. [9] discusses possibilities and usage of the sensor-based real-time applications using information from users. They propose multiple communication processes and management system for heterogeneous functions of such system. While there are numerous significant methods and proposals, there is none that would use user's context for application security.

Ho et al. [7] describes framework involving user's context in mobile devices to reduce the amount of communication from different devices. This work focuses more on timing of the messages and their aggregation. It uses innovative ways how to obtains user's context. However, security is not addresses in the paper.

Interesting way how to retrieve user's context is to integrate sensors with items of daily use. Farrington et al. [12] describes the usage of wearables, especially jacket, to retrieve real-time information for context awareness. The methods described in the paper addresses very well context retrieval, but it does not discuss the further usage of the context.

Context-aware security architecture for next generation applications is well described by Covington et al. [10] in his research. It describes all advantages of the context usage as well as its implementation. It only uses basic context that can be

obtained about the particular user through the application. Therefore, the context information is very limited and does not provide the big picture about user.

Another method for context-aware security describes Hu et al. [11]. This work proposes extending the role-based access control [13] model with context aware elements. Similar to beforementioned works, it does not address the issue with retrieving the context from the particular user.

3 Promising Solution

The notion Internet of Things is currently getting a lot of attention and the first real deployments are taking place in real-world scenarios. For instance, Gartner Inc. [3] predicts that by 2020 there will be 26 billion units installed in IoT products. Those devices can provide tremendous amount of information about the user's context. Especially the ones called "wearables". Nevertheless, even other forms of personal IoT devices, like smart homes, could provide us with plentiful of useful and valuable information.

Phone with GPS can provide precious location of its owner. Smart watches can do the same plus they can provide, for example, user's body temperature and pulse. First step of using those biometric information is to use them to form some kind of user's signature. For example, consider a car that would could measure weight and height of the owner. If someone with different body proportions would try to start the car, the car would require additional credentials (e.g. password entered through the entertainment system). This context-aware security system would solve the issue with passive keyless entry or keyless start that are vulnerable for theft [2].

Nevertheless, we can also use additional context data to alter security rules of the system. If we could measure blood pressure and pulse, we could guess the user moods e.g. stressed, angry, etc., and adjust the security of the system corresponding to it. Consider a very critical system, like stock trading or internet banking, if system would determine the user is nervous during performing the transaction with significant and unusual amount of money, it could ask for additional approval. For example, it could ask approval from a second trader or two-phase authorization to prevent wrong decisions based on actual emotions.

We focus on user context that is created based on near by devices to the user. This context helps the system to decide whether it should require additional approval or not. The reason is that most applications signs in or verify the user for the first time and then the session is maintained. An example is the OAuth protocol, when a token is created and assigned to the user. The token has a specific expiration time and when it expires then a new token is created based on the refresh token. However, the user is not asked to log in again. These approaches come with significant issues, for example: "How to decide whether the token was stolen?" or "How to decide whether this is really the user who was authorized in the first place?". The system could open sign in dialog and ask for username and password again or the system should use the two-factor verification (explained in case study), but this process should be initialized based on clues that alert the system. These clues might be user interaction with the system, or device that is used

or others devices that might be not directly involved in the interaction process between the device and software. The combination between the user interaction and IoT devices, that represent the indirect devices, are great choice for this type of situation.

We may observe IoT from several perspectives. We can focus on device itself or we can monitor the users, because every person has a specific set of behaviors and most of them has predictable time schedule. For example, the Google is able to decide where you work, where you park your car, etc. Based on that Google provides you morning traffic information and travel time estimation to the job. Your secretary knows when you usually come to work and what is your preferable restaurant, as well as she knows which car you use and what are your favorite hobbies, moreover she recognizes some of your friends. This implies the following: If somebody asks your secretary what are you doing in concrete time then she is able to predict what you are actually doing, because she knows you. In this section, we present a technique that helps machines to know you and based on your habits determine whether you should do additional verification when you want to use a specific part of software.

Unlike secretaries the IoT is not a human being and it does not pose prediction logic, but on the other hand it has access to sensors and devices on different places at same time. Your computer is connected to network via cable or Wi-Fi, therefore there is a specific device near to your location. The same applies for smart watch, fit bracer or another wearables device. When you are in a car then your mobile phone is connected with car via Bluetooth. Given the nature of the IoT, we can even use information from devices that are not connected directly to the user, but to one of his primary devices. We consider only devices that are connected with each other and we do not consider unconnected devices, because it is out of scope this paper.

Our solution represents connected devices as undirected graph. The edges connect two devices that interact with each other. There exist specific graphs for different situations. This means that the graph for office is different from a graph when the user is at home or when she or he moves from office to home in a car. The graph is also different

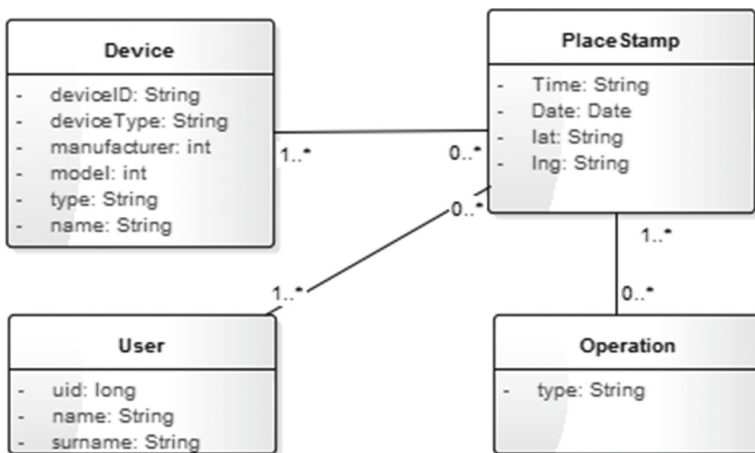


Fig. 1. Data structure to hold nearby devices

when the user practices any type of sport. The graph is created for the specific time and place, therefore it contains devices that the user usually uses in concrete time and place. Besides of graph we need meta-information of the graph. We created basic data structure that holds these types of information. It is represented in the Fig. 1. The data structure holds information about the device itself, time with place where it is used and the person who uses it.

The principle of this solution is to store data anytime when the user uses application. The user sends information about device that is used and he/she also sends information about other devices which are connected like smart watch, car Bluetooth or Wi-Fi. When the system has these types of information, then it is able to decide whether it is user's regular environment or it is not. The system stores signification amount of data, therefore it is critical to be aware of the time frame and aggregate data based on it. The basic aggregation and usage of this solution is demonstrated in a case study section; the detail information about how to implement this aggregation and how to make decision about user's confidentiality is matter of implementation and it should base on security rules in concrete usage.

4 Case Study

The proposed solution is demonstrated in the case study. We have chosen bank environment. The bank clients usually use internet banking. The software's task is to manage bank accounts. The user can work with transaction history, accounts, credit cards, loans, mortgages etc. For example, the user can change name of the accounts, create standing order, make request to offer, create payment, etc. The supported functionalities are different in each bank institution. However, most banks have one common functionality. It is two-factor verification when the user wants to create a payment. The payment could be created only by authorized user who is logged into internet banking. There exists a lot of login method for example: the user uses certificate with password or she/he uses login and password or combination with login, password and SMS authentication. We will consider only authenticated users and we will focus on the process of creation of a payment.

The process itself involves a lot of actions and preconditions. The user must be logged into internet banking, choose the source account, know the destination account, enter an amount and other details and finally, confirm the payment. There are a lot of processes that are triggered after the payment confirmation. The bank system has to verify whether it allows the user transfer the given amount from the source bank account, whether it is normal or suspicions operation and needs to authenticate the requesting user. The SMS two-factor verification is used to check the user's identity. This approach has various disadvantages. The user's identity and mobile phone could be stolen, unreachable, broken, or the provider is unreachable.

We simulated the data that could be obtained during the process in the Table 1. We store data from computer. The computer is connected via cable or Wi-Fi to Internet and it is also connected with mobile phone via Bluetooth.

Table 1. Examples of harvested data

DeviceID	Time	Lat	Lng	Operation	Source	User
Computer1	15:30:29	49.224	16.577	Login	Yes	mtomasek
Router1	15:30:29	49.224	16.577	NONE	No	mtomasek
Phone1	15:30:29	49.224	16.577	NONE	No	mtomasek
Computer1	15:32:15	49.224	16.577	Payment create	Yes	mtomasek
Router1	15:32:15	49.224	16.577	NONE	No	mtomasek
Phone1	15:32:15	49.224	16.577	NONE	No	mtomasek
Computer1	15:32:49	49.224	16.577	SMS verification	Yes	mtomasek
Router1	15:32:49	49.224	16.57	NONE	No	mtomasek
Phone1	15:32:49	49.224	16.577	NONE	No	mtomasek
Computer2	18:10:35	50.075	14.419	Login	Yes	mtrnka
Router2	18:10:35	50.075	14.419	NONE	No	mtrnka
Phone1	18:10:35	50.075	14.419	NONE	No	mtrnka
Computer1	15:25:58	49.228	16.577	Login	Yes	mtomasek
Router1	15:25:58	49.228	16.577	NONE	No	mtomasek
Phone1	15:25:58	49.228	16.577	NONE	No	mtomasek
Computer1	15:29:38	49.228	16.577	Payment create	Yes	mtomasek
Router1	15:29:38	49.228	16.577	NONE	No	mtomasek
Phone1	15:29:38	49.228	16.577	NONE	No	mtomasek

The table contains various information. It shows which device was used to access the network, the place where the user is and another device that she/he uses. The column Source indicates if the device is source of information that are in the table. The table is ordered based on date therefore, first twelve records are stored one day and the rest records are stored another day. The devices are the same when logging in, making the payment and verifying the payment. This is an initialization state and we require SMS verification in this state, because we do not know the user's behavior and environment. If the user creates another payment next day around 15:30 then we can compare connected devices with previous state in which was payment authorized. Moreover, we can compare place where the user is and if the place is the same, but the devices are different then the payment could be suspicious. We are able to create graph of devices that the user usually uses and their place in time. We can store any user who uses our application, therefore we can connect it together and compare their location in time to verify if their time schedule is usual.

The table represents another state. The second user (mtrnka) logged into internet banking at 6:10 PM. This is nothing special, however his computer is connected with the Phone1 that was used by another user mtomasek. If the phone is the authorization phone that belongs to mtomasek, then every payment creation by mtomasek should be suspicious. There is another case. It is the payment creation next day. We can compare connected devices, place and time and we can decide if the two-factor verification is

necessary or not. In this case, we might not to send the SMS on target device, because the phone is already somewhere around the computer, the user uses the same device, he is almost on the same position and he does the same action that he did yesterday in this time frame. We also could use these information as a fraud indicators and we can decide to use a different authorization method or to ban this transaction.

5 Conclusion

We presented an approach that targets the user rather than the system itself or actions in the system. Information from the user's nearby devices are used to obtain user context. The user's position, date, time and nearby devices by itself are critical parts of our method. The information is kept for future usage. When any decision about user behavior is needed, we can correlate current data with the historical data and tell whether the security rules should be altered. The basic data structure was presented in addition to the usage in an internet banking use case. We showed that our approach helps system to decide on the additional level of authorization necessity when the user's context is suspicious or unusual.

In future we would like to focus on the human health sensors. The sensor provides crucial data about the user, such as weight, hearth beat rhythm, etc. These data combined with nearby devices could provide more detailed information about user's context. Based on our approach we could decide more precisely whether the user exhibits some suspicious behavior. Integrating machine learning techniques in our decision scheme another direction we like to explore.

Acknowledgement. Research described in the paper was supported by the Grant Agency of the Czech Technical University in Prague, under grant No. SGS16/234/OHK3/3T/13 and by Technology Agency of the Czech Republic, under grant No. TH02010296.

References

1. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggle, P.: Towards a better understanding of context and context-awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999). doi:[10.1007/3-540-48157-5_29](https://doi.org/10.1007/3-540-48157-5_29)
2. Francillon, A., Danev, B., Capkun, S.: Relay attacks on passive keyless entry and start systems in modern cars. In: NDSS 2011 (2011)
3. Gartner Inc.: Hype Cycle for the Internet of Things (2013)
4. Harter, A., Hopper, A., Steggle, P., Ward, A., Webster, P.: The anatomy of a context-aware application. *Wirel. Netw.* **8**(2/3), 187–197 (2002)
5. Hong, J., Suh, E.-H., Kim, J., Kim, S.: Context-aware system for proactive personalized service based on context history. *Expert Syst. Appl.* **36**(4), 7448–7457 (2009)
6. Miroslav, M., Cerny, T., Slavik, P.: Context-sensitive, cross-platform user interface generation. *J. Multimodal User Interfaces* **8**(2), 217–229 (2014)
7. Ho, J., Intille, S.S.: Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2005), pp. 909–918. ACM, New York (2005)

8. Kranz, M., Holleis, P., Schmidt, A.: Embedded interaction: interacting with the internet of things. *IEEE Internet Comput.* **14**(2), 46–53 (2010)
9. Petriu, E.M., Georganas, N.D., Petriu, D.C., Makrakis, D., Groza, V.Z.: Sensor-based information appliances. *IEEE Instrum. Meas. Mag.* **3**(4), 31–35 (2000)
10. Covington, M.J., Fogla, P., Zhan, Z., Ahamad, M.: A context-aware security architecture for emerging applications. In: *Proceedings of the 18th Annual Computer Security Applications Conference*, pp. 249–258 (2002)
11. Trnka, M., Cerny, T.: On security level usage in context-aware role-based access control. In: *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC 2016)*. ACM, New York (2016)
12. Farrington, J., Moore, A.J., Tilbury, N., Church, J., Biemond, P.D.: Wearable sensor badge and sensor jacket for context awareness. In: *Proceedings of the Third International Symposium on Wearable Computers, Digest of Papers, San Francisco*, pp. 107–113 (1999)
13. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications (WMCSA 1994)*, pp. 85–90. IEEE (1994)

An Energy-Efficient Transmission Framework for IoT Monitoring Systems in Precision Agriculture

Peerapak Lerdsuwan and Phond Phunchongharn^(✉)

Department of Computer Engineering, Theoretical and Computational Science Center, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
{peerapak.l, phond.p}@mail.kmutt.ac.th

Abstract. Internet of Thing (IoT) technology has enabled efficient crop monitoring to support decision making in precision agriculture. The monitoring system collects environmental data in fields. A major challenge in the monitoring system is limited energy power of IoT sensor nodes. Consequently, we propose an energy-efficient transmission framework for IoT sensors in the monitoring system. Our proposed framework allows the sensor nodes adaptively collecting the data upon the environmental change. Furthermore, we propose an energy-efficient transmission algorithm for the proposed framework. The objective is to minimize the energy power at the sensor nodes while guaranteeing the transmission rate. A data-driven algorithm based on a greedy method is used to solve the problem with low complexity. We compare the performance of our algorithm with two traditional transmission protocols, called SPIN and ESPIN, through an experiment. From the results, our algorithm can provide better energy efficiency about 81.53% than SPIN and 36.84% than ESPIN.

Keywords: Energy efficiency · Internet of Thing · IoT sensor network · Monitoring system · Precision farming

1 Introduction

The development of agriculture is important for economic development in many countries, especially, those in the Southeast Asia. To improve the crop productivity, a monitoring system is introduced to apply in a farm field in order to collect the information of farm conditions (e.g., light intensity, humidity and temperature). This information can be later used in precision agriculture for improving crop productivity.

Nowadays, Internet of Things (IoT) technology has become more popular to employ in various fields, especially, in monitoring systems for agriculture [1, 2]. In [1], authors proposed an IoT as a monitoring system to sense soil moisture conductive for irrigation management. Furthermore, authors in [2] monitored the environmental data (e.g., temperature, carbon dioxide and light intensity) in a greenhouse by using an IoT technology. As a result, the operational efficiency could be improved. However, a monitoring system consists of several

sensor nodes which communicate together as a network, also called a sensor network. Although the sensor network can be both wired and wireless, wireless sensor networks are favor to used in IoT since the networks support mobility and easy to change the network structure. Particularly, the wireless sensor networks offer more advantage when they come to difficult-to-wire areas (e.g., across a river or farm fields that are physically separated but operate as one). One of the major challenges in the wireless sensor networks of IoT monitoring system is how to efficiently utilize energy in the network.

Recently, there are several researches about transmission protocols for IoT sensor networks in monitoring systems [3,4]. [3] introduced a traditional routing protocol called Sensor Protocol for Information via Negotiation (SPIN) which floods a negotiation message such as current resources to neighbor sensor nodes before performing a data transmission. SPIN can conserve the energy by choosing a resource-efficient route which calculate from the negotiation messages. On the other hand, the flood information will dissipate much energy. Therefore, authors in [4] illustrated an enhanced SPIN called Energy-efficient Sensor Protocol for Information via Negotiation (ESPIN) with the purpose of reducing redundant data and improving the network performance as well as decreasing energy consumption of the whole network. Although, ESPIN can decrease some consuming energy, the overall consuming energy is still high due to the use of multicast in the data transmission phase.

In this paper, we focus on the energy-efficient data transmission algorithm for an IoT sensor network in an IoT monitoring system. The proposed algorithm is divided into two main steps. The first step is data selection. Since each data transmission consumes most energy power of the sensor node, only useful information should be selected to transmit. The second step is energy-efficient data transmission. All the selected data will be transmitted by using our proposed data-driven transmission protocol. The objective of the algorithm is to find an optimal route for each sensor node to transmit the collected data to the server with lower transmission energy while the overall sensor node throughput is guaranteed. Finally, we evaluate and compare performance of our proposed algorithm with existing algorithms [3,4] by using an experiment.

2 A Data Collection Framework for an Energy-Efficient Monitoring System

To utilize the energy efficiently, the sensor nodes should be able to capture the important data adaptively to the change of environmental conditions. For example, a soil humidity sensor must work more frequently when the crop gets watering or raining while the sensor will rarely work when it is sunny. We consider that each sensor node has only one antenna. Therefore, it can either receive or transmit data at a time. Five important modes of a sensor node are (i) listening, (ii) collecting data, (iii) transmitting data, (iv) sleep, and (v) idle mode as shown in Fig. 1. The detail of each mode in the monitoring process is as follows:

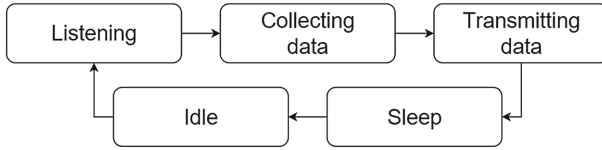


Fig. 1. Our proposed data collection framework for a monitoring system

- **Listening Mode:** Since the sensor nodes are sparsely located in a crop field, some sensor nodes cannot directly reach the server. These nodes need to relay the data through their neighbor nodes. As shown in Fig. 2, node 3, 4, and 5 must relay their data through either node 1 or 2. To successfully collect data from all the nodes, every node must start with this mode in order to help relaying data for their neighbors.
- **Collecting Data Mode:** Each sensor node can compose of various types of sensors (e.g., air temperature, air humidity, soil moisture, and light intensity). After the farm conditions are sensed, the data will be kept in the buffer before transmitting to the server. From our experiment, the data transmission consumes the highest energy power compared to other activities. To reduce the energy usage, only useful data should be transmitted to the server. Also, the buffer size of a sensor node is limited. Consequently, only useful data will be kept in the buffer. Otherwise, it will be removed.
- **Transmitting Data Mode:** When there is the sensing data in the buffer, the sensor nodes will try to transmit the data to the server. The detail of the proposed algorithm is presented in Sect. 3.
- **Sleep Mode:** Since the environmental conditions are slowly changed in most of the cases, the sensed data is slightly different from the recent sensed data. To save the energy power, the sensor nodes can fall asleep for a while. In this mode, the sensor node will disable communication ports and unnecessary operations.
- **Idle Mode:** After waking up from the sleep mode, the sensor node will enter to idle mode in order to set up the buffer, input pin, output pin, and other components in the sensor node being ready for working in other modes. The duration in this mode is less than 2s.

3 The Data-Driven Transmission Algorithm

3.1 Problem Formulation

In this paper, we focus on energy-efficient data transmission in an IoT sensor network for a farm monitoring system. The monitoring system consists of N IoT sensor nodes (as shown in Fig. 2).

From our experiment, the highest energy power is utilized for data transmission. Consequently, we focus on the energy efficiency of the data transmission in the monitoring system.

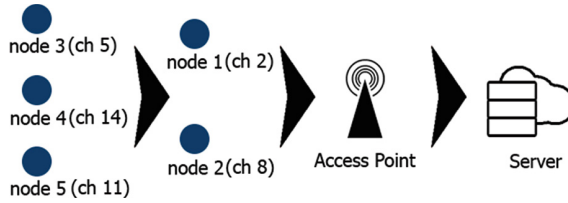


Fig. 2. An example of IoT sensor topologies in a monitoring system

The objective of our data transmission algorithm is to minimize the overall consuming energy power while the achievable data rate is guaranteed. Consequently, the optimization problem can be formulated as:

$$\min \sum_{n=1}^N \frac{1}{\mu_n} \tag{1}$$

$$\text{subject to } r_n \geq \tau_n, \forall n \in \mathcal{N} \tag{2}$$

where μ_n is the energy efficiency at node n . μ_n can be defined by $\frac{r_n}{E_n}$ where r_n is achievable bit rate of node n (bits/s) and E_n is electric energy at node n (J). τ_n is the threshold of the data transmission rate at node n that makes the buffer at node n not overflow. Note that this value can be calculated from the arrival rate of the sensing data, the buffer size, and the data removal rate (i.e., the rate that insignificant sensing data is removed from the buffer). \mathcal{N} is the set of sensor nodes in the monitoring system.

3.2 The Proposed Algorithm

Due to the complexity of the problem shown in (1)–(2), we propose a data-driven algorithm based on a greedy method. Since the buffer size is limited, when the data is available in the buffer, the sensor node will try to transmit the data as soon as possible to avoid the buffer overflow. This is so called data-driven algorithm.

The complexity of selecting the best route for each sensor node is an NP problem. This causes the long computation time. Consequently, we use a greedy method for selecting the route. Although the greedy method cannot provide the optimal solution, it can provide the approximate value close to the optimal value with low complexity. As a result, our algorithm can quickly adapt to the change in the monitoring system.

From the Shannon’s equation, the maximum bit rate depends on signal-to-noise ratio (SNR) as shown in Eq. (3).

$$B_{rate} = B_w \log_2(1 + SNR) \tag{3}$$

where B_{rate} (bits/s) is proportional to the bandwidth of specific channel B_w (Hz). The SNR can be obtained from the received signal strength indicator (RSSI).

From Eq. (3), an IoT sensor node will select the best signal route at the time to transmit the data in order to achieve the best transmission rate. The overall process of our algorithm is illustrated in Fig. 3. The proposed algorithm is executed at the sensor nodes as a distributed manner.

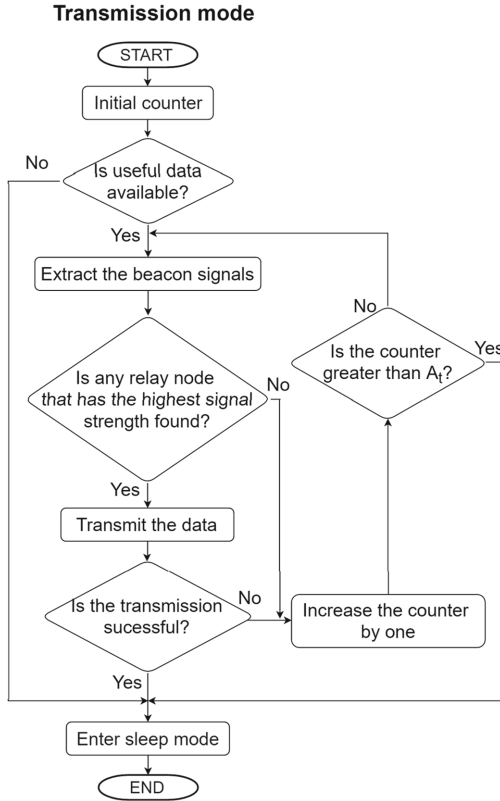


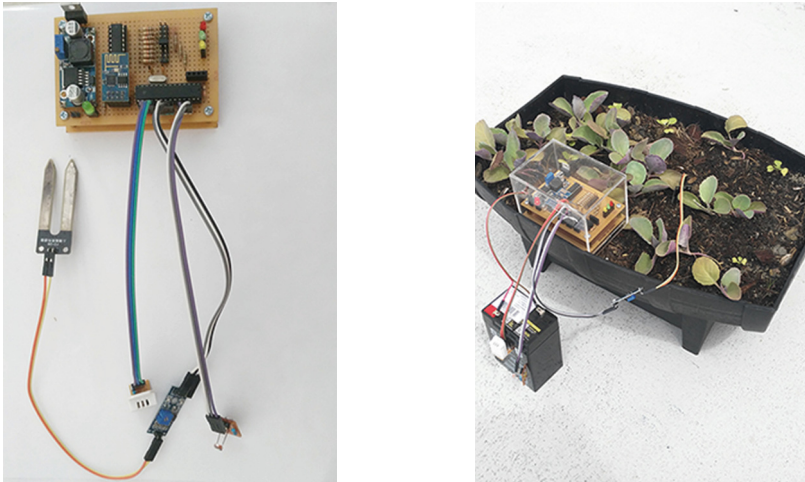
Fig. 3. The proposed transmission algorithm

When the useful data is available in the buffer of a sensor node, the node will find its neighbor nodes and evaluate their RSSIs by extracting and considering the received beacon signals. Then, the sensor node will select the neighbor node with the highest RSSI and closer to the sink node than itself as a relay node. Then, the sensor node transmits data to the selected relay node. However, the transmission can fail if SNR at the relay node is lower than a threshold. For example, there exists a simultaneous transmission from other nodes to the same relay node or there exists high interference at the relay node. If the transmission is unsuccessful, the sensor node will repeat the steps to find the new relay node. To avoid the wasting energy power with unsuccessful transmissions, we set up the maximum number of attempting transmission A_t . If the number of

contiguous unsuccessful transmissions is greater than A_t , the sensor node will stop transmitting data to the relay node and then go to the sleep mode.

4 Experimental Results

To evaluate our proposed framework, we set up an experiment by using Atmega328p micro-controller unit (MCU) [5] based on Arduino technology and ESP8266 WiFi [6] module to create IoT sensor nodes. Each node consists of a temperature and humidity sensor, a soil humidity sensor, and a light intensity sensor as shown in Fig. 4a and b. We deploy five sensor nodes in a crop field with one access point to transmit data to our remote server. A layer topology is used to locate our sensor nodes as shown in Fig. 2. In the experiment, we set listen duration and sleep duration as 120 and 600 s, respectively. The maximum number of attempting transmission (A_t) is set to 5. The total experimental duration is 8 h.



(a) IoT sensor node (b) Sensor node in farm field

Fig. 4. Sensor node in actual experiment

We compare our proposed transmission algorithm with SPIN and ESPIN in three performance metrics which are the average duration time until the transmission successes, the total number of successful transmission bits, and the average energy consumption for every 10 min. Figure 5 shows the average transmission time until the transmission successes for each protocol. We can see that our proposed algorithm can spent less transmission time than other algorithms. Also, our proposed algorithm can achieve the highest number of successful transmission bits as shown in Fig. 6.

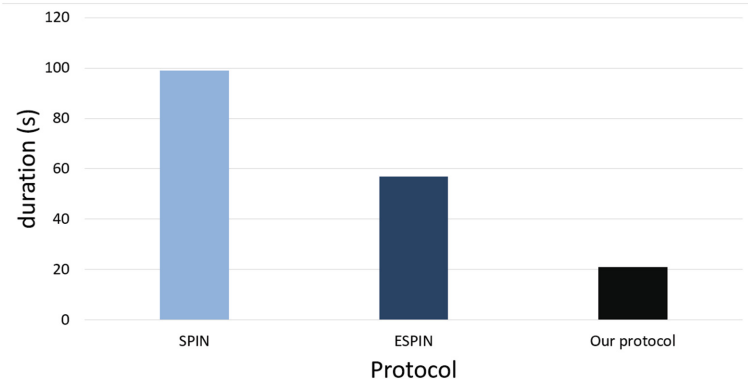


Fig. 5. The average duration time until the transmission successes

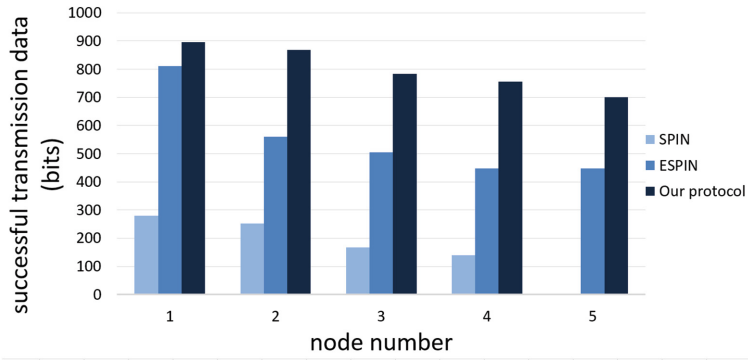


Fig. 6. The total number of successful transmission bits

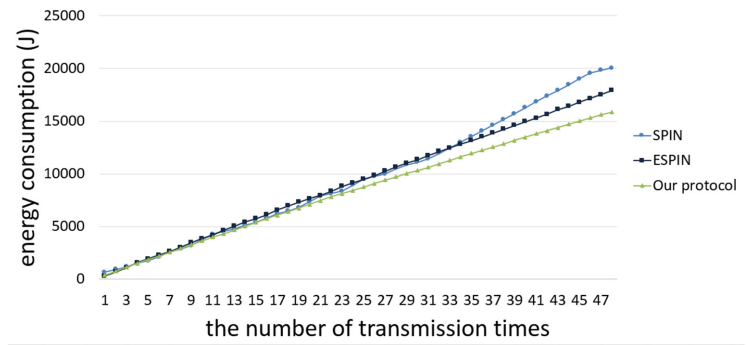


Fig. 7. The average energy consumption for every 10 min

Finally, we illustrated the average energy consumption for every 10 min over the overall experimental duration (as shown in Fig. 7). Our algorithm consumes energy lower than SPIN and ESPIN about 20.81% and 11.34%, respectively. Consequently, our protocol can outperform SPIN and ESPIN in term of energy efficiency about 81.53% and 36.84% for SPIN and ESPIN, respectively. This results from that our proposed algorithm does not need to multicast an advertisement to many other nodes every time before transmitting the data as they do in SPIN and ESPIN. Moreover, SPIN and ESPIN must wait for the request messages from the relay nodes before starting the transmission. This process wastes energy consumption and provides the big overhead in the transmission process.

5 Conclusion

This paper has proposed an energy-efficient transmission framework for an IoT monitoring system in a precision farming. The proposed framework consists of five modes which are listening, collecting data, transmitting data, sleep, and idle mode. For each mode, we focus on energy efficiency so that the energy power for the overall monitoring process is efficiently used. We have also proposed a data-driven transmission algorithm based on a greedy method to employ in the transmitting data mode of our proposed framework. From the experiment, the results have revealed that our proposed algorithm can achieve higher energy efficiency than SPIN and ESPIN protocols about 81.53% and 36.84%, respectively.

Acknowledgement. This work is supported by the Thailand Research Fund (TRF), under Grant No. TRG5780059 and the Higher Education Research Promotion and National Research University Project of Thailand (NRU), under Grant No. 59000399.

References

1. Baranwal, T., et al.: Development of IoT based smart security and monitoring devices for agriculture. In: Proceedings of the 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, pp. 597–602 (2016)
2. Dan, L., et al.: Intelligent agriculture greenhouse environment monitoring system based on IOT technology. In: Proceedings of International Conference on Intelligent Transportation, Big Data and Smart City, Halong Bay, pp. 487–490 (2015)
3. Pattani, K.M., Chauhan, P.J.: SPIN protocol for wireless sensor network. *Int. J. Adv. Res. Eng. Sci. Technol. (IJAREST)* **2**, 2394–2444 (2015)
4. Li, J., Shen, C.: An energy conservative wireless sensor networks approach for precision agriculture. *Electronics* **3**, 387–399 (2013)
5. ATmega328P. <http://www.atmel.com/devices/atmega328p.aspx>
6. ESP8266 Datasheet. <http://espressif.com/en/support/download/documents>

Piezoelectric Voltage Monitoring System Using Smartphone

Nazatul Shiema Moh Nazar^(✉), Suresh Thanakodi, Azizi Miskon,
Siti Nooraya Mohd Tawil, and Muhammad Syafiq Najmi Mazlan

Department of Electrical and Electronic Engineering, Faculty of Engineering,
National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
{nazatul.shima, suresh, azizimiskon, nooraya}@upnm.edu.my,
alongkp07@gmail.com

Abstract. This paper proposed to develop the voltage monitoring for piezoelectric system. The piezoelectric wireless monitoring system will enable voltage monitoring by utilizing a smartphone, piezoelectric sensor and Bluetooth to a device that installed with designated application. The Bluetooth system is the method used to connect the piezoelectric sensor and the smartphone. Thus, this research is aimed to monitor the voltage produced by the piezoelectric system wirelessly. The produced data can be monitored in real time as well as being extracted in excel data format for recording purpose. In the previous research, the piezoelectric were embedded in army boots for energy scavenging purpose to charge hand phone. Monitoring the voltage output utilizing multimeter not feasible at all, hence this research solves the challenges of monitoring the piezoelectric voltage output.

1 Introduction

The discovery and usage of piezoelectric materials dates back to the year 1880, when Curie brothers demonstrated that certain materials such as quartz, tourmaline and topaz, upon application of stress, exhibits accumulation of charges. This effect was later termed as “Piezoelectricity” [1]. In recent years, Wireless Sensor Network (WSN) has had a large increase in real applications which is main advantage over other peer technologies [2].

A wireless communication technology that provides semi-autonomous radio network connection, short range and to establish an ad hoc network. The most significant part of this project is the selection of the most developed monitoring system possible when analysing the data during experiment to ensure accurate results. In aspect of voltage monitoring system via smartphone, any kinetic energy exerted to the piezoelectric sensor will transfer the sensing result to the mobile phone through wireless transmission, namely the Bluetooth.

This research represents the significance of android phone application system development capable of monitoring the voltage reading produced by the piezoelectric sensor. This research uses sensors & smart phone technology to monitor wirelessly.

1.1 Smartphone

A mobile phone with an advanced mobile operating system which combines the useful features for mobile or handheld use and also features of a personal computer operating system with another is called smart phone. Usually smartphone has a high-resolution touch screen display, Web browsing capabilities, Wi-Fi connectivity and the ability to accept sophisticated applications [3]. A Smartphone is expected to have more storage space, more powerful CPU, larger RAM, greater connectivity options and larger screen compared to a regular cell phone.

1.2 Operating System

A Smartphones are operating with a mobile operating system that operate small handheld devices, have become an integral part of our lives [4]. To run the program, operating system software or 'OS' is used to communicate with its hardware [5]. System software, or the fundamental files your computer needs is the main element to boot up and function. An operating system provides basic functionality for the device, such as smartphones, desktop computer and tablet.

Android is one of the operating system developed by Google available for Smartphone. In 2007, android was unveiled along with the founding of the Open Handset Alliance (Google, HTC, Sony, Intel, Qualcomm) – a consortium of software, hardware, and telecommunication companies devoted to advancing open standards for mobile devices.

1.3 PIC Microcontroller

This research used PIC 16F767 as the main microcontroller to detect the voltage of piezoelectric and send to smartphone via Bluetooth. The data sheet of the PIC microcontroller gives detail about the microcontroller. There are two types of 28-pin microcontroller which is PIC16F737 and PIC16F767. There are several differences among these two PICs although these PICs have a common architecture.

PIC16F737 and PIC16F767 devices are 28-pin packages, whereas PIC16F747 and PIC16F777 devices are 40-pin and 44-pin packages respectively. The common architecture of PIC16F7X7 devices are same. Though with the same architecture, there are several differences between them.

- The total on-chip memory of the PIC16F747 and PIC16F777 are more than PIC16F737 and PIC16F767 by one-half.
- The 40/44-pin devices have five I/O ports, whereas 28-pin devices have three I/O ports.
- The 40/44-pin devices have 17 interrupts whereas 28-pin devices have 16 interrupts.
- The 40/44-pin devices have 14 A/D input channels, whereas 28-pin devices have 11 A/D input channels [5].

Figure 1 shows the PIC16F737/767 pin configuration and Table 1 shows the feature differences between PIC16F737 and PIC16F767.

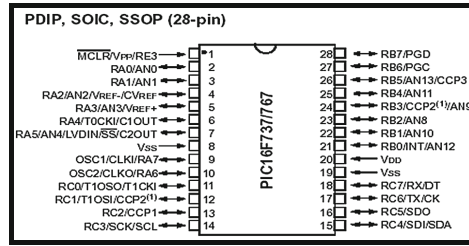


Fig. 1. PIC16F737/767 pin configuration

Table 1. Features of PIC16F737 and PIC16F767 [5]

Key features	PIC16F737	PIC16F767
Operating frequency	DC–20 MHz	DC–20 MHz
Resets and delays	POR,BOR(PWRT,OST)	POR,BOR(PWRT,OST)
Flash program memory (14-bit words)	4 K	8 K
Data memory (bytes)	368	368
Interrupts	16	16
I/O Ports	Ports A, B, C	Ports A, B, C
Timers	3	3
Capture/Compare/PWM Modules	3	3
Master serial communications	MSSP, AUSART	MSSP, AUSART
Parallel communications	–	–
10-bit Analog-to-Digital Module	11 Input Channels	11 Input Channels
Instruction set	35 Instructions	35 Instructions
Packaging	28-pin PDIP 28-pin SOIC 28-pin SSOP 28-pin QFN	28-pin PDIP 28-pin SOIC 28-pin SSOP 28-pin QFN

1.4 Bluetooth

A global wireless communication standard is using Bluetooth technology that connects devices together over a certain distance [6]. Radio waves are used instead of cables or wires to connect to a computer or phone to a Bluetooth device. It needs to pair each other when two Bluetooth devices want to interconnect. Short-wavelength UHF radio waves in the ISM band of 2.4 GHz is used to communicate between Bluetooth devices over a short-range. This is how mobile phones, computers and personal digital assistants (PDAs) can be easily interconnected using a short-range wireless connection [7].

2 Methodology

2.1 Overall Process and Components

The system consists of a PIC16F767 microcontroller on a circuit board with App Link Bluetooth module, piezoelectric voltage detector and an Android smartphone with Magnetcode application. All of this is the main components that assembled and ensured the objective of the research was achieved.

Basically the research works with a device assembled with a PIC16F767, App Link Bluetooth module and a Piezoelectric on a circuit board which was attached to an individual and connected wirelessly by Bluetooth connection to an Android smartphone. The android smartphone can orientate the data of the device by installing an application known as Magnetcode and programming the PICF767 to detect the output voltage produce by piezoelectric at the certain time. Figure 2 shows the flowchart system for this research.

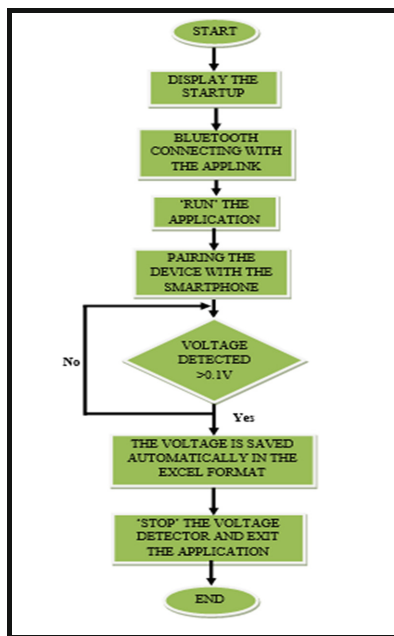


Fig. 2. Flowchart of the system

2.2 Assembling

This research required a software and hardware in the process of detecting the output voltage from the piezoelectric via Bluetooth. In the software, assembling are using PIC C Compiler to construct the program logic and PICKit 2 v2.55 to upload the coding into PIC16F767 microcontroller. The logic of the program can be shown in Fig. 3.

```

project16.c
1 #include <16f767.h> //use pic16f767
2 #define adc10 //use 10bit adc
3 #use delay(clock=6000000) //6mhz
4 #fuses INTRC_IO,noprotect,nowdt //fuse setting
5 #use rs232(baud=9600, xmit=PIN_C6, rcv=PIN_C7, parity=N) //rs232 setting
6
7 float bat;
8 int sta=1;
9 int run=1;
10 int data1, data2, data3;
11
12 //receive bluetooth data
13 #int_rda
14 void serial_isr()
15 {
16     data1=getch();
17     data2=getch();
18     data3=getch();
19     if (data3==0x0A)
20     {
21         run=2;
22     }
23 }
24
25 void main()
26 {
27     //set i/o for each pin
28     set_tris_a(0b00111111);
29     set_tris_b(0b11100000);
30     set_tris_c(0b10000000);
31     setup_oscillator(OSC_SHM5);
32     setup_port_a(AN0 TO AN4);

```

Fig. 3. The logic program coding using PIC C compiler

The logic program in Fig. 3 are installed into the PIC16F767 microcontroller using PICKIT 2 v2.55 software. The installation of the program is using PIC adapter. Thus, Fig. 4 shows the coding flowchart for this research.

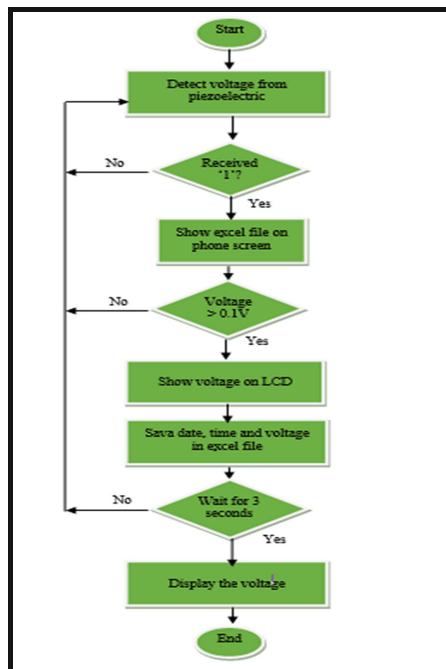


Fig. 4. The coding flowchart of the program

The assembling of hardware on this research which is fitted onto the main board includes the PIC16F767 microcontroller itself has shown in Fig. 5. In order to operate

the system, a 240 V power adapters was used to the main board. Figures 6 and 7 shows the connection and the pin configuration of the main board respectively.

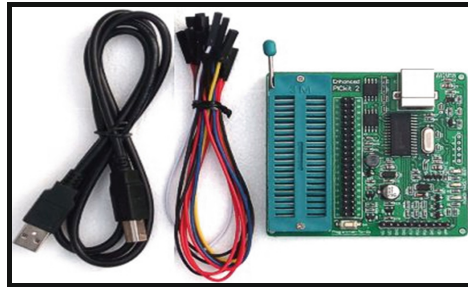


Fig. 5. PIC adapter

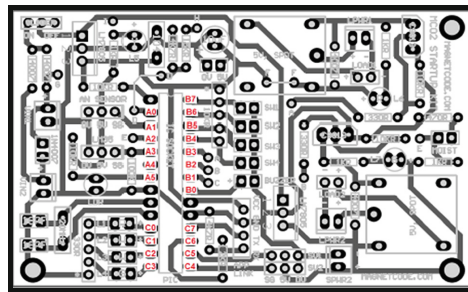


Fig. 6. The connection of the main board

PIN	I/O Status	Connect To	Function
A0	Analog Input	Voltage sensor	Detect voltage at point VINx (0V – 20.98V)
A1	Analog Input	Variable resistor, infrared sensor, Light dependent resistor (modified PCB)	Detect analog sensor value (0V-5V)
A2	Analog Input	Moisture sensor	Detect analog sensor value (0V-5V)
A3	Analog Input	Temperature sensor, Light dependent resistor (modified PCB)	Detect analog sensor value (0V-5V)
A4	X	X	X
A5	Analog Input	Light dependent resistor	Detect brightness (0V-5V)
B0	Digital Output	Buzzer	On / Off buzzer
B1	X	X	X
B2	Digital Output	Relay 1	On / Off AC / DC heavy duty load
B3	Digital Output	Relay 2	On / Off AC / DC heavy duty load
B4	Digital Input	Switch 4 (SW4)	Detect switch is open (5V) or close (0V)
B5	Digital Input	Switch 3 (SW3)	Detect switch is open (5V) or close (0V)
B6	Digital Input	Switch 2 (SW2)	Detect switch is open (5V) or close (0V)
B7	Digital Input	Switch 1 (SW1)	Detect switch is open (5V) or close (0V)
C0	Digital Output	LED (L1)	On / Off LED
C1	Digital Output	LED (L2)	On / Off LED
C2	Digital Output	LED (L3)	On / Off LED
C3	Digital Output	LED (L4)	On / Off LED
C4	Digital Output	Servo motor 2 (SV2)	Control servo motor angle (0 – 180)
C5	Digital Output	Servo motor 1 (SV1)	Control servo motor angle (0 – 180)
C6	Serial Output	AppLink Bluetooth module TX pin	Send data to Smartphone (9600bps)
C7	Serial Input	AppLink Bluetooth module TX pin	Receive data from Smartphone (9600bps)

Fig. 7. The pin configuration of the mainboard

2.3 Execution

Figure 8 shows the hardware for the piezoelectric voltage monitoring system consists of piezoelectric, monitoring system and android smartphone. The input of this system was the vibration from the piezoelectric. After that, the signal has been processed to the

monitoring system and the monitoring system has been sent the data to the Android smartphone. The data that have been appearing has voltage readings by the smartphone and appear in Microsoft Excel spreadsheet format.

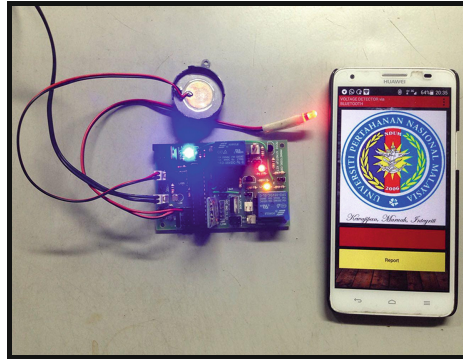


Fig. 8. Piezoelectric voltage monitoring system without casing

3 Analysis

3.1 Tabulation of Data

Table 2 depicts the voltage reading using the Applink Bluetooth and Multimeter for this research. This research has been taken for three readings for each device.

Table 2. Reading of voltmeter using applink bluetooth and multimeter

Reading	Types of devices	Voltage displayed (V)	Time taken for the voltage to return 0 V(ms)
Reading 1	Applink Bluetooth	0.13	300
	Multimeter	0.18	26
Reading 2	Applink Bluetooth	0.14	300
	Multimeter	0.12	30
Reading 3	Applink Bluetooth	0.16	300
	Multimeter	0.14	66

3.2 Analysis of Data

From the tabulated data, the Applink Bluetooth was compared to other monitoring device that is Multimeter. There are two significant differences that can be seen between the devices which are the time taken for the voltage to return to 0 V and the value of voltage output displayed when the piezoelectric applied with the mechanical pressure.

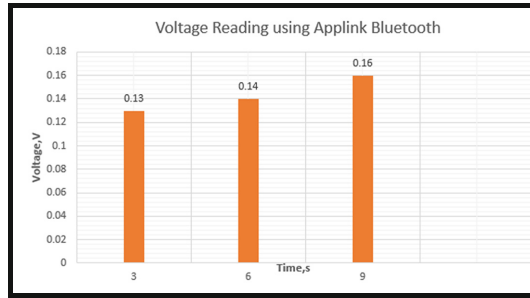


Fig. 9. The reading of voltmeter using applink bluetooth

Firstly, the time taken for the voltage to return to 0 V. The time taken for the Applink Bluetooth to display the voltage output has been set in the programming. The delay time is being set to 3 s. It provides time to the person to monitor the voltage output. Besides that, the values are directly recorded in the Excel format with the exact date and time. Compared with the Multimeter, the time taken to display the value of the voltages are very fast which is in millisecond(ms) as shown in Fig. 10. Thus, it will be difficult to monitor and capture the value of the output voltage as the value is too small. Moreover, the value cannot be automatically recorded by the Multimeter itself.

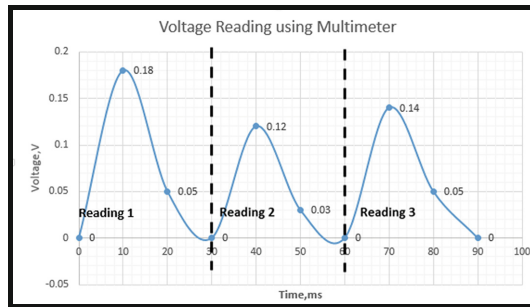


Fig. 10. Voltage reading using practical multimeter

Secondly, the value of voltage output displayed when the piezoelectric applied with the mechanical pressure. The value of voltage output by the Applink Bluetooth is discrete as shown in Fig. 9. The value will remain for 3 s, then it will automatically return to 0 V. When there is another mechanical pressure applied, the value displayed will stay for 3 s and it will return to zero instantaneously. When using the Multimeter, the value of voltage displayed is continuous. The value keeps on decreasing before return to 0 V. The continuously changing values of voltage displayed give the difficulties to monitor the value displayed by the Multimeter.

Although this research has achieved the main objective of the research, but there are few limitations in the research that could be improved in future. Throughout the experiment held, the data saved in the internal storage of the Smartphone depends on the delay time set

in the programming. When the delay time is too small, too much data being saved and it will cause disruption to the operating system of the Smartphone. But it still reliable to monitor the value of voltage output due to the program set in the PIC microcontroller.

Other than that, the durability of the piezoelectric sensor itself also part of the limitation of this research. This research only focused to monitor the output voltage produced by the piezoelectric. In order to increase the quality of this research, durability element is important to ensure that the device is sustained and can function in its best condition.

This monitoring system also has limited range of data connection through Bluetooth because Bluetooth ranges of connection are limited to 10 m or 33 feet. Improvements for these limitations need further researches so that it will be a monitoring device with no limit in data connection to a larger area and more durable in the future.

4 Application

In the military application, the soldiers often doing their exercise such as long marching for the training purpose before it comes to real situations. As it was a mandatory exercise, this system should be applied in order to improve and wider the application by using piezoelectric. Studies of piezoelectric in previous application shows that the embedding of piezoelectric itself into combat boot and how its harvest energy to charge a device. When a long marching exercise is conducted, walking with a combat boot will then produce the voltage as the piezoelectric is a voltage producing by the mechanical pressure. This monitoring system is an advance of the previous studies that monitor the voltage produce in every steps taken with time. Furthermore, this system is using wireless technology compared of using practical ways to collect data and record simultaneously.

5 Conclusion

The main objective of this research has been achieved to monitor the output voltage produced by the piezoelectric system wirelessly. There is no real time wireless monitoring system yet in the market for piezoelectric and many other small scale energy scavenging materials. This research suits very well in providing solution for monitoring voltage wirelessly and record it in real time data as well.

Acknowledgement. This research is fully supported by a short term grant, UPNM/2014/GPJP/TK05. The authors fully acknowledged Universiti Pertahanan Nasional Malaysia (UPNM) for the approved fund which makes this important research viable and effective.

References

1. Gupta, M.N., Suman, Yadav, S.K.: Electricity Generation Due to Vibration of Moving Vehicles Using Piezoelectric Effect. University Department of Engineering & Technology (SCRIT), Research India Publications (2014)

2. Messrs, V.B., Nazarenko, A., Sarian, V., Sushchenko, N., Lutokhin, A.: Application of Wireless Sensor Network in Next Generation Network. Telecommunication Standardized Sector of ITU-T (2014)
3. Smartphone: Technopedia Inc: <https://www.techopedia.com/definition/2977/smartphone>. Retrieved from 2 Apr 2016
4. Dar, M.A., Parvez, J.: Smartphone operating systems: Evaluation & enhancements. Srinaga: National Institute of Electronics & Information Technology (NIELIT) (2014)
5. Operating System. TechTerms: http://techterms.com/definition/operating_system. Retrieved from Jan 2012
6. Shu, Y.C., Lien, I.C.: Analysis of power output for piezoelectric energy harvester systems. *Smart Mater. Struct.* **15**, 1499–1512 (2006). Institute of Physics Publishing
7. Rocha, J.G., Goncalves, L.M., Rocha, P.F., Silva, M.P., Lanceros-Mendez, S.: Energy harvesting from piezoelectric material fully integrated in footwears. *IEEE Trans. Ind. Electron.* **57**(3), 813–819 (2010)

A System for Classroom Environment Monitoring Using the Internet of Things and Cloud Computing

Wuttipong Runathong¹, Winai Wongthai^{1,2(✉)}, and Sutthiwat Panithansuwan¹

¹ Department of Computer Science and Information Technology,
Naresuan University, Phitsanulok, Thailand
winaiw@nu.ac.th

² Research Center for Academic Excellence in Nonlinear Analysis and Optimization,
Naresuan University, Phitsanulok, Thailand

Abstract. A classroom environment monitoring system was developed as a demonstration to Computer Science and Information Technology undergraduate students to enhance their learning experience. Monitoring and controlling the classroom environment, including the lighting and temperature levels in real-time was the primary functionality of the system. Using data on the optimal light and temperature setting, the demonstration system was able to monitor and assess the environment to ensure the comfort of the students. The system demonstrated to the students the concepts and practices of the Internet of Things (IoT) and cloud computing can be beneficially applied and to provide services in specific application areas, this time in education, with simple system design and implementation. While such an application is not new in concept or implementation, the important features of similar systems discussed in previous systems related to the classroom environment monitoring were identified and analysed, and the best and most important features incorporated in our system, together with our own ideas, to provide the students with a significant learning experience based on a real application, which we implemented and presented as a prototype. The attributes of our system are discussed, and the success in providing a good learning experience for the students are discussed. We suggest that argue that more research is needed on this topic, and encourage other researchers to participate in the topic.

Keywords: Classroom environment monitoring · Internet of Things · Cloud computing

1 Introduction

Our purpose in this project was to build a prototype system using current technology to demonstrate to students a modern approach to system development, including the significant range of development tools now available within the technological context of the Internet of Things (IoT) and the Cloud. As such it is an exercise in project-based learning, which is an aspect of Teaching and Learning that was considered important to enhance student learning. For our purpose we selected a familiar environment, the classroom, and a useful system type, environment monitoring and control, to ensure the students could see an example of a real world, useful, potentially commercialisable system.

As students, the classroom is an important place in which they spend a significant amount of their time. It is essential that the classroom environment is conducive to studying and is comfortable and an optimal studying environment is created. As has been noted elsewhere, inappropriate environmental factors, particularly light and temperature levels can reduce students' ability to study [1]. The selection of this system type met all the education criteria we considered important, and therefore, as discussed in [7], an environment monitoring system is entirely appropriate for our educational purposes. Using technology to monitor the indoor environment of buildings has been an application of computing and communication technology for many years, but the quite recent advent of the Cloud and the Internet of Things has provided the opportunity to create more sophisticated systems, and these technologies are now an imperative part of ICT students' learning.

Our project therefore leveraged the combination of these two recent technologies. The Internet of Things or IoT refers to the ability to combine smart objects with the Internet and enable these objects to interact with other objects connected to the Internet [2]. Cloud computing provides on-line computing resources such as storage, operating systems, applications and infrastructure, allowing these resources to be accessed via the Internet [31], importantly without the need for expensive local infrastructure. We investigated previous reported development of similar systems to identify the essential elements of this type of system, and also to identify what aspects of our thinking had not been included. For example, in [4], two aspects which we consider important, measuring of ambient light levels and the availability of comparison data were not mentioned. Our comparison with various other prior studies is discussed more fully in the following Literature Review.

According to [5, 6], the optimal temperature range for studying is between 20°C and 23.33°C and the optimal light levels range between 400 lx and 600 lx, lux being the SI (International system of units) unit which denotes luminous density [10] (also stated as lx). A classroom environment monitoring system should be able to assess the classroom environment as being at these desirable levels and provide a feedback system with constant, or frequent, manipulation of these environmental factors. To simplify our system prototype we did not include this full feedback mechanism, but included a display of the current temperature and light level readings to provide this information to classroom caretakers who could then take action, manually, to adjust the settings. The further development of our system will include this more extensive feedback and control system.

Summary of Contributions: We consider that there are two main aspects to our contribution to the field. First, there is the educational contribution in that we demonstrate the effectiveness and success of our project-based learning approach. The students were introduced to the concept and practice of prototyping as a successful development approach, they were able to gain in-depth knowledge of the contemporary development environment including the Internet of Things and Cloud computing, and the principle of system usefulness was clearly embedded in their learning. As well, the student's knowledge of the marketplace for software development tools was significantly enhanced, and the symbiotic relationship between modern development tools and contemporary development methodologies, such as prototyping, was clearly demonstrated.

We do admit to the possible over-kill in the selection of tools, but this was done to ensure a wide understanding of the marketplace; this was not a commercial development demanding a lean approach. From a technology point of view we have applied these two relatively recent technologies, The Internet of Things and Cloud computing to a previously well understood application thereby enhancing that system type and extending our understanding of the applicability of these technologies, and the advance in development productivity offered by these tools. The proposed system, and the method of development, illustrate how the Internet of Things and Cloud computing benefit applications and services in the education area with simple system design and implementation.

2 Literature Review

2.1 Environment Monitoring Systems, the IoT, and Cloud Computing

An environment monitoring system includes one or more sensors and data storage [7]. To monitor the classroom environment, a system also requires sensors to continuously monitor the environment and send the data to a storage server. The classroom environment is the subject and focus of this paper. The IoT refers to the interconnectivity of smart objects over the Internet. The concept is to enable any smart object connected to the Internet to be able to interact with any other smart object or objects connected to the Internet [2]. In this paper, the smart objects being considered are environment sensors in the classroom which have the ability to connect to the Internet, and which are able to send data of changes in environmental variables (light and temperature level) to a cloud server, regardless of where the classroom is situated. Cloud computing offers computational resources to customers, such as networking, processing, and storage [32]. A server in the cloud allows remote access, and it includes both hardware and system software that can deliver services over the Internet [3]. The cloud can be remotely accessed by any Internet connected device at any time and from any location, and can send and receive data to and from the cloud through the Internet. Many previous work apply the cloud such as in [3]. With the benefits of IoT and the cloud, monitoring the classroom environment by using both is ideal. [8] state that IoT and cloud share their benefits to reduce IoT weaknesses. The IoT has four important problems of reliability, performance, security and privacy. To solve these problems, cloud computing provides at least a partial solution. One reason can be that the cloud has huge storage capacity, processing power and level of reliability. To deal with data generated by the IoT, cloud is the most convenient and cost effective solution to most information processing requirements and solves most of the problems inherent in IoT. This paper exploits the benefits of IoT and cloud, demonstrating ease of use, convenience and power of IoT coupled with cloud.

2.2 The IoT Technologies in Teaching, Learning and Basic Education Management

IoT technology is explored to assess its ability to improve learning, teaching and education management [9]. The application of IoT technology is classified under various

headings such as health in education, teacher education, learner support, social mobilization and support services, planning and delivery oversight, quality assessment, inclusive education, curriculum policy, support and monitoring, and administration [9]. In [9] the learner support classification is addressed. It focuses on a climate-controlled classroom environment and lighting factors, as part of learner support. This paper also focuses on the learner support category.

2.3 Previous Works Related to Environment Monitoring

[4] describes a system using the cloud and IoT for monitoring the classroom environment. That system monitors humidity and temperature through sensors and sends the data to Google Drive® which is a cloud computing service provided by Google. This service stores the data as an excel file. Graphs of the data can be generated and presented on a website. The objective of this study was to show a solution by using Google Drive® and the possibility of using it for both data storage and especially for charting. [4] did not discuss the optimum values of the environment or conditions conducive to study. According to [11] the cloud and IoT are technologies that have been used to monitor the saturation line, water levels and possible deformation of dam walls in a tailings dam at a mine site. The system remotely monitors these aspects and creates pre-alarm information automatically and in any kind of weather conditions. [12] used various sensors for detecting and monitoring temperature, humidity and CO₂ in an in-door environment. The system changes the colour of displayed pictures if these environmental conditions deteriorate to poor levels. This system did not connect to the cloud network, but used a local server.

The researchers in [13] give an example of using IoT, cloud and Near Field Communication (NFC) to control the environment in a classroom. NFC technology is used with the information being communicated over a radio frequency. The collected data are sent to the Internet and cloud. The outcome of this study is to allow the monitoring of classrooms and to display the status of each classroom graphically. This work did not discuss possible optimum values of environmental conditions conducive to study. [14] used IoT and cloud to monitor air quality of different classrooms at a university. Each classroom had a number of wireless nodes and each node had a number of sensors. This system monitored, stored and analysed the data collected. This work also did not discuss possible optimum values for the environment conducive to study. [15] Investigated the effect of temperature in call centres. Two call centres, each in a different time zone with different weather conditions, were investigated. This work did not involve light level measurement which is also an important factor.

2.4 Features Summarization and Comparison of Previous Systems

Regarding environmental monitoring systems, the appropriate features of some systems described in Sects. 2.1 to 2.3 are summarized. Then we will design and implement our system based on combination of these features, as this enables our proposed system to be applicable in the current situation of emerging technologies (such as IoT) and the meeting of educational needs. These features are as follows. (1) The proposed system

should be a monitoring system used to continually receive environment values as agreed by [7], without continually monitoring the environment the environmental values could

Table 1. Comparing table

Topic	(1) Monitoring system	(2) Focusing on the classroom	(3) Using IoT	(4) Incorporating cloud	(5) Measure temperature	(6) Measure light level	(7) Compare data to be suitable for studying
A cloud solution for monitoring classroom environmental conditions in a smart university [4]	/	/	/	/	/	x	x
The IoT and cloud computing based tailings dam monitoring and pre-alarm system in mines [11]	/	x	/	/	x	x	x
The IoT at school and at the CES in Las Vegas [12]	/	/	/	x	/	x	x
An IoT Example: Classrooms Access Control over Near Field Communication [13]	x	/	/	x	x	x	x
Indoor air quality monitoring through software defined infrastructures [14]	/	/	/	/	/	x	x
The effect of air temperature on labour productivity in call centres – a case study [15]	/	x	/	x	/	x	x

not be known in real time. (2) It should focus on a classroom environment to improve student learning because the classroom environment may affect student learning ability as argued by [1]. (3) The system should use IoT to detect environmental values and send the values to the Internet. This should enable the system to easily send the values to a cloud server as agreed by [2]. (4) The system should be incorporated into the cloud server to solve problems of IoT [8] and reduce costs [3]. (5) It should measure the temperature as inappropriate temperature conditions may reduce student learning [1]. (6) The system should measure light levels because unsuitable light levels may reduce student learning [1]. (7) The proposed system should compare the captured temperature and light values to accepted optimal values to ensure that the environment is not too hot or too cold, therefore suitable for studying. Regarding the environmental monitoring systems, Table 1 shows the appropriate features of some systems as described in Sects. 2.1 to 2.3. The notation ‘/’ in the table means that a system applies that particular feature, and ‘x’ is otherwise. From the table, there is no systems that achieve all these features. Our proposed system aims to meet all the features.

3 Design and Implementation of the Proposed System

3.1 System Architecture of the Proposed System

The proposed system is designed to meet the features described above. Figure 1 illustrates the architecture of the proposed system. There are 4 steps in the figure and each step is in a small circle with a number 1, 2, 3, or 4. Each step can comprise related components. Each component is in a pair of brackets such as (1). The system deploys a light sensor [17] as used by [18–20], in similar research. The system also deploys a temperature sensor [21] to detect temperature levels as used by [22–24] in similar research. The cloud server from [25] is used by researchers and system developers, including IBM, HP, MIT, and etc. The system is connected to a wireless USB adapter [26] for Internet connections. Lastly, Raspberry Pi 2 model B v1.1 which is a small low cost, computer [16] is used to control the sensors, and connected to the Internet via the USB adapter. Pi also has a program for sending data to the cloud server. The cloud has MongoDB [33] which is a free and open source database program.

Step 1, the light sensor (see (1)) and temperature sensor (see (2)) send data of the classroom environment to a Raspberry Pi, see (3). Pi receives the data and sends it to the cloud server (see (5)) through the adapter, see (4). Then, the wireless USB adapter receives the data. Step 2, the adapter forwards the data to the cloud server. Step 3, then the program in the server that received the data stores it in the database. Step 4, when users of the system (see (6), (7)) request a monitoring information webpage, the information can be transferred through HTTP (9) by the cloud server, regardless of the device or operating system; such as a personal computer running Windows (see (6)) or a mobile device running Android or IOS (see (7)).

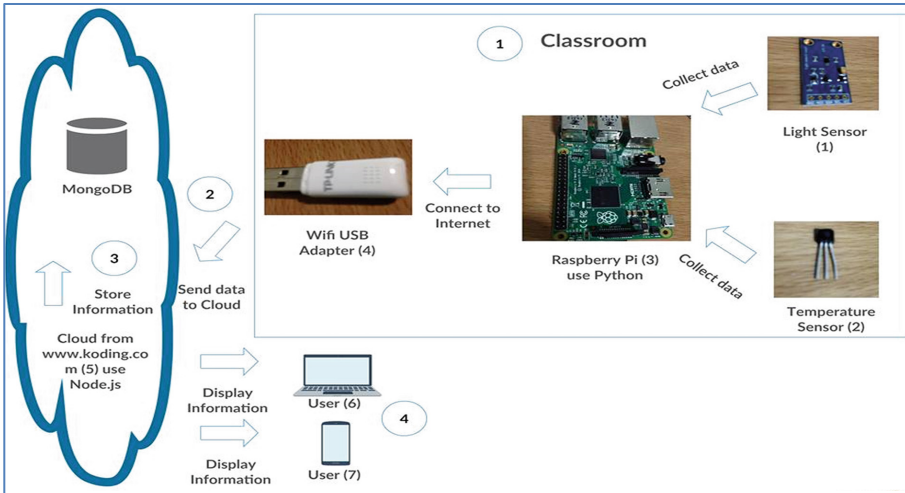


Fig. 1. System architecture

3.2 Implementation

Raspberry Pi is connected with the light and temperature sensors by following the instructions in [29, 30] respectively. A Python program to collect data from the sensors and send the data to the cloud server is created and in Pi. For this program, it is necessary to install Node.js to run JavaScript files and MongoDB program to store captured environmental data. Note that, Node.js is a JavaScript runtime built on Chrome’s V8 JavaScript Engine, and is used to create a web server in the cloud server (see (5)) in Fig. 1 Then we create three files as following. (1) Server.js, this file is used to create a server with functions to receive data from Pi and store it to MongoDB. The file also creates a HTTP server to serve the monitoring information website. (2) Index.html, this file allows users to see the structure of the classroom and its monitored information. (3) Function.js, this file is used to receive data from the cloud server and compare it to the optimal ranges. All the files can be found in [34].

4 Research Results and Discussion

4.1 Results

After running the proposed system, there are two main parts to the results, as illustrated in Fig. 2. Firstly, from the figure, A1–D3 represent student tables in a model of a typical classroom. Each table, such as the one labelled ‘A1’, has its own light and temperature data. Both data can be seen in the figure as 438.33 lx and 29.12°C respectively in the first line text under the labelled text. Again according to [5, 6], the optimal temperature and lighting ranges for study is between 20.00°C–23.33°C and between 400 lx – 600 lx respectively. These ranges are illustrated at the top right of the figure. Thus, A1 table has a ‘GOOD’ light level and a ‘TOO HIGH’ temperature level, see in the second line

text under the labelled text ‘A1’. In this case, the temperate level is not suitable for study and may reduce student learning according to [1].

Lighting average : 438.33 Overall status : GOOD		Optimal lighting range : 400 - 600	
Temperature average : 29.12 Overall status : TOO HIGH		Optimal temperature range : 20 - 23.33	
A1	A2	A3	
438.33 lx : 29.12 C	N/A	N/A	
GOOD / TOO HIGH			
B1	B2	B3	
N/A	N/A	N/A	
C1	C2	C3	
N/A	N/A	N/A	
D1	D2	D3	
N/A	N/A	N/A	

Fig. 2. The results

After running the proposed system, there are two main parts to the results, as illustrated in Fig. 2. Firstly, from the figure, A1–D3 represent student tables in a model of a typical classroom. Each table, such as the one labelled ‘A1’, has its own light and temperature data. Both data can be seen in the figure as 438.33 lx and 29.12°C respectively in the first line text under the labelled text. Again according to [5, 6], the optimal temperature and lighting ranges for study is between 20.00°C–23.33°C and between 400 lx–600 lx respectively. These ranges are illustrated at the top right of the figure. Thus, A1 table has a ‘GOOD’ light level and a ‘TOO HIGH’ temperature level, see in the second line text under the labelled text ‘A1’. In this case, the temperate level is not suitable for study and may reduce student learning according to [1].

Secondly, the summarized information of temperature and light averages and the status of all tables in this classroom can be seen at the two text lines on the top left of Fig. 2. In this case, this information is spurious due to the limitations of equipment and the fact that only table A1 was fitted with sensors in this experiment. Then only the table’s light and temperature level data is taken to calculate the overall averages and status as illustrated on the top left of the figure. Each table of A2–D3 shows the text ‘N/A’. This is because they do not have their own sensors. When all are fitted with appropriate sensors, all the texts will be changed to the correct ones in the same way as table A1. Then, the overall averages and statuses will be given according to the available recorded information. Thus, this information can be used to decide whether the classroom environment is suitable for students to study or not.

4.2 Discussion

Due to the system storing light and temperature level data, the data can be used to calculate the amount of energy consumed. Researchers in [28] studied energy consumption by placing sensors in classrooms to monitor indoor climate conditions. Their system can calculate the energy consumption of this classroom. We could enhance our system based on the guide lines from their research to calculate efficient energy consumption. This could plan an effective energy consumption in classrooms, while theses classrooms

are suitable for studying. Due to the system being able to monitor not only light and temperature factors, new sensors can be added to monitor other appropriate factors to improve student learning ability. According to [1], sound can reduce student learning efficiency, also [5] states that acoustic and air quality can reduce student learning. The proposed system can apply to monitor these new environment factors. This can enhance our system to collect all significant environment factors, enable students to study in comfortable and appropriate environment, then increase student learning ability. In Fig. 2, if the system can automatically adjust the temperature and light levels of table A1 to the optimal range for study, the student in A1 position may yield a higher learning ability. Additionally, when a classroom has a table layout that differs from the one in Fig. 2, the proposed system in this paper could apply the new layout. Lastly, we believe this paper can apply to the monitoring of the environment in other types of room such as meeting rooms, based on the rooms' conditions and the appropriate optimal ranges of levels of temperature, light, or, other essential factors.

5 Conclusion and Recommendations

This paper summarizes important features of previous systems used in similar research which can be considered as relating to classroom environmental monitoring. These features are important for applications and services in developing classroom monitoring environment systems. Based on these features, a prototype system was designed and tested. Lastly, we discuss the proposed system and its results in other aspects. For example, the proposed system can achieve all the important features, such as the measuring of light and temperature and the use of a cloud system. Moreover, with further research, the system also can be enhanced to yield more of its abilities such as automatic environment control to reduce energy consumption. The proposed system demonstrates how Internet of Things or IoT and cloud computing could benefit applications and services in education with simple system design and implementation. [27] discuss an air conditioning control system. Based on guide lines of this study, one of the future research directions could be to enable our system to automatically control the monitored classroom environment to be continuously suitable for studying in real time. This same possibility could be extended to other applications such as monitoring the conditions in libraries and meeting rooms.

Acknowledgement. Many thanks to Mr. Roy Morien and Mr. Kevin Roehl of the Naresuan University Language Center for his editing assistance and advice on English expression in this document.

References

1. Ryan, H.: *The Effect of Classroom Environment on Student Learning* (2013)
2. Medaglia, C.M., Serbanati, A.: An overview of privacy and security issues in the internet of things. In: Giusto, D., Iera, A., Morabito, G., Atzori, L. (eds.) *the internet of things*, pp. 389–395. Springer, New York (2010)

3. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., et al.: A view of cloud computing. *Commun. ACM* **53**(4), 50–58 (2010)
4. Mircea cel Batran. A cloud solution for monitoring classroom environmental conditions in a smart university (2015)
5. Cheryan, S., Ziegler, S.A., Plaut, V.C., Meltzoff, A.N.: *Designing Classrooms to Maximize Student Achievement* (2014)
6. European Committee for Standardization: *Light and lighting – Lighting of work places – Part 1: Indoor work places* (2002)
7. Boatman, J.F., Reichel, B.S.: *Environment monitoring system* (2006)
8. Liu, Y., Dong, B., Guo, B., Yang, J., Peng, W.: Combination of cloud computing and IoT in medical monitoring systems. *Int. J. Hybrid Inf. Technol.* **8**(12), 367–376 (2015)
9. Dlodlo, N.: *The IoT technologies in teaching, learning and basic education management* (2012)
10. Palmer, J.M.: *Radiometry and Photometry FAQ* (1999)
11. Sun, E., Zhang, X., Li, Z.: The IoT and cloud computing (CC) based tailings dam monitoring and pre-alarm system in mines. *Saf. Sci.* **50**(4), 811–815 (2011)
12. Weinberger, M.: *The IoT at school and at the CES in Las Vegas* (2015)
13. Palma, D., Agudo, J.E., Sanchez, H., Macias, M.M.: *An IoT Example: Classrooms Access Control over Near Field Communication* (2014)
14. Spachos, P.: *Indoor air quality monitoring through software defined infrastructures* (2016)
15. Niemela, R., Hannula, M., Rautio, S., Reijula, K., Railio, J.: The effect of air temperature on labour productivity in call centres - a case study. *Energy Buildings* **34**(8), 759–764 (2002)
16. Raspberry Pi: *Raspberry Pi 2 Model B* (2015)
17. ROHM: *Digital 16bit Serial Output Type Ambient Light Sensor LC* (2011)
18. Ding, D., Chen, H., Zhang, L., Chen, H., Lin, H., Gao, F.: *An intelligent and telecontrol environment monitoring equipment with extended interfaces* (2015)
19. Xianghong, K., Weiguo, Q., Kexiang, L., Xinlei, J., Xiang, P.: *The design of LED fish gathering lamp PC free multipoint photometer* (2015)
20. Shao, Y., Wang, F., Zhang, Y., Zan, P.: *Research of metro illumination control based on BP neural network PID algorithm*. In: Fei, M., Peng, C., Su, Z., Song, Y., Han, Q. (eds.) *Computational Intelligence, Networked Systems and Their Applications. LSMS/ICSEE 2014. Communications in Computer and Information Science*, vol. 462, pp. 1–8. Springer, Heidelberg (2014)
21. Maxim Integrated. *DS18B20* (2015)
22. Ping, L., Yucai, Z., Zeng, X., Ting-fang, Y.: *A Design of the Temperature Test System Based on Grouping DS18B20* (2007)
23. Pengfei, L., Jiakun, L., Junfeng, J.: *Wireless temperature monitoring system based on the ZigBee technology*, April 2010
24. Zhang, X., Fang, J., Yu, X.: *Design and implementation of nodes based on CC2430 for the agricultural information wireless monitoring* (2010)
25. Koding, Inc. San Francisco. www.koding.com (2016)
26. TP-LINK Technologies: *Mbps Mini Wireless N USB Adapter TL-WN723N* (2016)
27. Mochizuku, M., Sato, K., Kato, T., Isikawa, M., Sugiyama, T.: *Air conditioning control system* (1995)
28. Rattanongphisat, W., Suwannakom, A., Harfield, A.: *Indoor weather related to the energy consumption of air conditioned classroom: Monitoring system for energy efficient building plan* (2016)
29. <http://www.raspberrypi-spy.co.uk/2015/03/bh1750fvi-i2c-digital-light-intensity-sensor/>
30. <http://www.reuk.co.uk/wordpress/raspberry-pi/ds18b20-temperature-sensor-with-raspberry-pi/>

31. Wongthai, W., van Moorsel, A.: Quality analysis of logging system components in the cloud. In: Kim, K., Joukov, N. (eds.) Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 651–662. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_64](https://doi.org/10.1007/978-981-10-0557-2_64)
32. Wongthai, W., Van Moorsel, A.: Performance measurement of logging systems in infrastructure as a service cloud. ICIC Express Letters (2016)
33. <https://www.mongodb.com/>
34. <https://dl.dropboxusercontent.com/u/4620323/mce.rar>

4th Convergence of Healthcare and Information Technology

Research on Design of End Site Architecture to Connect LHCONE in KREONET

Chanjin Park, Wonhyuk Lee, Kuinam J. Kim, and Hyuncheol Kim^(✉)

KREONET Operation and Service Division of Supercomputing,
Korea Institute of Science and Technology Information,
245 Daehak-ro, Yuseong-gu, Daejeon, 34141, South Korea
{pcj0722, livezone}@kisti.re.kr, kuinamj@gmail.com,
hckim@nsu.ac.kr

Abstract. As the large hadron collider (LHC) community has requested a high-performance network for the transmission of massive physics data, the large hadron collider optical private network (LHCOPN) and the large hadron collider open network environment (LHCONE) have provided a support internationally. For the LHC research groups in the Republic of Korea as well, the connection with LHCOPN and LHCONE has been necessary. Therefore, this study attempted to briefly review the LHCONE linked to the backbone of the Korea Research Environment Open NETwork (KREONET) and propose basic site architecture to accept the LHC community.

Keywords: LHCONE · KREONET · End site architecture design

1 Introduction

The European Organization for Nuclear Research (CERN) has configured worldwide LHC computing grid (WLCG) for the analysis of physics data in a massive volume, which are generated by the LHC test and worked with nearly 200 global computing centers. [1] For data exchange with these global computing centers, the LHCOPN and LHCONE have been used. The KREONET has also connected the CERN with the global science experimental data hub center (GSDC) through the LHCOPN in 2015. In addition, it has constructed the LHCONE in 2016 for connection with the ESnet's LHCONE and acceptance of domestic LHC research groups. This study briefly reviews the LHCONE and suggests basic site architecture to introduce the LHC community.

1.1 Large Hadron Collider Open Network Environment (LHCONE)

The conventional WLCG computing model has shifted from the hierarchical to distributed structure. [2] As illustrated in Fig. 1 above, therefore, tier2-tier2, tier2-tier3 and tier1-tier3 connections have been required. To meet these needs, a new network is now needed. The LHCONE is a network specially designed for the HEP community which enables direction connection among tier1, tier2 and tier3. At present, the LHCONE is available in diverse R&D network providers such as ESnet, GEANT, Internet2,

SURFnet and NORDUnet. The KREONET has also established the LHCONE in its backbone network and connected it with ESnet’s LHCONE in Chicago. The LHCONE’s core services include virtual routing and forwarding (VRF)-based layer 3 VPN, bandwidth-guaranteed P2P service and perfSONAR designed to measure network conditions and performances.

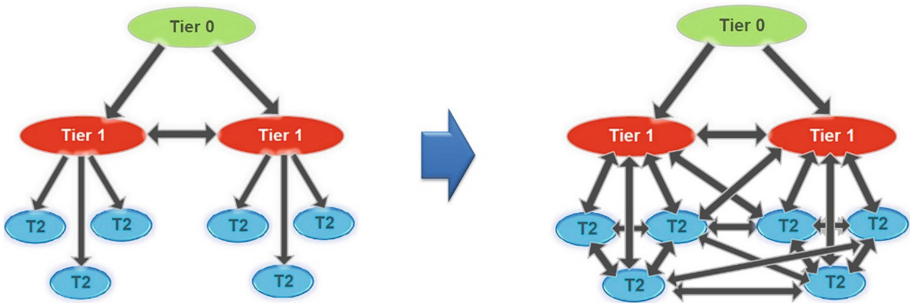


Fig. 1. Shift in the structure of WLCG computing model (from hierarchical to distributed)

1.2 L3 VPN Service in LHCONE

The L3 VPN service provides any-to-any connectivity among tier1, tier2 and tier3. It creates a virtual routing instance in the physical router, using VRF and forms a HEP-only network which transmits HEP data only by connecting each VRF. [2] As show in Fig. 2, tierX is connected to either national or continental VRFs. The national VRFs are connected via the continental VRF. In contrast, the continental VRFs are connected through trans-continental links.

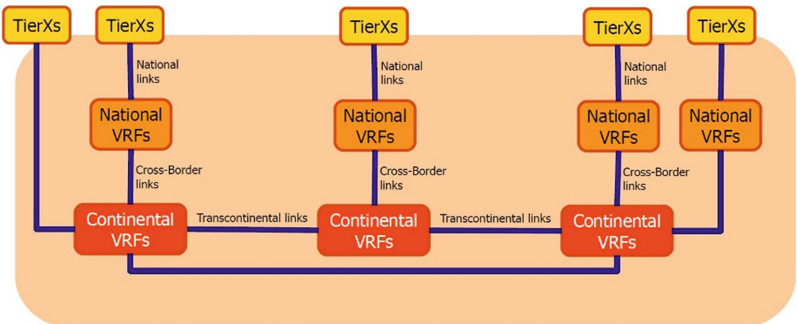


Fig. 2. LHCONE layer-3 VPN architecture

1.3 L3 VPN for LHCONE in KREONET

The KREONET is a research network which provides a high-performance network to high-tech science & technology researchers in the Republic of Korea. To build a virtual

overlay network on a physical network as illustrated in Fig. 3, the KREONET has formed VRF in Seoul, Daejeon, Daegu, Hong Kong and Chicago routers. Then, it connected iBGP peering among VRFs and eBGP peering with external VRFs such as TEIN. It is slated to be connected with LHCONE networks later such as GEANT and SRUFnet. According to the BGP filtering guide, the LHCONE does not have the BGP prefix obtained from the other LHCONE network exchanged [3].

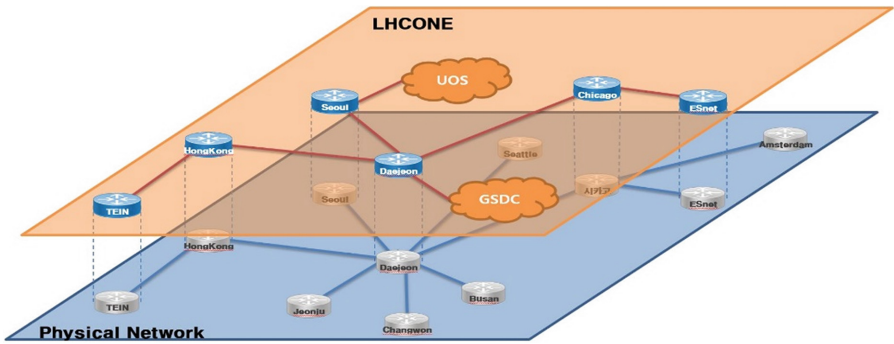


Fig. 3. LHCONE in KREONET

For example, the BGP prefix acquired from the TEIN in Fig. 4 is not sent to the ESnet. Basically, therefore, ESnet does not communicate with TEIN via the KREONET. In addition, whether or not the TEIN is connected with the KORENET is unknown. In other words, security is guaranteed by this kind of limited connection.

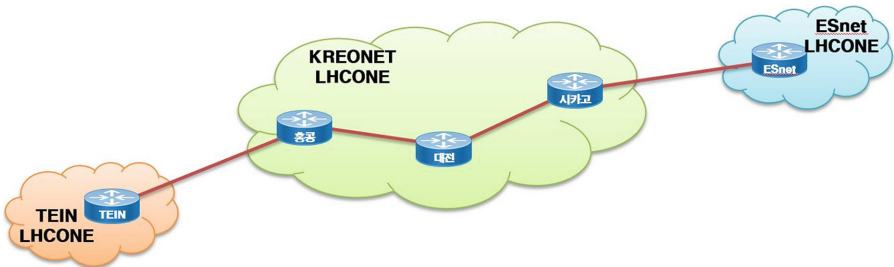


Fig. 4. BGP peering between KREONET and ESnet, TEIN LHCONE

2 LHCONE Basic Site Architecture

2.1 LHCONE Basic Site Architecture #1

This structure reveals the supply of VRF by the KREONET. As illustrated in Fig. 5, the KRONET’s routing instance is divided into the followings: general routing instance in which general traffic is handled through the VRF in router A, LHCONE VRF routing instance which handles the LHCONE traffic. In addition, the site is separated into two

subnets: LHC tier-X center and core network. Then, the LHC tier-X center’s traffic is forwarded to the LHCONE VRF while the core network’s traffic is sent to the general routing instance through two VLANs. A key to this structure lies in fast data transmission because the LHC tier-x center’s traffic transfers data by avoiding firewalls. This structure is similar to the science DMZ and safe because reliable IP bandwidths are only accessible through policy base routing [4, 5].

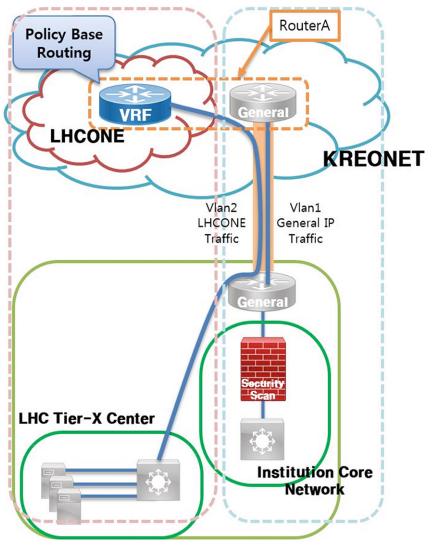


Fig. 5. Basic architecture #1

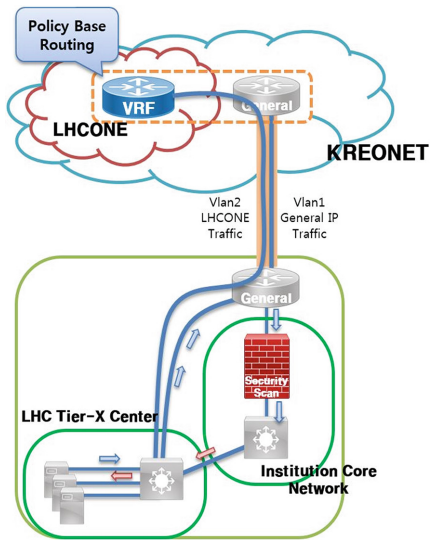


Fig. 6. Basic architecture #1-a

2.2 LHCONE Basic Site Architecture #1-a

Basically, there should be no connection between the core network and LHC tier-x center. Sometimes, however, there might be a demand for this connection. In this case as well, the traffic from the LHC tier-x center to the core can be forced to go through the security device as stated in a blue arrow in Fig. 6 to protect the core network. Then, the traffic from the core network to the LHC tier-x center should not directly pass through the security device as marked in a red arrow.

2.3 LHCONE Basic Site Architecture #1-b

Sometimes, the LHC tier-X center may ask for the connection with the general Internet, not with the LHCONE domain. In this case, the outgoing traffic should be forced to avoid the security device as stated in a blue arrow in Fig. 7, using policy base routing. In contrast, the incoming traffic is designed to pass through the security device to assure security.

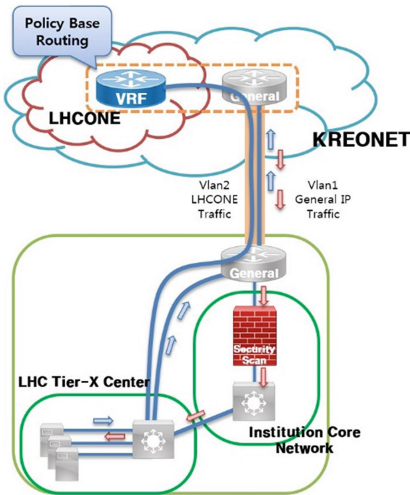


Fig. 7. Basic architecture #1-b

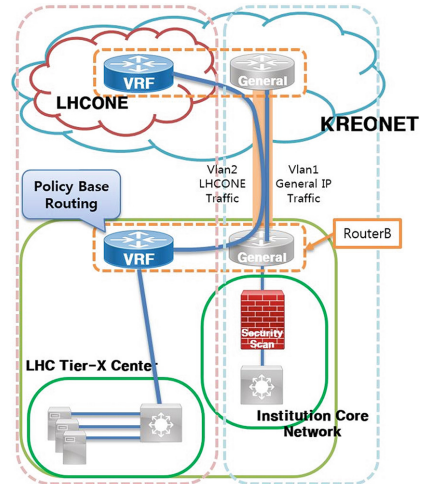


Fig. 8. Basic architecture #2

2.4 LHCONE Basic Site Architecture #2

This architecture configures VRF in the institute’s border router B, as shown in Fig. 8. Under this structure, an institute is able to directly control traffic flow with a right to control policy base routing. In this architecture as well, internal and external connections are enabled just like Architecture #1-a and Architecture #1-b.

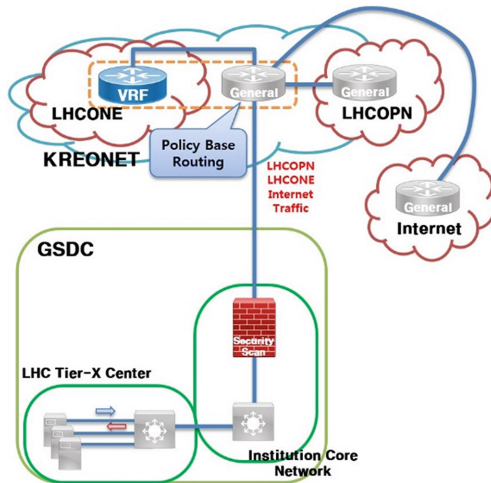


Fig. 9. LHCONE implementation design in GSDC

2.5 LHCONE Implementation Design in GSDC

Figure 9 reveals the network of the GSDC, an Alice tier1 center connected to the KREONET at present. The GSDC cannot transmit LHCONE, LHCOPN and the Internet traffic separately. Therefore, the three traffics are separated in the router A's general routing instance, using policy base routing and delivered to each network. It reveals that the architecture mentioned above reveals can be adjusted according to the situation of each site when constructed in the actual site as an example.

3 Conclusion

This study briefly introduced LHCONE, L3VPN service in the LHCONE and establishment of the LHCONE in the KREONET and proposed basic site architecture to accept the LHC community. In addition, it investigated the architecture of the GSDC. In the architecture mentioned in this study, increase in performances is expected because of the fact that it avoided the security device. However, it failed to present objective data for performance improvement. Therefore, there should be further studies on the measurement of performances by connecting the perfSONAR server on the LHCONE network with the one on the LHCONE network of the KREONET and improvement of performances in an objective measure [6].

References

1. LHC Open Network Environment. <http://lhcone.web.cern.ch/>
2. Martelli, E., Stancu, S.: LHCOPN and LHCONE status and future evolution. *J. Phys. Conf. Ser.* **664**, 1–7 (2015)
3. Martelli, E.: LHCONE AUP audit. LHCOPN-LHCONE Meeting (2015)
4. O'Connor, M.: LHCONE Operations Update. LHCOPN-LHCONE Meeting (2015)
5. O'Connor, M.: LHCONE Operations Update. LHCOPN-LHCONE Meeting (2016)
6. McKee, S.: perfSONAR for LHCOPN/LHCONE Update. LHCOPN-LHCONE Meeting (2015)

A Study of Children Play Educational Environment Based on u-Healthcare System

Minkyu Kim, Soojung Park, and Byungkwon Park^(✉)

Department of Kinesiology, Inha University, Incheon, Korea
{leisure.loisir,psj}@inha.ac.kr, zexrol@naver.com

Abstract. The objective of this study is to discuss the application of u-Healthcare System to play educational environments and its effects. It is expected that this application will contribute to addressing various problems among Korean children such as negative habitual behaviors, pressures in the education system, etc. In addition, positive effects are expected in terms of the changing meaning of wellness, educational aspects, and psychological aspects of leisure play for children. To this end, IT-based data collection and analysis was conducted on changes in behaviors and emotions among children during their physical play activities: Specifically, the brainwave bio information collecting technology and physical activity information collecting technology are utilized so that teachers can monitor them easily. It is expected that this method will contribute to the establishment of a new child play environment as well as physical, mental, and social health of children.

Keywords: U-healthcare system · Brainwave bio information collecting technology · Physical activity information collecting technology

1 Introduction

Originally, ‘wellness’ is a compound word of wellbeing and fitness, but in modern society, it means rather wellbeing + happiness: the pursuit of an optimal status in physical, emotional, social, mental, and intellectual areas [1]. As such, the concept of this term has changed to include space, behavior, and effort for healthy and lively activity [2]. Social interest in the changed meaning of wellness has been expressed with such words as convergence, complexation, smart, etc., and accordingly, there is a demand for convergence among various industrial sectors such as IT [3].

In the utilization of IT, the new paradigm of wellness affects child play educational environment as well. An approach in a psychological perspective on leisure play is as follows: In terms of leisure play, the utilization of new mechanisms affects leisure motivation, leisure flow, leisure satisfaction, and leisure continuation positively so much so that it sustains educational environments and prevents dropouts [4]. According to developmental psychology, cognitive skills in the preoperational stage of child classification can handle a far larger amount of symbols than in the traditional Piaget theory, and it is possible to represent characteristics of thinking in the concrete stage during which thinking and acting specifically are possible [5]. In other words, this theory suggests that

children are more capable than generally expected and that they can think specifically. Thus, if new tools are developed and utilized properly, they can help develop general aspects of children including various social and cognitive interactions [6–10].

It is reported that health-related habitual behaviors among Korean children are not positive [11]: The daily sleeping time of 0 to 3-year-old infants are 11 h and 53 min on average, which is shorter than that in Western (13 h and 1 min on average) and Asian (12 h and 19 min on average) countries [12], and 85.5% of 3 to 9-year-old children use the internet frequently [13]. Habitual behaviors include physical movements, sleep and sitting acts in which there is little or no energy consumption, physical activity of low intensity, and energy-consuming physical activity of middle or high intensity [14]. It is thought, therefore, that various factors are involved in the health deterioration among children and youths in Korea, and it turned out that the most outstanding factors are the competitive education structure, fixed priorities in the education fever, and pressure from such intense school work [15].

Thus, it is necessary to address various problems such as negative habitual behaviors among Korean children and various pressures from the education system. In addition, demands in academic and industrial circles and participant groups are increasing for new instruments in the area of play educational environment that are expected to contribute in terms of the changing meaning of wellness, education of children, and psychological aspects of leisure play. Accordingly, this study discusses the establishment of play educational environment for children based on u-Healthcare System and its expected effects.

2 Body

Play provides children with opportunities to act freely, develop essential factors for holistic growth such as imagination, and find joy from a process itself rather than intending a certain result [16].

Play is defined as ① a voluntary act that is distinguished from work; ② stimulating imagination among children; ③ aiming at behaviors themselves rather than certain results; ④ facilitating creativity; ⑤ having unique rules and likely to be demolished once the rules are violated; ⑥ likely to involve competitive elements; and ⑦ represented with active participation of players (Fein, 1983).

In particular, a play that involves physical exercise lets players generate energy and satisfy physiological desires through movements. According to Geum-ja Hong (1999), children aged 4 to 5 start expressing secondary needs based on their physiological desires with the society as a medium. These are high-level desires – social desire, personal desire, and self-actualization desire. The joy of participating in a group is another example.

The play is an act that combines physical and emotional characteristics. It would be helpful for child development if patterns of the act and emotional aspects of children can be grasped. Accordingly, this chapter presents the results of IT-based data collection and analysis on changes in behaviors and emotions during play that involves physical exercise. This method will make teachers' monitoring easier.

2.1 Brainwave Bio Information Collecting Technology

When it comes to technology trends in the area of biomechanics, such factors as heart rates, respiration, blood pressure, and calorie consumption are measured by means of special sensors on a body, and information on exercise intensity, distance, and quantity is provided to exercise participants by means of devices that make scientific sports activity possible [2]. Such technologies utilized in sports activity may be applied to grasping the play conditions of children. One of them is the brainwave technology.

It is likely that children’s emotions change significantly depending on the type of play. In a physical activity which requires a high level of concentration, for example, a group of children confident of their athletic abilities would display Mid-β waves that indicate the status of concentration and activity while a group of children with a relatively low level of athletic abilities would display γ waves that indicate intense stress, anxiety, and nervousness. Psychological states may be different depending on the place of physical activity: small or large areas. Such differences in mental states and interests can be indicators of psychological changes among those who participate in physical activity.

When the types of activity are matched with psychological states among children properly, it is possible to expect psychological states depending on each individual’s play preferences. Brainwave signals applicable to child play are presented in Table 1

Table 1. Brainwave signals

Frequency band	Frequency name	Characteristics
8–12	A	Relaxation and rest
12–15	SMR	Attention
16–20	Mid-β	Concentration and activity
21–30	B	Anxiety, excitement, and stress
30–50	γ	Intense stress – anxiety, nervousness, etc.

2.2 Physical Activity Information Collecting Technology

In the era of new media, play cultures for children are changing the modern society. While various types of play and physical activity were in broad areas such as playground and athletic field in the past, children are recently facing new play cultures in indoor areas due to safety problems and limited space. As a result, the areas and extent of physical activity are reduced, and it is possible to monitor how children move and how their emotional states change in a relatively small space real-time.

Because this experiment was conducted among children, indoor network digital cameras were installed on the ceiling instead of wearable devices. Children’s movements were traced and identified by means of these cameras: specifically, children’s movement distance and time, duration at a certain location, etc. This way, it is possible to grasp how much children move, on which location they stay long, and what kind of play they prefer.

2.3 Application

In this study, changes in children’s psychological states depending on their movement in a limited space and changes in movement depending on psychological states are examined. Figures 1, 2 and 3 show the diagrams of children’s movements and psychological states. While many children enjoy playing in each interesting zone of the play room, play elements that interest or do not interest them are grasped. For children not to feel stress or anxious feelings during the leisure play time, a teacher can monitor their psychological states constantly (Fig. 4).

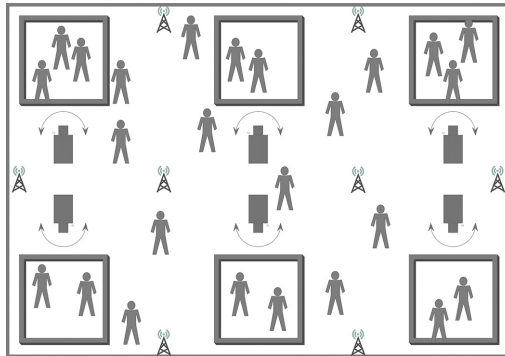


Fig. 1. Play room

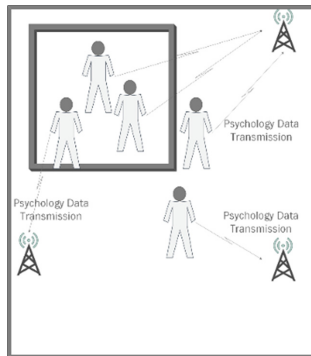


Fig. 2. Psychology data transmission

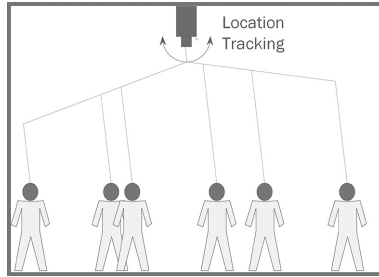


Fig. 3. Network camera location tracking

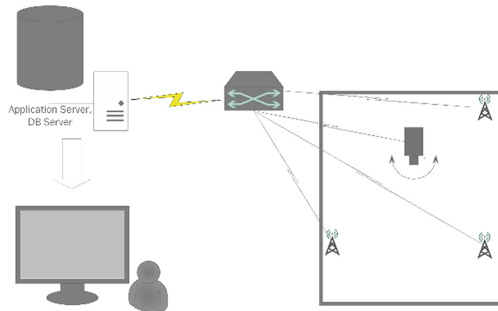


Fig. 4. System configuration

3 Conclusion

The objective of this study is to discuss the application of u-Healthcare System to play educational environments and its effects. It is expected that this application will contribute to addressing various problems among Korean children such as negative habitual behaviors, pressures in the education system, etc. In addition, positive effects are expected in terms of the changing meaning of wellness, educational aspects, and psychological aspects of leisure play for children. To this end, IT-based data collection and analysis was conducted on changes in behaviors and emotions among children during their physical play activities: Specifically, the brainwave bio information collecting technology and physical activity information collecting technology are utilized so that teachers can monitor them easily. It is expected that this method will contribute to the establishment of a new child play environment as well as physical, mental, and social health of children.

References

1. Han, T.-H., Min, K.-P., Son, J.-G.: A case study on device for wellness services. In: *Weekly Technology Trends of National IT Industry Promotion Agency*, vol. 1639, pp. 14–26 (2014)
2. Kim, M., Park, S., Park, B., Cho, Y.H., Kang, S.Y.: A new paradigm for the spread sport leisure culture focusing on the IT-based convergence interactive system. In: Kim, K., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016*. LNEE, vol. 376, pp. 1477–1485. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_142](https://doi.org/10.1007/978-981-10-0557-2_142)
3. Park, S.-H., Jang, D.-G.: IT convergence trends in wellness. *Commun. Korean Inst. Inf. Scientists Eng.* **31**(3), 61–72 (2013)
4. Kim, M.-K., Park, S.-J.: Grounded theoretical analysis on the formation of leisure addiction. *J. Leisure Recreation Stud.* **38**(3), 1–16 (2014)
5. Shade, D., Nida, R.E., Lipinski, J.M., Watson, J.A.: Microcomputers and preschoolers: Working together in classroom setting. *Comput. Schools* **3**, 53–61 (1986)
6. Clements, D.H., Nastasi, B.K., Swaminathan, S.: Young children and computers: Crossroads and directions from research. *Young Child.* **48**(2), 56–64 (1993)
7. Paris, C.L., Morris, S.K.: The computer in the early childhood classroom: Peer helping and peer teaching. In: *Paper Presented at the Meeting of the Microworld for Young Children Conference*, College Park, MD, March 1985
8. Perrin-Clermont, A.: *Social Interaction and Cognitive Development in Children*. Academic Press, New York (1980)
9. Muller, A.A., Perlmutter, M.: Preschool children’s problem-solving interactions at computers and jigsaw puzzles. *J. Appl. Dev. Psychol.* **6**, 173–186 (1985)
10. Vygotsky, L.S.: *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge (1978)
11. Lee, E.Y.: *Biological maturation, physical activity, and sedentary behaviour among Korean adolescents*. (Doctoral dissertation). Edmonton, Alberta: University of Alberta (2015)
12. Ahn, Y., Williamson, A.A., Seo, H.J., Sadeh, A., Mindell, J.A.: Sleep patterns among south korean infants and toddlers: global comparison. *J. Korean Med. Sci.* **31**, 261–269 (2016)
13. *Korean Statistical Information Service: Statistics on Internet use*. Statistics Korea, Daejeon, South Korea (2013)
14. Carson, V., Faulkner, G., Sabiston, C.M., Tremblay, M.S., Leatherdale, S.T.: Patterns of movement behaviors and their association with overweight and obesity in youth. *Int. J. Public Health* **60**(5), 551–559 (2015)
15. ICEF Monitor. High performance, high pressure in South Korea’s education system, 22 January 2014. <http://monitor.icef.com/2014/01/high-performance-high-pressure-in-south-koreas-education-system/>
16. Ahn, Y.I.: *The Effect of Sandtray Play on the Increase of Emotional Intelligence and Social Ability of Children*. Unpublished doctoral dissertation, Mokpo National University (2010)

Virtual Resources Allocation Scheme in ICT Converged Networks

Hyuncheol Kim^(✉)

Department of Computer Science, Namseoul University, Cheonan, Korea
hckim@nsu.ac.kr

Abstract. NFV enhancing the infrastructure agility, thus network operators and service providers are able to program their own network functions (e.g., gateways, routers, load balancers) on vendor-independent hardware substrate. One of the main challenges for the deployment of NFV is the efficient resource (e.g. virtual network function (VNF)) allocation of demanded network services in NFV-based network infrastructures. However, the effective mapping and scheduling of VNFs are essential to successfully provide NFV services. In this paper, we proposed revised online (dynamic) virtual network function allocation scheme to cope with successive network service (NS) requests. Unlike previous research on resource allocation, we assumed that each virtual node processes one or more functions at a time using multiprocessing technologies as in the real environment.

Keywords: Resource allocation · Virtual network · Network function virtualization

1 Introduction

With the advent of cloud and virtualization technologies and the integration of various computer communication technologies, today's computing environments can provide virtualized high-quality services. The network traffic has also continuously increased with remarkable growth. With such a huge trend, due to the flexibility and significant economic potential of these technologies, software defined networking (SDN) and network functions virtualization (NFV) are emerging as the most critical key enablers [1, 2]. As shown in Fig. 1, NFV enhancing the infrastructure agility, thus network operators and service providers are able to program their own network functions (e.g., gateways, routers, load balancers) on vendor-independent hardware substrate. They facilitating the design, delivery and operation of network services in a dynamic and scalable manner. NFV allows for the decoupling of physical network equipment from the services or functions that run on them, such that a given service can be decomposed into a set of virtual network functions (VNFs), which could then be implemented in software that can run on one or more industry standard computing nodes [3–5].

Funding for this paper was provided by Namseoul university.

© Springer Nature Singapore Pte Ltd. 2017

K. Kim and N. Joukov (eds.), *Information Science and Applications 2017*,

Lecture Notes in Electrical Engineering 424, DOI 10.1007/978-981-10-4154-9_87

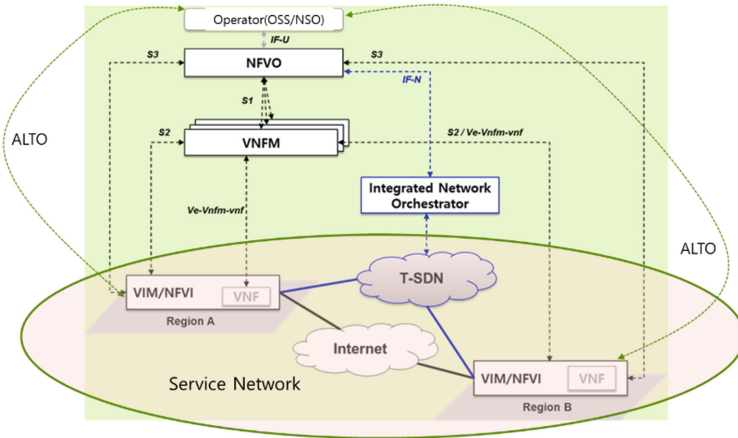


Fig. 1. Typical NFV network components and architecture

One of the main challenges for the deployment of NFV is the efficient resource (e.g. virtual network function (VNF)) allocation of demanded network services in NFV-based network infrastructures. However, the effective mapping and scheduling of VNFs are essential to successfully provide NFV services. In this paper, we proposed revised online (dynamic) virtual network function allocation scheme to cope with successive network service (NS) requests. Unlike previous research on resource allocation, we assumed that each virtual node processes one or more functions at a time using multiprocessing technologies as in the real environment [6–8].

The composition of this paper is as follows. We first highlight some mapping and scheduling challenges of VNFs in the NFV in Sect. 2. The proposed coordinated mapping and scheduling scheme of VNFs are presented in Sect. 3. Finally, the paper concludes in Sect. 4.

2 Related Works

Resource allocation in NFV requires efficient algorithms to determine on which high volume servers (HVSs) VNFs are placed, and be able to migrate functions from one server to another for such objectives as load balancing, reduction of CAPEX and OPEX, energy saving, recovery from failures, etc. [9].

In the NFV architecture framework the component that performs the resource allocation is the orchestrator. The orchestrator manages VNFs through the VNF manager (VNFM) and the virtualized infrastructure manager (VIM). The orchestrator evaluates all the conditions to perform the assignment of VNFs chains on the physical resources, leaning on the VNF managers and the virtualized infrastructure managers. The resource allocation in NFV has carried out in three stages: (1) VNFs Chain composition (VNFs-CC), also known in the literature as Service Function Chaining [1, 10] (2) VNF Forwarding Graph Embedding (VNF-FGE)2 and (3) VNFs Scheduling (VNFs-SCH). Next section deeply details the NFV-RA problem and its derived sub-stages.

Considering that NFV is still seen as a concept under investigation, little research has been conducted on VNFs scheduling. Riera et al. [4] provided the first formalization of the scheduling problem in NFV as a Resource Constrained Project Scheduling Problem. Recently, Mijumbi et al. [5] proposed an approach to tackle the online VNF-FGE and VNFs-SCH by proposing greedy and metaheuristic (tabu search) approaches aiming at reducing the flow execution time. The algorithms perform both mapping and scheduling at the same time (one-shot) resulting in high acceptance ratio, low average flow time and low embedding cost. This work considers a resource sharing approach that allows a given VM to process multiple VNFs, one after another (possibly) from a queue.

3 Proposed Virtual Network Function Allocation Scheme

The features of the proposed dynamic NFV-RA scheme in this paper are as follows. Unlike previous research on resource allocation, each virtual node processes one or more functions at a time using multiprocessing technologies.

- Support one or more VNF component instances (VNFCIs) that provide the same functionality within a VNF through scale out
- Online RA algorithm supporting integration of VNF-FGE and VNF-SCH
- Multi-tenancy support

To explain the proposed VNF allocation scheme, we define the following variables and functions.

- N : A set of all virtual nodes, $N = \{1, 2, 3, 4, \dots, n\}$
- n : The number of virtual nodes
- m : The number of VNFs
- S : NS, consists of m sequential VNF, $F = \{1, 2, 3, 4, \dots, m\}$
- $\rho_{i,j}$: Processing time of VNF i at node j
- δ_i : The buffer used by the node to which function i is mapped
- B_j : At some point, the available buffer size at node j
- $\beta_{i,j}$: 1 if node j can handle function i , 0 otherwise
- t_i : Deadline time for which service to process
- t_i : Completion time of VNF i
- t_c : Current time of VNF i
- t_a : The time at which mapping and scheduling requests for the service arrived on the physical network
- π_j : Expected completion time of the last function waiting for processing at the corresponding virtual node j
- μ_i : Start time of the first function waiting for processing at the corresponding virtual node j

The NFV-RA problem is to select a virtual node for $j \in N(i)$ each VNF i and to select a completion time t_i . Therefore, the NFV-RA problem is divided into two stages. First,

a virtual node $j \in N(i)$ (mapping problem) to which each VNF should be mapped is selected, and the order in which VNFs are executed in each node is determined (scheduling problem). Figure 2 shows the pseudo code of the proposed dynamic NFV-RA in this paper and Fig. 3 shows the VNFs supported by each node and each node constituting the network. Assume that the request for $S_1 = \{f8 - f2 - f3 - f6 - f5\}$ arrives at time $T1$. Then $S_2 = \{f6 - f8 - f4\}$ arrives. Figure 4 shows the scheduling map of the existing algorithm [5] and Fig. 5 shows the scheduling map proposed in this paper. Figures 4 and 5 show the scheduling map after S_2 has been allocated, respectively. For the proposed algorithm, $f8$ is performed at n_7 to reduce the queue latency at n_1 .

```

function vnf_resource_alloc (S,N,T) {
  //T = {speed, sharing}
  matrix[] []; //graph
  graph *[]; //pointer array of graph
  graph = setup_node (matrix)

  for (NF i ∈ S) {
    N' = {} //node initialization
    If (i == 1)  $t_{i-1} = t_a$  //time initialization

    for (Node j ∈ N) {
       $t_e = \rho_{i,j} + \max(\pi_j, t_{i-1})$ 
      If (( $\beta_{i,j} == 1$ )  $\wedge$  ( $B_j \geq \delta_i$ )  $\wedge$  ( $t_e \leq t_l$ )) then {
         $N' = N' \cup n$ 
        If (( $T == sharing$ )  $\wedge$  ( $t_i \neq \mu_j$ )) then {
          vnf_resource_alloc (NS( $\mu_j$ ),N,T)
        }
      }
    }
  }
  If (N' == 0)
    { Reset the substrate network; return1;}
  sort(N',T)
  mapping (i,select_top(N'))
  update ( $\pi_j$ )
   $t_i = \max(\pi_j, t_{i-1})$ 
  update ( $B_j, t_{i-1}$ )
}
}

```

Fig. 2. Pseudo code of the proposed VNF allocation scheme

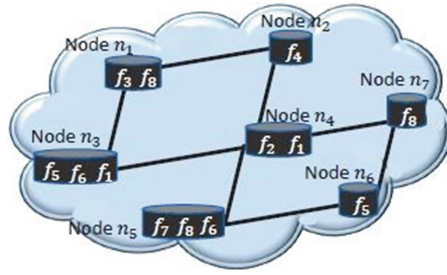


Fig. 3. Node capabilities and architecture of example node

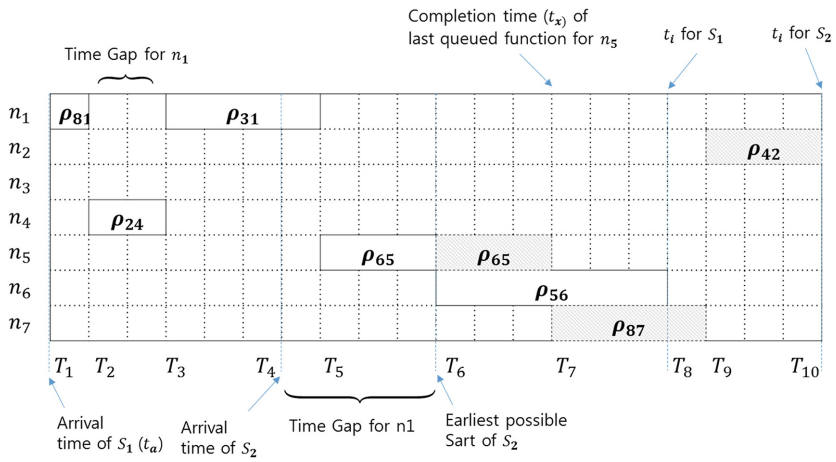


Fig. 4. Static virtual network function scheduling

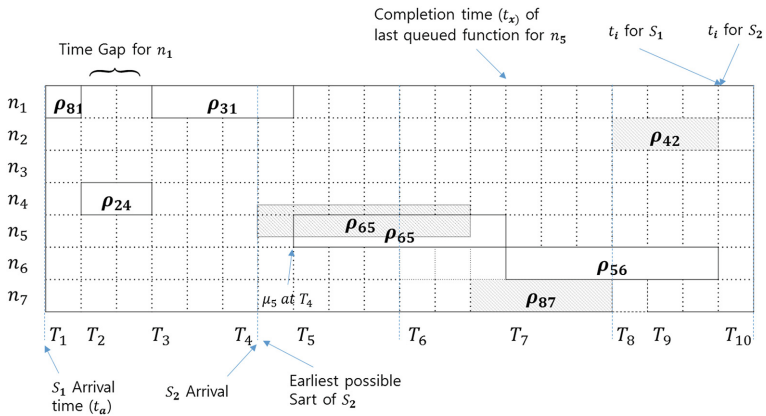


Fig. 5. Proposed virtual network function scheduling

Through the comparison of Figs. 4 and 5, it can be seen that the proposed algorithm reduces the execution time by one time unit.

4 Conclusion

One of the main challenges for the deployment of NFV is the efficient resource (VNF) allocation of demanded network services in NFV-based network infrastructures. However, the effective mapping and scheduling of VNFs are essential to successfully provide NFV services. In this paper, we proposed revised online (dynamic) virtual network function allocation scheme to cope with successive NS requests. Unlike previous research on resource allocation, we assumed that each virtual node processes one or more functions at a time using multi-processing technologies as in the real environment. Compared with existing algorithms, it can be seen that the proposed algorithm reduces execution time by 10% time unit.

References

1. Herrera, J.G., Botero, J.F.: Resource allocation in NFV: A comprehensive survey. *IEEE Trans. Netw. Serv. Manage.* **13**(3), 518–532 (2016)
2. Kim, H.: Network function virtualization (NFV) platform for wellness in high-speed network. In: Kim, K., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016*. LNEE, vol. 376, pp. 1459–1464. Springer, Heidelberg (2016). doi:[10.1007/978-981-10-0557-2_140](https://doi.org/10.1007/978-981-10-0557-2_140)
3. Mijumbi, R., Serrat, J., Gorricho, J., Bouten, N., De Turck, F., Boutaba, R.: Network function virtualization: state-of-the-art and research challenges. *IEEE Commun. Surv. Tutorials* PP(99), 1 (2015)
4. Riera, J.F., Hesselbach, X., Escalona, E., García-Espin, J.A., Grasa, E.: On the complex scheduling formulation of virtual network functions over optical networks. In: *International Conference Transparent Optical Networks (ICTON)*, pp. 1–5 (2014)
5. Beck, M.T., Botero, J.F.: Coordinated allocation of service function chains. In: *IEEE Global Communications Conference*, pp. 1–6 (2015)
6. Mijumbi, R., Serrat, J., Gorricho, J.-L., Latr, S., Charalambides, M., Lopez, D.: Management and orchestration challenges in network functions virtualization. *IEEE Commun. Mag.* **54**(1), 98–105 (2016)
7. Mijumbi, R., Serrat, J., Gorricho, J.-L., Bouten, N., De Turck, F.: Design and evaluation of algorithms for mapping and scheduling of virtual network functions. In: *IEEE Conference on Network Softwarization (NetSoft)*, pp. 1–9 (2015)
8. Kao, H.-Y., Yang, Y.-M., Huang, C.-H.: Dynamic virtual machines placement in a cloud environment by multi-objective programming approaches. In: *International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pp. 364–365 (2015)
9. ETSI: *Network Functions Virtualisation (NFV): Virtual Network Functions Architecture* (2014)
10. Munoz, R., Vilalta, R., Casellas, R., Martinez, R., Szyrkowicz, T., Autenrieth, A., Lopez, V., Lopez, D.: SDN/NFV orchestration for dynamic deployment of virtual SDN controllers as VNF for multi-tenant optical networks. In: *Optical Fiber Communications Conference and Exhibition (OFC)*, pp. 1–3 (2015)

Enhanced Metadata Creation and Utilization for Personalized IPTV Service

Hyojin Park¹, Kireem Han², Jinhong Yang^{3(✉)}, and Jun Kyun Choi²

¹ Department of Information and Communications Engineering, KAIST, Daejeon, South Korea
gaiaphj@kaist.ac.kr

² School of Electrical Engineering, KAIST, Daejeon, South Korea
hanrob@kaist.ac.kr, jkchoi59@kaist.edu

³ HECAS, Seoul, Republic of Korea
jinhong.yang@hecas.co.kr

Abstract. Metadata is the clue to connecting videos and individual users. For a successful personalization, collecting more detailed and concrete metadata is important. This paper proposes a method to create and utilize metadata more precisely as a result of research on how to do the personalization better. For that, we propose a way to consume and utilizing a video in a series of segmented images so that can get the user's taste on specific sections or points of the video. Proposed method is implemented on the web site and the effectiveness of proposed method is revealed by comparing the site stay time and number of video usage per session between new visitors and returning visitors.

1 Introduction

The role of IPTV service providers can be defined as connecting videos to users, and the success of connection can be judged as the number of video consumption. To promote the IPTV users' video consumption, personalization is appealing solution but the problem is how to do. Because providing videos tailored to the user's tastes can surely increase the video consumption, but improper personalization will make not just decreasing the video consumption but make users to leave the service.

To successfully display the videos custom to individual users, it is very important to catch the user's preference and taste on video contents both in general and for the moment correctly. Metadata, which contains additional, explaining, and helpful information around the videos, service, user, and the user's community activity, gives all the information required to analyze the user's appetite and generate the personalized program guide.

In this matter, this paper proposes an enhance metadata creation and utilization method, i.e., serving a video into an image carousel form. Extracting multiple key frame images with full script of the video and serving in image carousel form increases the potential and efficiency of the video linkage and usage especially in social network services. As shown in Fig. 1, by allowing users' to consume segmented image based video content and tracing their activity in image unit while saving them as specific

metadata, IPTV service provider can collect and utilize more detailed, concrete, and precise metadata.

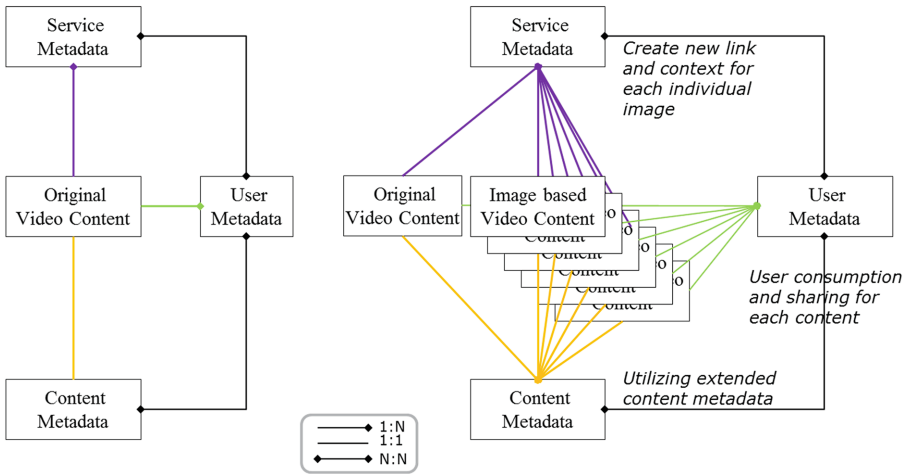


Fig. 1. Concept of the proposed method

Rest of this paper is composed as follows. Section 2 explains the proposed image carousel based video consuming method and the generation process. In Sect. 3, the list of newly created and extracted metadata elements by applying proposed method and their usage is introduced. Section 4 shows the implementation and service test results of the proposed method and concludes in Sect. 5.

2 Proposed Method for Generating Image Carousel

In this paper, we have studied how to divide video into several segments and package into a convenient form in order to understand user’s taste on video. In addition to consuming video in each video unit, we have invented a way to play, share, and utilize specific segments of video to create and reflect the new and more specific metadata while keeping the original video’s content.

Proposed method extracts keyframes representing each segment and to combine them into an image carousel. To ensure minimum loss in the original video content while saving playtime and data traffic, proposed method keeps the entire script automatically generated by audio mining which are actively researched owing to the development of the recent advances on audio recognition [1–3, 10, 12]. By doing so, users can grasp the content of the video without playing but reading and can share or play from any keyframe image on the image carousel. In addition, proposed method reduces the long play time and the demand for high data traffic, which are the disadvantage of existing video services.

Technically, serving video in image carousel can be regarded as one of the method in the video abstraction or video summary [4, 5, 11]. However, in the sense that proposed image carousel generation does not shorten the voice of the video but keeps the entire

story written in script, it is different from the existing approaches [6–8]. Also, generating image carousel can be regarded as one of the methods in the keyframe extraction [9]. However, proposed scheme extracts keyframes from the meaning unit which is distinguished from the script, not from the entire video or by image processing.

The process to generate the proposed image-carousel from a video is done by the following steps as shown in Fig. 2.

- Script generation: Generating script of the video by audio mining
- Script parsing: Time based script parsing and keyword extraction
- Video segmentation: Segmenting the video based on the maximum length of one script unit
- Keyframe extraction: Select a keyframe of each segment using key-word, expression, and extracting information
- Packaging: Overlay the script on the keyframe and packaging as image carousel format

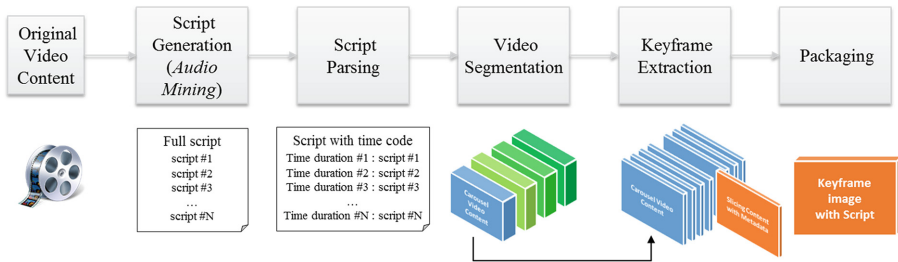


Fig. 2. Proposed image carousel generation process

3 Enhanced Metadata Creation and Utilization

When videos are served and consumed by proposed method, then new types of metadata elements are created from the video segmentation and extracted while the users are enjoying the image based video service. The list of new metadata elements are as follows:

- Newly created metadata elements from video segmentation
 - Script on video
 - Each script on time period
 - Keywords
 - Keywords on entire video
 - Number of times the keyword appeared in each segmented script
- Newly extracted metadata element from proposed service
 - Impressions by image
 - Inter-image exposure time
 - Image playback conversion frequency per image (conversion rate)
 - Number of shares per image

- Number of plays after conversion (conversion rate)
- Ad impressions by image
- Image-specific ad impression keywords
- Ad clicks by image (conversion rate)

Table 1 shows a comparison between the metadata provided by legacy service, i.e., YouTube and the metadata generated by the method of this paper. Also, the table is listing which new services are possible using the new metadata, not limited to but as typical examples including enhanced IPTV service personalization.

Table 1. Usage of newly created metadata

YouTube (Legacy service)	New metadata from proposed method	Advances and new service opportunities
Watch time (Estimated total audience watch time for videos)	<ul style="list-style-type: none"> ◆ Inter-image exposure time ◆ Keywords on entire video ◆ Number of times the keyword appeared in each segmented script 	<ul style="list-style-type: none"> ◆ Individual Ad service by enhanced advertisement matching ◆ Effective Ad insertion in the video
Views	<ul style="list-style-type: none"> ◆ Impressions by image ◆ Number of shares per image 	<ul style="list-style-type: none"> ◆ Effective Ad placement in service page
Average watch duration	<ul style="list-style-type: none"> ◆ Impressions by image ◆ Inter-image exposure time ◆ Each script on time period ◆ Number of times the keyword appeared in each segmented script 	<ul style="list-style-type: none"> ◆ Upgraded content recommendation ◆ Sophisticated service personalization
Average view rate	<ul style="list-style-type: none"> ◆ Impressions by image ◆ Image playback conversion frequency per image (conversion rate) ◆ Number of shares per image ◆ Number of plays after conversion (conversion rate) 	

When a video is played, the reason that the user likes the video can be different for actors, writers, directors, backgrounds, and so on. This difference in video consumption is hard to understand by tracking the consumption history on video units, but it can be better distinguished from the generation and consumption histories of the proposed specified metadata elements. The precisely seized user’s preference can be used to recommend videos and create a personalized IPTV service program guide that is more tailored to each user.

4 Implementation and Test Results

Figure 3 is the screenshot of the implemented web site to test the effectiveness of the proposed method. As shown in the figure, on each video has two overlaid buttons. When

a user chooses the left-hand side button, then the video is played, and when the right-hand side button is selected, then the video is served as the proposed image carousel. By running implemented service for one month, we could collect site stay time and content consumption information via google analytics. The drop-off rate of moving between contents is 21.55% based on the total users, 6.9 contents used during the visit, and the stay time is 5:57 min. When we divide this by the number of contents usage compared to the visit time, it can be understood that the use time for each piece of contents is about 51 s.

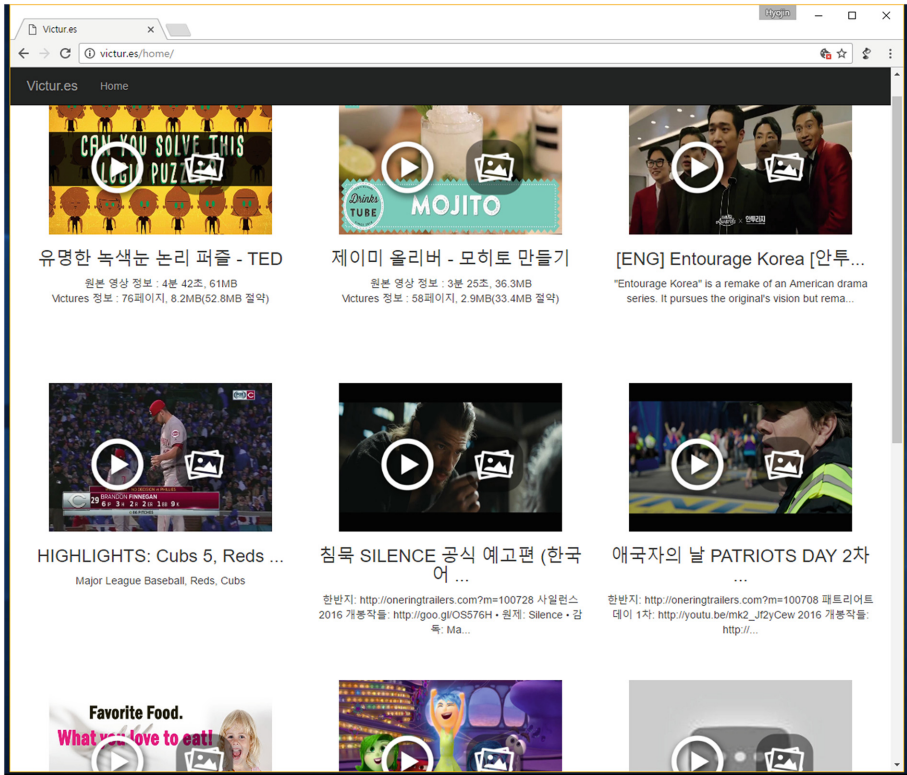


Fig. 3. Screenshot of the web site for running proposed method

The effectiveness of proposed method for making users to consume the more videos is revealed by comparing the site stay time and number of contents usage between new visitors and returning visitors. As shown in Table 2, from the aspect of drop-off rate, it can be seen that the drop-off rate of returning visitors is lower than that of new visitors. This implies that more content is consumed on the site by returning visitors, and that is confirmed by the fact that the number of content usage per session of returning users is more than 4 more than that of new visitors. Also, it was confirmed that the site stay time was very high as 9:01 min for returning visitors compared to the 3:45 min for new visitors. This data can directly or indirectly explain the satisfaction of return visitors to

the service and that shows the use of newly generated concrete metadata facilitates video consumption by video service visitors.

Table 2. Statistics on the implemented service with proposed method

	Session	Drop-off rate	Number of video usage per session	Average session time
All visitors	100%	21.55%	6.91	5:57 min
New visitors	58.16%	26.26%	5.2	3:45 min
Returning visitors	41.84%	15.00%	9.28	9:01 min

5 Conclusion

To link proper contents with different users, grasping the characteristics of contents and the interests of each user is very important. Proposed method for enhanced metadata creation and utilization for IPTV service personalization offers a new way of consuming video on IPTV. By providing multiple key frame images in carousel form with full script of the video, proposed method enables users can ‘read’ the content rather than ‘watch’. Segmenting the video and re-generate it in image carousel form allowed IPTV service providers to get the metadata on user’s preference and activity in specific sections or points and lead to the longer site stay time and the more video consumption as shown in the field test results.

Acknowledgement. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No: R-20160906-004163, Developing Bigdata Autotagging and Tag-based DaaS System).

References

1. Yi, H., Rajan, D., Chia, L.-T.: Semantic video indexing and summarization using subtitles. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 634–641. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-30541-5_78](https://doi.org/10.1007/978-3-540-30541-5_78)
2. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/script: alignment and parsing of video and text transcription. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5305, pp. 158–171. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-88693-8_12](https://doi.org/10.1007/978-3-540-88693-8_12)
3. Košút, M., Šimko, M.: Improving keyword extraction from movie subtitles by utilizing temporal properties. In: Freivalds, R.M., Engels, G., Catania, B. (eds.) SOFSEM 2016. LNCS, vol. 9587, pp. 544–555. Springer, Heidelberg (2016). doi:[10.1007/978-3-662-49192-8_44](https://doi.org/10.1007/978-3-662-49192-8_44)
4. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimed. Comput. Commun. Appl. **3**(1), 1–37 (2007)
5. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **41**(6), 797–819 (2011)
6. Kim, W., Kim, C.: A new approach for overlay text detection and extraction from complex video scene. IEEE Trans. Image Process. **18**(2), 401–411 (2009)

7. Wactlar, H.D., Kanade, T., Smith, M.A., Stevens, S.M.: Intelligent access to digital video: informedia project. *Computer* **29**(5), 46–52 (1996)
8. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. *Pattern Recogn.* **37**, 977–997 (2006)
9. Gianluigi, C., Raimondo, S.: An innovative algorithm for key frame extraction in video summarization. *J. Real-Time Image Process.* **1**(1), 69–88 (2006)
10. Liu, Y.-J., Ma, C.-X., Fu, Q., Fu, X., Qin, S.-F., Xie, L.: A sketch-based approach for interactive organization of video clips. *ACM Trans. Multimed. Comput. Commun. Appl.* **11**(1), 1–21 (2014)
11. Mei, T., Tang, L.-X., Tang, J., Hua, X.-S.: Near-lossless semantic video summarization and its applications to video analysis. *ACM Trans. Multimed. Comput. Commun. Appl.* **9**(3), 16–39 (2013)
12. Money, A.G., Agius, H.: ELVIS: entertainment-led video summaries. *ACM Trans. Multimed. Comput. Commun. Appl.* **6**(3), 17–47 (2010)

A Study of Teaching Plan for the Physical Activity Using ICT

Seung Ae Kang^(✉)

Department of Sport and Healthcare, Namseoul University, Cheonan, Korea
sahome@nsu.ac.kr

Abstract. Possibility of using ICT to the education of physical activities is becoming a reality due to the increased interest in the field of physical activities that applied technologies such as motion recognition or virtual reality. This study sought to present the possibility of using ICT and education model for the education on the physical activities. Education of physical activities faces a number of limitations. In particular, share of the time and space limitations is high. Virtual reality can be cited as the ICT technology that can complement these limitations. Sports virtual reality system that enables sports experience service with high sense of immersion since sports and fun elements get mixed together is likely to be used for the education of physical activities since it is possible to complement the fun element while complementing the time and space, and environment limitations. When physical activity task is delivered, the following process is performed in order; searching for information via web or VOD, setting up the strategy for learning sports techniques, executing virtual reality simulation and evaluating the accuracy level of motions, carrying out physical activities by using related virtual reality sports program, feedback. The teaching strategy by applying virtual reality technology will be effective for learning exercise function.

Keywords: ICT · Virtual reality · Physical activity · Education

1 Introduction

In case of the information society where knowledge and information are the core of individuals' and nation's competitiveness and that serve as the sources of value creation, individuals' ability to use information by searching and analyzing information by taking initiative, re-configuring information according to purpose and creating new information have become the basic capabilities required in life. Thus, there was a demand for making active attempts for the development of education method for the cultivation of competent human resources who are suitable for the information society [1]. Even in the education field where technological change's impact is relatively less significant, application of information communication technology is increasing

Funding for this paper was provided by Namseoul University.

© Springer Nature Singapore Pte Ltd. 2017

K. Kim and N. Joukov (eds.), *Information Science and Applications 2017*,

Lecture Notes in Electrical Engineering 424, DOI 10.1007/978-981-10-4154-9_89

incrementally. Numerous countries in the world are planning to strengthen support for the ICT education and striving to implement accordingly. Moreover, efforts are underway to develop and disseminate diverse contents for the use of the digital textbooks and learning materials. Transformation of the learning and teaching environment into a wireless one, and digitalization of the educational contents do not stop merely at using ICT devices. Instead, these trends can lead to the cultivation of the students' creativity, and problem-solving, communication and cooperation abilities through the two-way interaction between teachers and students. In case of the education on the physical activities, physical experiences take place through movements at the locations outside of the classrooms such as playground or gym. Thus, this was a field where use of ICT may be difficult [2]. However, possibility of using ICT to the education of physical activities is becoming a reality due to the increased interest in the field of physical activities that applied technologies such as motion recognition or virtual reality. Accordingly, this study seeks to present the possibility of using ICT and education model for the education on the physical activities.

2 Education and ICT

Information communication technology, ICT (Information & Communication Technology) is the compound word of Information Technology and Communication Technology. When interpreted narrowly, it means, "hardware and software for searching, collecting and delivering information." However, when interpreted broadly, it refers to "all the methods related to the collection, production, processing, preservation, delivery and use of information by using hardware and software technologies and all these technologies." [3]. Diverse changes are in demand to adapt to the rapid changes of the knowledge information society in the entire society. Use of the ICT is increasing worldwide in the field of education as well to cultivate competent human resources. Use of the ICT in the Korean education field entered into the era of smart education after undergoing the infra development and e-learning generalization process, starting from the 'informatization of education' in 1996. Figure 1 shows the development history of the education information using ICT in stages.

In the beginning, ICT was applied to the combination of education contents such as Internet lectures and e-learning, and lecture system. Later, services such as Clouding Computer, Big Data analysis and Artificial Intelligence technology and flipped learning that strengthened interaction were enabled (Fig. 2) [2].

	Stage 1 (1996~2000)	Stage 2 (2001~2005)	Stage 3 (2006~2010)	Stage 4 (2010~2014)
Purpose	Infra development	Vitalization of the education using ICT	Infra development	Country with powerful human resources, education science and technology
Direction pursued	<ul style="list-style-type: none"> Establishment of the stable foundation for the pursuit of education informatization business Development of the world-class education informatization infra 	<ul style="list-style-type: none"> Acceleration of the ICT use and generalization of the e-learning in the teaching and learning activities Adoption of the informatization system in the field of the higher education's administrative field 	<ul style="list-style-type: none"> Support for the education customized to individuals Use of ubiquitous technology Expansion of the informatization extension into the higher education and lifelong education fields 	<ul style="list-style-type: none"> Software power cultivation Adoption of the digital learning ecology system Development of the communication and converged governance Active liberalization and sharing
Key performance	<ul style="list-style-type: none"> Start of the Edunet, which is the education information service system Spread of the educational computer and Internet 	<ul style="list-style-type: none"> Expansion of the education informatization's scope (elementary, middle and high schools → higher education, lifelong education, special education and education for the gifted children) 	<ul style="list-style-type: none"> E-learning → U-learning Education informatization, international consulting, and export of e-learning education 	<ul style="list-style-type: none"> Development of the strategy to pursue after smart education

Fig. 1. Transition of education informatization using ICT (Source: KERIS, 2014) [4]

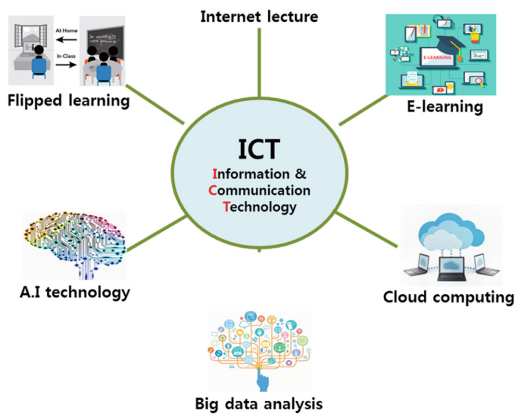


Fig. 2. Type of ICT application in education

Flipped learning is the education method in which teacher records the contents for the students to study in the form of VOD, and the VOD is provided to the students. Then, the students watch the VOD at home or school prior to the class time via Internet or TV. Then, it is an education method for carrying out discussions, experiments and in-depth

Q&A centered learning regarding the contents of the VOD during class time. As such, this can be used extensively in school education.

3 Value of Using ICT in the Education on the Physical Activities

In the sports curriculum, reality of the ICT use is at a very elementary stage. Some of the root causes are teachers' lack of the ability to use information devices and lack of the structured training. Compared to this environment reality, PE teachers' expectation on the use of the ICT is very high. Value and impact of using ICT for the education of physical activities are as follows [1].

- Learning accurate motions: Accurate motions are very important for carrying out physical activities effectively. Sequence of physical activities appears instantly. Thus, it is not easy to recognize the accurate motions by observing with the naked eyes. Accordingly, use of the image devices such as digital camera enables observation and analysis of multi-dimensional motions.
- Supply of detailed and instant feedback: Result of undesired execution ability results when the students practice continuously using an incorrect method when it comes to the learning about physical activities. Motion correction and learning are made easy through students' motion analysis, and visual feedback via specific and instant feedback.
- Teaching proxy enabled at the situation when the demonstration cannot be carried out in actuality: Hazard of unexpected situation that may result during physical activities is difficult to produce or to show in actuality. This can be demonstrated or produced by searching the Web, and by using graphic and simulation software.
- Diverse changes in the environment and conditions for physical activities: Main spaces for the physical activities such as playground, gym and other space that is needed depending on the type of physical activities is sensitive to weather change. Use of the ICT is resourceful for overcoming this environment and condition limitations. Physical activity simulation in the virtual space can overcome limitations of time and space. Reduction in the labor cost and expenses results since dynamic environment and physical activities can be provided.
- Motivating and increase in the learning speed: Since use of ICT enables supply of vital data to the students, enables recognition of the detailed physical activity related tasks, and instant feedback, this can lead to interest and fun, which helps to improve learning ability.

4 Physical Activity Education Model Using Virtual Reality

As examined above, education of physical activities face a number of limitations. In particular, share of the time and space limitations is high. In fact, some of the key reasons that the PE classes are not carried out effectively in elementary, middle and high schools are weather, place and decrease in students' interest. Virtual reality can be cited as the ICT technology that can complement these limitations. Adoption of virtual reality

technology is becoming active in the sports field. Sports virtual reality system can be classified into four areas, and can be applied in a diverse manner according to the education goal (Table 1).

Table 1. Product classification of sports virtual reality system

Major field	Major field	Major field	Major field
Virtual reality sports system	Virtual reality sports of the sensory game type	Golf simulator, Baseball simulator, Tennis simulator, Shooting simulator, Virtual martial arts, Virtual clay shooting, Realistic table tennis game, Dance simulator	Virtual reality system that realized all types of sports into game from with the sensory type of interaction based on the recognition of users' motions
	Virtual reality on sports training	Running simulator, Cycling simulator, Rowing simulator, Archery simulator, Cyber fitness, Physical fitness estimation simulator, Cyber sport lesson	Virtual reality system that enables user to verify the effect on the body parts during exercise by experiencing science principles such as somatology/kinematics that is hidden in the sports, and that enables finding the optimized training method through sports players' sports ability and pattern analysis
	Rider type virtual sports	Bike simulator, Ski simulator, Snowboard simulator, Racing sports, Riding simulator, Bungee jump simulator	sensory type of sports simulation, developed based on the motion base for delivering sense of exercise to the users
	Functional virtual reality sports	Training and sports system, Customised exercise rehabilitation assistant system, Exercise assistance system for the disabled	Virtual reality sports for using for special purpose that entails adding on the diverse resourcefulness (education, PR, treatment and rehabilitation, military drill, and mind control) in addition to the entertainment's fun element

According to the research conducted by Son (2001), students are interested in the use of the ICT for the practical PE training class and that their interest increases more

than textbook oriented education [5]. Sports virtual reality system that enables sports experience service with high sense of immersion since sports and fun elements get mixed together is likely to be used for the education of physical activities since it is possible to complement the fun element while complementing the time and space, and environment limitations.

When physical activity task is delivered, students search information via web or are provided with information via VOD, and set up the strategy for learning sports techniques by carrying out team discussions. During this process, it is advised that the teachers provide feedback on the information’s appropriateness and cooperative behavior. Later, accuracy level of motions is evaluated by executing virtual reality simulation that enables interaction for accurate motion learning. Physical activities by individuals or teams are carried out by using related virtual reality sports program, and improved exercise function learning is enabled with provided feedback (Fig. 3).

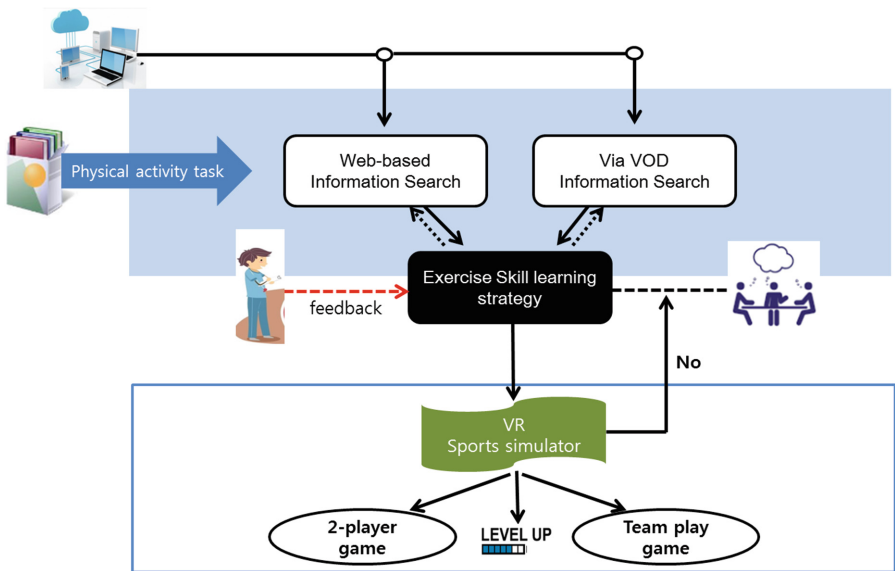


Fig. 3. Model of physical activity education using VR

5 Conclusion

As the smart world is emerging, use of the ICT is expanding in diverse fields. Education field is no exception. In particular, adoption of ICT technology is effective for the sports or exercising to complement environment limitations since there are many environment limitations such as space and weather. When virtual reality is applied to the physical activities to provide adequate feedback, it is possible to benefit from the educational effect by inducing active participation along with the fun element. This study searches for the strategy for learning exercise function after physical activity assignment is notified. Moreover, simulation is carried out by applying ICT technology. Possibility of

applying ICT during the PE class or physical activity education process was suggested via the model for executing processes such as feedback supply, team assignment and game.

References

1. Korea education and research information service. Training material for scholarship support staff on ICT utilization (2001)
2. Kang, S.Y., Kang, S.A., Jung, H.S.: Study on ICT utilization contents for physical education. *J. Inf. secur.* **16**(5), 17–22 (2016)
3. Cha, B.K.: The convergence of ICT and education. The Korean Institute of Communications and Information Science in Autumn Conference, pp. 297–298 (2015)
4. Korea education and research information service. Study on the establishment of ICT convergence digital learning ecosystem (2014)
5. Son, J.G.: The applicability of information communication technology in elementary school physical education class. *Korean J. Elementary Phys. Educ.* **7**(2), 69–78 (2001)

Design and Implementation of Headend Servers for Downloadable CAS

Soonchoul Kim¹, Hyuncheol Kim², and Jinwook Chung¹✉

¹ Department of Computer Engineering, Sungkyunkwan University, Suwon, Korea
{choulsim, jwcheong}@skku.edu

² Department of Computer Science, Namseoul University, Cheonan, Korea
hckim@nsu.ac.kr

Abstract. This paper presents the design and implementation of headend servers for a downloadable conditional access system (DCAS) that can securely transmit CA code via a broadband channel. To design DCAS headend server, we define core functions to be performed in the headend and categorize them into four sections such as authentication, provisioning, personalization and key management. In order to verify the stability and effectiveness of the implemented headend servers, we construct a testbed using them and a legacy cable headend system in a laboratory. The experimental results show that the DCAS headend servers are well designed.

Keywords: DCAS · Downloadable CAS · DCAS network protocol

1 Introduction

The Conditional Access (CA) application modules, which execute methodology to extract the secured keys for descrambling and decryption, are locked-in into STBs as unchangeable elements. DCAS uses a secure microprocessor (SM) soldered onto a circuit board instead of a removable. DCAS technology can remove the lock-in issue for CAS or STB vendors, and it is also much more flexible and easier to manage and distribute a CAS module onto an STB [1]. As shown in Fig. 1, the DCAS headend server communicates with a trusted authority (TA) to authenticate an STB which accesses to a cable network. The DCAS host includes a secure micro (SM) and a transport processor (TP) for supporting DCAS [4]. The SM performs a DCAS protocol and stores a CA code transmitted from a download server. The TP is used for decryption of video encrypted by the CA code [2].

In this paper, we improve the previously proposed protocol for reducing the processing time of DCAS messages without degrading a security level and apply it to a DCAS headend server. In Sect. 2 we analyze the security vulnerability about the previous proposed DCAS network protocol and describe the improved protocol to solve known problems. Also, it includes an efficient session control management. In the Sects. 3 and 4, the design and implementation of DCAS headend server are described and experimental results are provided respectively. Finally, the conclusion is followed.

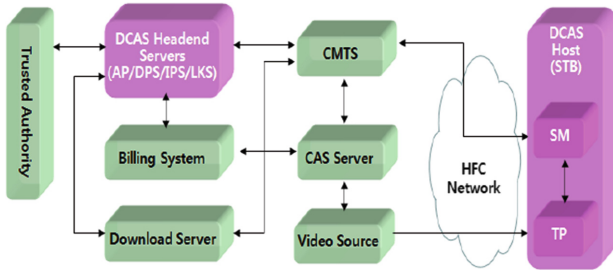


Fig. 1. System architecture for DCAS in HFC network.

2 Improved DCAS Protocol and Session Control Management

A DCAS network protocol requires the mutual authentication between DCAS headend server and DCAS host so that two parties establish a session key securely. The session key is used for content encryption and key transmission. The designed protocol is based on pre-shared key and related in 3rd party authentication among TA, AP, and SM. Accordingly, it is very important that the needed key materials among them should be delivered without an exposure. We estimated the secrecy of the key establishment phase on a formal security analysis tool of Burrows-Abadi-Needham logic [6].

The designed DCAS protocol include the initial assumptions which both TA and SM share a pre-shared key(K_i), and Session_ID, NONCE_SM, RAND_TA, K_c are randomly generated every session, and an SSL session key(KBS) is applied between TA and AP.

$$A \equiv A \xleftrightarrow{K_{AB}} B \quad B \equiv A \xleftrightarrow{K_{AB}} B \tag{1}$$

$$A \equiv B \equiv A \xleftrightarrow{K_{AB}} B \quad B \equiv A \equiv A \xleftrightarrow{K_{AB}} B \tag{2}$$

Let's generalize the protocol messages related to session key generation in order to verify (1) and (2). In (3) B can decrypt the message using B's private key as shown in (9). We can get (10) by message syntax rule from the initial assumption and (9). In (10) B cannot confirm the freshness of RA, therefore logical postulates cannot be applied to (3) no more. Because (4) and (5) are communicating over a known SSL protocol and KBS is secured, they are verified for themselves.

In (6), A can decrypt the message using A's private key as shown in (11). We can get (12) by message syntax rule from the initial assumption and freshness rule. Also, we can get (13) from nonce verification rule, and additional logical postulates. We can get (14) by belief rule in (13). In the same method, (7) and (8) can be concluded in (15) and (16) respectively.

$$\text{AuthRequest message: } A \rightarrow B \tag{3}$$

$$\{R_A, \{ID_A\}_{K_S}\}_{K_B}, \{R_A, \{ID_A\}_{K_S}\}_{K_A^{-1}}$$

$$\text{AuthRequest message: } B \rightarrow S \quad (4)$$

$$\{\{ID_A\}_{K_S}\}_{K_{BS}}$$

$$\text{AuthResponse message: } S \rightarrow B \quad (5)$$

$$\{R_{S1}, R_{S2}\}_{K_{BS}}$$

$$\text{AuthResponse message: } B \rightarrow A \quad (6)$$

$$\{R_A, R_{S1}\}_{K_A}, \{R_A, R_{S1}\}_{K_B^{-1}}$$

$$\text{SKeyShare message: } A \rightarrow B \quad (7)$$

$$\{R'_A, R_A\}_{K_B}, \{R'_A, R_A\}_{K_A^{-1}}$$

$$\text{SKeyShareConfirm message: } B \rightarrow A \quad (8)$$

$$\{R_A, A \xleftrightarrow{K_{AB}} B\}_{K_{AB}}, \{R'_A, \{R_A, A \xleftrightarrow{K_{AB}} B\}\}_{K_B^{-1}}$$

$$B \triangleleft R_A, \{ID_A\}_{K_S} \quad B \triangleleft \{R_A, \{ID_A\}_{K_S}\}_{K_A^{-1}} \quad (9)$$

$$B| \equiv A| \sim (R_A, \{ID_A\}_{K_S}) \quad (10)$$

$$A \triangleleft R_A, R_{S1} \quad A \triangleleft \{R_A, R_{S1}\}_{K_B^{-1}} \quad (11)$$

$$A| \equiv \#(R_A, R_{S1}) \quad A| \equiv B| \sim (R_A, R_{S1}) \quad (12)$$

$$A| \equiv B| \equiv (R_A, R_{S1}) \quad A| \equiv B| \sim R_{S1} \quad (13)$$

$$A| \equiv B| \equiv R_{S1} \quad (14)$$

$$B| \equiv A| \sim R'_A \quad (15)$$

$$B| \equiv \#(A \xleftrightarrow{K_{AB}} B) \quad (16)$$

But because B cannot confirm that A trusts K_{AB}, logical postulates cannot be progressed further. By applying BAN logic to the designed protocol, it has been found that AP server needs the fact that SM module trusts the session key (K_{AB}). Accordingly, the proposed DCAS protocol need be improved about known vulnerabilities. To solve it, a SKeyShare message includes a hash value of the session key which is generated by SM module. The SM's session key is regarded as a temporary key until it is confirmed from AP server.

AP server verifies the own hash value of the session key which generates together key materials extracted from the SKeyShare message. The DCAS network protocol is to accomplish session key share in order to deliver SM client for safe keeping onto STB. Figure 2 shows the timing that session keys are generated, validated, and destroyed. Firstly, the announcement phase (DCASAnnounce, DPRRequestInfo) is to inform MSO's DCAS region and information needed to invoke terminals.

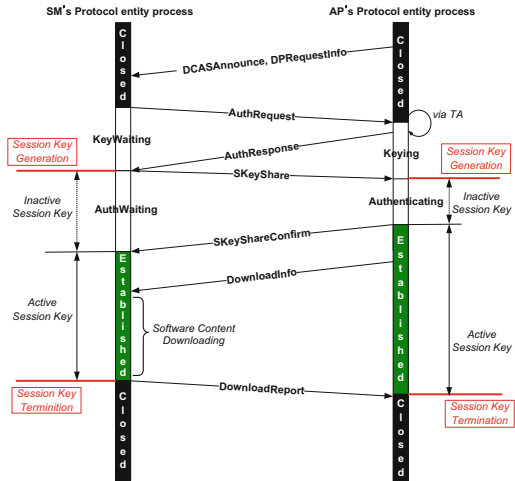


Fig. 2. Session request and response in DCAS network protocol.

The AP periodically sends out announcement messages to STBs including SM. Secondly, the keying phase (AuthRequest/AuthResponse) is to request, register, and pair keying information among SM, AP, and TA. The trust for keying information is guaranteed via TA. Thirdly, the authentication phase (SKeyShare/SKeyShareConfirm) is to establish the secured channel.

3 Design and Implementation of DCAS Headend Servers

A configuration of DCAS headend is shown in Fig. 3. The DCAS headend consists of four servers, authentication proxy (AP), DCAS provisioning server (DPS), the local key server (LKS) and integrated personalization server (IPS). The AP authenticates SM in a DCAS host according to DCAS protocol and generates several encryption keys. The DPS manages an SM configuration and distributes a policy for CA download. The LKS stores all keying information in the headend and provides a secure interface for the

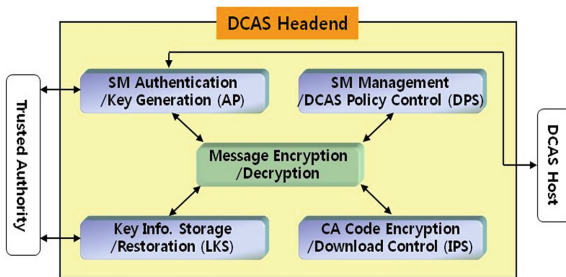


Fig. 3. The configuration of DCAS headend.

retrieval of key records. The IPS encrypts a CA code using the key delivered from AP and controls a mechanism for downloading.

The AP server includes five function blocks as shown in Fig. 4. The DPPB performs DCAS protocol for the mutual authentication and for downloading a CA code. To reduce the processing time of DCAS messages without degrading a security level, we apply an HMAC instead of an RSA signature for message authentication and add several parameters to strengthen a security level. The SCB controls a creation and deletion of session related to a DCAS host. The ABM authenticates a DCAS STB with TA using information transferred from SCB and transmits the result to an SCB. The KMB stores keying information into DB and simultaneously transmits it to LKS. The DCB delivers a download command to IPS using download policy received from DPS when an authentication process is finished.

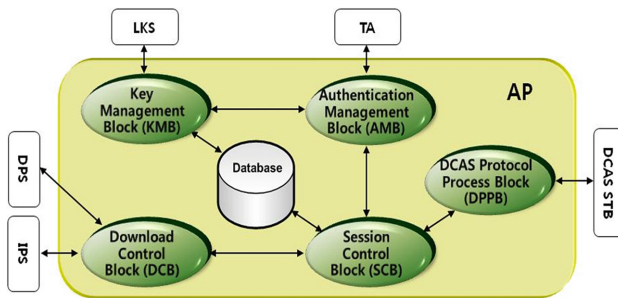


Fig. 4. Block diagram of AP server

4 Experimental Results

By testing various conditional environments, we verified that any SM or AP does not generate any unexpected state under even random message flow. The experimental conditions are as follows:

- Condition – 1: the STB with virgin state SM attaches into DCAS network, and then authenticates with AP. Successfully the STB downloads and runs SM client.
- Condition – 2: the STB with authenticated SM receives update info of SM client from AP, and then authenticates with AP.
- Condition – 3: the STB with authenticated SM receives update info of SM client from AP, and then stops in process of authentication with AP.
- Condition – 4: the STB with unauthorized SM tries authentication request in order to obtain SM client, and then sends a fake info to AP.

Figure 5 shows four DCAS headend servers. In the experiment test, a CA code was downloaded successfully from the DCAS headend to the DCAS host according to the DCAS protocol and it was installed in the SM automatically.

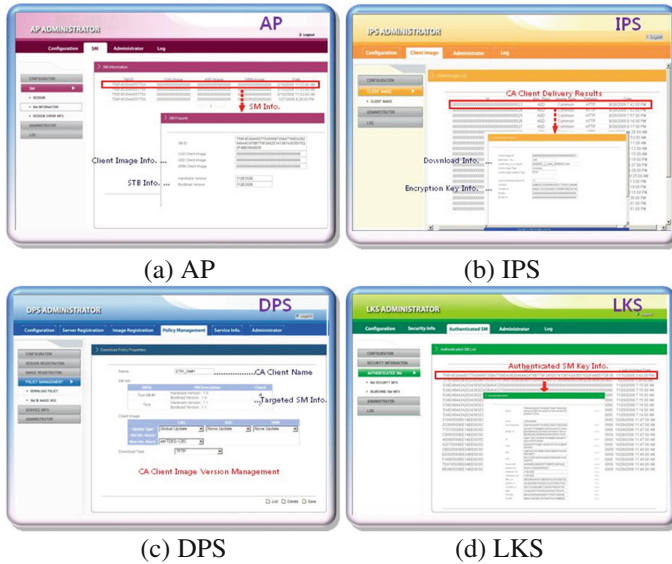


Fig. 5. Implemented DCAS headend servers

5 Conclusion

In this paper, we designed and implemented DCAS headend servers for downloading a CA code securely via a broadband channel. According to the categorized core functions into four sections, four headend servers for authentication, provisioning, personalization and key management was designed and implemented. In order to reduce the processing time of DCAS messages without degrading a security level, we improved the previously proposed DCAS protocol and applied it to a DCAS headend server. For verifying the stability and effectiveness of the implemented headend servers, we construct a test-bed using them and a legacy cable headend system in a laboratory. Through experimental results, we confirmed that our enhanced DCAS protocol and the implemented DCAS headend servers can operate stably, securely, and effectively.

References

1. Lookabaugh, T., Fahrny, J.: Openness and secrecy in security systems: polycipher downloadable conditional access. In: The Cable Show Conference, May 2007
2. Jeong, Y., et al.: A noble protocol for downloadable CAS. *IEEE Trans. Consum. Electron.* **54**(3), 1236–1243 (2008)
3. Xu, J., Feng, D.: Security flaws in authentication protocols with anonymity for wireless environments. *ETRI J.* **31**(4), 460–462 (2009)
4. Koo, H.-S., et al.: Key establishment and pairing management protocol for downloadable conditional access system host devices. *ETRI J.* **32**(2), 204–213 (2010)

5. Chen, T.-H., Shih, W.-K.: A robust mutual authentication protocol for wireless sensor networks. *ETRI J.* **32**(5), 704–712 (2010)
6. Burrows, M., Abadi, M., Needham, R.M.: A logic of authentication. *Proc. Roy. Soc. London A* **426**, 233–271 (1989). A preliminary version appeared as Digital Equipment Corporation Systems Research Center report No. 39 (1989)

A Method of Modeling of Basic Big Data Analysis for Korean Medical Tourism: A Machine Learning Approach Using Apriori Algorithm

Jun-Ho Huh¹, Han-Byul Kim², and Jinmo Kim^{1(✉)}

¹ Department of Software, Catholic University of Pusan, Busan, Republic of Korea
{72networks, jmkim11}@cup.ac.kr

² Department of Computer Engineering,
Pukyong National University, Busan, Republic of Korea
knb6513@naver.com

Abstract. The Republic of Korea (ROK) has emerged as a country of superior medical tourism in the last decade among the people of China, Japan, Southeast Asia, Russia and the Middle East for the plastic surgery or others requiring precision skills. Although the ROK's medical tourism industry grew quantitatively in its revenue and the number of visitors, the report from the 2015 World Economic Forum concerning the competitiveness of ROK's tourism including the medical tourism showed that its rank had dropped to 29th position, a drop of 4 places from 25th in 2013. Thus, it is about time to improve the situation by investigating the actual conditions of tours taken by the foreign tourists to establish new strategy, which is the main contribution of this research. As a research method, a big data analysis has been performed on the basis of machine learning and using R-studio. During the analysis process, there were some relevant regularities which were difficult to be found in the big data and based on these findings, we have attempted to find the solutions for the bad images that foreign visitors had shared in common. The result of the big data analysis showed that their purpose of visit was different from each other depending on the age groups and the details of their experience of inconvenience varied as well.

Keywords: Big data · R · Machine learning · Medical tourism · Apriori algorithm

1 Introduction and Background of Analysis

The Republic of Korea (ROK) has emerged as a country of superior medical tourism in the last decade among the people of China, Japan, Southeast Asia, Russia and the Middle East for the plastic surgery or others requiring precision skills [1–6]. Although the ROK's medical tourism industry grew quantitatively in its revenue and the number of visitors, the report from the 2015 World Economic Forum concerning the competitiveness of ROK's tourism including the medical tourism showed that its rank had dropped to 29th position, a drop of 4 places from 25th in 2013. Thus, it is about time to improve the situation by investigating the actual conditions of tours taken by

the foreign tourists to establish new strategy. For this purpose, a big data analysis has been performed on the basis of machine learning and using R-studio. During the analysis process, there were some relevant regularities which were difficult to be found in the big data and based on these findings, we have attempted to find the solutions for the bad images that foreign visitors had shared in common.

2 Bigdata

The big data used in this study was made by referring [7] to ‘The Survey on the Public Tours’, ‘The Survey on Foreign Tourists’ and ‘A basic statistics on Tourist Service Companies’ provide by the Korea Culture and Tourism Institute. Individual data indicates the responses for the survey carried out for the foreign tourists from 2009 to 2015, where the questions (variable contents) and responses are represented as variable names and variable values, respectively. Figure 1 shows the big data of 2015 survey on the foreign tourist.

Variable Name	Variable Contents/Questionnaire	Variable Value/Responses
q1	Q1. Revisiting Korea?	1: First time 2: Revisiting
q1a	Q1-1. Number of Revisits	1: Once 2: Twice 3: Three times 4: More than three times 9: Not sure/No response
wq1a	Q1-1 Number of Revisits_original Data	Original Data
q2a	Q2. Visiting Korea only	1: Korea only 2: Right before Korea 3: Right after Korea
q2b1	Q2. Country visited right before Korea	1. Japan, 2. China, 3. Hong Kong, 4. Singapore, 5. Taiwan, 6. Australia, 7. Thailand, 8. USA, 9. Canada, 10. UK, 11. Germany, 12. France, 13. Russia, 997. other
q2c1	Q2. Country visiting right after Korea	
q3	Q3. How early did you make a decision to visit Korea?	1. 1 month, 2. Two months, 3. 3-4 months, 4. 5-9 months, 5. More than 9 months
mq3	Q3. How early did you make a decision to visit Korea?	
q4	Q4. Did you consider visiting other country(s) before the trip?	1: Korea was the primary destination 2: Compared with others and finally chose Korea

Fig. 1. The big data of 2015 survey on the foreign tourist.

2.1 Data Integration

Figure 2 shows the integrated variables after Variable-conversions, the data in 2014 & 2015 have been integrated. The 2014 Survey on Foreign Tourists’ is a big data consisting 205 variables and survey size of 12,888 and the same survey in 2015 consists of 209 variables and 12,02 survey size. Their variables are changed and the categories are integrated to analyze these data acquired in both years.

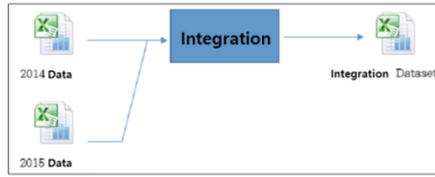


Fig. 2. The integrated variables after variable-conversions (data in 2014–2015).

2.2 Data Analysis and Results

Figure 3 shows the statistics where a represents the number of visits by foreign tourists and b is the proportion of foreign tourists’ visits in each year, while c is the number of visits of foreign tourists by country and d represents the main purpose of foreign tourists’ visits.

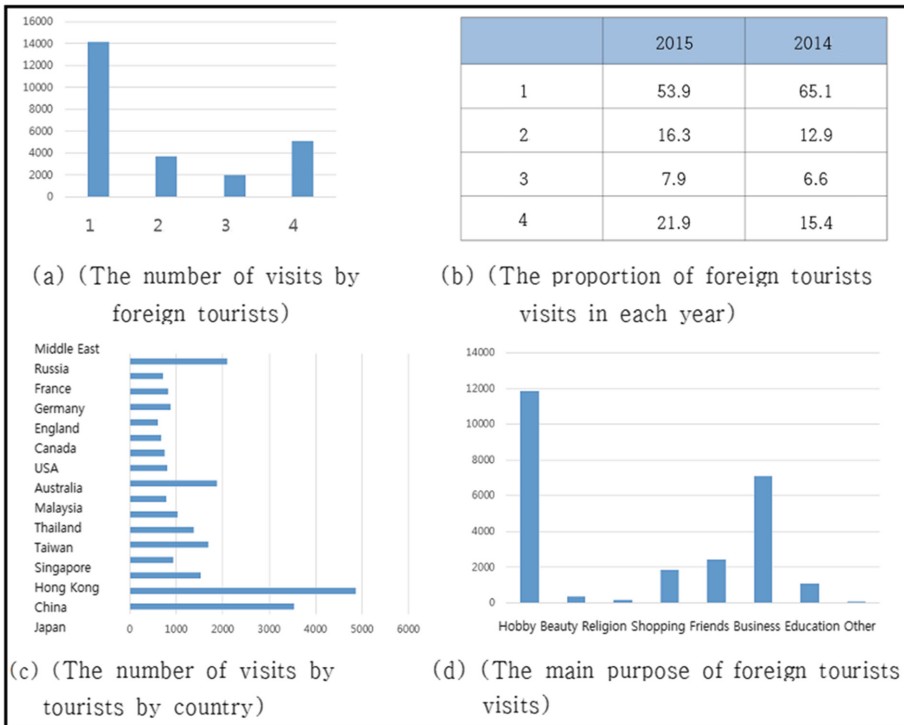


Fig. 3. Statistical analysis.

3 A Machine Learning Approach Using Apriori Algorithm

The explosive growth of data (Big Data) makes the data analysis technique more sophisticated and require additional computing power. As a result, a larger and more interesting data can be gathered and it largely affects the machine learning field. Thus, in this data analysis, we have applied the machine learning process using R-Studio.

3.1 Apriori Algorithm

This algorithm is a kind of machine learning algorithm which is to find and learn the association rules between the data. Analysis of association rules is used to find the some a number of interesting associations between a large number of variables. That is, this algorithm is quite suitable to the big data-oriented analysis so that one can obtain the results that are easily comprehensible. Also, the algorithm is useful for the data mining operations and able to find some unexpected knowledges in the database.

3.2 Data Gathering

As in Fig. 4, an integrated dataset (2014–2015) has been retrieved using R-Studio. In machine learning it is required to exclude the first variable 'id' as this variable can cause an erroneous prediction in the machine learning process so that the model that includes identifiers could become 'Overfitting' and do not generalize with other data.

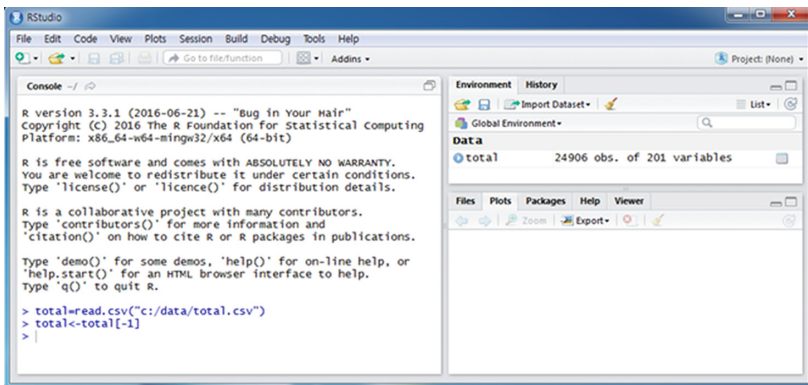


Fig. 4. An integrated dataset (2014–2015) has been retrieved using R-Studio.

3.3 Data Preparation and Exploration

As in Fig. 5, R reads data as an ordinary matrix. Several variables will be created to store data in a matrix but this method could cause handling of unintended rules in the machine learning process. Such a problem can be solved by the data structure called 'Sparse Matrix'.

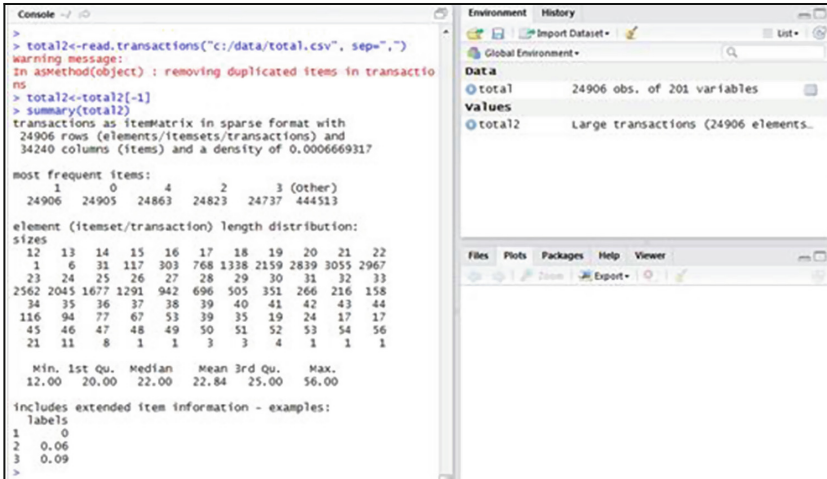


Fig. 5. Generation of a sparse matrix for the data.

3.4 Data Preparation and Exploration

As in Fig. 6, the 'Arules Package' provided by R-Studio was used for the machine learning. The Apriori () function, which was to find the association rules, read the data in a 'total2' sparse matrix and generated the rules based on the respective minimum values of support and confidence. As the difference in the number of rules that can be generated based on the values of support and confidence could be quite large, some adequate values were used for support and confidence.

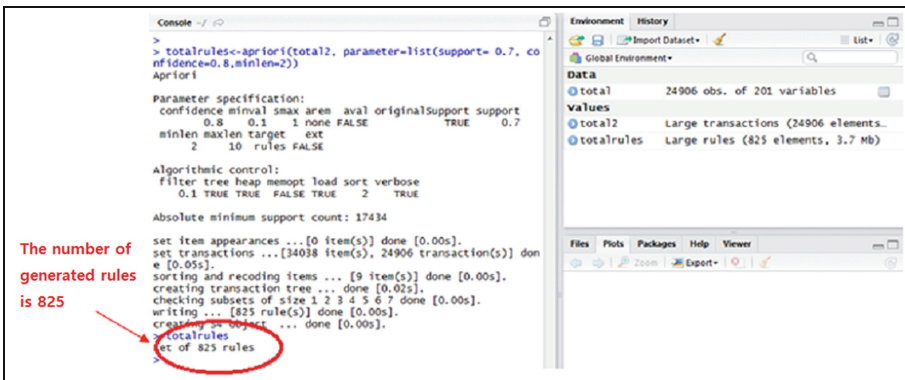


Fig. 6. Training with data-applied model

3.5 Performance Evaluation for the Model

As in Fig. 7, the summary statistics of an object ‘totalrules’ can be checked with Summary () function. There were 825 association rules. The specific rules can be viewed with Inspect () function. For example, inspect (Totalrules [10:20]) will show 10th to 20th rules. Here, the 10th rule has approx. 0.79 support level and 0.99 confidence level.

```

> summary(totalrules)
set of 825 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5  6  7
54 165 260 225 102 19

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.000  3.000  4.000  4.258  5.000  7.000

summary of quality measures:
  support      confidence      lift
Min.   :0.7093   Min.   :0.8129   Min.   :0.9964
1st Qu.:0.7675   1st Qu.:0.9928   1st Qu.:1.0000
Median :0.7939   Median :0.9976   Median :1.0000
Mean   :0.8170   Mean   :0.9817   Mean   :1.0004
3rd Qu.:0.8107   3rd Qu.:0.9999   3rd Qu.:1.0006
Max.   :1.0000   Max.   :1.0000   Max.   :1.0061

mining info:
 data ntransactions support confidence
total2      24906      0.7      0.8
    
```

```

> inspect(totalrules[10:20])
  lhs  rhs support confidence lift
10 {6} => {4} 0.7921384 0.9982291 0.9999555
11 {6} => {0} 0.7935036 0.9999494 0.9999896
12 {6} => {1} 0.7935437 1.0000000 1.0000000
13 {8} => {5} 0.8010520 0.9853806 1.0020778
14 {5} => {8} 0.8010520 0.8146258 1.0020778
15 {8} => {3} 0.8079981 0.9939250 1.0007154
16 {3} => {8} 0.8079981 0.8135182 1.0007154
17 {8} => {2} 0.8106882 0.9972342 1.0005686
18 {2} => {8} 0.8106882 0.8133989 1.0005686
19 {8} => {4} 0.8116117 0.9983701 1.0000968
20 {4} => {8} 0.8116117 0.8130153 1.0000968
    
```

Fig. 7. Performance evaluation for the model

4 Conclusion and Future Work

Due to the explosive expansion of database size, the probabilistic techniques for analyzing the large-scale data have developed and the needs for additional computing power have increased as well. This leads to a virtuous circle of technical revolution in gathering the larger, precise and interesting data and largely affects machine learning practices. Thus, the authors used an Apriori algorithm to analyze the big data as they believe that it works better than other ones and can find some unnoticeable association rules in the target databases. The result of analysis showed that the foreign visitors’ purpose of visit varied depending on the age groups, as well as their inconvenient experiences. In the future research, the authors plan to present several novel tourism strategies and policies based on the analysis of such uncomfortable experiences shared by the foreign tourists from both nearby Asian and Western countries who are unfamiliar with Asian or Korean culture.

References

1. Huh, J.-H., Kim, H.-B., Seo, K.: A preliminary analysis model of big data for prevention of bioaccumulation of heavy metal-based pollutants: focusing on the atmospheric data analyses. *Adv. Sci. Technol. Lett. SERSC* **129**, 159–164 (2016)
2. Huh, J.-H., Je, S.-M., Seo, K.: Design and configuration of avoidance technique for worst situation in zigbee communications using OPNET. In: Kim, K.J., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016*. LNEE, vol. 376, pp. 331–336. Springer, Heidelberg (2016). doi:10.1007/978-981-10-0557-2_33
3. Kim D., Kim J.: Procedural modeling and visualization of multiple leaves. *Multimed. Syst.* 1–15 (2016). Springer, Berlin, Heidelberg

4. Birkenmeier, G.F., Park, J.K., Tariq, R.S.: Ring hulls of semiprime homomorphic images. In: Brzeziński, T., Gómez Pardo, J.L., Shestakov, I., Smith, P.F. (eds.) *Modules and Comodules*. TM, pp. 101–111. Springer, Birkhäuser Verlag Basel, Heidelberg, Basel (2008)
5. Huh, J.-H., Otkonchimeg, S., Seo, K.: Advanced metering infrastructure design and test bed experiment using intelligent agents: focusing on the PLC network base technology for Smart Grid system. *J. Supercomput.* **72**(5), 1862–1877 (2016). Springer, USA
6. Huh, J.-H., Je, S.-M., Seo, K.: Communications-Based Technology for Smart Grid Test Bed Using OPNET Simulations. In: Kim, K.J., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016*. LNEE, vol. 376, pp. 227–233. Springer, Heidelberg (2016). doi: [10.1007/978-981-10-0557-2_23](https://doi.org/10.1007/978-981-10-0557-2_23)
7. Korea Culture and Tourism Institute.: *The Code Book of Survey on the Foreign Tourists* (2015). (In Korean)
8. Li, Z., Pan, H., Liu, W., Xu, F., Cao, Z., Xiong, G.: A network attack forensic platform against HTTP evasive behavior. *J. Supercomput.* 1–12 (2016). Springer, USA
9. Sung, Y., Jeong, Y.-S., Park, J.-H.: Beacon-based active media control interface in indoor ubiquitous computing environment. *Cluster Comput.* **19**(1), 547–556 (2016). USA
10. Birkenmeier, G.F., Park, J.K., Rizvi, S.T.: Principally quasi-baer ring hulls. In: Van Huynh, D., López-Permouth, S.R. (eds.) *Advances in Ring Theory*. TM, pp. 47–61. Springer, Birkhäuser Basel, Heidelberg, Basel (2010)

The Study of Application Development on Elderly Customized Exercise for Active Aging

YoungHee Cho¹, SeungAe Kang², SooHyun Kim², and SunYoung Kang³(✉)

¹ Department of Sport Education, Kookmin University, Seoul, Korea
aga02@hanmail.net

² Department of Sport and Healthcare, Namseoul University, Cheonan, Korea
{sahome, shkim001}@nsu.ac.kr

³ Department of Physical Education, Korea University, Seoul, Korea
1010kang@hanmail.net

Abstract. The purpose of this study is to examine the current status and problems pertaining to the applications on the Korean elderly's exercise in the Korean android market using smart phone, for the development of the exercise program application customized for the elderly and to provide services. Moreover, improvement measures are analyzed to provide the base data to provide resourceful information to the elderly, leaders, trainees interested in the elderly and institutions and to develop customized elderly exercise application for the Active Aging. When the elderly exercise applications in Korea were examined, there is only one exercise program with standardized contents. Only the program called the "Health Ewha" present convergence of food, nursing and sports, but this application cannot be searched easily with keyword. Thus, this study planned customized elderly exercise for Active Aging, and this is the base data for developing application that can input information, measuring current state (exercise strength level setting, physical strength via heart rate measurement), and customized exercise program (cardio, muscular strength, flexibility, equilibrium, and coordination exercise for healthy elderly and exercise by disease).

Keywords: Health management · The elderly · Active aging · Application

1 Introduction

A society is referred to as an aging society when the share of people who are 65 years old or older exceeds 7% of the total population. It is called the aged society if the share is over 14% and the super aged society when the share is 20%. As of today, Korea had already entered into the aged society stage in 2000. Likewise, people who are at least 65 years old comprise 13.53% of the total population with 2.91 million men and 3.97 million women [1].

Increase in the elderly population is causing problems such as lack of health management, decrease in labor productivity, decrease in activities compared to the extension of lifetime, imbalance in the working population due to early retirement, increase in the 1

person households, severance of the social relations due to the loss of the family's and society's role, and mental issues such as depression and suicide.

Active Aging which is the concept that includes human beings' state and behavior and effort for the healthy and active activities during the senescence is the concept defined by the World Health Organization (WHO). It means leading of comfortable and comforting old age everyday life by participating actively and continually in the effort to maintain health, everyday life and life in the society by actively coping in order to purchase after the optimal state in the physical, mental, emotional, social and intellectual domains [2]. In other words, this signifies importance of healthy lifetime more than average lifetime. Interest in the contents that increase healthy management during aging and on the program development for Active Aging are emerging as the individuals, society and nation's interest. Thus, effective research is needed. To back up this research, development of the contents and programs for increasing health management during aging with our country's superb IT technological capability in this age of fusion and convergence age in which IT industry is active, is considered to be effective.

Accordingly, purpose of this study is to examine the current status and problems pertaining to the applications on the Korean elderly's exercise in the Korean android market using smart phone, for the development of the exercise program application customized for the elderly and to provide services. Moreover, improvement measures are analyzed to provide the base data to provide resourceful information to the elderly, leaders, trainees interested in the elderly and institutions and to develop customized elderly exercise application for the Active Aging.

2 Exercise Application for the Elderly in Korea

2.1 Current Status on the Elderly Exercise Application in Korea

Applications produced as of today are diverse in terms of the types such as the contents on the elderly health, health care and others. However, elderly exercise program is planned with the elderly exercise movement that is provided with Information Bank for Technology & Standards in Korea (<http://www.ibtk.kr>) [3] and Korean Agency for Technology and Standards.

Exercise program for the elderly for each application is the VOD that is based on the Korean Agency for Technology and Standards, and it presents the same exercise contents by suggesting five exercise programs by each domain in the following order; warm-up, cardiovascular endurance, muscular strength strengthening, flexibility strengthening, strengthening of the sense of equilibrium/physical strength and wrap-up exercise (Fig. 1).

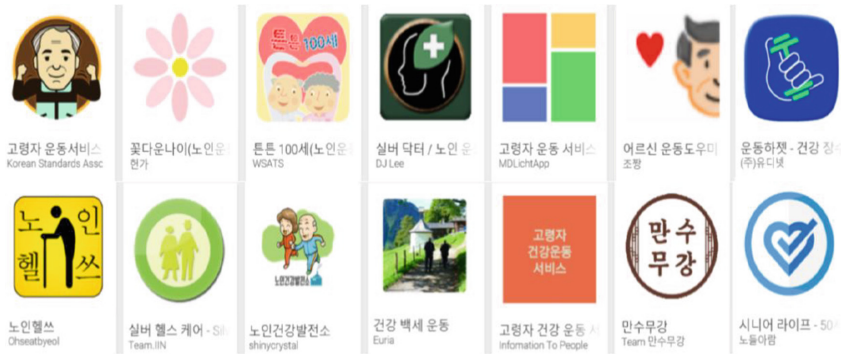


Fig. 1. Elderly exercise related application in the Korean Android market

In addition to the basic exercise, customized exercise program, “Health Ewha” application (convergence of Ewha Womans University’s Food Nutrition, Nursing and Sports departments) was produced to carry out as three year-long business (Fig. 2).

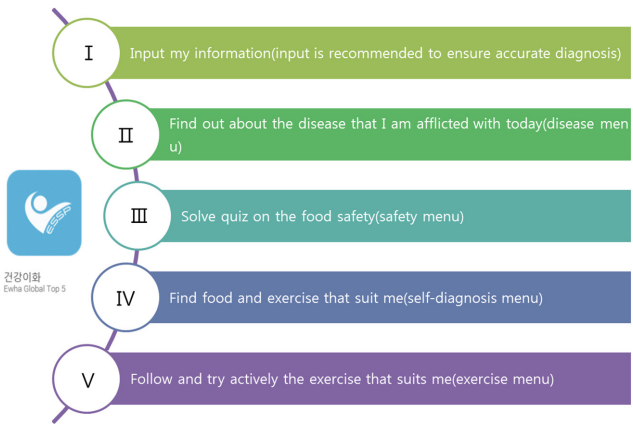


Fig. 2. Elderly exercise application in Korea (Health Ewha) [2]

2.2 Problem on the Elderly Exercise Application in Korea

Applications produced today provide the official elderly exercise program with the elderly exercise motions that are provided by the Korean Agency for Technology and Standards. However, customized exercise program cannot be expected by each elderly age, elderly disease, each posture and each tool with the even, general and standardized exercise program.

Accessibility of the application called Health Ewha is difficult to approach when searching with the elderly exercise, silver exercise, aging exercise and others since it is comprised of food, and disease control in addition to exercise program.

3 Development of Customized Exercise Application for the Elderly

It is deemed that the Korea’s elderly exercise related application will help to build the foundation for the elderly health business due to the drastic increase in the elderly population as the country that is becoming a aging society fast. Thus, this paper seeks to assess current status on the application in the elderly exercise in Korea and to present the direction for development system.

3.1 Research Timeline

This study is carried out in Stage 1 and Stage 2 as shown on Fig. 3, and research execution process is planned in a total of five stages.

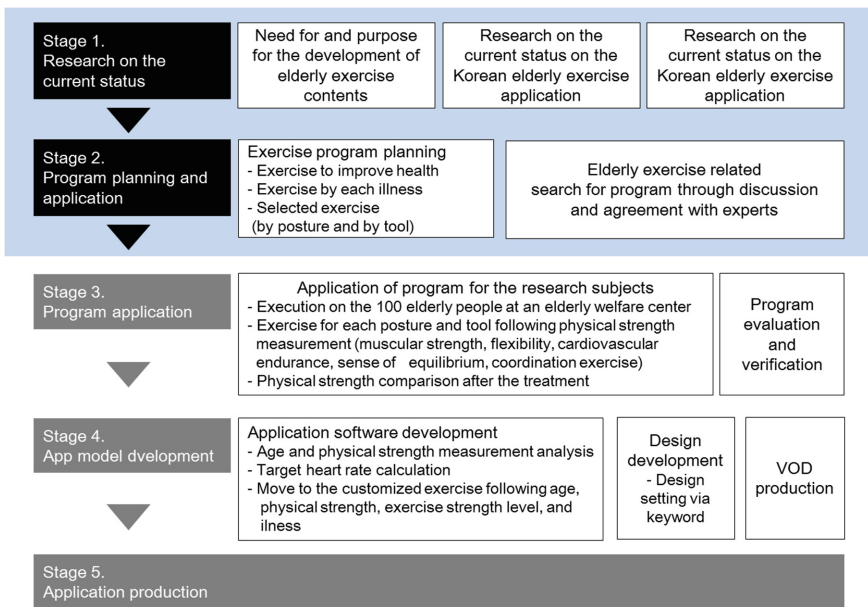


Fig. 3. Research timeline

3.2 Direction for the Composition of Application

- There is a need to develop differentiated program following physical strength characteristics by each elderly age.
- There is a need to develop program following the precautions by each disease.
- There is a need to plan customized exercise program for the elderly by each posture and each tool.
- Even when selecting title for the elderly exercise application for the energetic senescence, it is planned with the keyword related to the ease of access by elderly, leaders, and elderly related subjects. Accordingly, title of this application is “customized

elderly exercise for health” since it entails planning for the customized elderly exercise for Active Aging (Fig. 4).

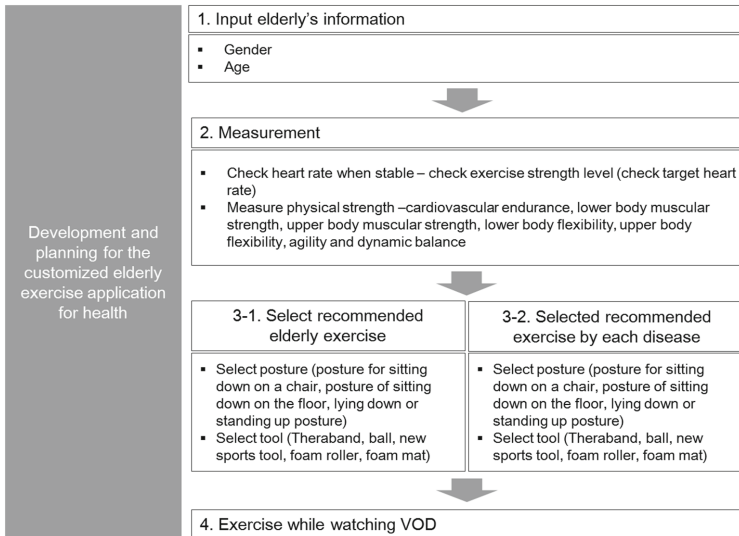


Fig. 4. Elderly customized exercise application development and planning

3.3 Contents of the Application Composition

- 1) In put information
 - Select gender (male, female)
 - Input age (select the standard that suits standard for the elderly physical strength in Korea suggested by Korean Citizen Physical Strength 100 by classifying into 60–64 years old, 65–69 years old, 70–74 years old, 75–79 years old, 80–84 years old, 85–89 years old, and at least 90 years old)
- 2) Measurement
 - Measure heart rate: Measure heart rate when stable, calculate target heart rate by setting up the exercise strength level
 - Target heart rate (THR) = exercise strength level (%) * (maximum heart rate- heart rate while stable) + heart rate while stable, [Maximum heart rate (MHR) = 220-age]
 - Exercise strength level adjustment (beginner: 40%, medium strength level: 50-60%, super high strength level: 70%)
 - Measure physical strength (elderly physical strength test)
 - Lower body muscular strength: Stand up from the chair during 30 s (lower body muscular endurance)/Frequency of repeating the following process; standing up from the chair after sitting down on the chair for 30 s while the two arms are gathered together in front of the chest

- Upper body muscular strength: grasping power come-back
 - Cardiovascular endurance: Step test during 2 min/Frequency of complete steps carried out during two minutes while lifting up the right and left knees to the middle of each kneecap and crista iliaca (at least knee high)
 - Upper body's flexibility: Touch the back with the hand/Distance between middle finger while having one hand touch the top of the shoulder and having the other hand extend out to the center of the back (cm)
 - Lower body's flexibility: sitting trunk flexion/Bend the upper body towards the front while the knee and two legs are stretched out
 - Agility and dynamic balance: Returning to the 3 m target while sitting on a chair/Time required to return to the sitting down posture after getting up from the chair and after walking for 2.44 m (second)
- 3) Customized elderly exercise plan for health for Active aging
- Move to customized exercise after checking the physical strength that suits the listed age when the information was input (Check physical strength: Set up the standard for elderly physical strength carried out by Korean Citizen Physical Strength 100)
 - Elderly customized exercise program plan
 - Exercise program to increase health: Plan into five exercises including the four exercises (plan for the exercise program to increase health by age based on the muscular strength, cardio, flexibility and lower body balance) needed for the elderly as suggested by the National Institutes of Health of the US and coordination exercise needed by the elderly for carrying out everyday life activity.
 - Exercise program by disease: Compose by incontinence, diabetes, arthritis, high blood pressure, elderly depression and dementia and plan exercise by listing down the characteristics by each disease and by listing up the precautions by exercise
 - Plan elective exercise by posture and by tool:
 - By posture: Posture of sitting down on a chair / standing state / state of sitting down on the floor or lying down
 - By tool: Theraband / new sports tool / foam roller / foam mat (Fig. 5).

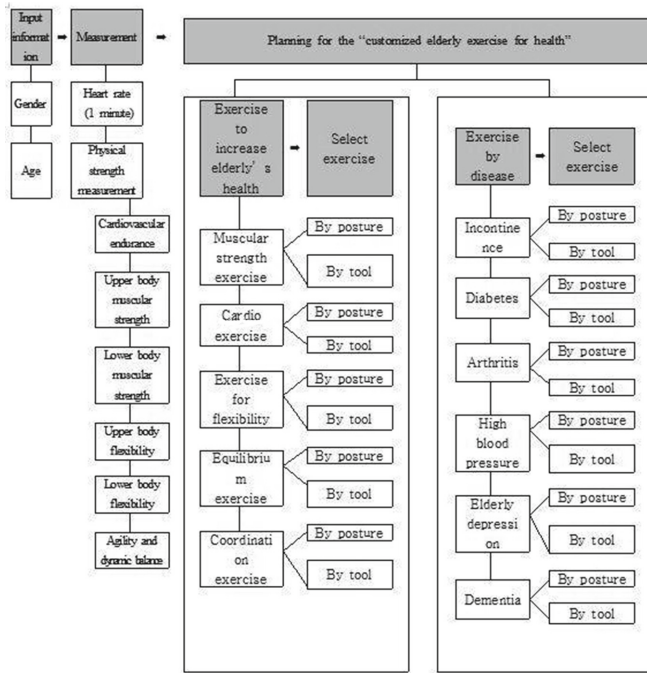


Fig. 5. Flow chart of Elderly customized exercise application

4 Conclusion

This study is the planning stage for developing application for elderly using smart phone that is easy to access amidst the IT industry that can be accessed fast and diverse ways now that the issue regarding healthy elderly is expanding along with the increase in the elderly population in Korea that is entering into the aging era fastest. Increase in the number of elderly in Korea who are not prepared is leading to physical, emotional and social problems during the senescence. Although need for elderly exercise is highlighted to reduce medical cost to ensure active senescence, elderly in Korea are not accessing exercise due to the lack of the awareness on the need for elderly exercise, lack of information, and lack of specialized and diverse programs.

When the elderly exercise applications in Korea were examined, there is only one exercise program with standardized contents. Only the program called the “Health Ewha” present convergence of food, nursing and sports, but this application cannot be searched easily with keyword. Thus, this study planned customized elderly exercise for Active Aging, and this is the base data for developing application that can input information, measuring current state (exercise strength level setting, physical strength via heart rate measurement), and customized exercise program (cardio, muscular strength, flexibility, equilibrium, and coordination exercise for healthy elderly and exercise by

disease). Moreover, this data is the base data for planning program for each posture and tool, and for developing elderly related exercise application to the leaders, public health clinic and related institutions in addition to program diversity and the elderly.

References

1. National Statistical Office (2015). <http://kostat.go.kr/portal/korea/index.action>
2. Chang, H.S., Bahn, T.H., Chang, S.C., Chung, H.G.: Elderly care system in the U-health environment. *J. Korea Inst. Inf. Commun. Eng.* **17**(11), 2693–2698 (2013)
3. Information Bank for Technology & Standards in Korea (2015). (<http://www.ibtk.kr>)
4. Lee, G.k., Cho, M.S., Koh, G.S., Chung, D.Y., Cha, J.Y.: Development of the application system with the elderly health management convergence model. 2014 The Korean Society of Aging and Physical Activity's Academic Seminar Collection, Elderly and Exercise (2014)

Improving Jaccard Index for Measuring Similarity in Collaborative Filtering

Soojung Lee^(✉)

Gyeongin National University of Education,
155 Sammak-ro, Anyang, Republic of Korea
sjlee@gin.ac.kr

Abstract. In collaborative filtering-based recommender systems, items are recommended by consulting ratings of similar users. However, if the number of ratings to compute similarity is not sufficient, the system may produce unreliable recommendations. Since this data sparsity problem is critical in collaborative filtering, many researchers have made efforts to develop new similarity metrics taking care of this problem. Jaccard index has also been a useful tool when combined with existing similarity measures to handle data sparsity problem. This paper proposes a novel improvement of Jaccard index that reflects the frequency of ratings assigned by users as well as the number of items co-rated by users. Performance of the proposed index is evaluated through extensive experiments to find that the proposed significantly outperforms Jaccard index especially in a dense dataset and that its combination with a previous similarity measure is superior to existing measures in terms of both prediction and recommendation qualities.

Keywords: Recommender system · Similarity measure · Collaborative filtering · Memory-based collaborative filtering · Jaccard index

1 Introduction

Internet users are very often overwhelmed by the amount of information provided by the web. A popular method to solve this problem is the recommender system. This system is usually utilized in commerce to recommend products that might be preferred by customers. Collaborative filtering (CF) is a well-known type of implementation of a recommender system. It refers to other likeminded users and recommends items which have been highly rated by them. This filtering method has been successful in many systems such as GroupLens, Ringo, and Amazon.com [1].

As CF basically performs by incorporating ratings of similar users, determination of similar users is a critical aspect of the CF system. Several approaches have been developed to calculate similarity. They are usually classified into correlation-based and vector cosine-based [1, 2]. However, similarity calculation is based on the history of ratings of users, thus insufficient amount of rating data

in the system often producing unreliable similar users. This problem, known as *data sparsity*, is fundamental due to the principle of CF. Detailed analysis of drawbacks resulting from the data sparsity problem of traditional similarity measures can be found in [3, 4].

Various techniques in literature have addressed the data sparsity problem of CF, while the simplest technique to compute similarity may be the incorporation of Jaccard index into the previous similarity measure [4–8]. Jaccard index reflects the number of common items rated by two users. As Jaccard index becomes an important component of measuring similarity and reportedly improves performance of CF, this paper focuses on this index and proposes a novel improvement. To verify its novelty, we conducted extensive experiments using two datasets with very different characteristics. The results state that the proposed index outperforms Jaccard index on both datasets. Especially, the degree of improvement is higher on a denser dataset. Furthermore, the proposed index is incorporated into a previous measure to produce a new similarity measure which proves to perform the best among or comparably to existing measures experimented.

2 Related Work

Jaccard index measures the proportion of the number of items commonly rated by two users out of the total number of items rated by them [7]. Let I_u be the set of items rated by user u and $|I_u|$ be its cardinality. Then Jaccard index between users u and v is calculated as follows.

$$Jaccard(u, v) = \frac{|I_u \cap I_v|}{|I_u \cup I_v|}.$$

As seen from the above formula, Jaccard index does not take the ratings into account, but only considers relative number of common items. Hence, it seems improper to use the index to estimate similarity between two users.

Several researchers propose new similarity measures that incorporate Jaccard index into traditional measures [4, 5, 9]. These approaches surely compensate the defects of previous similarity measures which are mainly caused by data sparsity or cold-start users. Bobadilla et al. proposed a new similarity measure that combines mean squared differences with Jaccard index [5]. Saranya et al. calculate similarity by incorporating Pearson correlation and Jaccard index which is reported to achieve a little improvement in recommendation quality [4]. In the meantime, a measure named UOD (Uniform Operator Distance) is suggested by Sun et al. [9]. This measure is combined with Jaccard index to better estimate similarity between two users. The authors report that their combined similarity measure called *JacUOD* leads to better prediction quality than when Jaccard index is combined with Pearson correlation. We examined performance of *JacUOD* through several experiments, whose results are presented in Sect. 4. The formula of *JacUOD* is defined as follows. Let $r_{u,i}$ be the rating of item i given by user u and $r_{u,max}$ be the maximum rating given by u . Also let $m = |I_u \cap I_v|$.

Then

$$UOD(u, v) = \begin{cases} \frac{\sqrt{m(r_{u,max}-r_{u,min})^2}}{0.9 + \sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - r_{v,i})^2}}, & \text{if } r_{u,i} = r_{v,i} \text{ for all } i \\ \frac{\sqrt{m(r_{u,max}-r_{u,min})^2}}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - r_{v,i})^2}}, & \text{otherwise.} \end{cases}$$

$$JacUOD(u, v) = Jaccard(u, v) \times UOD(u, v)$$

3 Proposed Index

3.1 Motivation

The idea of our study is based on the work by [5]. This work examines the mean and deviation of the ratings of MovieLens (<http://www.movielens.org>) and NetFlx datasets (<http://www.netflixprize.com>). Within the integer range of [1..5] allowed in these datasets, it is found that users tend to give ratings higher than the median and avoid the extreme values. The highest frequency of ratings is associated with the rating of four, followed by the rating of three. It is also found that the standard deviation of ratings is seldom larger than 1.2. This result implies that two users giving a same extreme rating can be treated as more similar than those giving a more common rating.

The above observation motivated our research to improve Jaccard index. This index calculates the number of commonly rated items by two users, regardless of the rating values. However, as discussed above, it is worth considering that the rating of a common item is normal or extreme.

3.2 Formulation of the Index

We are interested in how many items are commonly rated with normal or extreme values. Hence, in our index, the rating range allowed in the system is divided into three sub-intervals, within each of which Jaccard index is computed separately. Specifically, let L_{bd} and H_{bd} be boundaries of the sub-intervals, where $L_{bd} < H_{bd}$. That is, when $[r_L, r_H]$ represents the range, $r_L < L_{bd} < H_{bd} < r_H$. We divide the set of items rated by user u , I_u , into three sets as follows, based on the rating values assigned by u .

$$I_{L,u} = \{i \in I_u | r_{u,i} \leq L_{bd}\}, \quad I_{M,u} = \{i \in I_u | L_{bd} < r_{u,i} < H_{bd}\}, \quad I_{H,u} = \{i \in I_u | r_{u,i} \geq H_{bd}\}.$$

Then three types of Jaccard indexes between users u and v are defined as follows.

$$Jac_L(u, v) = \frac{|I_{L,u} \cap I_{L,v}|}{|I_{L,u} \cup I_{L,v}|}, \quad Jac_M(u, v) = \frac{|I_{M,u} \cap I_{M,v}|}{|I_{M,u} \cup I_{M,v}|}, \quad Jac_H(u, v) = \frac{|I_{H,u} \cap I_{H,v}|}{|I_{H,u} \cup I_{H,v}|}$$

Finally, our metric, named as $JacLMH$, is calculated as an arithmetic average of the three Jaccard indexes as follows.

$$JacLMH(u, v) = \frac{1}{3}(Jac_L(u, v) + Jac_M(u, v) + Jac_H(u, v))$$

Table 1. Characteristics of the datasets

	Matrix size (users \times ratings)	Rating scale	Sparsity level
MovieLens	1000 \times 3952	1~5 (integer)	0.9607
Jester	998 \times 100	-10 ~ +10 (real)	0.2936

4 Performance Experiments

4.1 Experiments Plan

We conducted extensive experiments using two popular datasets with very different characteristics, as presented in Table 1. Sparsity level represents how sparse the dataset is. It is defined by $1 - (\text{total number of ratings} / \text{matrix size})$.

The baseline similarity measures of our experiments are Jaccard index (Jaccard), Pearson correlation (PCC), JacUOD, UOD, the proposed JacLMH, and JacLMH \times UOD (JLMHUOD). The last measure is experimented to compare the degree of improvement made by incorporating JacLMH instead of Jaccard into UOD. We adopted five-fold cross validation [10] to obtain more reliable results, where the ratio of training and testing data is set to 80:20 for each experiment.

Performance is evaluated based on two well-known standards in related studies, prediction quality and recommendation quality. MAE (Mean Absolute Error) is usually used to measure prediction quality, which is the mean difference between the predicted rating of an unrated item and its corresponding real rating. The rating prediction is typically made by referring to ratings of users similar to the current user, while weights are imposed according to the degree of similarity. Recommendation quality is usually measured by precision and recall metrics or their harmonic mean F1 [11]. We employed only F1 due to the space constraint.

4.2 Effect of Bounds

To examine how L_{bd} and H_{bd} parameters used in our index affect performance, we measured MAE with various combinations of these parameter values. Figure 1 shows the results with two datasets. With MovieLens, it seems that (1,3) is definitely worst, while others are almost competitive. In particular, MAE results using (2,4), (2,5), and (3,4) are virtually no different with one another. Hence, we chose (2,5) for ensuing experiments of our metric. With Jester, the results using different parameter values are more distinguishable than with MovieLens. We used (-3,3) yielding the obviously lowest MAE in our further experiments.

4.3 Performance Results

Figure 2 shows performance results with MovieLens with varying number of nearest neighbors (topNN). To view any performance improvement of our index

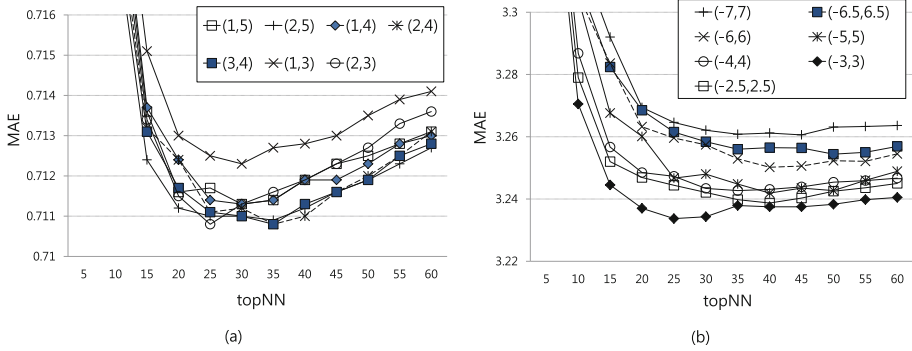


Fig. 1. MAE results with varying bounds of (L_{bd}, H_{bd}) pairs: (a) MovieLens (b) Jester

over Jaccard index more clearly, two metrics, Jaccard and JacLMH, are provided together, separately from the others. It is observed that JacLMH yields better MAEs consistently, implying that our idea of separate application of Jaccard index to sub-ranges of ratings proves successful with respect to prediction accuracy.

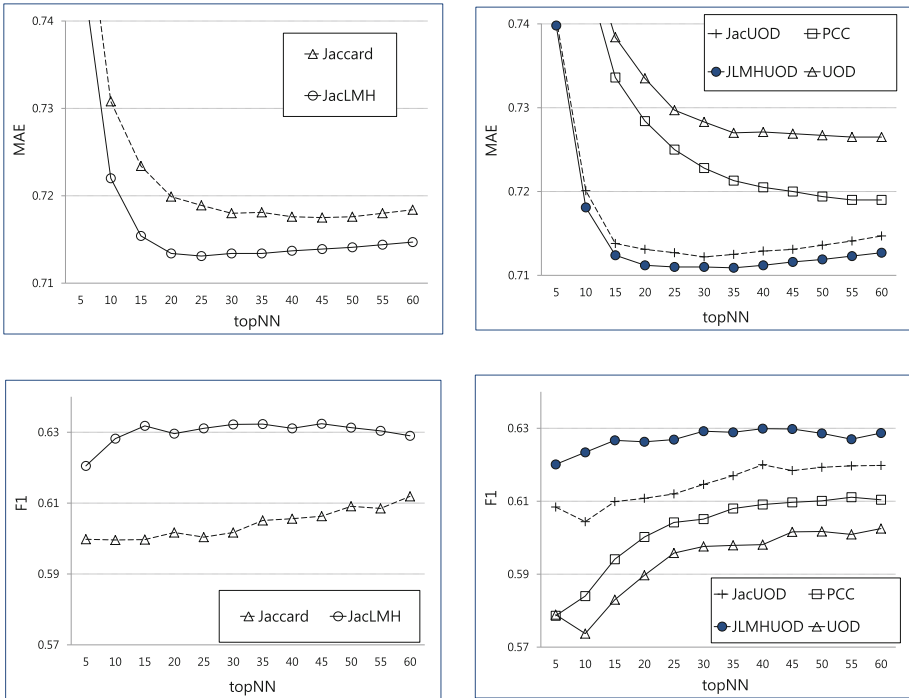


Fig. 2. MAE and F1 results with MovieLens dataset

Note that PCC performs very poor compared to JacUOD and JLMHUOD, although it consults specific ratings of neighbors, which is not the case for the latter two metrics. The reason for this poor performance of PCC comes mostly from sparseness of the dataset, as many researchers discussed the resulting drawbacks of traditional measures [3, 4]. JLMHUOD, somewhat against our expectation, outperforms JacUOD only slightly, compared to the performance difference between Jaccard and JacLMH shown in the left figure. Nevertheless, it is notable that JLMHUOD performs best among all the metrics experimented, even though it divides the rating range and so may be disadvantageous with a sparse dataset such as MovieLens. F1 results of the proposed metric are vastly superior to the others, better achievements than MAE results. One thing to note is that JacLMH is slightly better than JLMHUOD, although it does not reflect any ratings but only the number of ratings.

Performance of metrics with Jester dataset is presented in Fig. 3. As seen, Jaccard and JacLMH differ greatly in MAE performance, where its difference is much bigger than with MovieLens. This is because Jester dataset is much denser, thus providing much more meaningful Jaccard indexes in sub-intervals of JacLMH. JacLMH improves about 4.35 to 7.1% of Jaccard results with Jester.

MAE of JLMHUOD is also the lowest and even better than that of PCC overall, especially with lower topNNs. This result is surprising, since one of most

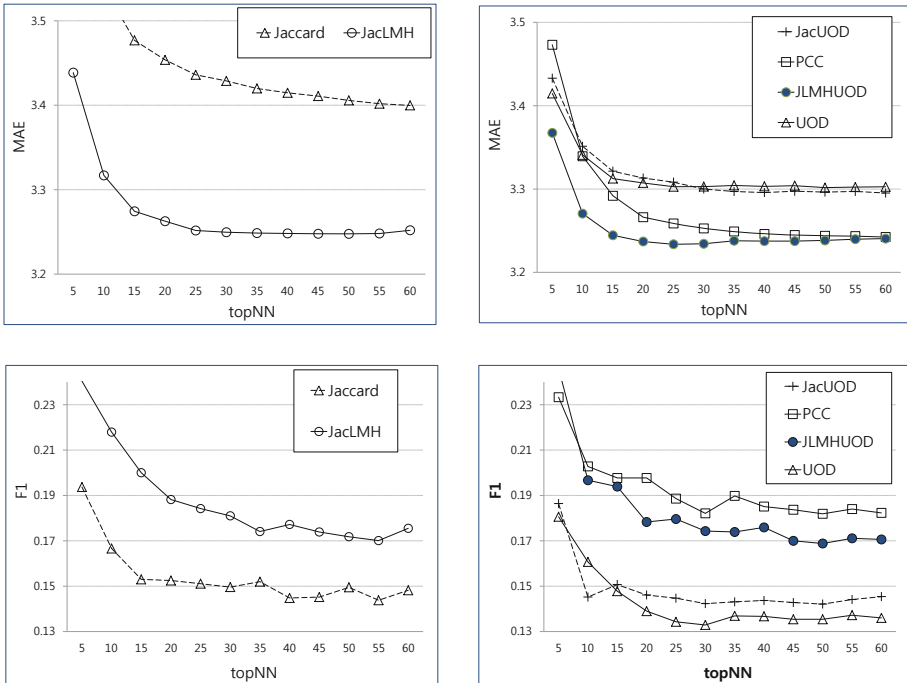


Fig. 3. MAE and F1 results with Jester dataset

popularly used similarity measures such as PCC is thought to perform better. The other two metrics, JacUOD and UOD are very competitive throughout topNNs. This means that considering the number of common ratings as in Jaccard index is no longer effective with sufficient number of ratings data provided. Comparison of F1 performance between metrics is analogous to MAE, as observed in the figure. Roughly, the metrics are grouped into two, in terms of performance. Note that different from MAE results, PCC, followed by JLMHUOD, yields the best F1 results. In conclusion, the proposed metric and its combination with UOD are proved to yield the best overall prediction and recommendation qualities regardless of the rating data density.

5 Conclusion

This study proposed a novel improvement of Jaccard index. The proposed idea takes the frequency of rating values assigned by users as well as the number of common items into consideration. We investigated the performance of the proposed index when used for collaborative filtering and found that it outperformed Jaccard index especially on a dense dataset. Furthermore, the combination of the proposed index with a previous measure is used as a similarity measure for collaborative filtering, whose experimentation results are found superior to those of previous measures, regardless of the data sparsity of datasets. One possible limitation of the proposed index is determination of the boundaries of sub-intervals, on which extensive experiments are conducted with various boundaries and their performance results are provided in the text.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **17**(6), 734–749 (2005)
2. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. In: *Advances in Artificial Intelligence 2009* (2009)
3. Ahn, H.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Inf. Sci.* **178**(1), 37–51 (2008)
4. Saranya, K.G., Sadasivam, G.S., Chandralekha, M.: Performance comparison of different similarity measures for collaborative filtering technique. *Indian J. Sci. Technol.* **9**(29), 1–8 (2016)
5. Bobadilla, J., Serradilla, F., Bernal, J.: A new collaborative filtering metric that improves the behavior of recommender systems. *Knowl. Based Syst.* **23**(6), 520–528 (2010)
6. Bobadilla, J., Ortega, F., Hernando, A., Bernal, J.: A collaborative filtering approach to mitigate the new user cold start problem. *Knowl. Based Syst.* **26**, 225–238 (2012)
7. Koutrica, G., Bercovitz, B., Garcia-Molina, H.: FlexRecs: expressing and combining flexible recommendations. In: *The 2009 ACM SIGMOD International Conference on Management of Data*, pp. 745–758. ACM (2009)

8. Liu, H., Hu, Z., Mian, A., Tian, H., Zhu, X.: A new user similarity model to improve the accuracy of collaborative filtering. *Knowl. Based Syst.* **56**, 156–166 (2014)
9. Sun, H.-F., et al.: JacUOD: a new similarity measurement for collaborative filtering. *J. Comput. Sci. Technol.* **27**(6), 1252–1260 (2012)
10. Bengio, Y., Grandvalet, Y.: No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* **5**, 1089–1105 (2004)
11. Bobadilla, J., Ortega, F., Hernando, A., Gutierrez, A.: Recommender systems survey. *Knowl. Based Syst.* **46**, 109–132 (2013)

Temperature Recorder System

Suresh Thanakodi¹(✉), Nazatul Shiema Moh Nazar¹, Azizi Miskon¹, Ahmad Mujahid Ahmad Zaidi², and Muhammad Syafiq Najmi Mazlan¹

¹ Department of Electrical and Electronic Engineering, Faculty of Engineering, National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
{suresh,nazatul.shima,azizimiskon}@upnm.edu.my,
alongkp07@gmail.com

² Department of Mechanical Engineering, Faculty of Engineering, National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
mujahid80s@yahoo.com

Abstract. A temperature recorder system is based on the changes of the patients' temperature over a fixed time that uses the advantages of a smartphone. This paper proposes a wirelessly controlled system to achieve reliability and mobility of the user. The wireless system was achieved by utilizing a smartphone, PIC Controller and Bluetooth to a device that's been installed with the temperature sensors. This system can be used for medical purpose to monitor and record 24/7 of the patient's temperature consistently which can systematically save manpower, time and lives.

1 Introduction

Body temperature is affected due to many reasons that give us a sign of the condition of a person. Heart beat rate, blood pressure, sugar level and so on are also an indicator that gives general information on the condition of a person [1]. In this paper, body temperature is measured and recorded as the main indicator to monitor a certain patient. This body temperature readings and patterns can indicate the severity of a patient's illness.

Dengue fever is a common disease that easily leads to death. Research already came with a common pattern of dengue fever and effects to the patient body temperature. The sequence of rise and decline of body temperature gives us as a sign either the patient is in a severe condition or not [2]. This also allows the doctors to give the exact amount medicine need of the particular patient.

The main problem is when the body temperature is monitored manually, the medical assistants or nurses would have to contact with the patient directly and numerously that would affect the patient's quality time of recovery physically, mentally and emotionally. If the body temperature could be monitor wirelessly, it would defiantly give an opportunity to the patient to have quality time to recover. Thus, this project is tested to overcome this matter.

As the Bureau of Labor Statistics (2012) report that the medical field and health care careers with the most projected employment are increasing. Some of these jobs need less than one year of medical field education. Based on the Fig. 1, medical field education

reduces the period of training and learning to fulfil the high demands of people in social market these days.

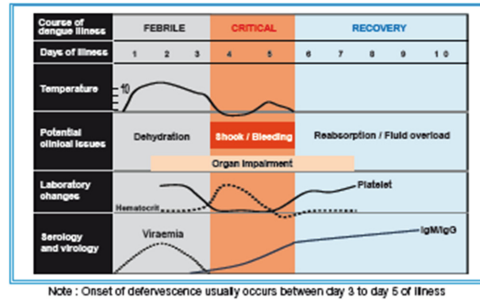


Fig. 1. Clinical course of DHF12 [2]

This points out the reason for the medical field carrier expected to rise at a yearly rate of 2.6%, accumulating 5.0 million jobs between the year 2012 and 2022 stated by the Bureau of Labor Statistics (2012) [4]. Form this statement is true that employees in the medical field are a high demand, but the quality of employees in this field is very essential. It’s unreasonable to produce employee that is trained for 1 year to have a standard quality. Thus tools and devices are developed to overcome errors occurring in treating a patient.

2 Technologies

2.1 Smartphones

A smartphone is a portable phone with an innovative mobile operating system which syndicates features of an individual computer operating system with other useful features for handheld or mobile use. Smartphone usually combines the features of a telephone with those of other famous mobile devices, such as a media player, personal digital assistant (PDA), and GPS navigation unit [3]. Most of the smartphones have a touchscreen user interface, can access the Internet and run third-party apps, camera phones and music players. Furthermore, smartphones that’s been produced from 2012 ahead have high-speed mobile broadband which called as the 4G LTE internet, mobile payment mechanisms, and motion sensors.

2.2 Mobile Operating System

Android is known as an open-source platform by Andy Rubin, founded in October 2003 and supported by Google, along with major hardware and software developers such as HTC, Intel, Motorola, LG, and Samsung and ARM that create the Open Handset Alliance.

IOS is a portable working framework created by Apple Inc. and appropriated only for Apple equipment. It is the working framework that powers the organization’s i-Devices. In 2007, Apple presented the iPhone, the primary gadget to utilize iOS and one of the principal cell phones to utilize a multi-touch interface. The iPhone was remarkable for its utilization of a vast touch screen for direct finger contribution as its principal method for cooperation, rather than a stylus, console, or keypad as regular for cell phones at the time. Windows Phone this product stage runs the Microsoft Mobile cell phones, and has gotten some positive gathering from the innovation press and been lauded for its uniqueness and separation.

2.3 PIC Microcontroller

In this paper, the controller for the development is PIC microcontroller. This paper contains gadget particular data about the accompanying gadgets. PIC16F737/767 devices are available in 28-pin packages only. Meanwhile, PIC16F747/777 devices are available either in 44-pin and 40-pin packages. All the devices in the PIC16F7X7 family share common architecture and ideas with the following differences as shown in Table 1. Thus, (Fig. 2) shows the pcb schematic circuit for temperature recorder system.

Table 1. PIC16F7X7 Device Features [5]

Key features	PIC16F737	PIC16F747	PIC16F767	PIC16F777
Operating frequency	DC–20 MHz	DC–20 MHz	DC–20 MHz	DC–20 MHz
Resets (and Delays)	POR, BOR (PWRT, OST)	POR, BOR (PWRT, OST)	POR, BOR (PWRT, OST)	POR, BOR (PWRT, OST)
Flash program memory (14-bit words)	4 K	4 K	8 K	8 K
Data memory (bytes)	368	368	368	368
Interrupts	16	17	16	17
I/O Ports	Ports A, B, C	Ports A, B, C, D, E	Ports A, B, C	Ports A, B, C, D, E
Timers	3	3	3	3
Capture/ Compare/PWM modules	3	3	3	3
Master serial communications	MSSP, AUSART	MSSP, AUSART	MSSP, AUSART	MSSP, AUSART
Parallel communications	–	PSP	–	PSP
10-bit analog-to-digital module	11 Input channels	11 Input channels	11 Input channels	11 Input channels
Instruction set	35 Instructions	35 Instructions	35 Instructions	35 Instructions
Packaging	28-pin PDIP 28-pin SOIC 28-pin SSOP 28-pin QFN	40-pin PDIP 44-pin QFN 44-pin TQFP	28-pin PDIP 28-pin SOIC 28-pin SSOP 28-pin QFN	40-pin PDIP 44-pin QFN 44-pin TQFP

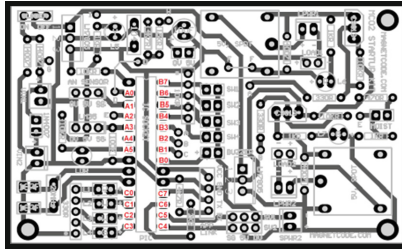


Fig. 2. Schematic circuit diagram

2.4 Bluetooth

Bluetooth is a wireless innovation standard for trading data over short distances (using short-wavelength UHF radio waves in the ISM band between 2.4 to 2.485 GHz) from settled and cell phones, and building personal area networks (PANs). Developed by telecom merchant Ericsson in 1994, it was initially considered as a remote other option to RS-232 data cable. It can interface a few gadgets and thus, overcoming issues of synchronization [7].

Furthermore, the Bluetooth Special Interest Group has managed more than 25,000 members of an organization in the areas of computing, telecommunication, consumer electronics, and networking. Bluetooth has been standardized using IEEE standardized as IEEE 802.15.1. However, the Bluetooth is no longer followed and maintains the standard. The Bluetooth Special Interest Group supervisory manages the qualification program, development of the specification, and protects the trademarks. The manufacturer must create a device that meets the Bluetooth Special Interest Group standards to commercialize it as the Bluetooth device. Besides that, a system of patents has utilized to the technology, which are registered to personal qualifying devices.

3 Methodologies

3.1 Overall Process and Components

In this paper, the system produced consists of a PIC-16f767 micro controller on a circuit board with a heartbeat detector, App Link Bluetooth module, and an Android smart-phone using the Magnet code application. All these components are the most important aspect to ensure that the aims of the paper were attained.

If observed in Fig. 3, basically the project works with a device assembled with a PIC-16f767, App Link Bluetooth module and a temperature detector on a circuit board which attached to a person and connected wirelessly by Bluetooth connection to an Android smart phone. The android smart phone can orientate the data of the device by installing an application known a Magnet code and programming the PIC-f767 to track either the heart is beating using two program which are PIC compiler and PIC kit 2 v2.55. Figure 4 summarized the overall process.

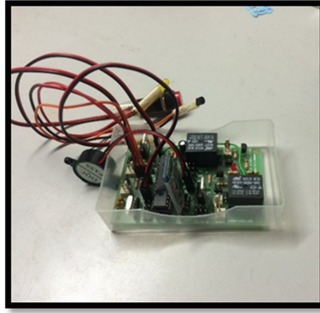


Fig. 3. Temperature recorder device without casing

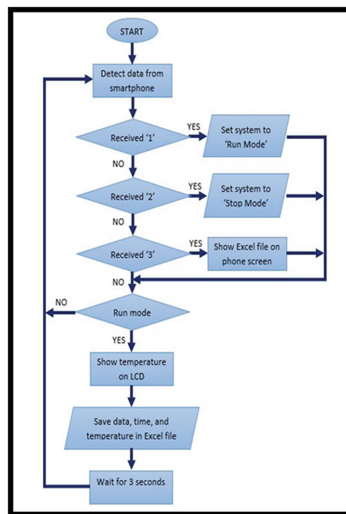


Fig. 4. Flow chart

3.2 Assembling

Assembling the project are based two elements that are hardware & software. In the hardware assembling process the components of the device were collected & fitted onto the mother board [5].

Meanwhile, the assembling process in software is planned, checked & installed into the PIC micro-controller using two Software Program which are 'PIC C Compiler' to construct the program logic & PIC kit 2 v2. 55 to install the program logic into the PIC [6]. The logic of the program can be viewed in Fig. 5.

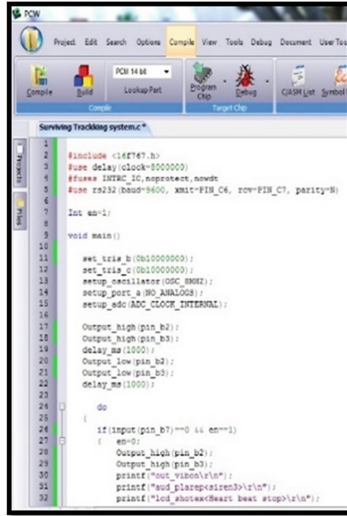


Fig. 5. Program logic in PIC C compiler

3.3 Testing Method

The test was tested out on the assembled temperature recorder attached to a user. Meanwhile the smart phone is held by another user to examine the reliability of the system based on connection, responds, and accuracy of the system reacting to the different variable and situation conducted [8]. The data sheet provided by the software is also verified to obtain reliable results.

4 Discussion and Analysis

Table 2 shows the data that were gained from the test. The data for the communication part assembled firstly and continued by detection of the body temperature. After that, the device will respond by record the temperature detected.

Table 2. Test results

	Yes	No	Comment
Connection			
Communicate	✓		
Lagging	✓		0.2 s
Accuracy	✓		0.5–1°C

Although we've archived promising results, facing the facts, there a few restrictions on this project that need to be improved in future. All the way through the experiment, the most suitable place for the device to detect a body temperature. But still, it's reliable

to calculate one’s body temperature due to the sensitivity of the sensor used in the device itself.

Based on the data sheet shown in Fig. 6, there are a few adjustments that could be made to improve to optimize the capability of the project. One of the issues is the fixed time laps can only be adjusted by changing the setting in the PIC software. If the fixed time laps can be adjusted by the smartphone would be much friendly user.

	C	D	E	F	G
7	27-04-2016 13:50:43				28C
8	27-04-2016 13:50:47				30C
9	27-04-2016 13:50:50				30C
10	27-04-2016 13:50:53				40C
11	27-04-2016 13:50:56				38C
12	27-04-2016 13:50:59				38C
13	27-04-2016 13:51:02				35C
14	27-04-2016 13:51:06				30C
15	27-04-2016 13:51:09				30C
16	27-04-2016 13:51:12				31C
17	27-04-2016 13:51:15				30C
18	27-04-2016 13:51:18				29C
19	27-04-2016 13:51:21				28C
20	27-04-2016 13:51:24				27C
21	27-04-2016 13:51:28				27C
22	27-04-2016 13:51:31				26C
23	27-04-2016 13:51:34				26C
24	27-04-2016 13:51:37				25C
25	27-04-2016 13:51:40				25C
26	27-04-2016 13:51:43				24C
27	27-04-2016 13:51:46				24C
28	27-04-2016 13:51:49				24C
29	27-04-2016 13:51:53				23C
30	27-04-2016 13:51:56				23C
31	27-04-2016 13:51:59				23C
32	27-04-2016 13:52:02				23C
33	27-04-2016 13:52:05				22C
34	27-04-2016 13:52:08				22C
35	27-04-2016 13:52:11				22C
36	27-04-2016 13:52:15				22C
37	27-04-2016 13:52:18				21C
38	27-04-2016 13:52:21				21C
39	27-04-2016 13:52:24				21C
40	27-04-2016 13:52:27				21C
41	27-04-2016 13:52:30				21C
42	27-04-2016 13:52:33				21C
43	27-04-2016 13:52:37				21C
44	27-04-2016 13:52:40				21C
45	27-04-2016 13:52:43				21C
46	27-04-2016 13:52:46				20C
47	27-04-2016 13:52:49				20C
48	27-04-2016 13:52:52				20C
49	27-04-2016 13:52:55				20C
50	27-04-2016 13:52:58				20C
51	27-04-2016 13:53:02				20C
52	27-04-2016 13:53:05				20C

Fig. 6. Screen shot on data sheet

Finally, the temperature recorder system has limited range of Bluetooth connection, this is basically because the budget provided. Bluetooth range of connection is limited due to the type of inserted.

5 Conclusion

The designed temperature recorder system was capable to function either some minor glitches along the duration of taking body temperature. Moreover, this paper also succeeds to recognize the limitations to the design and recommendations in enhancing it. This early design shows the possibility of the usage which have the opportunity to make this research proceed to another level in nanotechnology. In addition, the device could enhance the capabilities by inserting other input sensors such as heart beat, blood pressure and other detectors that would increase various information on the condition of the patient. The design could be improved with more researches on the subject to gain high possibilities of success. This research also would benefit the health and care industry and have high potential due to the nature leveraging on the smartphone technology able to reduce the components used. The recorded data in real time would give many insights for any drugs impact introduce by pharmaceutical industry.

References

1. Broomhead, D.H., Lowe, D.: Multivariable functional interpolation and adaptive networks. *Complex Syst.* **2**, 321–355 (1988)
2. The Richrd Group of Charities: Vital Signs, March 2015
3. Ministry of Health, Academy of Medicine Malaysia: Management of Dengue Infection In Adults (revised 2nd Edition)
4. Yorozu, Y., Hirano, M., Oka, K., Tagawa, Y.: Electron spectroscopy studies on magneto-optical media and plastic substrate interface. *IEEE Transl. J. Magn. Japan* **2**, 740–741 (1982). August 1987 [Digests 9th Annual Conference on Magnetics Japan, p. 301, 1982]
5. Young, M.: *The Technical Writer's Handbook*. University Science, Mill Valley (1989)
6. Verle, M.: *PIC Microcontrollers - Programming in Basic*, 1st edn. mikroElektronika, Virginia (2010)
7. Verle, M.: *PIC Microcontrollers - Programming in C*, 1st edn. mikroElektronika, Virginia (2009)
8. Huang, A.S., Rudolph, L.: *Bluetooth Essentials for Programmer*, 1st edn. Cambridge University Press, Cambridge (2007)
9. Clerk-Maxwell, J.: *A Treatise on Electricity and Magnetism*, vol. 2, 3rd edn, pp. 68–73. Clarendon, Oxford (1892)

7th International Workshop on ICT Convergence

The Emergence of ICTs for Knowledge Sharing Based on Research in Indonesia

Siti Rohajawati^(✉), Boy Iskandar Pasaribu, Gun Gun Gumilar,
and Hilda Rizanti Putri

Department of Information System, Bakrie University, Jakarta, Indonesia
{siti.rohajawati, boy.pasaribu, gungun.gumilar,
hilda.rizanti}@bakrie.ac.id

Abstract. This paper presents organizations in applying ICTs (Information, Communication, and Technologies) contribution for supporting knowledge activities, in order to gain organizational competitiveness. Through ICT, knowledge workers in different fields are empowered to contribute and share their knowledge effectively and efficiently. In Knowledge Sharing (KS), organization must focus on results of research in the development of Knowledge Management System features, adoption, and use of ICTs itself. Creating a knowledge repository and providing best practices via ICTs tools enable the starting of knowledge sharing application. The study found that the features to support mechanism, content, and process of KS are mostly applied based on identification of knowledge from structural and functional areas. Moreover, it needs to share in perspective as well as an important asset that must be managed efficiently for organizational success.

Keywords: Information technology and communications · Knowledge sharing

1 Introduction

Emerging of Knowledge Management (KM) has to a great extent result to industries since it has been launched in economy era. KM is an important escalating interest for organizations, to prove the concept and theory, getting the best practices, and to achieve organization in competition through ICT. KM also is widely becoming a core competence that companies must develop in order to succeed and exist in dynamic global economy [1]. The importance of intellectual capital beyond knowledge will be able to increase efficiency and effectiveness within the organization and now widely better acknowledged in several large companies and/or small business organizations. Valuable skills and competencies have become knowledge resources, and they will be wasted unless management supports the efforts to capture, collect, gather, store, transform, and share knowledge among others.

ICT allows the process of the movement of data and information at increasing speeds and efficiencies, and thus facilitate sharing as well as organization needed for the growth of knowledge. Specifically, [2] mentioned internet and web site through WWW (World Wide Web) that becomes unlimited sources of knowledge available for all of us. Also, [3] reported that ICTs provided a major impetus for implementing of

KM applications. Moreover, as learning has accrued in the area of social and structural mechanisms, through mentoring that enable effective KS, it is possible to develop KM applications that best leverage by deploying sophisticated technologies. Also, she contend that using ICTs (e.g., Web-based conferencing) to support KM mechanisms in ways not earlier possible (e.g., interactive conversations or exchange of voluminous documents among individuals which is located at long distances) enables dramatic improvement in KM.

The top and middle managers in organizations are consistently looking for better ways to improve performance, to sustain the existence, and to increase business results by gaining new understandings better. Mechanisms of knowledge and KS application features for coordinating organization effectiveness are complex. Indeed, it has been acknowledged that requirement to represent activities into application in business process is broad and complex to be covered. Organizations recognize the important of knowledge asset belong to their employees. Therefore, it needs to develop KS application in order to help and support the easiest communication and collaboration. Depending on the KM process that is most directly supported, KM systems can be classified into four types: knowledge-discovery systems, knowledge-capture systems, knowledge-sharing systems, and knowledge-application systems [4, 5]. This paper will explore and describe the potential of ICT used by the organization for increasing effectiveness and efficiency process through the deployment of knowledge sharing system and developed specific feature application.

2 ICTs for Knowledge Sharing

Technology plays a fundamental role in creating an organizational culture (OC) for KS and an infrastructure to stimulate and enable access to internal knowledge and expertise existing in the organization. According to [10], the access to internal organizational knowledge sources was predicted as intranets. It will also play a dominant role in supporting internal knowledge due to cost-effective technical capabilities including: *“access to the legacy systems, platform independence, access to multimedia data formats, a uniform and easy-to-use, point-and-click interface, and capability for easy multi-media publication for knowledge sharing”*.

According to [11], KS is defined as the exchange or transfer process of facts, opinions, ideas, theories, principles and models within and between organizations include trial and error, feedback and mutual adjustment of both the sender and receiver of knowledge. In order to improve KS, generating new knowledge depends on the OC, and generating an interactive context must be developed and maintained. It happens when OC allows and encourages change, participation, expression of ideas, communication, and dialogue, then learning and KS are possible. For an increased efficiency, an online discussion forum may be created, which certainly will improve the processes sharing of knowledge and expertise. Managers must involve and take appropriate measures so that KS takes place, perhaps leading to the establishment of organizational changes and trust to encourage greater use of online discussion forum [12]. In terms of KS, ICTs (portals) can be useful for knowledge application. Portals (internet base) are well suited for publishing and sharing collection of documents based on the intellectual products of many subgroups.

Further, because of their flexibility in combining a variety of tools and services, portals can be customized to create a rich KS environment. In addition, [3] mentioned that knowledge-sharing systems also utilize mechanisms and ICTs that facilitate exchange. Some of the mechanisms that facilitate exchange are document minute of meeting, news, memos, manuals, progress reports, letters, and presentations. ICTs that facilitating exchange include groupware and other team collaboration mechanisms, Web-based access to data, databases, and repositories of information, including best-practice databases, lessons-learned systems, and expertise-locator systems.

3 Organizational Trust

The organization consists of a number of people connected to each other in different ways (business unit, departments, structural hierarchies, tasks, role, rules etc.). The willingness of individuals to share their knowledge in an organization heavily depends on the OC and trust. As mentioned [6], definitions of trust are various and sometimes confusing, it depends on each discipline viewing trust and from its own perspective. Organization trust will influence in KS frequency and sharing effort, that is the factor of successful applied ICTs as claimed of [7]. They also identified 14 KMS success factors, one of which is “*An organizational culture that supports learning and the sharing and use of knowledge*”.

According to [8], the facilities of infrastructure are suggested to invest on smart people and providing incentives for sharing information, besides providing enough unstructured time to communicate to each other (talk face to face). In motivating people towards KS, the according activities must be encouraged and rewarded from the top and middle management level (highest hierarchical). It is clear that sharing of knowledge is something important for the whole organization. Without this, the natural tendency will become barriers to the flow of knowledge. Since the success of a KM relies heavily on people to share their knowledge, it can be assumed that if the level of organizational trust is high, then people will have fewer barriers to share their knowledge, and consequently, the level of success of the KM should be highest [6].

Enabling factors that facilitating KM activities, should be existed on sharing of knowledge assets among individuals. Specifically, for the enabler is organizational trust which becoming attitude and mindset of people, and also is critical for facilitating KS and learning motivation in organization. Refer to [9] KM effectiveness is an integration of people relationship and technology. He states that employees’ enthusiasm and trust in others have direct influence on the ability of ICT to transfer knowledge across various departments.

4 Research Methods

The main focus of this research is to explore the underlying typical of ICTs which is used to KS application and identified the organizational categories. This study used in-depth literature review and methods to collect document report of practices KM

implementation in organizations. The analyzed object of document reports are as following: the organizational categories, field of KM areas, KM analysis and approach, and ICTs enabler as solution. All of those were conducted to diagnose the causes of organizational preference to adopt ICTs and what kind of requirements to apply the possible directions of KM's needed programs. Discussion and clarification were made with the expertise in iterative rounds in order to develop a common discourse on KM issues. The discussion results in respect to knowledge categorization and ICTs solution were to be features in KS applications. The objects of document including paper, dissertation, thesis, etc. were collected and classified. They were used to analyze the concept and theory aligned with KM in organization. Taking into account the complexity of the issues, we sought insights from the document in various purposes and scopes. From August 2012 to November 2014, we had 108 documents collection. Based on the guidance in conducting benchmarking and qualitative research methods, the data were transcribed and analyzed to identify similarities and varieties of ICTs adoption.

5 Findings

We have collected several tools applications related to KM sharing. Its offers valuable supports for sharing of knowledge in organization. Some researcher mentioned and lists ICTs technologies in varying KM life cycle phase, here we presented the KS phase.

- (a) Indeed, many ICTs contributing to KS activities are e-mail and video conferencing, virtual whiteboard and brainstorming tools, content management system, personalization tools, visualization tools and automatic recommendation tools, e-learning environment, authoring tools, technologies for automatically generating new content, mind mapping, bibliography management, artificial intelligence, networking technologies, format and standards for file transfer formats and meta data standards, and hardware by providing the necessary infrastructure for all the above mentioned [13].
- (b) Like [14] presents an alternative of enabling technologies, from decision support tools to database tools, that can be used to enable various phases of the KM life cycle. These technologies provide the connectivity needed to efficiently transfer information among knowledge workers. Authoring, interface, data capture, decision support, simulations, professional database, pattern matching, groupware, controlled vocabularies, graphic, application specific, web, cataloging, and infrastructure.
- (c) Further [15] classify the KMS technologies, specifically IT/ICTs tools being implemented, based on the Knowledge Life Cycle stage. This model has 4 stages, i.e. knowledge creation, knowledge storage/retrieval, knowledge transfer, and knowledge application. It is expected that the KMS will use specific technologies to support each stage for which the KMS was created to support.
- (d) Next [16] depicts four layers of KM. One of layers is supporting and enabling technologies including knowledge representation, semantics and ontology, unstructured

data indexing and storage, software agents, networks, knowledge organization and indexing, data mining, information retrieval, meta-knowledge and metadata, knowledge discovery, storage and retrieval, mobility, presentation and application integration, computational experimentation, artificial intelligence, data mining, security, computer mediated communication, networks, portals, encryption access control, interface, human factors, and other specific technologies impacting KM.

- (e) Moreover [17] stated internet features or technologies that support KM are common architecture and interface, easy to use front-end systems (browser user interface), internet based processes, XML wrapping of documents and other data, back-end systems that provide database access to users, Search Engines, and Virtual Private Networks.
- (f) Agreed with [18], three major KM technologies classification are acquisition and application phase, creation and capture phase, sharing and dissemination phase. KS and dissemination phase includes communication and collaboration technologies (i.e. telephone, fax, videoconferencing, chat rooms, instant messaging, internet telephony, e-mail, discussion forum, groupware, wikis, and workflow management) and networking technologies (i.e. intranets, extranets, web servers, browsers, knowledge repository, and portal).

After all, we summarized and categorized ICTs tools related to our work, and create synthesis matrix across organization KM sharing deployment. From in-depth literature review of documents, we have collected 108 and created from 2001 to 2014. All topics are representing KM implementation in organization. Concerning the object of study, Table 1 depicts the organization categories and focus on KM areas. The organization categories include government, association, industries, media, services, education, IT companies, and services, banking, manufacture, and etc.

Table 1. Percentages of organizational categories and KM areas

Organization categories	KM areas:
- Industry (33.33%)	- KM Initiatives (6.48%)
- Government (17.59%)	- KM Sharing (56.48%)
- Education (16.67%)	- KM Repository (23.15%)
- IT service (7.41%)	- KM Evaluation (5.56%)
- Media (6.48%)	- KM Distribution (2.78%)
- Banking (4.63%)	- KM Transfer (0.93%)
- Insurance (3.70%)	- Unidentified (4.63%)
- Association (0.93%)	
- Communication (2.78%)	
- Hospital (0.93%)	
- Property (0.93%)	
- Pharmacy (0.93%)	
- Service (3.70%)	

Table 1 shows the complete organization categories in Indonesia, and the relevant of KM areas based on title in the document and journals published. It depicted the total number of document collected and analyses over the period in this study. The organization categories are banking, association, communication, educational, government, hospital, industry, insurance, IT, media, property, pharmacy, services. Thus, KM areas are dominated by initiatives, sharing, repository, evaluation, distribution, transfer, and unidentified. It reflects the amount of the documents for specific organization and KM construction. The percentage of number shows on the coverage of each organizational categories and KM areas constructed. The organization is classified to industry if initial letter is “PT. XYZ”, and unidentified KM areas represented for non ICTs such as barriers of KM etc.

Table 2 shows top seven of organizations that is applying of KS applications. It is dominated with common applications, and network technologies. Several advanced technologies applied to e-learning system, content management system, decision supports, personalization tools, and limited of artificial intelligent in seeking expertise. The applications are classified to common, network, and advanced technologies. Based on Table 2, advanced technologies for sophisticated applications are still rare. It is caused by infrastructure and the development of KMS that is low and moderate. Moreover, the KM field study in Indonesia is a new concept and theory.

Table 2. Top seven organizations with knowledge sharing applications

Organizations/ applications	Ind. (33%)	Gov. (17%)	Educ. (16%)	IT (7,4%)	Med. (6.4%)	Bank (4.6%)	Ins. (3.7%)
Common applications							
Telephone/Fax	v	v	v	v	v	v	v
E-mail & video conferencing	v	v	v	v	v	v	v
Chat rooms	v		v				
Instant messaging	v	v	v	v	v	v	v
Discussion forum	v	v	v	v	v	v	v
Groupware	v		v	v		v	
Wikis	v	v	v			v	
Workflow management	v		v	v	v	v	v
Networking technologies							
Intranets	v	v	v	v	v	v	v
Extranets	v	v	v	v	v	v	v
Web Servers	v	v	v	v	v	v	v
Browsers	v	v	v	v	v	v	v
Knowledge repository	v	v	v	v	v	v	v
Portal	v	v	v	v	v	v	v
Advanced technologies							
Virtual whiteboard & brainstorming							
Content management system	v	v	v	v	v	v	v
Personalization	v		v	v		v	v
Visualization tools & automatic recommendation	v					v	v

(continued)

Table 2. (continued)

Organizations/ applications	Ind. (33%)	Gov. (17%)	Educ. (16%)	IT (7,4%)	Med. (6.4%)	Bank (4.6%)	Ins. (3.7%)
E-Learning environment	v	v	v	v		v	
Authoring tools							
Automatically generating new content	v					v	
Mind mapping							
Bibliography management							
Artificial intelligence	v	v	v			v	
Format & standards for file transfer	v			v		v	v
Meta data standards							
Authoring							
Interface	v	v	v	v	v	v	v
Data capture	v		v			v	v
Decision support	v					v	v
Simulations							
Professional database	v		v			v	v
Pattern matching							
Groupware							
Controlled vocabularies							
Graphic							
Application specific							
Cataloging	v					v	v
Infrastructure	v	v	v	v	v	v	v
Knowledge discovery							
Mobility						v	
Computer mediated communication						v	
Security	v		v	v		v	
Encryption access control	v						
Human factors							
Data mining							
Software agents							
Semantics and ontology							
Unstructured data indexing and storage							
Knowledge representation							
Information Retrieval							
Meta knowledge							

Note: Ind. = Industry; Gov. = Government; Educ. = Education; IT = IT; Med = Media; Bank = Banking; and Ins. = Insurance.

The reason for organization to adopt ICTs is to get easier in collecting data, manipulating, restoring and distributing individual knowledge. Most of all are preparing to keep and maintain individual knowledge to get solution on their existing problems. The purpose of applications is for communicating among individual or employee. The commonly used ICTs in KS application are dominated by networking

Table 3. ICTs features mostly applying in knowledge sharing applications

Knowledge sharing features:	ICTs enabler for knowledge sharing:
<ul style="list-style-type: none"> - Forum discussion - Minute of meeting - Question and answer - Email (mailing lists) - Find expertise - Chat room - News - Documentation managements (procedure, material training etc.) - Reports - Voting/monitoring - Library/Repository (employee, products, e-book, etc.) 	<ul style="list-style-type: none"> - Intranets - Network - Blog - Portals - Email - Forum Group (chat, seminar, discussion, monitoring, alert etc.) - Search/seek expertise - Artificial intelligent - Data warehouse - Information repositories (video, content document management, voice conferencing etc.)

technologies (i.e. portal, web server, intranets, and knowledge repository). Table 3 shows the summary of KS application features. It mostly preferred to apply in organizations, due to condition, situation, and available of infrastructure.

6 Discussion and Conclusion

Several points can be drawn from the documents which are:

- (a) KM is emerging in Indonesia to improve their business processes and competitiveness, even though it is starting with KS using ICTs. It represents the field of KM that has been growing since academic intensity to research and its application in organization were evident.
- (b) Most of the documents present analysis and design of KM implementation, which are expressed by various methods and approach. However, in the most cases authors do not have enough collection to represent individual knowledge in fully meeting for knowledge requirement of the work they do. Of course, in an ideal situation, each employee in a unit of organization would know which activities and tasks to all members are involved. The larger the organization, the more complex to this process becomes.
- (c) In subsequent KM process, it covers by initiation, analysis, and design phase. In most of the initiation phase, the authors would like to improve business process, how to make possible using ICTs based on KM. The analysis phase identifies the requirements of individual knowledge mostly dominated; afterwards it analyzes by knowledge mapping within tools, and the last phase the authors will design and develop application.
- (d) Requirement analysis is dominated evaluation factors externally and internally, then it leveraged by SWOT (strengthen, weakness, opportunities, threat) method for mapping core knowledge and classifying individuals knowledge needed.

- (e) Web based application are preferred to develop in order to spread on location organization and extended sophisticated technology.
- (f) It is important to develop metrics to assess benefits of application in the future, with focus on organization, human resources, technologies, culture, and trust.

This article explored the possibilities and limitations of the ICTs in supporting KS in Indonesia organization. As shown above, the ICTs can support such management in solving problems through online discussion, disseminating of documents of tasks or procedures, trusting among employees, understanding of the organization needed to increase, and establishing against to competitiveness.

References

1. Skyrme, D.J., Amidon, D.M.: New measures of success. *J. Bus. Strategy* **19**(1), 20–24 (1998)
2. Dalkir, K.: *Knowledge Management in Theory and Practice*, pp. 1–21. Elsevier Butterworth-Heinemann, Amsterdam (2005)
3. Becerra-Fernandez, I., Sabherwal, R.: ICT and knowledge management systems. In: Schwartz, D.G. (ed.) *Encyclopedia of Knowledge Management*, pp. 230–236. IDEA Group Reference (2006)
4. Becerra-Fernandez, I., Gonzalez, A., Sabherwal, R.: *Knowledge Management: Challenges, Solutions and Technologies*. Prentice Hall, Upper Saddle River (2004). 386 p.
5. Becerra-Fernandez, I., Sabherwal, R.: ICT and knowledge management systems. In: Jennex, M.E. (ed.) *Knowledge Management: Concepts, Methodologies, Tools, and Applications*, pp. 1042–1050 (2008)
6. Ribie`re, V., Tuggle, F.D.: The role of organizational trust in knowledge management: tool & technology use & success. In: Jennex, M.E. (ed.) *Knowledge Management: Concepts, Methodologies, Tools, and Applications*, pp. 1137–1154 (2008)
7. Jennex, M.E., Olfman, L.: Assessing knowledge management success/effectiveness models. In: *Proceedings of the 37th Hawaii International Conference on System Sciences* (2004)
8. Kucza, T.: Knowledge management process model. VTT Electronics. Technical research centre of finland ESPOO 2001, 10 December 2012. <http://www.inf.vtt.fi/pdf/>
9. Chan, I., Chau, P.Y.K.: Knowledge management gap: determined initiatives, unsuccessful results. In: Jennex, M.E. (ed.) *Knowledge Management in Modern Organizations*, pp. 354–370 (2007)
10. Alavi, M., Leidner, D.: Knowledge management system: issues, challenges and benefits. *Commun. Assoc. Inf. Syst.* **1**(7), 2–41 (1999)
11. Szulanski, G.: Exploring internal stickiness: impediments to the transfer of best practice within the firm. *Strateg. Manag. J.* **17**, 27–43 (1996)
12. Madge, O.L.P.: Creating a culture of learning and knowledge sharing in libraries and information services. In: Hou, H.-T. (ed.) *New Research on Knowledge Management Models and Methods*, pp. 245–268. Techopen, Croatia (2012)
13. Marwick, A.D.: Knowledge management technology. *IBM Syst. J.* **40**(4), 814–830 (2001)
14. Bergeron, B.: *Essentials of Knowledge Management*, pp. 111–132. Wiley, Hoboken (2003)
15. Alavi, M., Leidner, D.: Knowledge management and knowledge management systems: conceptual foundations and research issues. *MIS Quart.* **25**(1), 107–136 (2001)

16. Schwartz, D.: A bird-eye view of knowledge management: creating a disciplined whole from many interdisciplinary parts. In: Jennex, M.E. (ed.) *Knowledge Management in Modern Organizations*, pp. 18–29 (2007)
17. Jennex, M.E.: Internet support for knowledge management systems. In: Jennex M.E. (ed.) *Knowledge Management: Concepts, Methodologies, Tools, and Applications*, pp. 564–570 (2008)
18. Dalkir, K.: *Knowledge Management in Theory and Practice*, vol. 2, pp. 267–306. Massachusetts Institute of Technology Press, Cambridge (2011)

Quality of Transformation of Knowledge as Part of Knowledge Management System

(Research in Private University in Jakarta)

Dyah Budiastuti^(✉) and Harjanto Prabowo

Bina Nusantara University, Jakarta, Indonesia
{dbudiastuti, harprabowo}@binus.edu

Abstract. The main activity in education is the transfer of knowledge from educators with learners and among learners themselves. The knowledge can be in the form of explicit or implicit. On the other hand, the implementation of Knowledge Management conducted in almost all business and social activities, including in the field of education. This study aimed to get an idea of the condition of the implementation of the transformation of knowledge and develop knowledge management model of private university in Jakarta. Data were analyzed using descriptive tool “Radar”.

Keywords: Knowledge transformation · Model KM

1 Introduction

Indonesian public and governmental awareness of the importance of higher education continues to rise, it can be seen from the increase in the number of graduates and an increase in the percentage of educated labor. This awareness course is very good because it will increase productivity and ultimately improve competitiveness.

In the era of knowledge-based economy, competitiveness associated with increased knowledge that is intangible, such as brand recognition, patents, customer loyalty, etc., which is a manifestation of creativity and innovation which is based on knowledge, which is the outcome and be the primary responsibility of the college.

On the other hand, the development of Knowledge Management was rapid and implementation is done in nearly every business and social activities, including in the field of education.

Knowledge Management is the organization’s activities in managing knowledge as an asset, the distribution effort is needed right knowledge to the right people and in a short time, so they can interact, share knowledge, and apply it in their daily work in order to improve organizational performance.

2 Literature Review

Knowledge Management works to increase the organization’s ability to learn from their environment and incorporate knowledge within an organization to create, collect,

preserve and disseminate knowledge of the organization. Information technology plays an important role in knowledge management as a business process aimed at enabling that aim to create, store, preserve and disseminate knowledge (Loudon, 2002). Meanwhile, according to Debowski (2006), knowledge has two dimensions, explicit knowledge and tacit knowledge [1]. Explicit knowledge is knowledge that can be or has been codified in the form of documents or other tangible form so it can be easily transferred and distributed using various media. While Tacit knowledge is the knowledge that dwell in the human mind in the form of intuition, judgment, skills, values, and belief is very difficult formalized and shared with others.

Key activities that support successful implementation of knowledge management is the transformation of knowledge goes. Nonaka and Takeuchi (1995) proposed four modes of knowledge transfer with SECI model, that Socialization, externalization, Combination, and Internalization [2].

In the study of higher education, knowledge in addition to the elements forming a sustainable competitive advantage, knowledge is also the value created by the college to be delivered to consumers. (Rowley, 2000). Thus, the concept of knowledge in higher education is extracting knowledge internally and externally, both as a resource as well as the output of the process of developing knowledge management which is run by the university. Davenport (1998) divides the implementation of knowledge management in four main processes, namely (1) provides a place to store knowledge, (2) improving access to knowledge, (3) promote environmental knowledge and (4) manage knowledge as an asset [3].

In the process of creating a store of knowledge, universities need to provide a printed or electronic document, such as thesis, dissertation, research and publications, and the results of operations of other academic services. For ease of storage and retrieval, the presence of information and communication technology (ICT) is necessary. The last part is to manage knowledge as an asset, meaning that knowledge can be assigned the same value even higher than the asset value of a building, facilities, and other tangible assets.



Fig. 1. Main process of universities utilizing KM

According Jillinda J. Kidwell, Knowledge Management in college used in 5 main processes, namely (1) the product development process and curriculum, (2) poses research, (3) the process of administrative services, (4) the process of student services and alumni, (5) in the service process community (Fig. 1) below [4].

Forms of Tacit Knowledge and Explicit Knowledge Higher education by Jillinda J. Kidwell described as follows (Fig. 2).

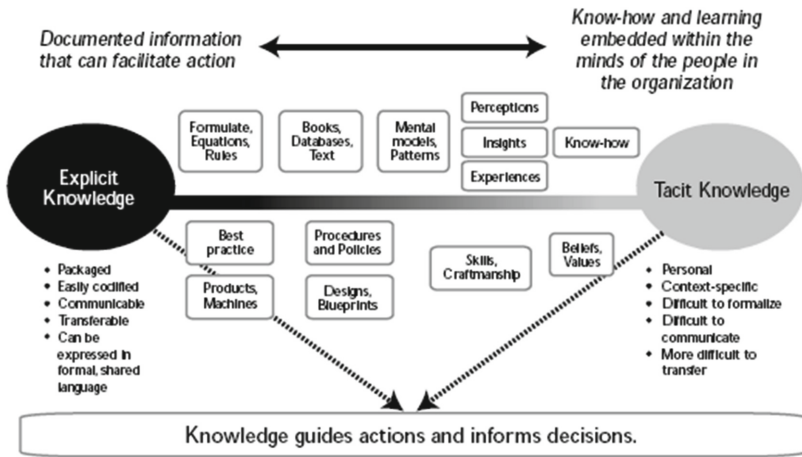


Fig. 2. Tacit and explicit knowledge

3 Methods

3.1 Samples

The population in this study were 21 private universities in Jakarta accredited institution, with respondents are 5 permanent lecturers or as much as 105 per university professor. Selected University accredited institution with the consideration that the accredited university has already been implementing knowledge management.

3.2 Samples

Data were collected through questionnaires and focus group discussion. The questionnaire aims to examine the use of knowledge management and conditions of any kind of transformation of knowledge in college.

FGD made to the leadership of the university in order to obtain confirmation of KM implementation in higher education, particularly for knowledge transformation activities and KM models.

3.3 Analysis

The data obtained were analyzed descriptively by using the tool Radar and gap Analysis.

4 Result

4.1 The Conditions of Implementation of the Transformation of Knowledge Private University in Jakarta

Implementation of the transformation of knowledge is measured by parameters:

(a). The direction and policy of the use of KM, (b). Implementation of the transformation of knowledge, (c). Benefits to transform knowledge, and (d). Award after transforming knowledge. The direction and policy of the use of KM see the clarity of the policy issued by the university to use KM. Clarity of procedures for the use of KM in Jakarta Private Universities is essential to good performance. Implementation of the transformation of knowledge Private Universities in Jakarta is very important but the condition is not good. This means that the transformation of knowledge are important but private universities in Jakarta have not run well. Knowledge transformation activities is beneficial, but the result of the transformation of knowledge is not used in full. Award after transforming knowledge in Private Universities Jakarta is very important but the condition is not good.

The following figure shows the utilization of Private Universities Knowledge Management in Jakarta (Fig. 3).

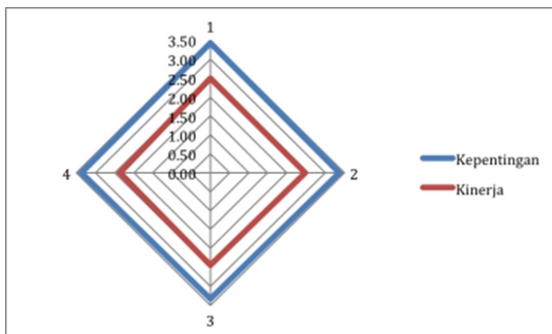


Fig. 3. KM user in private universities in Jakarta

4.2 Condition Knowledge Transformation Socialization

Regularly Private University in Jakarta has conducted seminars and workshops in academic associated with the development of the duties as a lecturer, training related to the development duties as a lecturer, invited external expert in accordance with the development of the duties as a lecturer, involving lecturers and the external experts in the development program of lecturers, providing a meeting place for the lecturers social interaction and exchange of knowledge and encourage social activities and others who

made a lecturer in off-campus, such as outbound, social service, etc. so that the sharing of experiences in order to be able to create tacit knowledge has been going well.

4.3 Condition Knowledge Transformation Externalization

Private University in Jakarta documenting ideas / experiences lecturer in learning, in research and publication, in community service activities, documenting the sequence of activities lecturer in SOPs, and documenting the competence of lecturers in accordance with science and expertise of lecturers, so that the process of articulating tacit knowledge in the form of explicit knowledge is already well underway.

4.4 Condition Knowledge Transformation Combination

Private University in Jakarta classify data / information into a file, database, and reports are easy to understand, using the knowledge they have to develop programs, procedures, and other activities associated with the lecturers, utilizing the knowledge they have to communicate with professors, and use the knowledge that developed a lecturer for the development of programs, procedures, and academic decision making. Thus the process of systematization concepts into a knowledge system by combining different bodies of explicit knowledge is well documented.

4.5 Condition Knowledge Transformation Internalization

Regularly Private University in Jakarta have already explained all the rules and procedures for faculty to be able to perform tasks tri dharma college, encourage lecturers to access all sources of knowledge are developed, providing a means of (media) to be used by lecturers in accessing the knowledge that exist, and encourages lecturers to improve their knowledge through further studies education. Thus the process of conversion of explicit knowledge into tacit knowledge that is closely related to learning by doing (internalization) has been running well.

4.6 Knowledge Management Model in DKI Jakarta Private Universities

The results of the research model developed consists of four main sections, namely (1) Leadership Approach, (2) Strengthening human resources, (3) Strengthening Process Transformation, and (4) Use of Information Technology and Social Media, as shown in Fig. 4 below.

Leadership approach is needed so that improvements and changes to the direction and policies in the implementation of KM Private Universities and Private Universities Leaders support in the implementation of KM and KM use in decision making.

Strengthening the process is carried out in accordance with the KM process itself, i.e. starting from activity (1) Creating Knowledge, (2) Saving Knowledge, (3) Divide (sharing) Knowledge, (4) Using the Knowledge, and (5) Enriching Knowledge.

Step strengthening knowledge transformation performed on each type of transformation (SECI) so it can run better. The use of information technology and social media is done to get a knowledge management system that is easier to use safely, support user mobility, and changing user behavior in the use of knowledge management.

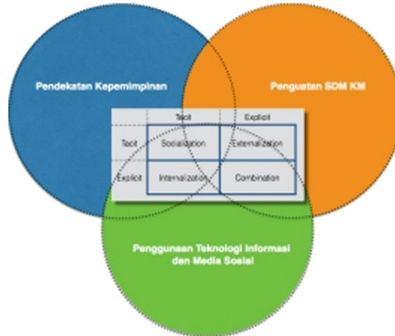


Fig. 4. KM model in private universities

Approach onto the proposed model, university leaders agree and support with the model, with some emphasis that: (1). Leadership should not be to eliminate the characteristics of academic autonomy possessed by the faculty and study program. It is they who should take the initiative to establish a system of academic KM. (2). KM become an important part of human resource development, so any kind of training followed by faculty and staff include material KM. (3). On strengthening the transformation of knowledge, activity was focused more on the establishment of explicit knowledge, so in addition there is a new culture (sharing knowledge) there is also a knowledge that can be managed more easily. (4). On the use of ICT and social media, support for investing in procurement, to be aware of the use of social media.

5 Conclusion

The condition of each type of transformation of knowledge that includes the sharing of experiences and thereby creating tacit knowledge (socialization), the process of articulating tacit knowledge in the form of explicit knowledge (externalization), the process of systematizing concepts into a knowledge system by combining different bodies of explicit knowledge (combination) and the process of converting explicit knowledge into tacit knowledge and is closely related to learning by doing (Internalization) occurring in Jakarta Private Universities is already well underway.

Model Knowledge Management in accordance with the conditions in Jakarta Private Universities is a model that was developed on 4 main sections, namely (1) Leadership Approach, (2) Strengthening human resources, (3) Strengthening Process Transformation, and (4) Use of Information Technology and Social Media.

This research is mainly focusing on Private Universities in Jakarta. For wider usage of this research on worldwide universities shall require a further research to develop.

References

1. Debowski, S.: Knowledge Management. Wiley, Australia (2006)
2. Ichijo, K., Nonaka, I.: Knowledge Creation and Management: New Challenges for Managers. Oxford University Press, Oxford (2007)
3. Davenport, T.H., De Long, D.W., Beers, M.C.: Successful knowledge management projects. *Sloan Manage. Rev.* **39**(2), 43–57 (1998)
4. Kidwell, J.J., Linde, K.M.V., Johnson, S.L.: Applying corporate knowledge management practices in higher education. In: Bernbom, G. (ed.) *Information Alchemy: The Art and Science of Knowledge Management*. EDUCAUSE Leadership Series #3. Jossey-Bass, San Francisco, pp. 1–24 (2001)

Author Index

A

Abdullah, Muhamad Taufik, [467](#)
Abelha, António, [433](#)
Abou-Tair, Dhiah el Diehn I., [698](#)
Abushaikha, Ismail, [698](#)
Ahmad Zaidi, Ahmad Mujahid, [807](#)
Ahmad, Johanna, [615](#)
Ahmed, Mudassar Adeel, [476](#)
Ahn, Shihyun, [131](#)
Ahsan, Imran, [476](#)
Al Ani, Ismail I.K., [577](#)
Ali, Fakariah Hani Mohd, [347](#)
Aljunid, Syed Ahmad, [347](#)
Alshreef, Abed, [84](#)
Ang, Jacqueline Lee Fang, [688](#)
Anwar, Muhammad Waseem, [476](#), [485](#)
Azam, Farooque, [476](#)
Azam, Shoaib, [114](#)
Azman, Azreen, [467](#)
Azuma, Godai, [533](#)

B

Bae, Min-Ho, [191](#)
Baek, Nakhoon, [139](#)
Baharom, Salmi, [615](#)
Becker, Matthias, [76](#)
Bhati, Abhishek Singh, [652](#)
Bora, Joyatri, [105](#)
Bora, Popi, [105](#)
Budiastuti, Dyah, [827](#)
Bures, Miroslav, [585](#), [594](#)
Butt, Wasi Haider, [476](#), [485](#)

C

Cam, Nguyen Tan, [298](#)
Cerny, Tomas, [585](#), [594](#), [706](#)
Chen, Cai-Jin, [625](#)
Chen, Chien-Ming, [282](#)

Chen, Fangjiong, [22](#)
Chen, Guoyue, [171](#)
Chen, Tzu-Yang, [625](#)
Chiang, Chia-Chu, [377](#)
Chiba, Takahiro, [202](#)
Chiou, Piao-Yi, [504](#)
Chmielewski, Leszek J., [512](#)
Cho, YoungHee, [791](#)
Choi, Hak-Yeol, [315](#), [331](#), [339](#), [358](#)
Choi, Jaeyeong, [97](#)
Choi, Jun Kyun, [763](#)
Choi, Sunghee, [315](#)
Chuah, Chai Wen, [231](#), [266](#)
Chung, Jinwook, [38](#), [777](#)
Chung, Tai-Myoung, [61](#)
Chung, Tsai-Yuan, [645](#)
Cui, Lijuan, [443](#)
Cui, Xingmin, [396](#)
Cuthbert, Laurie, [3](#)

D

Danpakdee, Nontachai, [553](#)
Dawood, Hassan, [114](#)
Dawson, Edward, [231](#)
Do, Van Thuan, [221](#)

E

Eng, Bah Tee, [660](#)
Engelstad, Paal, [221](#)
Eum, Jun-Ho, [191](#)

F

Feng, Boning, [221](#)

G

Guan, Quansheng, [22](#)
Gumilar, Gun Gun, [817](#)
Guo, Zhenhua, [494](#)

Guyeux, Christophe, 404
Gwon, Gyeongjae, 97

H

Ham, Young-hwan, 38
Han, Kireem, 763
Hata, Mohsen Mohamad, 347
He, Ruiyi, 396
Heinsen, Rene Ivan, 668
Heng, Swee-Huay, 306, 366
Hiari, Omar, 698
Ho, Chih-Chiang, 504
Hogrefe, Dieter, 274
Huang, Hui, 155
Huh, Eui-Nam, 668
Huh, Jun-Ho, 519, 784
Hui, Lucas C.K., 396
Hussain, Md. Anwar, 105
Huynh, Cong Viet-Ngu, 519
Hwang, Min-Shiang, 645
Hwang, Mintae, 164

I

Ibrahim, Naseem, 577

J

Jagwani, Priti, 12
Jang, Han-UI, 315, 331, 339, 358
Jang, Woo Sung, 603, 609
Janowicz, Maciej, 512
Jeon, Byung Kook, 603
Jeong, Jongmun, 164
Ji, Fei, 22
Ji, Sang-Keun, 323
Jiang, Linhua, 145, 155, 544
Jindal, Poonam, 249
Jung, Euihyun, 426
Jung, Hyo-taeg, 38

K

Kadir, Rabiah Abdul, 467
Kang, Changsoon, 164
Kang, Ji-Hyeon, 323
Kang, SeungAe, 770, 791
Kang, SunYoung, 791
Kaushik, Saroj, 12
Keong, Phang Keat, 69
Khan Khattak, Muazzam A., 485
Khan, Abdullah Aman, 114
Kim, Chang-Geun, 679
Kim, Cheolhwan, 131
Kim, Dongkyu, 315, 331, 339, 358
Kim, Han-Byul, 784
Kim, Hyoungshick, 388

Kim, Hyuncheol, 38, 745, 757, 777
Kim, Jinmo, 784
Kim, Jongmin, 519
Kim, Kuinam J., 745
Kim, Kyoung Min, 417
Kim, Minkyu, 751
Kim, Nam-Uk, 61
Kim, R. Young Chul, 603
Kim, Seunggho, 164
Kim, SooHyun, 791
Kim, Soonchoul, 777
Kitakoshi, Daisuke, 533
Klima, Matej, 585
Koh, Wen Wen, 266

L

Lai, Yi-Horng, 504
Latif, Muhammad, 476
Lazarova-Molnar, Sanja, 459
Lee, Buhm, 417
Lee, Cheng-Chi, 282
Lee, Garam, 53
Lee, Heung-Kyu, 323, 331, 339, 358
Lee, Hyunjin, 53
Lee, Manhee, 213
Lee, Soojung, 799
Lee, SooKyung, 191
Lee, Wai Kong, 688
Lee, Wonhyuk, 745
Lee, Yun-Jung, 131
Lee, Yunli, 179
Lerdsuwan, Peerapak, 714
Li, Cheng-Yi, 645
Li, Chun-Ta, 241, 282
Li, Lin, 84
Li, Mengqin, 124
Li, Qin, 494
Li, Wanli, 443
Lim, Kyung-Gyun, 679
Limpiyakorn, Yachai, 637
Lin, Xiao, 145, 544
Lin, Zhuosheng, 404
Liu, Cong, 155
Liu, Juntao, 451
Liu, Xiaolei, 46
Lopez, Cindy Pamela, 668
Lu, Dang-Nhac, 30
Lu, Haofang, 526
Ly, Hoang Tuan, 290

M

Ma, Lin, 544
Ma, Zhengyu, 22
Machado, José, 433

Mat Deris, Mustafa, 231
 Mazlan, Muhammad Syafiq Najmi, 722, 807
 Mehmood, Zahid, 114
 Miskon, Azizi, 722, 807
 Moh Nazar, Nazatul Shiema, 722, 807
 Mohamed, Nader, 459
 Mohd Tawil, Siti Nooraya, 722
 Morgado, Pedro, 433

N

Nasharuddin, Nurul Amelina, 467
 Nazir, Farhana, 485
 Neves, João, 433
 Neves, José, 433
 Ngo, Thi-Thu-Trang, 30
 Nguyen, Duc-Nhan, 30
 Nguyen, Ha-Nam, 30
 Nguyen, Tan Cam, 290
 Nguyen, Thi-Hau, 30
 Nguyen, Tri D.T., 668
 Nguyen, Tuan, 298

O

Ochnio, Luiza, 512
 Oh, Sangyoon, 191
 Ooi, Boon Yaik, 688
 Ooi, Thomas Wei Min, 688
 Orłowski, Arkadiusz, 512

P

Panithansuwan, Sutthiwat, 732
 Park, Bo Kyung, 603, 609
 Park, Byungkwon, 751
 Park, Chanjin, 745
 Park, Hyojin, 763
 Park, Jaewoo, 388
 Park, Soojung, 751
 Park, Yongtaek, 131
 Pasaribu, Boy Iskandar, 817
 Payakpate, Janjira, 257
 Peng, Yuxing, 443
 Pham, Van-Hau, 290, 298
 Phunchongharn, Phond, 714
 Phuntusil, Natthaphong, 637
 Ping, Guo, 114
 Prabowo, Harjanto, 827
 Putri, Hilda Rizanti, 817

R

Rajeh, Wahid, 84
 Rehman, Saad, 114
 Ren, Yongji, 46
 Rohajawati, Siti, 817

Røset, Connor, 377
 Runathong, Wuttipong, 732

S

Saruta, Kazuki, 171
 Sheng, Chai Yit, 69
 Shih, Dong-Her, 241
 Shirogane, Junko, 565
 Sinha, Rupali, 249
 Sirisom, Pongsak, 257
 Sohn, Dohee, 417
 Sohn, Kyung-Ah, 53
 Son, Hyun Seung, 603, 609
 Son, Jeongho, 315, 331, 339, 358
 Son, Minju, 97
 Song, Gibeom, 213
 Song, Insu, 652
 Songpan, Wararat, 553
 Suzuki, Masato, 533

T

Tamura, Yuji, 202
 Tan, Kelwin Seen Tiong, 179
 Tan, Ludan, 443
 Tan, Syh-Yuan, 306, 366
 Terata, Yuki, 171
 Thanakodi, Suresh, 722, 807
 Thi, Doan Truong, 202
 Tomasek, Martin, 706
 Trnka, Michal, 706

V

Vadivel, Sithira, 652
 van Do, Thanh, 221
 Vicente, Henrique, 433

W

Wang, Changjian, 443
 Wang, Chun-Cheng, 241
 Wang, Eric Ke, 396
 Wang, Guangyuan, 46
 Wang, Qianxue, 404
 Wang, Xi, 274
 Wang, Yapeng, 3
 Warren, Van, 377
 Weng, Chi-Yao, 282
 Wongthai, Winai, 257, 732
 Wu, Caihua, 451
 Wu, Tsu-Yang, 282

X

Xiao, Lin, 3
 Xu, Xiaofeng, 46

Y

Yan, Zhixun, [145](#)
Yang, Chao-Tung, [625](#)
Yang, Cheng-Ying, [645](#)
Yang, Jinhong, [763](#)
Yang, Xu, [3](#)
Yao, Jenq-Foung JF, [645](#)
Yeow, Kin-Woon, [306](#), [366](#)
Yi, Keunsang, [609](#)
Yiu, S.M., [396](#)
Yokoyama, Takanori, [202](#)
Yoo, Myungryun, [202](#)
Yoon, Jiyoung, [131](#)
You, Jane, [494](#)
Young Chul Kim, R., [609](#)
Yousaf, Rehan Mehmood, [114](#)

Yu, Hua, [22](#)
Yu, In-Jae, [358](#)
Yu, Simin, [404](#)
Yu, Zhihao, [155](#)
Yue, Guiyang, [155](#)

Z

Zhang, Hang, [274](#)
Zhang, Tiankui, [3](#)
Zhang, Xingguo, [171](#)
Zhang, Zi-Ke, [526](#)
Zhao, Ruohan, [494](#)
Zhong, Xiaopin, [124](#)
Zhou, Gang, [396](#)
Zhou, Ying, [526](#)
Zhu, Yiming, [155](#)