Lecture Notes in Electrical Engineering 448

James J. (Jong Hyuk) Park
Shu-Ching Chen
Kim-Kwang Raymond Choo
Editors

Advanced Multimedia and Ubiquitous Engineering

MUE/FutureTech 2017

Springer

Lecture Notes in Electrical Engineering

Volume 448

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

James J. (Jong Hyuk) Park
Shu-Ching Chen · Kim-Kwang Raymond Choo
Editors

Advanced Multimedia and Ubiquitous Engineering

MUE/FutureTech 2017

Springer

Editors
James J. (Jong Hyuk) Park
Computer Science and Engineering
Seoul National University of Science and Technology
Seoul
Korea (Republic of)

Kim-Kwang Raymond Choo
Department of Information Systems and Cyber Security
The University of Texas at San Antonio
Adelaide
Australia

Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL
USA

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-10-5040-4 ISBN 978-981-10-5041-1 (eBook)
DOI 10.1007/978-981-10-5041-1

Library of Congress Control Number: 2017940720

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

Message from the FutureTech2017 General Chairs

FutureTech2017 is the 12th event of the series of international scientific conference. This conference will take place on May 22–24, 2017 in Seoul, South Korea. The aim of the FutureTech2017 is to provide an international forum for scientific research in the technologies and application of information technology. FutureTech 2017 is the next edition of FutureTech2016 (Beijing, China), FutureTech2015 (Hanoi, Vietnam), FutureTech2014 (Zhangjiajie, China), FutureTech2013 (Gwangju, Korea), FutureTech2012 (Vancouver, Canada), FutureTech2011 (Loutraki, Greece), FutureTech2010 (Busan, Korea, May 2010), which was the next event in a series of highly successful the International Symposium on Ubiquitous Applications & Security Services (UASS-09, USA, January 2009), previously held as UASS-08 (Okinawa, Japan, March 2008), UASS-07 (Kuala Lumpur, Malaysia, August, 2007), and UASS-06 (Glasgow, Scotland, UK, May, 2006).

The conference papers included in the proceedings cover the following topics: Hybrid Information Technology, High-performance Computing, Cloud and Cluster Computing, Ubiquitous Networks and Wireless Communications, Digital Convergence, Multimedia Convergence, Intelligent and Pervasive Applications, Security and Trust Computing, IT Management and Service, Bioinformatics and Bio-inspired Computing, Database and Data Mining, Knowledge System and Intelligent Agent, Game and Graphics, and Human-centric Computing and Social Networks. Accepted and presented papers highlight new trends and challenges of future information technologies. We hope readers will find these results useful and inspiring for their future research.

We would like to express our sincere thanks to Program Chairs: Ching-Hsien Hsu (Chung Hua University, Taiwan), Yunsick Sung (Dongguk University, Korea), Kwang-il Hwang (Incheon National University, Korea), Houcine Hassan (Universitat Politecnica de Valencia, Spain), Jin Wang (Yangzhou University, China), all Program Committee members, and all reviewers for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

FutureTech 2017 General Chairs
Kim-Kwang Raymond Choo
Gangman Yi
Jiannong Cao

Message from the FutureTech2017 Program Chairs

Welcome to the 12th International Conference on Future Information Technology (FutureTech 2017), which will be held in Seoul, Korea, on May 22–24, 2017. FutureTech2017 will be the most comprehensive conference focused on the various aspects of information technologies. It will provide an opportunity for academic and industry professionals to discuss recent progress in the area of future information technologies. In addition, the conference will publish high-quality papers which are closely related to the various theories and practical applications in multimedia and ubiquitous engineering. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in these important subjects.

For FutureTech2017, we received many paper submissions, and we accepted only articles with high quality for the FutureTech2017 proceedings after a rigorous peer-reviewed process. All submitted papers have undergone blind reviews by at least two reviewers from the technical program committee, which consists of leading researchers around the globe. Without their hard work, achieving such a high-quality proceeding would not have been possible. We take this opportunity to thank them for their great support and cooperation. We would like to sincerely thank the following invited speakers who kindly accepted our invitations, and, in this way, helped to meet the objectives of the conference: Prof. Weijia Jia, Director of Cyber-Space Intelligent Computing Lab, Shanghai Jiaotong University, China, and Prof. Han-Chieh Chao, President of National Dong Hwa University, Taiwan.

Finally, we would like to thank all of you for your participation in our conference, and also thank all the authors, reviewers, and organizing committee members. Thank you and enjoy the conference!

FutureTech 2017 Program Chairs
Ching-Hsien Hsu
Yunsick Sung
Houcine Hassan
Kwang-il Hwang
Jin Wang

Organization

Honorary Chair

Doo-soon Park	SoonChunHyang University, Korea

Steering Chairs

James J. Park	SeoulTech, Korea
Young-Sik Jeong	Dongguk University, Korea

General Chairs

Kim-Kwang Raymond Choo	The University of Texas at San Antonio, USA
Gangman Yi	Dongguk University, Korea
Jiannong Cao	The Hong Kong Polytechnic University, Hong Kong

Program Chairs

Ching-Hsien Hsu	Chung Hua University, Taiwan
Yunsick Sung	Dongguk University, Korea
Houcine Hassan	Universitat Politecnica de Valencia, Spain
Kwang-il Hwang	Incheon national University, Korea
Jin Wang	Yangzhou University, China

Workshop Chairs

Chunho Jang	Keimyung University, Korea
Neal N. Xiong	Northeastern State University, USA

International Advisory Committee

James J. Park	SeoulTech, Korea (Steering Chair)
Doo-soon Park	SoonChunHyang University, Korea
Laurence T. Yang	St Francis Xavier University, Canada
Jason C. Hung	Oversea Chinese University, Taiwan
Jianhua Ma	Hosei University, Japan
C.S. Raghavendra	University of Southern California, USA
Philip S. Yu	University of Illinois at Chicago, USA
Hai Jin	Huazhong University of Science and Technology, China
Yi Pan	Georgia State University, USA
Shu-Ching Chen	Florida International University, USA
Victor Leung	University of British Columbia, Canada
Vincenzo Loia	University of Salerno, Italy
Hsiao-Hwa Chen	National Cheng Kung University, Taiwan

Publicity Chairs

Yi Pan	Georgia State University, USA
Victor Leung	University of British Columbia, Canada
Vincenzo Loia	University of Salerno, Italy
Hsiao-Hwa Chen	National Cheng Kung University, Taiwan
Laurence T. Yang	St Francis Xavier University, Canada
C.S. Raghavendra	University of Southern California, USA
Philip S. Yu	University of Illinois at Chicago, USA
Hai Jin	Huazhong University of Science and Technology, China
Qun Jin	Waseda University, Japan
Yang Xiao	University of Alabama, USA

Program Committee

Abdelbadeeh Salem	Ain Shams University, Egypt
Alfredo Cuzzocrea	University of Trieste, Italy
Bing Chen	Memorial University of Newfoundland, Canada

Caldelli Roberto	University of Florence, Italy
Chi-Fu Huang	National Chung-Cheng University, Taiwan
Epaminondas Kapetanios	University of Westminster, UK
Guillermo Diaz-Delgado	Autonomous University of Queretaro, Mexico
Huang Kuo-Chan	National Taichung University of Education, Taiwan
Homenda Wladyslaw	Warsaw University of Technology, Poland
Irene Chang	Shih-Hsin University, Taiwan
Joel Rodrigues	National Institute of Telecommunications, Brazil
Jeng-Shyang Pan	Fujian University of Technology, China
Jinli Cao	La Trobe University, Australia
Jiqiang Lu	Institute for Infocomm Research, Singapore
Kyungbaek Kim	Chonnam National University, Korea
Maumita Bhattacharya	Charles Sturt University, Australia
Mudasser Wyne	National University, USA
Niketa Gandhi	University of Mumbai, India
Qiang He	Swinburne University of Technology, Australia
Raylin Tso	National Chengchi University, Taiwan
Ren-Song Ko	National Chung Cheng University, Taiwan
Roberto Caldelli	University of Florence, Italy
Ruben Rios	University of Malaga, Spain
Ryszard Tadeusiewicz	AGH University of Science and Tech., Poland
Schahram Dustdar	Technical University of Vienna, Austria
Seng Loke	La Trobe University, Australia
Soon M. Chung	Wright State University, USA
Shu-Ching Chen	Florida International University, USA
Suren Byna	Lawrence Berkeley National Lab, USA
Tatjana Davidovic	Serbian Academy of Sciences and Arts, Serbia
Weichao Wang	University of North Carolina at Charlotte, USA
Wookey Lee	Inha University, Korea
Yu-Chen Hu	Providence University, Taiwan
Yunquan Zhang	Chinese Academy of Sciences, China
Zoubir Mammeri	Paul Sabatier University, France

Message from the MUE2017 General Chairs

MUE 2017 is the 11th event of the series of international scientific conference. This conference will take place on May 22–24, 2017 in Seoul, South Korea. The aim of the MUE2017 is to provide an international forum for scientific research in the technologies and application of multimedia and ubiquitous engineering. Ever since its inception, International Conference on Multimedia and Ubiquitous Engineering has been successfully held as MUE-16 (Beijing, China), MUE-15 (Hanoi, Vietnam), MUE-14 (Zhangjiajie, China), MUE-13 (Seoul, Korea), MUE-12 (Madrid, Spain), MUE-11 (Loutraki, Greece), MUE-10 (Cebu, Philippines), MUE-09 (Qingdao, China), MUE-08 (Busan, Korea), and MUE-07 (Seoul, Korea).

The conference papers included in the proceedings cover the following topics: Multimedia Modeling and Processing, Multimedia and Digital Convergence, Ubiquitous and Pervasive Computing, Ubiquitous Networks and Mobile Communications, Ubiquitous Networks and Mobile Communications, Intelligent Computing, Multimedia and Ubiquitous Computing Security, Multimedia and Ubiquitous Services, and Multimedia Entertainment. Accepted and presented papers highlight new trends and challenges of multimedia and ubiquitous engineering. We hope readers will find these results useful and inspiring for their future research.

We would like to express our sincere thanks to Steering Chair: James J. (Jong Hyuk) Park (SeoulTech, Korea). Our special thanks go to the Program Chairs: Naveen Chilamkurti (La Trobe University, Australia), Jungho Kang (Soongsil University, Korea), Ka Lok Man, Xi'an(Jiaotong-Liverpool University, China), Joon-Min Gil(Catholic University of Daegu, Korea), all Program Committee members, and all reviewers for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

MUE2017 General Chairs
Shu-Ching Chen
Yi Pan
Jianhua Ma
Young-Sik Jeong

Message from the MUE2017 Program Chairs

Welcome to the 11th International Conference on Multimedia and Ubiquitous Engineering (MUE 2017), which will be held in Seoul, South Korea, on May 22–24, 2017. MUE2017 will be the most comprehensive conference focused on the various aspects of multimedia and ubiquitous engineering. It will provide an opportunity for academic and industry professionals to discuss recent progress in the area of multimedia and ubiquitous environment. In addition, the conference will publish high-quality papers which are closely related to the various theories and practical applications in multimedia and ubiquitous engineering. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in these important subjects.

For MUE2017, we received many paper submissions, and we accepted only articles with high quality for the MUE2017 proceedings after a rigorous peer-reviewed process. All submitted papers have undergone blind reviews by at least two reviewers from the technical program committee, which consists of leading researchers around the globe. Without their hard work, achieving such a high-quality proceeding would not have been possible. We take this opportunity to thank them for their great support and cooperation. Finally, we would like to thank all of you for your participation in our conference, and also thank all the authors, reviewers, and organizing committee members. Thank you and enjoy the conference!

MUE 2017 Program Chairs
Naveen Chilamkurti
Jungho Kang
Ka Lok Man
Joon-Min Gil

Organization

Honorary Chair

Doo-soon Park	SoonChunHyang University, Korea

Steering Chair

James J. Park	SeoulTech, Korea

General Chairs

Shu-Ching Chen	Florida International University, USA
Yi Pan	Georgia State University, USA
Jianhua Ma	Hosei University, Japan
Young-Sik Jeong	Dongguk University, Korea

Program Chairs

Naveen Chilamkurti	La Trobe University, Australia
Jungho Kang	Soongsil University, Korea
Ka Lok Man	Xi'an Jiaotong-Liverpool University, China
Joon-Min Gil	Catholic University of Daegu, Korea

Workshop Chairs

Kwangman Ko	Sangji University, Korea
Shuo Xu	ISTIC, China
Jun-Ho Huh	Catholic University of Pusan, Korea

International Advisory Committee

Han-Chieh	Chao National Ilan University, Taiwan
Weijia Jia	Shanghai Jiaotong University, China
Borko Furht	Florida Atlantic University, USA
Thomas Plagemann	University of Oslo, Norway
Roger Zimmermann	National University of Singapore, Singapore
Stephan Olariu	Old Dominion University, USA
Koji Nakano	University of Hiroshima, Japan

Publicity Chairs

Kehua Guo	Central South University, China
Zhi Li	Guizhou University, China
Ruisheng Shi	Beijing University of Posts and Telecommunications, China
Jaehwa Chung	Korea National Open University, Korea
Ayaz Ahmad	COMSATS Institute of Information Technology, Pakistan
Qingchun Chen	Southwest Jiaotong University, China
Junbo Wang	University of Aizu, Japan

Program Committee

Afrand Agah	West Chester University of Pennsylvania, USA
Akihiro Sugimoto	National Institute of Informatics, Japan
Angel D. Sappa	Universitat Autonoma de Barcelona, Spain
Bin Lu	West Chester University, USA
Chao-Tung Yang	Tunghai University, Taiwan
Changsun Shin	Sunchon National University, Korea
Daewon Lee	SeoKyeong University, Korea
Dakshina Ranjan Kisku	National Institute of Technology Durgapur, India
Dongkyun Kim	KISTI, Korea
Guillermo Camara Chavez	Universidade Federal de Minas Gerais, Brazil
Hari Om	Indian School of Mines University, India

HaRim Jung	Sungkyunkwan University, Korea
Helen Huang	The University of Queensland, Australia
Joel Rodrigue	National Institute of Telecommunications (Inatel), Brazil
Jungong Han	Northumbria University, UK
Kilhung Lee	Seoul National University of Science and Technology, Korea
Kwang Sik Chung	Korea National Open University, Korea
Lukas Ruf	Consecom AG, Switzerland
Marco Cremonini	University of Milan, Italy
Mario Döller	University of Applied Sciences, Austria
Paisarn Muneesawang	Naresuan University, Taipei
Pascal Lorenz	University of Haute Alsace, France
Quanqing Xu	Data Storage Institute, A*STAR, Singapore
Rachid Anane	Coventry University, UK
Seunghae Kim	KISTI, Korea
Sokratis Katsikas	University of Piraeus, Greece
Yong-Yoon Cho	Sunchon National University, Korea
Zheng-Jun Zha	National University of Singapore, Singapore

Contents

Single Password Authentication Protocol

Pramote Kuacharoen(✉)

Department of Computer Science, Graduate School of Applied Statistics,
National Institute of Development Administration, Bangkok 10240, Thailand
pramote@as.nida.ac.th

Abstract. Internet users usually subscribe to a few online services. Remembering a different password for each service becomes a burden and a challenge for some. As a result, many Internet users frequently use the same password for multiple accounts. This kind of practice is risky since each service has a different security level. For example, an online community site has a weaker security measure than an online bank site. If an attacker has compromised a lower security service and obtained the user's password, the attacker may be able to identify other accounts and use the stolen password. Therefore, reusing passwords becomes a security risk, and is not generally recommended. This paper tries to mitigate the risk of reusing an identical password for multiple accounts by implementing a single password authentication protocol. The proposed protocol does not expose the user's password in the event of the server or the communication line has been breached.

Keywords: Single password authentication · Mutual authentication · Authentication

1 Introduction

Authentication is the process of verifying the identity of the subject by comparing one or more factors against a database of valid identities, such as user accounts. There are three categories of authentication factors of verifying a user's identity [1, 2, 3]:

- Knowledge factors are something the individual knows which include a password and a personal identification number (PIN).
- Possession factors are something the individual has, such as a smart card and cryptographic token.
- Inherence factors are some individual characteristics or something the individual does which include static and dynamic biometrics such as fingerprint, retina, face, voice pattern, handwriting characteristic, and keystroke dynamics.

The knowledge factors are the most commonly used form of authentication. The user is required to supply a secret such as a password, which matches the one stored in the server in order to be authenticated. Due to its simplicity and convenience, password-based authentication is widely adopted.

Many online services require users to register and provide the credentials in the form of username and password. The users can later authenticate themselves using the

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_1

registered user credentials. Internet users usually subscribe to more than one online services. For security reasons, the users should use a different password for each service. However, remembering a different password for each service becomes a burden and a challenge for some. As a result, many Internet users frequently use the same password for multiple accounts. This kind of practice is risky since each service has a different security level. For example, an online community site has a weaker security measure than an online banking site. Even worse, a lower security site may store passwords in clear text form and may run a buggy web application. Attackers may be able to use an SQL injection attack to obtain the user's password [4].

If an attacker has compromised a lower security service and has obtained the user's password, the attacker may be able to identify other accounts and use the stolen password. Therefore, reusing passwords becomes a security risk, and is not generally recommended.

In this paper, a single password authentication protocol is proposed to mitigate the risk of reusing an identical password for multiple accounts. The proposed protocol does not expose the user's password in the event of the server or the communication line has been breached. Although the protocol helps reduce the risk, the user should be cautious using an identical password for multiple accounts. There are other attack techniques which can be deployed to steal password. For instance, an attacker may perform a shoulder surfing attack while the user is authenticated in a public place.

2 Background and Related Work

This section provides background and related work which includes password authentication and existing single password protocols.

2.1 Password Authentication

Password authentication is the simplest form of authentication. Because of its simplicity, the password authentication is commonly used in authenticating the user over the Internet. The user registers a username and a password as credentials. The server needs to store the user's credentials in order to verify the user. When the user logs into the server, the user simply presents the username and the password to the authenticating entity [5, 6]. The server checks the user's credentials against the database of valid identities. If the values match, the user is authenticated. Otherwise, the authentication fails.

The passwords stored at the server must be protected. Storing passwords in clear text is inadvisable. This is because a compromised user database file reveals all passwords. The attack can use the obtained passwords on other systems. The attacker may be able to obtain the password database through an SQL injection attack. Instead of storing passwords in clear text, the hash values of the passwords should be saved. However, a weakness of using the hash value is that two users with the same password have an identical hash value. Furthermore, the attacker can use dictionary attack against the entire user database. The best way to protect password is to employ salted password

hashing. The server randomly generates a random number called salt and calculates the hash of the salt and the password. Therefore, two users with the same password have different salted passwords. The server stores the salt and the salted password along with other information. Upon logging in, the user supplies the username and the password. The server computes the hash value of the salt and the received password, and then compares the resulting value to the stored hash value. If the two values are identical, the user is authenticated.

Both username and password are sent to the server. Sending password in clear text is vulnerable to eavesdropping. Using an SSL connection helps protect the conversation during transit. However, it does not prevent phishing attacks.

2.2 Single Password Protocols

Rose et al. present a method to improve password security and to defend against password phishing [7]. The server stores the hash of the user's password and the domain name. When the user enters the username and the password is prefixed with two escape characters, the browser extension applies a cryptographic hash function to a combination of the password and the domain name. The username and the hash value are sent to the server. The domain name is automatically obtained. If the user enters the credential on the phishing page, the phisher cannot obtain the clear text password. Moreover, the hash value is different from the one stored on the actual server since the phishing site is in a different domain name. Using domain name as salt has a drawback. The attacker may compromise a server under the same domain and may set up a phishing page. The correct salted password can be captured. Since password and domain name remains unchanged, the salted password is the same, making it susceptible to replay attacks.

Gouda et al. propose a protocol that allows a client to securely use a single password across multiple servers and prevents phishing attacks [8]. The client can be authenticated without revealing the password to the server at any point. The protocol employs a one-time ticket technique. The client sets the next authentication ticket. The ticket consists of the hash of the random number, the password, and the server name. The client identifies himself by sending the identity. The server challenges with previously stored random numbers. Subsequently, the client computes the ticket using the received number. The client also randomly generates a number and uses it to create the next authentication ticket. The client responds with the current ticket, the next challenge random number, and the hash value of the next authentication ticket. Although the clear text password remains unchanged, the ticket changes each time the client is authenticated. This is equivalent to changing the password at the server every time the user signs in, which makes the protocol susceptible to message modification attacks. Consider the scenario where an attacker intercepts the response from the client. The attacker then can create a ticket using his password and replaces the hash value with the one generated from his ticket. The server has no way to verify the authenticity of the hash value. Hence, the attacker can log in.

Acar et al. also propose several single password authentication protocols which can allow a user to use a single password to authenticate to multiple services securely [9]. The proposed protocols employ blind signatures.

3 Proposed Authentication Scheme

One of the objectives of the proposed single password mutual authentication protocol is to help mitigate the risk of using the identical password for multiple accounts while achieving user authentication without revealing the password to the server at any point. Another objective is to be resilient against known attacks such as password database attacks, server spoofing attacks, and denied of service attacks. The protocol also allows multiple authentication sessions.

3.1 Storing Passwords

It is crucial that the user's credentials are protected even though the user database has been compromised. The attacker should not gain knowledge from it. Therefore, the password should not be stored or sent as clear text. Traditionally, for each user, the server stores the username, the salt, and the hash value of the salt and the user's password. This protects the user's credentials and defends against dictionary attacks and pre-computed rainbow table attacks. However, both username and password are sent to the server to be authenticated.

To achieve the design objectives, the authenticator must be derived from the user's password and must be server specific. Figure 1(a) shows the user database. Each row consists of username, salt, and masked secret. The salt is credential specific. In other words, each user is randomly assigned a salt. This prevents a dictionary attack on the entire database. The attacker must pick an individual to attack. The masked secret is calculated by XORing the user's secret and the mask. The user's secret is derived from the username, the password, and the Fully Qualified Domain Name (FQDN), i.e., Hash (username ‖ password ‖ FQDN). The mask is the hash value of the server's secret and the user's salt, HMAC(K_S, salt). The server's secret is not stored on disk. It is inputted when server starts. Therefore, a compromised database does not reveal the server's secret or the user's secret.

3.2 The Protocol

The challenges in authentication over an insecure network are confidentiality and timeliness. To prevent compromise of the user's secret and masquerades, the user's password must not be sent in clear text. The second issue, timeliness, is important because of the threat of message replays. An attacker may be able to impersonate the user by replaying the authentication exchanges.

Figure 1(b) summarizes the basic authentication dialog. First, the client authenticates itself to the server. The authenticator is generated using HMAC with the key derived from the user's password. The information includes the user's ID, the address

of the client (AD_C), the address of the server (AD_S), a random value N_1, and the times. Times are a valid period which consist of start time and end time. Second, the server verifies the client's authenticator. If it is valid, the server authenticates itself to the client with a challenge N_2, to prevent message replay. Session Authentication Code (SAC) is added so that the server does not have to keep state information. Third, the client verifies the identity of the server and response to the challenge.

username	salt	masked secret
u_1	$salt_1$	$K_{c1} \oplus HMAC(K_s, salt_1)$
u_2	$salt_2$	$K_{c2} \oplus HMAC(K_s, salt_2)$
u_3	$salt_3$	$K_{c3} \oplus HMAC(K_s, salt_3)$

(a)

1. C→S: $ID_C \| AD_C \| AD_S \| N_1 \| Times \| Authenticator_{C1}$
2. S→C: $ID_C \| AD_C \| AD_S \| N_1 \| N_2 \| Times \| Authenticator_S \| SAC$
3. C→S: $ID_C \| AD_C \| AD_S \| N_2 \| N_3 \| Times \| Authenticator_{C2} \| SAC$

$Authenticator_{C1} = HMAC(K_C, ID_C \| AD_C \| AD_S \| N_1 \| Times)$

$Authenticator_S = HMAC(K_C, ID_C \| AD_C \| AD_S \| N_1 \| N_2 \| Times)$

$Authenticator_{C2} = HMAC(K_C, ID_C \| AD_C \| AD_S \| N_2 \| N_3 \| Times)$

$SAC = HMAC(K_S, ID_C \| AD_C \| N_2 \| Times)$

(b)

Fig. 1. (a) User database. (b) Summary of the authentication exchanges.

4 Security Analysis

4.1 Compromised Password Attack

A compromised password attack can be conducted by eavesdropping the authentication exchange or obtaining password database. Since the password is not sent to the server, an eavesdropper will not be able to obtain the password. A compromised database will not reveal the password. As mentioned earlier, the password is converted into a secret key and is stored as masked secret by XORing with hash of the server secret and user's salt. Security of the password relies on the security of HMAC algorithm.

4.2 Replay Attack

A replay attack involves capturing the authentication exchanges and subsequent retransmitting the captured message to impersonate the user. The protocol employs both timestamp and challenge/response approaches to thwart replay attacks. If an attacker replays the first message, the attacker must use the same address as the user and must be done within a valid period. Verification at the server yields valid. However, the server sends a new challenge. Without knowledge of the user's password, the attacker will not able to generate a valid response.

The protocol is designed in such a way that the server is stateless. The attacker may replay the third exchange from the same address as the user and within a valid period. However, the server stores unexpired SACs of successful authentication. Therefore, the server knows that the received message is a replay.

4.3 Phishing and Man-in-the-Middle Attacks

Phishing is a type of internet fraud that attempts to acquire a user's credentials by deception. The attacker sets up a website that masquerades as a trustworthy site. Phishing messages are sent to lure to the user to visit the site. Unknowingly, the user enters the username and password which are captured by the phisher.

Consider the phishing attack scenario where the attacker entices the user to log into a bank's site through the attacker's site. The attacker's address is used as a component to derive the key to create an authenticator. Therefore, the attacker cannot use the authenticator.

Suppose that the attack captures the first authentication exchange and uses it at another location. Authenticator verification will fail because authenticator is client address dependent. The server can obtain the client's address from the connection, not the one indicated in the message.

5 Conclusion

The proposed single password mutual authentication protocol helps mitigate the risk of using the identical password for multiple accounts while achieving user authentication without revealing the password to the server at any point. Therefore, a compromised password database does not reveal passwords. A user's secret is server dependent since it is derived from the username, the password and the server name. Hence, the attacker cannot use a compromised user's secret on another server. The analysis shows the proposed protocol can reduce the risk of using the same password for various services.

References

1. Kim, J.J., Hong, S.P.: A method of risk assessment for multi-factor authentication. J. Inf. Process. Syst. **7**(1), 187–198 (2011)
2. Stallings, W.: Cryptography and Network Security: Principles and Practice, 7th edn. Prentice Hall, Upper Saddle River (2016)
3. Kennedy, E., Millard, C.: Data security and multi-factor authentication: analysis of requirements under EU law and in selected EU member states. Comp. Law Secur. Rev. **32** (1), 91–110 (2016)
4. Kuacharoen, P.: A practical customer privacy protection on shared servers. In: International Conference on Information Theory and Information Security (ICITIS), pp. 525–529 (2010)
5. Internet Security Glossary, IETF RFC 2828-2000
6. Liao, I., Lee, C., Hwang, M.: A password authentication scheme over insecure networks. J. Comput. Syst. Sci. **72**, 727–740 (2006)
7. Ross, B., et al.: Stronger password authentication using browser extensions. In: 14th Usenix Security Symposium, pp. 17–31 (2005)
8. Gouda, M.G., et al.: SPP: an anti-phishing single password protocol. Comput. Netw. **51**(13), 3715–3726 (2007). Elsevier
9. Acar, T., et al.: Single password authentication. Comput. Netw. **57**(13), 2597–2614 (2013). Elsevier

Performance Analysis of Congestion Control for Massive MTC Networks

Yi-Yen Chen, Yu-Chee Tseng, and Jyh-Cheng Chen(✉)

Computer Science Department, National Chiao Tung University, Hsinchu, Taiwan
yiyen.cs00g@nctu.edu.tw, {yctseng,jcc}@cs.nctu.edu.tw

Abstract. This paper describes scheduling and processing of congested packets in massive Machine Type Communications (MTC). When there are too many uplink packets, it may cause congestions in the Radio Access Network (RAN) and the Evolved Packet Core (EPC). To solve this problem, we propose a Critical Random Early Detection (CRED) method to compute the current average queue length for each process to determine whether a packet needs to be discarded or not to prevent from congestions. The proposed method is evaluated through NS-2 simulations.

Keywords: Machine Type Communications (MTC) · Cellular network · Congestion

1 Introduction

The term, Machine Type Communications (MTC), comes from the 3GPP (Third Generation Partnership Project) specification [1]. Nowadays, sensors, actuators, and RFID/NFC, are used to collect information and to bring convenience to our daily life, making the number of these devices increases exponentially in recent years. Once generated sensed data, those devices are scheduled to report their data to an evolved NodeB (eNB). Given the number of the devices is large, it may cause significant congestions and latency in the network.

In this paper, we focus on designing the management mechanism of scheduling the MTC devices transferring the packets to eNBs and Packet Data Network Gateways (P-GWs). To achieve and solve the congestion problem, we propose the Critical Random Early Detection (CRED) Method to decide if an MTC packet should be discarded for releasing network congestion. Our method is based on three steps to prevent the network congestion: (i) According to the MTC devices at the fixed position periodically sends data to eNB and P-GW, the sender packet time must match service time at eNB and P-GW; (ii) According to the data urgent and real-time priority to give higher priority; and (iii) use Early Detection Gateways for Congestion Avoidance [4] to migrate the MME/S-GW by drop congestion MTC packets at the radio access network (eNBs). It drops congestion packets by computing the average queue length of the arriving packets.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_2

1.1 Related Works

About the scheduling related mechanism, we know that in Loss type queuing system queue space is zero and does not allow queuing, when the MTC device arrives OM2M gateway. If all channels (channel) are occupied, that is, the MTC device transmitted the data packet to be discarded, and assume no re-transmission.

Another scheduling related mechanism is waiting type queuing system; we list them as follows:

FCFS (First Come First Service): According to the order of the queue, the system first serves the jobs in front of the queue.

LCFS (Last Come First Service): According to the order of the queue, the system last serves the jobs in front of the queue.

PS (priority): The system serves the jobs based on their priority.

SIRO (Service in Random Order): The jobs are served randomly.

We also introduce congestion relate work as follows.

Backoff window method has some issues when the backoff time may lead to retransmit the information time and cause congestion.

ACB (Access Class Baring) [2] mechanism provides a way to bar a range to access class user. However, packets' priority cannot be identified by the ACB so that high priority packets are also banned by the ACB.

There are four drawbacks when adopting CAAC (Congestion-Aware Admission Control) [3]. First, we need to find new reference recalculated as a benchmark to calculate congestion level. Secondly, the method of calculation of required sample time of the first N times and Random Early Detection only need to calculate the percentage of the average size of a Queue. Thirdly, it will be put forward to reduce congestion level reject need to spend more time and computational complexity to calculate congestion level. Fourthly, from the calculation of the average length of Queue it will be able to determine the current congestion length and therefore do not take the time to calculate congestion level.

Mass MTC devices for the LTE network characteristics, taken in conjunction mixed queuing system, loss type and wait for type queuing system is an ideal network model, the main principle of this mechanism is to allow queued but not infinitely long queue time, so do not allow infinite queue space, mainly in the following ways:

Queue length is limited, limited waiting space. When the number of queued packets length of service or transmission of the MTC device exceeds a predetermined, based on the MTC device statistics do good Average Queue length packets discarded, and continue to wait for the next delay time to uplink the data.

In addition, MTC's sensors can sometimes detect some urgent data, the packet such as ETWS signal by priority, so the data for the characteristics of the urgency of the needs in the schedule to give a higher queue weights.

Longer needs about time for continuous transmission of real-time video streaming data given long service time, so the data for the characteristics of the needs in the schedule to give a higher queue weights. Therefore, Sect. 2, we adopt [4] to take the basic idea and adjust the function to suit the 3GPP-based cellular network environments, we called it Critical Random Early Detection Method (CRED).

2 Our Critical Red Method

In this section, we introduce our CRED method. We first describe where the congestion may happen in a 3GPP-cell network environment in Sect. 2.1. Then we will illustrate how to implement the method in Sect. 2.2.

2.1 Congestion Point Due to Massive MTC Devices

One big challenge is that a 3GPP-cellular network needs to process the bursty situation when there is a large number of data packets arriving at the same. It may cause congestion in Radio Access Network (RAN) and the Evolved Packet Core (EPC) [5] as shown in Fig. 1 and [6, 7].

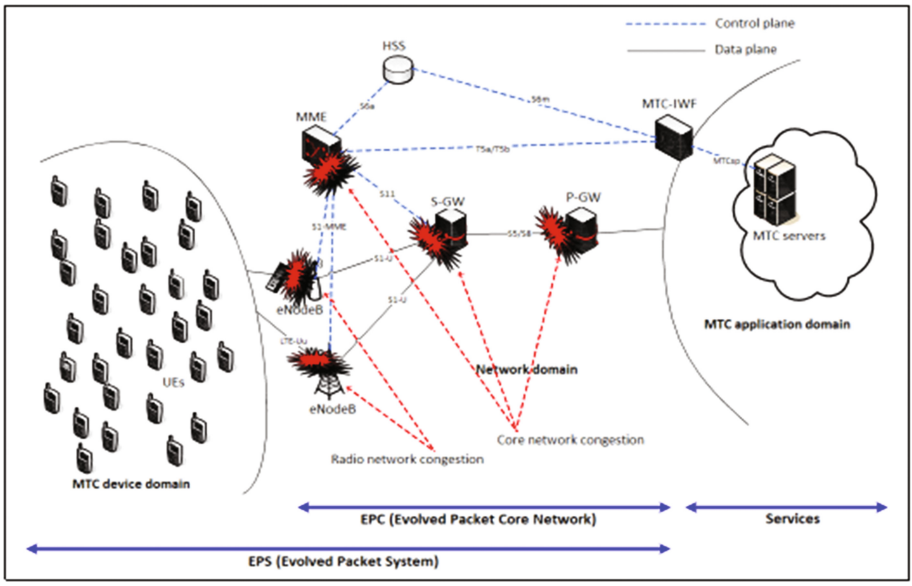


Fig. 1. Congestion in Radio Access Network (RAN) and the Evolved Packet Core (EPC)

2.2 Critical Random Early Detection (CRED) Method

MTC devices at the fixed position periodically send data to eNB and P-GW. In order to avoid congestion, the packet arrival rate should be less than the service rate of eNB and P-GW. CRED drops packets before the queue is full, according to the data urgent and real-time priority. It reduces queue delay and increases utilization. CRED also uses average queue length to predict impending network congestion and randomly discard packets. CRED calculates the average queue length using exponential weighted average approach, that is

$$MTCq_{avg} = (1 - w_q) \times MTCq_{avg} + w_q \times q \tag{1}$$

The following notation explains the meaning of (1) formula,

$MTCq_{avg}$: Average Queue Length
q : Current Queue Length
w_q : for mMTC device emerengency data and Realtime data Weights, range is $0 < w_q < 1$

Also, we decide to discard packets based on two thresholds $\min_{th}, \max_{th}$. If $MTCq_{avg} < \min_{th}$, all packets are allowed to enter queue and if $MTCq_{avg} > \max_{th}$, all packets are discarded. If $\min_{th} < MTCq_{avg} < \max_{th}$, packets are discarded by Eq. (2)

$$P_{real} = P_{now} \div (1 - \sum Packet \times P_{now}) \tag{2}$$

The following notation explains Eq. (2),

$$P_{now} = P_{operator} \times (MTCq_{avg} - \min_{th}) \div (\max_{th} - \min_{th})$$

P_{now} : Current Drop Packet Probability
Max_p : Operator Default Packet Probability
P_{real} : $P_{now} \div (1 - \sum Packet \times P_{now})$
$\sum Packet$: After last Packet Drop, the total paket number of entering queueing

3 Simulation

3.1 Simulation Settings

We implemented the CRED method by using the NS-2 simulator. In the simulation, the network topology consists of one MME, one S-GW, one eNB, 100 Mb Bandwith and 75 ms delay. We considered two scenarios with 1000 and 10000 devices, respectively. For the simulation, the huge packets come at the same time, when 1000 devices under one eNB, send packets range from 4 to 9 packets. Our simulations performance for 1000 devices is shown in Table 1.

3.2 Throughput

Table 1. Simulate 1000 devices throughput

No.	Sent packet	Lost packet	Throughput	Retry	Utilization
996	8	4	0.0016005722792832168	3	0.020007153491040208
997	8	4	0.0016005722792832168	3	0.020007153491040208
998	4	2	0.00053352409309440562	2	0.01333810232736014
999	9	4	0.0021340963723776225	3	0.023712181915306916
1000	6	3	0.0010670481861888112	3	0.017784136436480188

4 Conclusion

In this paper, we proposed CRED to avoid network congestion. The CRED takes network statistics, e.g., average queue length into consideration and then design algorithm to drop packets. Throught extensive NS-2 simulations, the results show that CRED can reduce network congestions.

References

1. 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Service requirements for Machine-Type Communications (MTC); Stage 1 (Release 12)," 3GPP, Tech. Spec. 3G TS 22.368, December 2012
2. Leyva-Mayorga, I., Tello Oquendo, L., Pla, V., Casares-Giner, V.: Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks. In: 2016 IEEE International Conference on Communications, ICC 2016, May 2016
3. Amokrane, A., Ksentini, A., Hadjadj-Aoul, Y., Taleb, T.: Congestion control in the context for machine type communication. In: IEEE International Conference on Communication (ICC), Ottawa, Canada (2012)
4. Floyd, S., Jacobson, V.L.: Random early detection gateways for congestion avoidance. IEEE/ACM Trans. Netw. (1993)
5. Bouallouche, D.: Congestion control in LTE-based machine type communication. Bibliography report, University of Rennes, January 2012
6. Sasada, K., Tanaka, I., Koshimizu, T.: Core network infrastructure and congestion control technology for M2M communication. NTT DOCOMO Tech. J. **15**(2), 31–36 (2013)
7. 3GPP "Architecutre Enhancements to Facilitate Communications with Packet Data Network and Applications version 11.0.0," 3G TS 23.682, March 2012

How to Train People to Increase Their Security Awareness in IT

Agata Niescieruk[1], Bogdan Ksiezopolski[1,2], Radoslaw Nielek[1], and Adam Wierzbicki[1(✉)]

[1] Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
agata.niescieruk@pjwstk.edu.pl
[2] Institute of Computer Science, Maria Curie-Sklodowska University, Pl. Marii Curie-Sklodowskiej 5, Lublin, Poland

Abstract. One of the key issues concerning IT systems is Information Security Management. Among the security objectives in the ISO/IEC 27002:2013 standard refers to information security awareness, education and training. In this area there are many important aspects but in this paper authors focus on people, their knowledge and their security awareness. Authors introduce a model that could illustrate organization members, their relations and knowledge about security. Results of simulations can be used to create plans of training to increase their security awareness. Finally authors present few cases where different strategies of teaching people are tested and the analysis is presented. If knowledge does not change under the influence of co-workers, it is better to train those with smallest knowledge.

Keywords: Security management · Security awareness · Social aspects of security

1 Introduction

When considering IT systems the Information Security Management [3] is a crucial aspect. Growing popularity of online applications increases the number of computer criminals [6]. One of the most threatened things is privacy [10].

The best approach to introducing security governance to the organization is to apply one of the security standard such as ISO/IEC 27001. which requires a systematic examination of information security risks [4], taking into account threats, vulnerabilities, and impacts. It is a very difficult task to comply with regulations and recommendations from the standard [2]. The problem is that some of aspects - among them improvement of security awareness of the employees through training - are not given enough attention.

Knowledge of privacy policy, security standards and consequences of noncompliance encourages and enables employees to behave responsibly in terms of IT system security. In [1] authors show that conceptual knowledge and procedural knowledge can increase security awareness of users in context of phishing threats.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_3

Employee education is important, however who and when should be taught is not so obvious. The ISO/IEC 27002 only suggests it should be scheduled in time and repeated regularly. In this paper authors propose a model, which abstracts the security awareness of employees in order to aid training planning with regard not only to costs but also relations between employees and knowledge of individuals. After this authors present an implementation of the model in the NetLogo tool and analyse different cases with different conditions, proposing strategies which should be used.

2 Related Work

Security awareness education is an important issue in IT, widely described in literature. Authors believe that there is no one universal solution [9] and specific needs for an organization have to be considered when choosing how to train coworkers. The problem is rather multi-dimensional (assessment, identification and education phases). However different scientists specify other layers/dimensions. According to [7] a variety of aspects has to be considered while talking about information security and though teaching this matter: strategic, policy, best practice and many others. Unfortunately, organizations still do not appreciate training employees. Certainly, security awareness is revelant part of Information Security Management and should consider many different aspects as mentioned in ISO/IEC 27002 (e.g. different delivery media, periodicity).

3 Model

To describe problem and possible solutions authors propose the following model. It is a simplification of a real word. There are five main steps required for configuration of the model:

- 1 - define social network of co-workers;
- 2 - define the required costs of the training;
- 3 - define method of teaching;
- 4 - define desired results;
- 5 - evaluate employees' security awareness.

3.1 Step 1 - Defining Social Network of Co-workers

Given company consists of different departments in which employees (x_1, ... x_n, $|n|$ - number of employees) work and cooperate. When two separate workers collaborate we have some specific connection between them (relation R_{xy} - worker x collaborates with y and x treats y as his colleague; relation R is neither symmetric nor transitive). And hereby a social network of co-workers is obtained. For all people in that network we can state their levels of security awareness (s_i for worker i) and list of all employees they work with ($F_i = \{x_j : j \in \{1..n\}, i\ 6 = j, R_{ij}\}$). The organization could gather such data using even a simple questionnaire on privacy policy, threats and vulnerabilities. To sum

up, in this step a description of social network should be created, containing information about security awareness of each employee and connections between employees.

3.2 Step 2 - Defining the Required Costs of the Training

The goal is to increase employees security awareness through education. Training programmes require funding, the amount of funding will be determined by how many people will be trained (a) and how long/how often (b). The desired outcome is a situation in which every employee has at least certain level of security awareness. Nevertheless often the goal will not be possible to achieve within the available budget and therefore a maximum number of trainings (B) will be set.

3.3 Step 3 - Defining Method of Teaching

When the description of social network and planned costs are available, the next step is the choice of training method. There are 4 strategies of teaching available - random, teaching those with lowest security awareness, teaching those with lots of connections, teaching those with few connections. Authors assume that employee with well informed colleagues will also change his/her attitude towards security. However this attitude could change for other external reasons (one can participate in other training/conference, one can improve his/her knowledge due to personal situation, etc.). Authors believe that importance of social ties and external factors may differ depending on organization. It should be considered in this step. Another aspect that should be taken into consideration is effective method of teaching. There is a variety of techniques, taking advantage of different learning styles, however discussing them is not the goal of this paper.

3.4 Step 4 - Defining Desired Result

The goal of training has to be set (S). All employees need to reach at least some certain level of security awareness ($\forall_{i \in \{1 \ldots n\}} s_i \geq S$), as the weakest link in the organization should be strong enough to ensure adequate security. This is the moment to choose that level. Employees will be trained in order to obtain this level.

3.5 Step 5 - Evaluation of Workers Security Awareness

At the beginning all workers in the network have their own values of security awareness (s_i). In each step a chosen number of people will be trained. Their security awareness increases - authors assume it holds in accordance with the square root function

$$s_i := 100\sqrt{s_i/100}$$

and has values in [0, 100]. However not only teaching influences security awareness. Each time for each single person all employee's friends need to be examined ($F_i = \{x_j j \in \{1, \ldots, n\} j \neq i : R_{ji}$). Importance of their connections (I_f) and how significant is randomness (I_r; there could be external factors that increase the coefficient) may also increase security awareness

$$\left(s_i = max \left\{ s_i, \frac{100 I_f}{I_f + I_r} \sqrt{\frac{\sum\limits_{k \in F_i} s_k}{|F_i|}} + \frac{100 I_r}{I_f + I_r} \right\} \right).$$

3.6 Visualisation

Authors used NetLogo to implement their model. NetLogo can be downloaded from [8] and used to run one's own simulation. Workers are called agents here (x_1, .., x_n). As there are connections between employees - here their relations (friendship, R_{xy}) are denoted by grey lines. Red background is proper for agents with low security awareness, green is for those whose coefficient at least equals `minimal-level-of-security-awareness` (S). If `network-from-file` is switched on then whole social network structure (relationships and value of security awareness for each agent) is taken from file. Otherwise one has to set `agents-number` with proper `avg-friends` to create network where each agent has `avg-friends friends` on average. His/her security awareness is randomly chosen with normal distribution taking as a distribution parameters both `security-mean` and `security-sd`. If we upload structure of this network from file then first line should contain number of agents and rest of the lines consist of 3 numbers - first is an agent, second - his/her friends and third one - security awareness (s_i) of an agent. In each step (`max-ticks` (B) is an upper limit for number of steps) some set percent of agents (`percent-of-educated` $a/b \cdot 100\%$) is trained. Those that are being taught are chosen based on of the strategies `who-is-educated`: `random, most-friendly, less-friendly` or `with-lowest-secure-awareness`. In each step we check for every agent how his/her friends are educated and according to friends-importance (I_f) and random-importance (I_r) security awareness s_i could be increased. The whole simulation ends either after `max-ticks` (B) steps or when each single agent i in the network reaches security awareness.

(s_i) at least equal to value of `minimal-level-of-security-awareness` (S). During that process we can observe (`low-level-of-security-awareness`) how many agents have too low value of coefficient and what is total cost of education (calculated as product of number of trained agents in each step and number of steps).

4 Simulations

To run a simulation initial parameters have to be set, either via a file with the network structure or setting the following parameters: `agents-number` (n), `avg-friends`, `security-mean` and `security-sd` (or leaving all these parameters with their initial values). Simulation will stop either when all agents will be educated enough or after finite number of steps in case reaching goal is impossible or will last too long. Below examples with different values of input parameters are presented. As parameters can be easily changed almost any existing organization can describe its employees and goals and run the simulation. Authors take a social network structure from a file [5] and randomly assign agents to security awareness. It is an example of an explored network of 36 workers in a small hi-tech computer firm, where there are 147 connections between co-workers. Authors present two cases chosen from all ran simulations as the most representative.

4.1 Case 1

All the parameters and their initial values are listed on Table 1. This time agents with the biggest number of friends were trained. As a result (simulation was run 100 times), after 250 ticks only few agents appear in green area (their security awareness is at least equal to 75). This mean - goal is not achieved. What can be noticed is a fact that in each step the same agents are trained and they have bigger and bigger security awareness. Remaining agents stay with fixed value of the coefficient or it grows but very slowly. Changing `percent-of-educated` only in limited way improves results.

Table 1. Inital values of parameters - Case 1 and 2

Case	1	2
network-from-file	On	On
friends-importance	100	100
random-importance	20	20
who-is-educated	most-friendly	with-lowest-secure-awareness
percent-of-educated	10	10
minimal-level-of-security-awareness	75	75
max-ticks	250	250

4.2 Case 2

It is the most interesting case. Still the same social network as before is used. The only difference to previous case is that here agents with the lowest value of security awareness are trained (Table 1).

It is visible that this method of teaching applied to chosen initial values of parameters gives better results. After running simulation 100 times, after around 28 ticks (min. 26, max 29, avg 27.7) all agents are located in the green area, so with desired value of coefficient. As one can notice, results here are slightly better than in Case 1.

5 Conclusions

Company, on the one hand, has limited funds and time but on the other its employees differ in regards to their security awareness and relations with other co-workers. Authors propose simple model which includes all these factors. Four basic ways of teaching co-workers are available. In both presented cases authors described simulations and results for existing social network of workers. and shown that the best strategy is to focus on those workers characterized by low value of security awareness coefficient.

What is surprising is a fact that changing some of the parameters do not influence results much.

Worth mentioning is the parameter `percent-of-educated`. It is obvious that for higher values of this number goal is achieved faster. However not only the number of ticks matters. Each tick costs. When more co-workers are educated at the same time the single tick cost is higher.

Authors can not claim that only training people with lowest values of security awareness is the best approach. However, if employees' knowledge does not change under the influence of co-workers they work with (either formal communication prevents sharing knowledge or there is a strong competition between employees), it is preferable to focus on those with smallest knowledge in this particular area.

In the future authors would like to further analyse external factors that could influence co-workers knowledge. They would like to test mixed ways of teaching and more sophisticated module for counting costs.

References

1. Arachchilage, N.A.G., Love, S.: Security awareness of computer users: a phishing threat avoidance perspective. Comput. Hum. Behav. **38**, 304–312 (2014)
2. Bylica, W., Ksiezopolski, B.: On scalable security audit for web application according to ISO 27002. In: Computer Networks. CCIS, vol. 160, pp. 386–397. Springer (2011)
3. Mazur, K., Ksiezopolski, B., Kotulski, Z.: The robust measurement method for security metrics generation. Comput. J. **58**(10), 2280–2296 (2014). Oxford Press
4. Niescieruk, A., Ksiezopolski, B.: Motivation-based risk analysis process for IT Systems. In: AsiaARES, vol. 8407, pp. 446–455 (2014)
5. de Nooy, W., Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek (2004). http://vlado.fmf.uni-lj.si/pub/networks/data/esna/hiTech.htm
6. Rogers, M.K., Seigfried, K., Tidke, K.: Self-reported computer criminal behavior: a psychological analysis. Dig. Invest. **3**, 116–120 (2006)
7. von Solms, B.: Information security - a multidimensional discipline. Comput. Secur. **20**, 504–508 (2001). Elsevier Science Ltd.
8. The web page of the QoP-ML project (2015). http://qopml.org/
9. Valentine, A.J.: Enhancing the employee security awareness model. Comput. Fraud Secur. **2006**(6), 17–19 (2006). Cybertrust's ICSA Labs
10. Westin, A.F.: Social and political dimensions of privacy. J. Soc. Issues **59**(2), 431–453 (2003). Columbia University

Advanced Data Communication Framework for Cloud Computing from CDMI

Jae-Yun Jeong and Jae-Woong Jeong(✉)

Kyung Hee High School, 26, Kyungheedae-ro, Dongdaemun-gu,
Seoul 02447, Republic of Korea
3500jjy@gmail.com, bestamicus@naver.com

Abstract. The expansion of cloud service has brought about the change of communication method between computer and user. Therefore, the significance of cloud data management system has increased. However, the current cloud data exchange method lacks in consideration regarding the utilization of big data. This thesis proposes a framework for cloud data management system considering the big data processing. For this, the thesis used CDMI standard structure to conduct separation and compression of cloud data to propose an efficient data management system.

Keywords: Cloud computing · Data communication · CDMI · Big data

1 Introduction

The introduction of cloud service has brought about the computer communication to change from using S/W to services that the data center provide in cell phone or other mobile devices. Cloud computing is rapidly changing into high quality computing technology [1]. which is a computing method which provides relevant functions in service type. This had led users who do not know have professional knowledge regarding the technology infra structure that supports to be able to use the service from the internet. In cloud computing environment, users can share information, and can be provided with application or program type of service anywhere anytime. In IEEE, it is defined as "a paradigm which permanently save information on internet server, and temporarily save information on clients such as desk top, table computer, lap top, wall-held computer, or mobile devices. Before long after the cloud service was introduced, the standard of cloud data server was not set. However, it is average to separate each storage in PC, and use LAN line to move data server and store to data center.

Cloud computing service is classified in to 3 by its module. It is classified by how to SW or HW is used which is Infra as service (IaaS), platform as service (PaaS), and software as service(SaaS). IaaS makes access to basic computer resource possible, which means the user has no need to build personal cluster or data center to use cloud service. It is flexible in deleting additional resource. PaaS is providing platform related to software distribution environment, programming tool, and OS used for cloud infra. SaaS enables to use applications of service providers operating in cloud service. Following the development of cloud computing, the use of numerous computing resources,

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_4

data, information, and application increased and the amount of data has also increased. Due to the features of cloud computing, data is in infinite produce, and the form is diversified, which keep on producing data, and the data server to store data was in need. The big data was proposed to efficiently classify and analyze the massive data and information by its feature and category. This thesis proposes a framework to efficient cloud data management system. Also, this thesis proposes efficient data management system based on separation and compression of cloud data and will conduct comparison with other existing systems to verify the efficiency of proposed system.

2 Cloud Storage and Data

Various cloud service suppliers such as Amazon, EMC, IBM, HP and NetApp provide cloud storage service. Also, cloud storage platforms such as HDFS, GFS, EMC Atoms, Data ONTAP, HP Upline, Cloud NAS, Amazon S3 are increasing. In order to standardize such cloud storage platform, SNIA has adopted a proposal to set CDMI international standard in 2009. Structure of CDMI international standard system adopted by SNIA is as described in Fig. 1 [2].

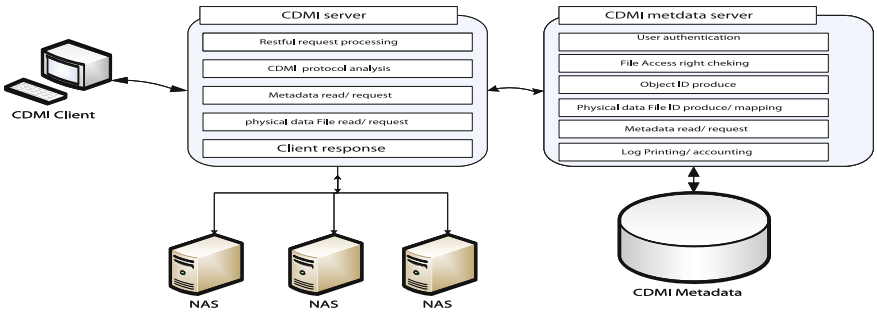


Fig. 1. A framework for SNAI CDMI international standard system

This system is composed of CDMI client, server and metadata. CDMI client is a user program which request actual data fit to CDMI standard to transmit request to CDMI server. The CDMI server responds to request received from CDMI client and transmit. CDMI metadata server processes and manages information used to provide CDMI from inner file system. When file creation, revision and delete order is requested from client, the server checks metadata and review if the order process requested from client is possible. If it is possible, it renews the information about metadata and process with the actual contents. To process contents, it revises the file and searching function to store the contents. CDMI server delegates CDMI metadata server for metadata processing, and directly mount NAS on CDMI server to process contents. CDMI metadata server manages information used to provide CDMI function in inner file system. With the development of cloud, immensive amount of data has been created. Such data is called big data. The characteristics of big data can be largely classified into

3 types which are also known as 3V. 3Vs are volume, velocity, variety which refers to the amount of data, input and output sped of data, variety of data types. IoT (internet of the things) which analyzes and utilizes data from sensor is becoming popular. For example, in America, sensors are attached to plane engines to receive and analyze information for error prediction, which the data produced from engine is about 2.5 billion terabyte per year. Due to the increase in big data, cloud data management system in computing became more significant. Google Map Reduce is the representative form of cloud data management. It is essential for big data to separate process data through multiple servers. However separate processing is the major technology of cloud, which makes big data and cloud technology complementary.

3 Cloud Computing Based Data Management Structure

3.1 Data Processing in the Existing Cloud Computing

The past cloud management system mostly use Map-Reduce. MapReduce is composed of multiple workers under one master. The master sends and receives ping messages regularly and classify on and off workers to divide work.

In Map Reduce, the master takes care of work scheduling, task distribution and procedure monitoring. Workers simultaneously conduct the task set at maximum. When users hand in operation, the master is it into divides Map task. And the divided tasks are distributed to workers. To equally divide the works to workers, the master sends ping messages to understand non-operating task number of workers and receives responses. Master divides the operation into Map task and Reduce task to distribute to on/off workers. Then the Map task is distributed to workers non-operating workers in priority. In other words, the tasks are distributed to conduct possible local disk data as Map task in input data pieces. This is to reduce the bandwidth of network. After distributing Map tasks, the master sends ping message to non-operating workers and distributes Reduce tasks. The Map workers after finishing the work stores result to local disk, and transmit the Map operation result information to master. The master stores results received from map workers to reduce task and sends it to Reduce workers (see Fig. 2) [3].

There is also CDMI international standard of SNIA, the structure managing cloud data. This system increased efficiency by separating the metadata server and data server. Also, because most of the cloud data is atypical, the data server and metadata server interacts in the form of RESTFUL which follow REST method. This thesis examines the performance of proposed Cloud Data Management Framework by comparing with CDMI international standard of SNIA. CDMI Data server receives requests from CDMI clients. CDMI server sets manager for each object to conduct requests. When the contents regarding metadata are sensed, the CDMI data server delegates operation to CDMI metadata as RESTFUL. CDMI metadata server also has manager for each object to enable efficient operation. After completing operation, CDMI metadata transmits result to CDMI data server. CDMI data server completes request from CDMI client and transmits result [4].

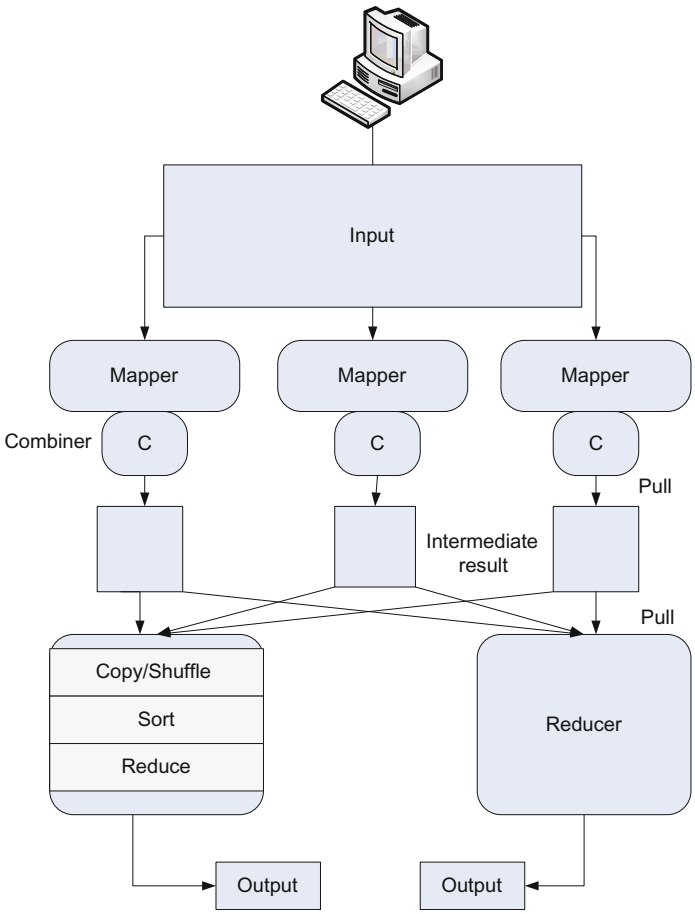


Fig. 2. Process structure of Map-Reduce

3.2 CDMI Framework Proposal

This thesis mainly focuses on connection to big data in cloud data management to design framework. Cloud data and big data technology has direct connection. Processing speed is very important in big data technology. If the result is already happened in the situation, the data becomes useless. The framework proposed by this thesis is as Fig. 2.

The framework directly connects CDMI client and metadata server to simultaneously operate on CDMI data server and metadata server. Also, by keeping separate NAS, cloud data server and cloud metadata server mounts simultaneously and store information, to simultaneously use information from data server and metadata server with the information stored in main NAS. This is to enable easier processing of request when clients make requests. Also, there is no need to save CDMI metadata client to cloud data server. Data server has no need to delegate works to metadata server, and the metadata examination separately sets server receiving requests from client. This server

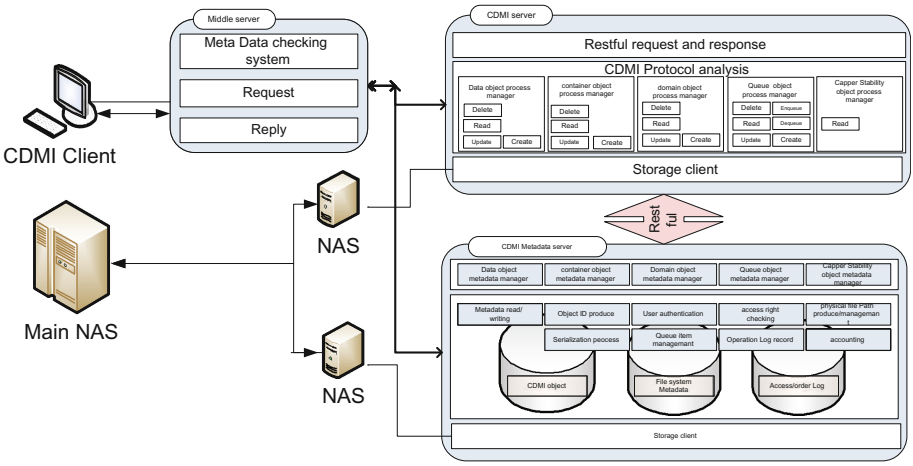


Fig. 3. Proposed framework using CDMI

receives direct requests from client, and only conducts metadata examination of separate request. Data server and metadata server conducts each request and transmit to middle server, and the middle server collects the result and transmit the result to client (Fig. 3).

4 Evaluation

When comparing the framework proposed by this thesis, such results followed. PCS proposed by Remi Cura, Julien Perret, Nicolas Paparoditis [4] received request from metadata server initially, and there was no general data server. With only metadata might lack in accuracy. Also, the system circulates within RDBMS. In this system, if an error occurs in one part, it influences the whole system which makes it unstable. However, it is able to conduct requests simultaneously as it conducts in circulatory moves. However, it has less connection of the utility of big data. The CDMI proposed by SNIA has separated metadata and general data server, which secured its accuracy and system stability by conducting the request twice. However, the client only interacts with data server which makes metadata non-operating during the time. Therefore, it lacks simultaneity and had no efficiency. On the same basis, it deducts result late, so the connection to big data also decreases. The proposed framework has separated data and metadata server to process the request twice to increase system accuracy, has separated main NAS and assisting NAS to increase system stability. Also, the client simultaneously transmits the request to both data server and metadata server for processing, which secured the simultaneity of system and improved efficiency. Also, by simultaneous processing, the velocity of big data is secured and the main and assisting NAS enables efficient use of storage to secure the volume of big data. As for the request transmission method, it uses JAVA's RESTFUL method supporting REST to secure variety (Table 1).

Table 1. Comparison between existences and proposed framework

Item	PCS [4]	CDMI [5]	Proposed framework
Accuracy	X	O	O
Stability	X	O	O
Simultaneity	O	X	O
Big data correlation	X	X	O

5 Conclusion

Recently, the increased use of SNS and cloud led to the increased attention to big data technology. The proposed framework is designed focusing on big data. This thesis compared the proposed framework with other frameworks and learned about the system. As the result, the proposed framework was verified with its accuracy regarding the request, system stability. Also, it was verified that it directly and simultaneously receive requests from clients through data server and metadata server, designs request processing method to secure the simultaneity and efficiency of the system. Also, it secured the 3V, features of big data to improve the connection.

References

1. Rehman, Z., Hussain, O.K., Hussain, F.K.: User-side cloud service management: state-of-the-art and future directions. J. Netw. Comput. Appl. **55**, 108–122 (2015)
2. Ahn, M., Jeon, I., Son, I., Li, H., Park, Y., Lim, J., Bok, K., Yoo, J.: Design and implementation of a cloud data management interface (CDMI) system. J. Korea Contents Soc. **13**(8), 28–35 (2013)
3. Lee, K.-H., Choi, H., Chung, Y.D.: Massive data processing and management in cloud computing: a survey. Korean Inst. Inf. Sci. Eng. Database **38**(2), 104–125 (2011)
4. Cloud Data Management Interface (CDMI). SNIA (2016). http://www.snia.org/

A Study of AI Based E-learning System and Application

Min-Gyu Shim(✉) and Jae-Woong Jeong

Kyung Hee High School, 26, Kyungheedae-ro,
Dongdaemun-gu, Seoul 02447, Republic of Korea
{mk071041,bestamicus}@naver.com

Abstract. It could be said that AI (Artificial Intelligence) is the most prominent field within the short history of computers and it is recognized as an important element deciding upon future human life. The study aims to explore the exact concept of AI and environments used in e-learning. Through this, the study aims to review learning combined with Artificial Intelligence (AI), thus, Internet self-directed learning which is recently receiving much interest and explore actual fields where AI is currently applied. As a result, the study aims to suggest the base data for providing students a more efficient and abundant education environment using artificial intelligence.

Keywords: Cloud computing · Data communication · CDMI · Big data

1 Introduction

Many think that AI (Artificial Intelligence) is like a supernova with a history of less than 10 years but it is not entirely so. in fact, the meaning of artificial intelligence was born in the 1950s by Alan Turing and simultaneously, research was conducted [1]. Artificial intelligence is a large topic that is gradually receiving interest from the masses in modern society. Even mobile phones utilize artificial intelligence, and recently the range of availability has largely broadened, being suggested as a method to bring a better learning effect in the education field, utilizing e-learning system where students learn through computers [2]. Therefore, development, application scope, and availability is rapidly increasing [3]. Artificial intelligence had already been studied and introduced long ago. However, artificial intelligence first introduced could not be actually applied, there was process where the concept was formulated, and it could be said that the period where it started to be applied to reality such as actual robot artificial intelligence and education-based artificial intelligence is recent [4].

Application of AI is becoming more active in fields such as the financial market and science. Like this, the fact that AI is closely applied to human life even though it has not been long since it has been introduced into our lives can be attributed to the efficiency and convenience of artificial intelligence. Also, in the future, artificial intelligence will be used as an important interface that has direct interactive relation with human life. However, among AI that is utilized in various fields, the field that has especially wider usage scope which has human and machine interaction is artificial intelligence based education system [5]. Especially in students of the East Asian region

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_5

with high scholastic ambitions, it is easy to find examples of e-learning, or Internet-based self-directed learning system [5]. However, there are suggestions of shortcomings in learning systems that apply e-learning. Especially, students spend more time in learning through e-learning done at home rather than concentrating on the time with their teachers at school, and there are suggestions of worries that the process of e-learning through artificial intelligence could be less effective. Also, there are doubts about if e-learning service conducted online rather than as offline classes are more effective, and it is not clear if all learners are satisfied by the learning process through the e-learning system. In our modern society, rather than considering only the development of artificial intelligence or market economy size, we need to come up with clear answers to questions such as, 'Is the development of artificial intelligence being done properly?' and 'Is it justifiable to introduce e-learning systems while toppling over the balance of society and economy?' [6]. Therefore the study analyzed the awareness about the application of artificial intelligence utilized in the education field through surveys on teachers and students. Especially in the suggested research model, surveys were conducted dividing into 2 groups; one with students and teachers accustomed to e-learning, and one with students and teachers unfamiliar with e-learning. The composition of the study is as follows. Section 2 investigates relevant previous studies, Sect. 3 explains research method and procedures, Sect. 4 suggests results about the survey analysis to prove the effectiveness of the study, and Sect. 5 concludes the study.

2 Relevant Previous Studies

Following the development of information and communication technology, relevant social field is also impressively growing. It is not an overstatement to say that the area with the largest change among them is the education field. Especially the new education system called e-learning, created through combination of the education field and characteristics of artificial intelligence, among information communication technology, is sitting place as a new education method to supplement or replace existing face-to-face off-line education in schools, education institutions, and even corporations, and it has the characteristic that from anywhere and anytime, anyone can receive customized learning according to level [7]. E-learning system which is a technology combining these great characteristics of artificial intelligence into the education field has a very wide range of research including development direction, prediction of usage scope, and frequency of actual usage of e-learning systems. However, rather than further asking development potential or direction of the system, there is need to thoroughly inspect information and communication technologies used in the education field, considering if the system could be effective for everyone, if there is enough value to proceed while toppling the balance with the previous face-to-face method off-line lessons, and if e-learning is actually giving users satisfaction and if it looks proper in the ice of people who are not users of the e-learning system. On the other hand, there is research underway to figure out the effective side of e-learning, exploring the shortcomings of off-line lessons.

Also, there is active research about specific exploration about e-learning and especially Lee Yongtak [7] showed the necessity of all artificial intelligence including education to go through systemized evaluation and certain procedures, studying methods to evaluate the validity of e-learning, outside of the abstract concept. Also, looking at the table above, studies mainly focus on interaction between e-learning and humans receiving education. For e-learning to give enough effect, trust, and satisfaction without influencing social and economic balance, it is determined that all fields about artificial intelligence rather than just the education field needs to be studied.

3 Research Methods and Procedures

The purpose of the study is based on a number of questions. They include, 'are teachers and students satisfied with e-learning?', 'do they think e-learning is related to actual studies?', 'what does the current education look like to nonusers of e-learning?' The comprehensive purpose is to inspire a sense of purpose through inspecting if education techniques combining artificial intelligence is being reasonably conducted, if there is value in further development, and most of all, if it does not affect society, as people respond to the survey. First, the method of this study is a simple survey. Like how there is a division between male and female on OMR cards, respondents are first divided into users or nonusers of e-learning, and it is further divided into detailed survey questions according to characteristics. Table 1 represents examples of these survey questions.

Table 1. The contents for questionnaire

Number	e-learning user question	e-learning nonuser question
No. 1	Are you satisfied with the e-learning you are currently using?	While it is shown statistically that students invest more time on the current online learning compared to off-line learning, do you think that this is appropriate?
No. 2	Was the e-learning you currently use actually helpful to grade performance?	If you start online learning, between online learning and off-line learning, will you invest more time in off-line learning?
No. 3	Do you think e-learning is better than off-line lessons?	Do you feel satisfied with the off-line lessons you take currently?
No. 4	Are you aware not roughly but specifically that artificial intelligence?	Do you think there is need for improvements on current smart phone or official intelligence?
No. 5	Developed information communication technology can replace the work of humans?	Do you think there is need for future new technologies?

Also, in common questions between the two groups, there are questions such as, "Do you know about modern computer information communication technology?" And

it was presented to investigate the scope of awareness of current teachers and students and then to grant more specific validity compared to the above three questions.

3.1 Research Panel Configuration

The population for the survey were teachers and students currently working or enrolled in high schools in Korea. Especially among teachers, it was divided between people who had experience of using e-learning, and those who had not.

The reason why teachers were included in the subject of the survey was because they are in the position of teaching the students who are trying to provide a quality learning environment as information and communication technology develops, and because they are people who are most interested in the utilization and effectiveness of e-learning. Another reason is that teachers, as examiners, can more effectively show the effect of e-learning. The reason why students are subjects of the survey is because they are the principal users of e-learning and subjects who actually participate in the learning. The population consisted 100 high school students in 3 classes, and 20 teachers (see Table 2).

Table 2. Comparison between existences and proposed framework

Item	e-learning user	e-learning nonuser
Teacher	6	14
Student	73	21

3.2 Data Processing

After collecting adequate data from a total of 120 subjects on blueprints (OMR cards), data was analyzed using Microsoft Excel. The procedure for research model is shown in Fig. 1.

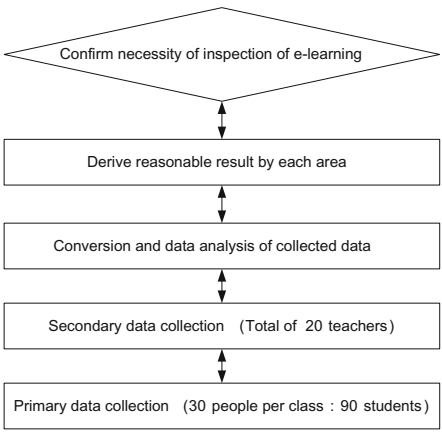


Fig. 1. Research model process

4 Results and Analysis

Largely, the subjects were classified into 2 groups; users and nonusers of e-learning, and the user group was divided into teachers and students.

As shown from Figs. 2, 3, 4 and 5, in the two groups, e-learning users and e-learning nonusers, three questions out of the total of five questions were differentiated. Among subjects that were nonusers of e-learning, the most frequent response to the question "While it is shown statistically that students invest more time on the current online learning compared to off-line learning, do you think that this is appropriate?" was that they disagreed. This shows that they are aware of the slight seriousness of artificial intelligence learning emphasis phenomenon. In question 2 (If you start online learning, between online learning and off-line learning, will you invest more time in off-line learning?), there were almost no differences between those who agreed and disagreed, but overall, unlike the characteristics of question 1, there were many that responded that they agreed. Next, in question 3 (Do you feel satisfied with the off-line lessons you take currently?), there were clearly more people who responded that they agreed. This is a complete opposite result from question 1. However, because question 4 and 5 were common questions, they were not classified separately.

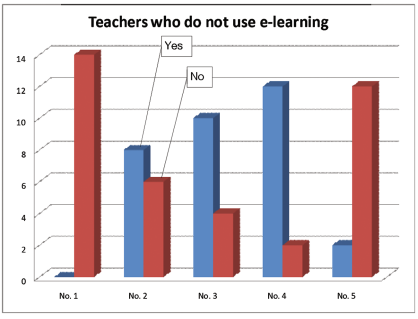


Fig. 2. Teachers who do not use e-learning

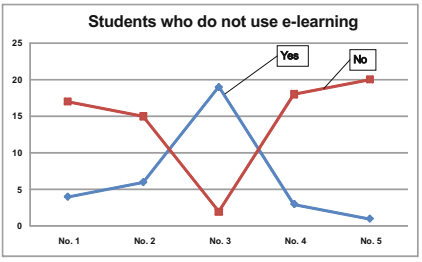


Fig. 3. Students who do not use e-learning

As shown in Figs. 3 and 4, the analysis results of e-learning users is as follows. The first questions for users was "Are you satisfied with the e-learning you are currently using?" and while they are people in the same classification, there was large deviation between teachers and students. It was found that teachers were not satisfied but in the group of students, an opposite result was shown. This result shows that "Artificial intelligence that has the same effect give different sense of achievement to people according to situation." A factor that can be considered here is the correlation according to occupation. Thus, because teachers are accustomed to off-line (face-to-face) lessons rather than utilization of online Smart learning, a result showing a relatively high ratio of disagreement was shown. On the other hand, because their position can be that what they can see itself is important, the result was that they were relatively accustomed. Question 2 was "Was the e-learning you currently use actually helpful to grade

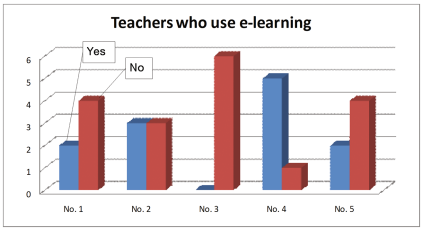


Fig. 4. Teachers who use e-learning

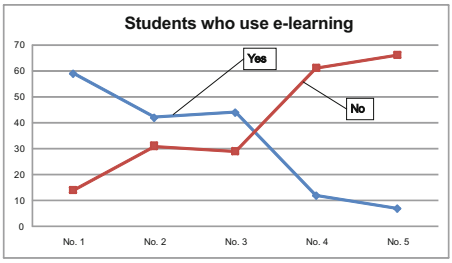


Fig. 5. Students who use e-learning

performance?" and the subjects were teachers and students. The study results showed that there were many people who had positive stances. Question 3 was "Do you think e-learning is better than off-line lessons?" and the analysis results were quite shocking. This is because while they were two groups of the same classification, the differences were extreme. While there was not a single teacher who had a positive response, there were more students that had positive responses then negative responses. This shows that work correlation, thus information communication science technology or artificial intelligence is already influencing our society.

If the research subjects were only students rather than teachers, the satisfaction, perception of smart learning, and its effects would have been positive. This is because smart learning has fewer limitations in space and time compared to face-to-face method lessons and large advantages such as formidable curriculum are directly reflected on the results. however, this was shown differently in the teacher group. This is because it shows confrontational aspects between the positions of students who are receiving education and the teachers who are directly communicating with society. This is also a clue that scientific technology is already showing profound impact on our lives. Students and teachers are in an interactive relation. Thus, if one of the two groups state that it is effective and if the other does not, this goes against the concept of interaction itself and it cannot be good education. It was said that an educator without the formula of purpose is unhappy. Therefore, good education can be done if interaction is kept and if the foundational source of the destruction of this linkage, artificial intelligence, could be more deeply understood and applied, the effect of smart learning utilizing artificial intelligence can suggest a very effective education method for both students and teachers.

5 Conclusion

According to analysis results, because there were relatively more people who thought that self-directed learning utilizing e-learning had relation with their grade performance, it could be said that there is significant result in learning effect through e-learning. Also in satisfaction, 6:4 or 7:3 ratio of high satisfaction result was derived. The reason is that, relatively, compared to off-line face-to-face lesson methods, the

content was vast and diverse which enabled more fun progression of learning, and that there was fewer restrictions in place and time compared to off-line. Therefore, effect or satisfaction was higher in e-learning than in off-line.

In an era of technological development, one must not focus only on taking advantage and must have knowledge in the position of using it to be worthy of a user. Also as it can be seen in the above table, the current education status of smart learning has a very systematic structure. however for smart education to continue development to give satisfaction to people in its the development, it is absolutely necessary to inspect the fundamental problems rather than seeing the institutional side. This is the criticism about information and communication science technology. Because technology is made by humans, there can be mistakes and there can be side effects. Therefore researchers and users must have a sense of purpose in using artificial intelligence and rather than engaging in the advantages of education or test scores in schools and institutions, there needs to be mutual synchronization so that equal educational opportunities can be given whether they are users or nonusers of smart learning.

References

1. Monthly App. 21 February 2016. http://withpress.co.kr/archives/1952
2. Monthly App, 16 February 2016. http://withpress.co.kr/archives/1865
3. Newsis, 18 February 2016. http://www.newsis.com/ar_detail/view.html?ar_id=NISX20160316_00-13961273&cID=10201&pID=10200
4. http://image.chosun.com/sitedata/image/201603/08/2016030801051_2.jpg
5. Daily UNN (Korea University News), 9 December 2015. https://news.unn.net/news/articleView.html?idxno=154140
6. Duration, Domestic issue, 4 October 2016. http://justcuration.tistory.com/174
7. Lee, Y.: Korea Institute of Professional Competence, 2005.7, Research about effectiveness evaluation method of e-Learning, pp. 65–88 (2015)

A Study on the Serious Issues in the Practice of Information Security in IT: With a Focus on Ransomware

Junhak Lee(✉) and Jae-Woong Jeong

Kyung Hee High School, 26, Kyungheedae-ro, Dongdaemun-gu,
Seoul 02447, Republic of Korea
{gygh75, bestamicus}@naver.com

Abstract. Contemporary society has been seeing rapid progression IT since the information revolution that was launched with the invention of the computer in the 20th century. Everyone living in contemporary society is affected directly or indirectly by IT in a wide variety of areas including medical services, transportation, fashion, business management, administrative management and art. With the rapid penetration of the internet, information in the cyber world is also on a dramatic increase [1]. Since the economic value and importance of such information are also on the increase, there have also been much increase in the risk factors that target highly valued information. Even though the risk factors are on the rise in tandem with the progress made in IT that make our life more convenient, there is a lack of awareness of information security among the public. Therefore, this study seeks to explain the concept of computer viruses and how they developed, the risk posed by ransomware that have newly emerged as a threat and citizens' ransomware that is in contrast to that. It also seeks to analyze the results of a survey on the perception and understanding of the importance of cyber information security, thereby promoting the risks of ransomware and raising awareness. In doing so, the study seeks to present a basic set of data that can serve as reference for future measures to be taken.

Keywords: Ransomware · Information security · IT · Computer viruses · Understanding of the importance of cyber information security

1 Introduction

Starting in the 20th century the information revolution that was caused by the invention of the computer directly and indirectly affects everyone's lives in various areas including medical services, transportation, fashion, business management, administrative management and art. Due to the rapid penetration of the internet, we are handling rapidly increasing amounts of data that we are using to improve our comfort. Such progress in IT is fundamentally changing human society including our lifestyles [1].

With the progress in IT, the importance and value of information, documents, files and programs in the cyber world are also increasing. For example, programs that are developed by corporations can carry enormous economic value as a product of that corporation. As such, while those that threaten the security of precious information

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_6

with great value are also seeing progress by hackers, the awareness of such threats and measures to address the issues are still lacking.

This paper discusses one of the leading computer viruses among malicious codes with a criminal intent that threaten precious information of contemporary people. It also discusses the concept of worms, its pattern of progress and 'ransomware' [2] which is a hacking method that forces a user's document, spreadsheet, file or program by invading into the user's computer, forcefully encrypting them and asks for money in return for decrypting it. Some cases of victims are presented to emphasize the risks of ransomware and to raise awareness of the risks. A survey was also conducted to analyze the perception and understanding of the importance of information security and ransomware to raise awareness among the laymen of the importance of information security and to present a basic set of data that can be a reference point for measures to address ransomware.

2 Computer Virus

A computer virus refers to malicious computer programs where an invasion occurs without the user's knowledge and the requests programmed by the hacker are carried out. A computer virus is also able to self-replicate and reproduce itself. It is similar to biological viruses in this aspect and was thus named after them. Computer viruses were first reported in 1988 [3] and since then there have been many mutants. New types of computer viruses that had not existed before are consistently emerging in a way to help hackers conduct more efficient attacks on more targets.

A worm refers to a malicious program that invades into the PC through the internet, network or various types of communication networks without authorization to deplete system resources or slow down the processing speed of the computer to conduct the requests by the hacker. However, it goes beyond this to self-replicate in the infected computer to rapidly spread itself to other systems through various communication networks such as the internet and other networks [5].

As you can see in Fig. 1, when a hacker first spreads a worm, he attaches it to a board on a certain website or to a blog or email. This allows for the virus to spread through the internet and infect PCs. It self-replicates within the PC to infect other PCs. Because of this the number of infected PCs increases dramatically and the virus spreads extremely quickly. When worms first appeared, they were distinguished from other computer viruses because they did not infect other programs within the infected PC. But recently, the distinction between worms and viruses is being blurred. The reason is because when malicious codes divides into computer viruses, worms or Trojan horses once, they do not have a uniqueness but they continue to evolve infinitely as they combine or converge themselves with other existing malicious programs. If malicious codes of the past had the purpose of boasting the technical capabilities of hackers, recently with the dramatic progress of IT, many cases involve financial value attached to information, documents and files in the cyber world. Because of this, malicious codes have come to be used for a malicious purpose and as a tool of hackers for financial gain.

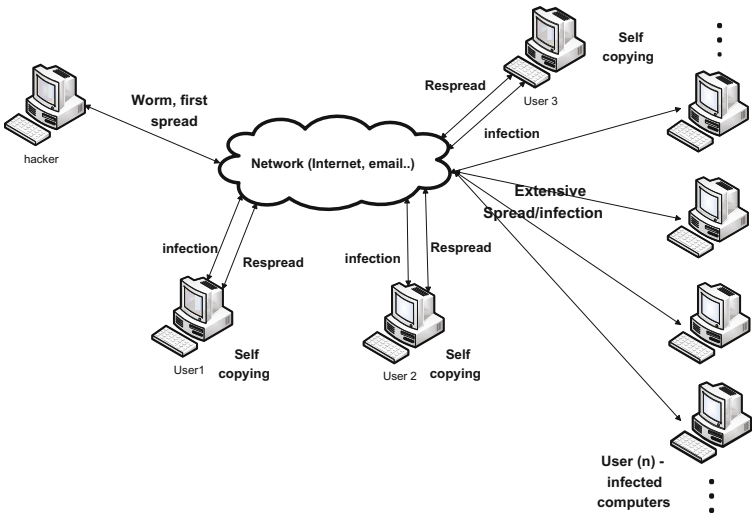


Fig. 1. The process and environment for worm virus

3 Ransomware and Its Risk

3.1 Ransomware

Ransomware first appeared in 1998. It invades the user's computer without authorization to forcefully encrypt the user's documents, photos, spreadsheets, files or programs so that they cannot be used. In exchange for decrypting them, it asks for a financial return [2]. The term ransomware is a combination of two words, ransom and ware. It infects the computer through various channels including emails and boards of online sites. When a computer is infected by ransomware, the ransomware starts going through the computer's files. At this point, there is no noticeable symptom of infection other than that the computer slows down, so users do not realize that the computer is infected by ransomware. Then the files start to be encrypted. For hackers to efficiently encrypt as many files within a short period of time, it brings in as much system resources as possible. This makes the speed of the computer to slow down drastically. Once the encryption of files and documents is complete and the computer is rebooted, a notification window pops up asking the user to wire money if she wishes to have the encrypted files restored. The documents, photos, files and programs on the infected computer cannot be used.

'The total number of worldwide users that have come into contact with various forms of ransomware including encryption attack methods increased by 17.7% from 1.967784 million people between April, 2014 to March, 2015 to 2.315931 people for the period of April, 2015 to March, 2016' [4]. According to the above data, over the course of a mere one year from April, 2014 to March, 2015 and from April, 2015 to March, 2016, the number of people who had come into contact with ransomware increased by 340,000 people from 1.97 million to 2.31 million. Given that the first ransomware had emerged less than 20 years ago, the speed at which it is spreading is

all the more astonishing [4]. Ransomware is widely affecting the general population, too, and is expected to cause significant damage going forward.

3.2 The Risk of Ransomware

Once infected by ransomware, the only choices available is to either make a payment as requested by the hacker or re-format the computer. But even if you wire money as per the hacker's request, there is no guarantee that the hacker would restore the encrypted documents and files to their original state. Director Changhun Lee of Casperski Lab Korea notes that handing over a key for decryption opens the possibility for the hacker to be traced so even when the victim wires money it is questionable whether the files would be encrypted. If the victim follows the hacker's order, it gives the hacker even more incentive to develop new forms of ransomware as he has seen evidence that it is profitable. He adds that for these reasons, we should not pay the hacker despite his requests [4].

Since the very purpose of developing ransomware is to invade unauthorized into a user's computer to forcefully encrypt the user's documents, photos, spreadsheets, files and programs and then to ask for financial pay for decryption, ransomware causes direct financial damage unlike existing malicious programs that had caused indirect financial damage by simply boasting the hacker's technical skills or collecting information illegally. Because of this, the psychological and material damage caused by ransomware is much more significant compared to that by existing malicious programs.

As mentioned earlier, ransomware is a hacking method where a user's computer is invaded without authorization, his documents, photos, spreadsheets, files and programs are forcefully encrypted so that they cannot be used, and where the hacker then asks for money to be wired to him in exchange for decryption. One might question whether there is a way to track the hacker down based on the address given for the wiring. But since hackers usually take money through Bitcoins that guarantee anonymous transactions, it is impossible to track the hacker.

Existing hacking programs were used to gain indirect financial gains by stealing high value governmental information or corporate information, but ransomware seeks direct financial gains and therefore the target is not limited to specific corporations or countries but the general public.

As mentioned earlier, since ransomware was first discovered, to date (as of March, 2016), the number of people who became victims of ransomware reached a whopping 2.31 million [4]. Given these numbers, it is clear that ransomware is rapidly spreading. If it spreads widely through networks in the way that worms spread, then the damage could be unthinkable.

Malicious codes that had threatened healthy IT life of users have constantly undergone evolution to better serve the malicious goals of hackers. Therefore, the possibility of ransomware to develop into a worm must be considered.

4 Questionnaire and Evaluation

In July of 2016, 33 students of Kyounghee High School in Seoul, Korea aged from their teens to their 20s were surveyed offline on questions such as “Do you know what ransomware is?”, or “Have you experienced ransomware?”.

Figure 2 shows the results of responses to the question, “Do you know ransomware?” Those who answered yes accounted for 30.3% or 10 people out of a total of 33 people. There were four people who answered they “don’t knot it very well”, which accounted for approximately 12.1%. A total of 19 people answered they “don’t know” which accounted for approximately 57.6% of total respondents. This indicates that those who do not know about ransomware count twice as much than those who do know about it. This indicates that laymen’s understanding of ransomware is very lacking. As we have seen through the risks posed by ransomware, since laymen can also become the victim of ransomware, it is important to understand what ransomware is and what is required to prevent their invasion. Also given that the survey results came from those in their teens and 20s who will become more influential to the IT industry in the future and the reality of a future where individuals’ efforts to prevent such infection will become more important, this figures are very worrisome.

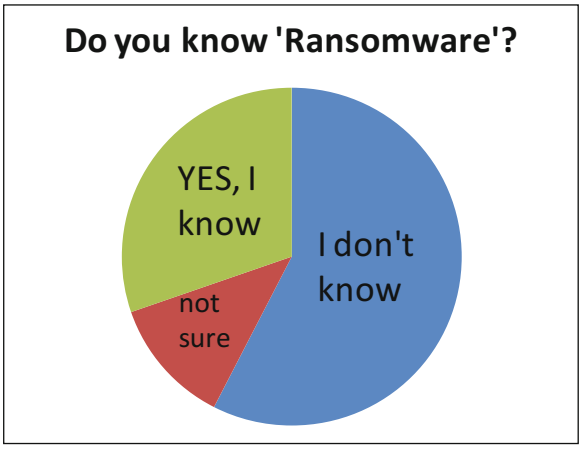


Fig. 2. Survey on the understanding of ransomware

Figure 3 shows the results of a survey asking “Are you aware of the importance of information security?” Those who are aware of information security and who also practice it were six people out of a total of 33, or 18.1%. There were 18 people or 54.5% who responded that they ‘are aware but are not practicing’. This was the highest percentage of response. Those who responded that they ‘are not aware’ were five people or 15.2% of total respondents, while those who answered that they ‘are not interested’ accounted for 12.2%. The survey results show that less than 20% of total respondents proactively take measures to safeguard against his own information. This also means that the general public is critically exposed to the threats of various malicious codes such as ransomware.

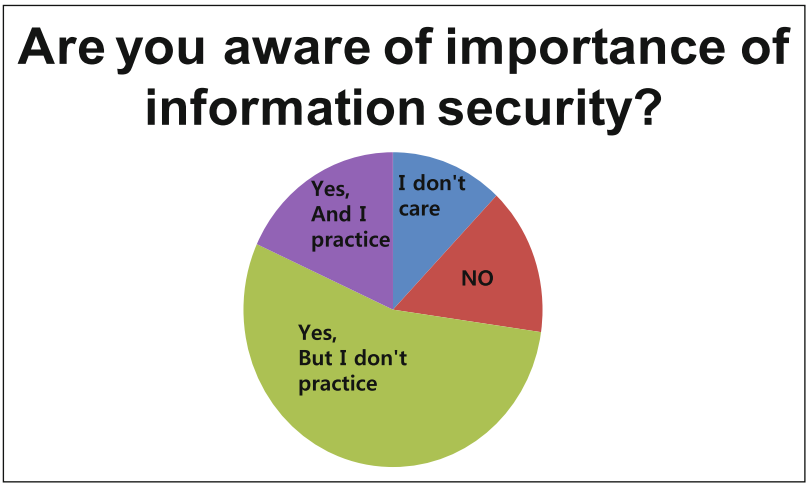


Fig. 3. Survey on the awareness of information security

5 Conclusion

Jihun Park, a Korean, mentioned that "information security is not about technology but about culture" [6]. Information security cannot be achieved simply through the efforts of minorities or a handful of security programs developed by some experts. It is something that must be paid attention to by all members of society on an ongoing basis where they recognize the importance of such security. This paper raised the awareness of the threats posed by ransomware and analyzed how high the level of understanding is by the general public on ransomware and the importance of information security. The analysis shows that the lacking of ransomware and measures that are actually practiced for information security are lacking. Therefore, mass media and preventive measures that can improve awareness of people who can become victims of malicious programs such as ransomware, as well as protect from cyber threat are much needed.

References

1. Luia, A.K.H., Ngaib, E.W.T., Lo, C.K.Y.: Disruptive information technology innovations and the cost of equity capital: the moderating effect of CEO incentives and institutional pressures. Inf. Manag. 53(3), 345–354 (2016)
2. Everett, C.: Ransomware: to pay or not to pay? Comput. Fraud Secur. **2016**(4), 8–12 (2016)
3. Khanh, N.H., Huy, N.B.: Stability analysis of a computer virus propagation model with antidote in vulnerable system. Acta Math. Sci. **36**(1), 49–61 (2016)
4. KASPERSKY Lab, 28 June 2016. http://news.kaspersky.co.kr/news2016/06n/160628.htm
5. https://en.wikipedia.org/wiki/Computer_worm
6. Park, J.H.: Penta Security system, standard procedure of IT security, vol. 7 (2015). https://www.pentasecurity.com

Mobile App for Analyzing Environmental Visual Parameters with Life Logging Camera

Hyeonsang Hwang[1], Daejune Ko[1], Mincheol Whang[2], and Eui Chul Lee[1(✉)]

[1] Department of Computer Science, Sangmyung University, Hongjimun 2-Gil 20, Jongno-Gu, Seoul 03016, Republic of Korea
eclee@smu.ac.kr

[2] Department of Media Software, Sangmyung University, Hongjimun 2-Gil 20, Jongno-Gu, Seoul 03016, Republic of Korea

Abstract. In this paper, a method of quantitatively extracting human emotions by analyzing the surrounding environment images obtained through smartphone cameras in real time is proposed. In the area psychology, it has been known that visual elements such as colors and complexity affect human emotions. Based on the foregoing, we developed an application for the extraction of emotions in real time using the colors and spatial complexity of images obtained through android smart phones. Among the color components of images, hue components that indicate colors were extracted as color elements and the spatial complexity was extracted through the quantities of the high frequency and low frequency components out of the frequency components of images, respectively.

Keywords: Mobile App · Life logging · Visual parameter · Color · Spatial complexity

1 Introduction

As developing from information to knowledge and then from knowledge to intellectualization stages, the importance of technologies for efficient management of data has been increasing [1]. In addition, following the popularization of smartphones and wearable devices, the amount and type of user information that can be collected in daily life has been rapidly increasing [2]. Demands for Life Logging based technologies that collect, analyze, and provide the information collected as such have been also increasing [3, 4]. The term Life Logging mentioned here is a compound word of "Life" and "Log" meaning "the log of life" referring to continuously recording those pieces of information that can be obtained from individuals' daily routine and analyzing and utilizing the information [5].

Currently, life logging devices that are widely utilized because they can be easily attached to the user's body and take images automatically without any separate manipulation are being developed in diverse forms [6–8]. However, most life logging devices focus on camera's hardware-wise functions. Furthermore, existing life logging devices have no App for classification of filmed images or extraction of elements that have effects from the viewpoint of emotions and those Apps that are interlocked with

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_7

these devices include only simple functions such as remote capturing. Therefore, to usefully use those pieces of information that are collected, intelligent technologies for analysis of the effects of surrounding environments on emotions are necessary. In general, visual parameters such as colors and spatial complexity are used to analyze the emotions felt by humans about images or environmental elements [9, 10]. In this paper, an Android Application was developed for collection and analysis of information on the effects of visual elements in surrounding environments based on the images of surrounding environments obtained in real time using Smart Phone Cameras.

2 Visual Parameters

2.1 Color Analysis

Many studies that quantitatively schematized diverse human emotions and those that examined the effects of many colors on human emotions from the viewpoint of color psychology have been conducted [11]. Those studies include Russell's emotion model that defines diverse human emotions with a two-dimensional plane of "pleasant-unpleasant"-"arousal-relaxation" and a model called Geneva Emotion Wheel that mapped this model with colors to define which colors affect which emotions [12].

In the present study, the axis that corresponds to "arousal-relaxation" in Geneva Emotion Wheel was mapped to the axis that corresponds to "pleasant-unpleasant" to use a one-dimensional "pleasant-unpleasant" color axis like Fig. 1.

Fig. 1. Mapping from colors to "Unpleasant-Pleasant" axis

To map the hue values of the HSI color model expressed in colors to the "pleasant-unpleasant" axis thereby showing the values quantitatively, values (P) in a range of −1–1 should be obtained using cosine functions. Since 0° corresponds to red color in hue, to apply the hue values to the color model mentioned earlier, θ should be rotated by 80°. Equation 1 as shown below was used to rotate θ in order to apply cosine functions.

$$\mathrm{P} = \cos(\theta + 80) \quad -1 \leq P \leq 1 \tag{1}$$

2.2 Analyzing Spatial Complexity

To calculate spatial complexity, surrounding environment images were transformed by using Fast Fourier Transform (FFT). However, because the result of FFT is depend on original spatial complexity of surrounding environment, quantification and normalization is difficult. To solve an above problem, the radius value(R) of the band in which 99% of power is included in the radial direction from the center of the FFT spectrum image should be obtained. In this point, general environmental image does not include only extremely high or low frequency component. Therefore, the max-min values of R should be obtained based on surrounding environment images with diverse complexity levels collected in advance and the obtained values should be normalized into values in a range of −1–1.

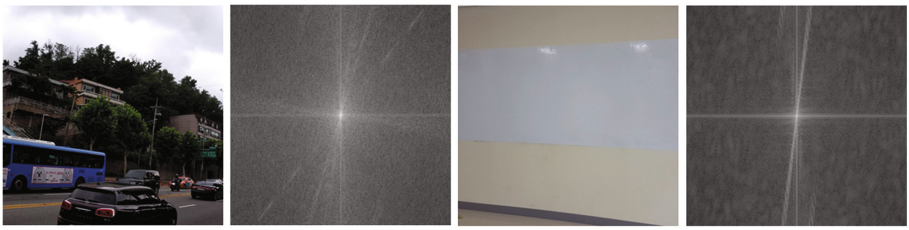


Fig. 2. FFT spectrum images in cases of the lowest and the highest Rs, respectively (a) original image of the highest spatial complexity (b) FFT result of (a) (c) original image of the lowest spatial complexity (d) FFT result of (c)

In Fig. 2(a) and (c) are reference images that were used in our application for calculating max and min values of R, and (b) and (d) are those frequency spectrum images, respectively. Based on the above obtained max and min values of R, normalization is performed in a range of −1–1 by using Eq. 2.

$$A = 2\left(\frac{R - R_{min}}{R_{max} - R_{min}}\right) - 1 \tag{2}$$

3 Image Processing with Android

The android device used in the driving test was Nexus 5X, Android version Marshmallow (6.0.1). In the application, the OpenCV for Android (2.4.11) was used to perform image processing related calculations and for performance optimization, the C/C++ native method was prepared to call image processing. A flowchart of the program is shown in Fig. 3.

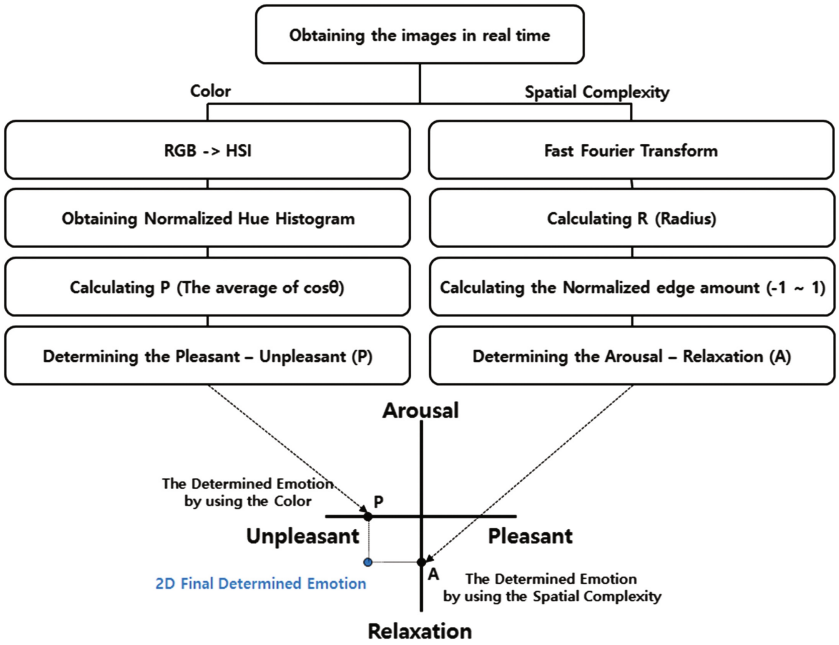


Fig. 3. Flowchart for feature extraction procedure of implemented App

The hue components of images calculated to perform Sect. 2.1 are separated to obtain histograms. For all θs, P can be obtained through Eq. 3 to obtain averages by multiplying histogram components by the cosine values of θs.

$$\mathrm{P} = \frac{\sum_{\theta=0}^{359} \cos(\theta + 80) \times histo[\theta]}{width \times height} \tag{3}$$

In addition, the images obtained to perform Sect. 2.2 are transformed into FFT spectrum images. Using the transformed FFT spectrum images, the radius R of the circle that contains 99% of energy and the radius R is normalized to calculate A.

Finally, using P and A that are the results of calculations through the two characteristics, colors and spatial complexity, the emotions of the current image are mapped on the two-dimensional plane of "pleasant-unpleasant" and "arousal-relaxation" with (P, A) coordinates.

4 Implementation and Feasibility Test

Converting to HSI and extracting hue are implemented by using OpenCV functions such as "cvtColor" and "extract", and P is calculated by Histogram of extracted hue.

Performing FFT and calculation of A are implemented by using OpenCV functions such as "merge", "dft", "split", and "magnitude". Since the FFT of original images at

Smart Phones takes long periods of time and the edge components used in the analysis are strong edge components, images resized into 64 × 64 pixels used in FFT for performance optimization. The actual results of execution of the program implemented through the foregoing are as shown in Fig. 4.

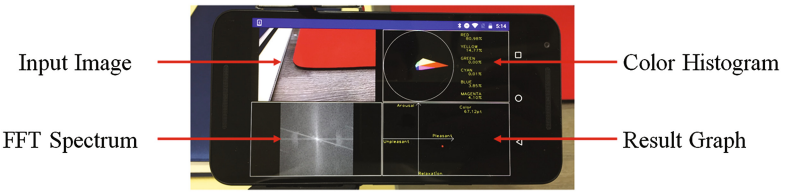


Fig. 4. Results of program execution

The results of mapping of human emotions on the two-dimensional plane of "pleasant-unpleasant"-"arousal-relaxation" using the colors and spatial complexity of images through the foregoing are as shown in Fig. 5.

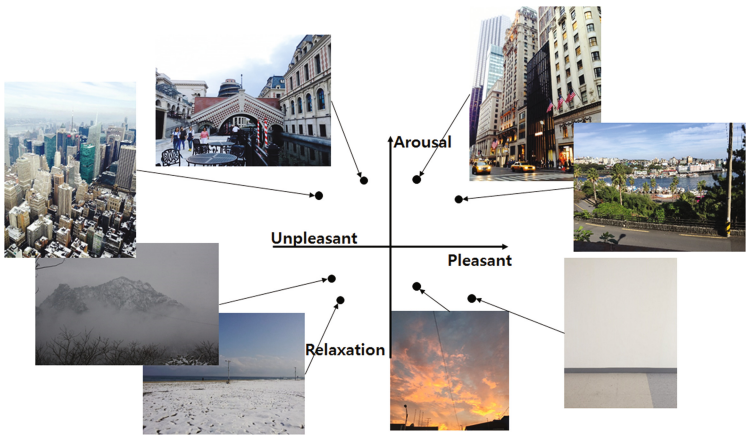


Fig. 5. Results of image classification

5 Conclusion

In the present study, a method was proposed through which surrounding environment images obtained in real time through smart phone cameras are analyzed to quantitatively extract human emotions from the two perspectives; colors and spatial complexity and map the emotions on a two-dimensional plane of "pleasant-unpleasant"-"arousal-relaxation." In addition, the method was implemented on actual Android smart phones to identify that the obtained images were reasonably analyzed. In future studies, whether the proposed method shows emotions completely identical to emotions felt by actual users will be

verified through experiments. In addition, considering the diversity of emotions felt by individuals, models that are applicable to more people will be developed to improve the method so that emotions can be more accurately extracted.

Acknowledgement. This work was supported by the ICT R&D program of MSIP/IITP [R0126-15-1045, the development of technology for social life logging based on analysis of social emotion and intelligence of convergence content].

References

1. Sellen, A.J., Fogg, A., Aitken, M., Hodges, S., Rother, C., Wood, K.: Do life-logging technologies support memory for the past?: An experimental study using sensecam. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 81–90. ACM (2007)
2. Mann, S., Nolan, J., Wellman, B.: Sousveillance: inventing and using wearable computing devices for data collection in surveillance environments. Surveill. Soc. **1**, 331–355 (2002)
3. Hair, J.F.: Multivariate Data Analysis. Pearson College Division, Upper Saddle River (2010)
4. Freedman, D., Thornton, A., Camburn, D., Alwin, D., Young-DeMarco, L.: The life history calendar: a technique for collecting retrospective data. Sociol. Method. **18**, 37–68 (1988)
5. Im, J., Park, M.W., Whang, M., Lee, E.C.: Environmental feature extraction method in terms of emotion. Asia-Pacific J. Multimedia Serv. Convergent Art Human. Sociol. **5**, 107–115 (2015)
6. Caprani, N., Piasek, P., Gurrin, C., O'Connor, N.E., Irving, K., Smeaton, A.F.: Life-long collections: motivations and the implications for lifelogging with mobile devices. Int. J. Mob. Hum. Comput. Interact. (IJMHCI) **6**, 15–36 (2014)
7. Wolf, K., Schmidt, A., Bexheti, A., Langheinrich, M.: Lifelogging: you're wearing a camera? IEEE Pervasive Comput. **13**, 8–12 (2014)
8. Paro, J.A., Nazareli, R., Gurjala, A., Berger, A., Lee, G.K.: Video-based self-review: comparing Google Glass and GoPro technologies. Ann. Plast. Surg. **74**, 71–74 (2015)
9. Kaya, N., Epps, H.H.: Relationship between color and emotion: a study of college students. Coll. Stud. J. **38**, 396 (2004)
10. Bradley, M.M., Hamby, S., Löw, A., Lang, P.J.: Brain potentials in perception: picture complexity and emotional arousal. Psychophysiology **44**, 364–373 (2007)
11. Valdez, P., Mehrabian, A.: Effects of color on emotions. J. Exp. Psychol. Gen. **123**, 394–409 (1994)
12. Scherer, K.R.: What are emotions? And how can they be measured? Soc. Sci. Inf. **44**, 695–729 (2005)

Fake Fingerprint Detection Based on Statistical Moments

Yosep Park[1], Unsoo Jang[1], Jiwon Im[1], Woohyuk Jang[2], Daejune Ko[1], and Eui Chul Lee[1(✉)]

[1] Department of Computer Science, Sangmyung University, Hongjimun 2-Gil 20, Jongno-Gu, Seoul 03016, Republic of Korea
eclee@smu.ac.kr
[2] Department of Sports ICT Convergence, Sangmyung University, Hongjimun 2-Gil 20, Jongno-Gu, Seoul 03016, Republic of Korea

Abstract. Fingerprint recognition is a biometric method. Recently, attempts for spoofing of fingerprint recognition systems through fake fingerprints have been frequently reported. Most existing fake fingerprint detection methods require either additional sensors or complicated calculations. In the present study, a new fake fingerprint detection method implemented through the combinations of six simple statistical moment features is proposed. The six statistical moments mean the deviation, variance, skewness, kurtosis, hyperskewness, and hyperflatness. Average brightness, standard deviation, and differential feature are additionally used. The multi-dimensional features were combined through the Support Vector Machine. According to the results of experiments, the proposed method showed classification accuracy of about 98%.

Keywords: Fake fingerprint detection · Statistical moments · Support vector machine

1 Introduction

Fingerprint recognition is a biometric method that is the most widely used to verify individuals of identify a person among many persons. However, cases of spoofing through fake fingerprints have been reported from many fingerprint recognition systems. Existing studies for fake fingerprint detection are as follows. The first type is methods in which whether sweat comes out from finger surfaces is measured [1]. These methods require additional hardware such as sensors that have a problem of possible insanitary conditions. The second type is methods in which where the skin has been deformed is observed [2]. However, these methods have problems such as the necessity to analyze time series data and high time complexity for analysis. The third type is methods for judgment based on fingerprint image quality [3]. However, since existing methods extract features dependent on the positions of pixels such as LBP and Gabor, these methods have problems of long processing time. The fourth type is methods in which sweat pores are observed [4]. However, this type of methods cannot be applied to cases where the resolution of images is low. To solve these problems, in the present study, a fake fingerprint detection method using statistical moment features is

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_8

proposed. In the proposed method, features are calculated from the PDF (Probability Density Function) of levels of the entire area of the finger print area. Six statistical moments; mean, variance, skewness, kurtosis, hyperskewness, and hyperflatness are used as features and these features are fused with additional three features (average brightness, standard deviation, difference) to classify real and fake fingerprints.

2 Proposed Method

The proposed method is implemented in the procedure shown in Fig. 1.

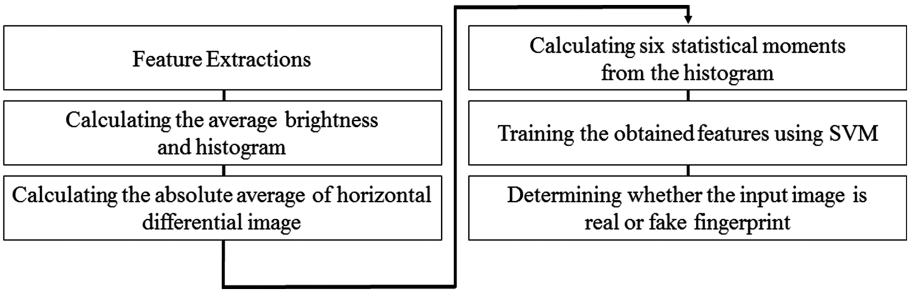


Fig. 1. Flow diagram of proposed method

2.1 Feature Extractions

Statistical Moments

Examples of real fingerprint and fake fingerprint images obtained from fingerprint sensors are as shown in Fig. 2.

Fig. 2. Examples of real (left) and fake (right) finger images, respectively

As features to distinguish between real and fake finger images, in the proposed method, statistical moments as shown by Eq. (1) are used [5].

$$u_n(\mathrm{v}) = \sum_{j=0}^{A-1} (v_i - m)^n p(v_i) \tag{1}$$

In Eq. (1), $p(v_i)$ refers to PDF and can be obtained through dividing intensity frequency by number of pixels. Values in a range of 0–253 are used as i by excepting 254 and 255. Example of PDFs for real fingers and fake fingers are as shown in Fig. 3.

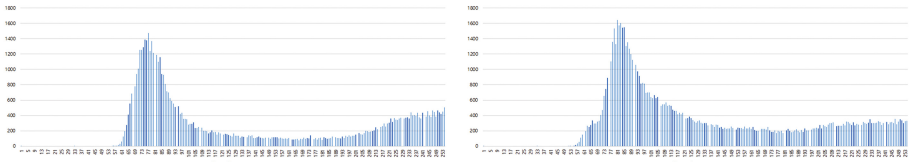


Fig. 3. Comparison of real and fake fingerprint PDF. (a) Real fingerprint image. (b) Fake fingerprint image

In general, statistical moments for PDF are known to reflect those features that are shown in Table 1 [6].

Table 1. Features of n_{th} statistical moments.

Moment number	Raw moment	Central moment	Standardized moment	Raw cumulant	Standardized cumulant
1	Mean	0	0	Mean	N/A
2	-	Variance	1	Variance	1
3	-	-	Skewness	-	Skewness
4	-	-	Kurtosis	-	Excess kurtosis
5	-	-	Hyperskewness	-	-
6	-	-	Hyperflatness	-	-

In Table 1, six statistical moments (moment number: 1–6) are used as features in order to determine real and fake fingerprints (*F#1–F#6*).

Obtaining Differential Feature

In addition to the mentioned six features, we used three additional features such as the average brightness (*F#7*), standard deviation (*F#8*), and the differential image feature (*F#9*). The average brightness and standard deviation can be easily obtained by calculating the statistical moment features as explained in previous section. The difference values of the entire area are calculated in pixel units using equation.

$$F\#9 = \frac{1}{wh}\sum_{i=0}^{i=h}\sum_{j=0}^{j=w-1}|a(i,j) - a(i,j+1)| \tag{2}$$

In Eq. (2), $a(i, j)$ means the intensity value at the position of i_{th} row and j_{th} column. Also, h and w mean height and width of the input image, respectively. Consequently, the absolute average of horizontal difference image is used the other feature (*F#9*) for detecting fake fingerprint detection.

2.2 Fusing Features Using SVM

In the proposed method, the SVM method is used to fuse the nine features at the feature level. SVM is a learning algorithm based classification method, which finds the optimal hyperplane to divide data into 2 categories at the maximal margin [7]. SVM uses the boundary data of each group. SVM finds two vectors from the boundary of each group. Then, a vector having a maximal margin with the two vectors is determined as a support vector. SVM can determine a nonlinear classifier by using various kernels. Consequently, the SVM can calculate a curve-shaped classifier in the case of a 2-dimensional feature space.

3 Experimental Results

Fingerprint data from a total of 17 persons were used to make and use a database consisting of 16 real fingerprint images and 16 fake fingerprint images per person, amounting to 272 real fingerprint images and 272 fake fingerprint images. The used data (ATVS-FFp DB) is given by Biometric Research Group – ATVS. The numbers of both training data and test data was 136 real data and 136 fake data, respectively. As the first experiment, the real and fake fingerprint classification performance of each of the six features was checked. Figure 4 shows the distributions of real and fake fingerprints for individual features.

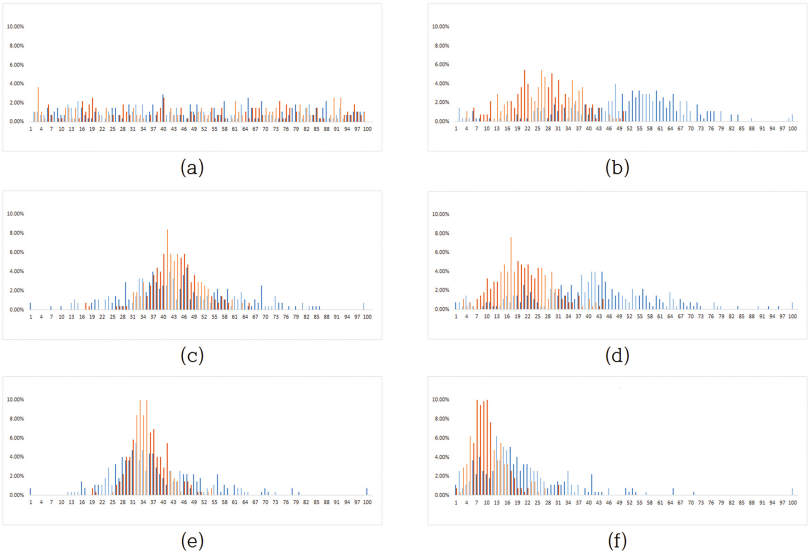


Fig. 4. Distributions of real (blue) and fake (orange) fingerprints in cases of using six features. (a) 1st moment (*F#1*). (b) 2nd moment (*F#2*). (c) 3rd moment (*F#3*). (d) 4th moment (*F#4*). (e) 5th moment (*F#5*). (f) 6th moment (*F#6*).

As the second experiment, the real and fake fingerprint classification performance of each of the three features was checked. Figure 5 shows the distributions of real and fake fingerprints for individual features.

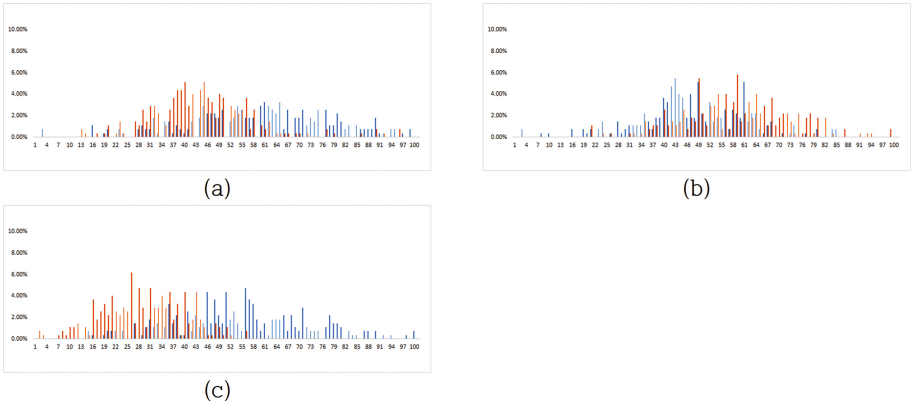


Fig. 5. Distributions of real (blue) and fake (orange) fingerprints in cases of using three features. (a) Average brightness (*F#7*). (b) Standard deviation (*F#8*). (c) Difference (*F#9*).

In the third experiment, the nine features were learned using SVM and the accuracy of classification was tested using the remaining data. RBF kernels were used and the classification accuracy based on gamma values is as shown in Table 2.

Table 2. Two kinds of error rate such as FAR and FRR according to using several Gamma parameters.

Features	Kernel, Gamma	Training set		Test set	
		FRR	FAR	FRR	FAR
Nine feature set	RBF 0.000001	8.82% (12/136)	1.47% (2/136)	16.18% (22/136)	0% (136/136)
	RBF 0.000006	**0% (0/136)**	**0% (0/136)**	**2.94% (4/136)**	**0% (0/136)**
	RBF 0.00001	0% (0/136)	0% (0/136)	7.35% (10/136)	1.47% (2/136)
	RBF 0.0001	0% (0/136)	0% (0/136)	1.47% (2/136)	1.47% (2/136)

In Table 2, the FAR (False acceptance Rate) is the rate of wrong recognition of fake fingerprints as real fingerprints and the FRR (False Rejection Rate) is the rate of wrong recognition of real fingerprints as fake fingerprints. Based on Bayesian classification theory, FAR and FRR have trade-off relationships based on changes in the threshold. To determine the optimal Gamma value of RBF kernel, least mean square approach using Lab-made program was used in terms of the minimum EER (Equal Error Rate).

In Table 2, even though the average error rate in cases of 2nd and 4th rows were best, the 2nd row case was regarded as the best one in terms of security. Actually, FAR is more critical than FRR in terms of security.

4 Conclusion

In the present study, a method of identifying real and fake fingerprints using statistical moments with six features; mean, variance, skewness, kurtosis, hyperskewness, and hyperflatness three additional features; average brightness, standard deviation and difference without any separate equipment was proposed. When the proposed method was tested, the overall accuracy was shown to be 98.53%(=100% − 1.47%), FRR was shown to be 2.94%, and FAR was shown to be 0%.

Acknowledgements. This research was supported by the Ministry of Science, ICT and Future Planning (MSIP), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2016-H8501-16-1014) supervised by the Institute for Information & Communications Technology Promotion (IITP). Also, this research was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (2016M3A9E1915855).

References

1. Parthasaradhi, S.T., Derakhshani, R., Hornak, L.A., Schuckers, S.A.: Time-series detection of perspiration as a liveness test in fingerprint devices. IEEE Trans. Syst. Man Cybern. Part C. **35**, 335–343 (2005)
2. Reddy, P.V., Kumar, A., Rahman, S., Mundra, T.S.: A new antispoofing approach for biometric devices. IEEE Trans. Biomed. Circuits Syst. **2**, 328–337 (2008)
3. Nikam, S.B., Agarwal, S.: Ridgelet-based fake fingerprint detection. Neurocomputing **72**, 2491–2506 (2009)
4. Ghiani, L., Hadid, A., Marcialis, G.L., Roli, F.: Fingerprint liveness detection using binarized statistical image features. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6. IEEE (2013)
5. Gonzalez, R.C., Richard, E.W.: Digital Image Processing. Nueva Jersey (2008)
6. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of 5th Annual Workshop on Computational Learning Theory (COLT 1992), New York, pp. 144–152 (1992)
7. 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**, 311–321 (2012)

Path Planning Method for Collision Avoidance of Multiple UAVs

Hyeok Kim[1], Jeonghoon Kwak[2], Guichang Sim[1], and Yunsick Sung[2(✉)]

[1] Faculty of Computer Engineering, Keimyung University, Daegu, South Korea
{hyeok, guichang}@kmu.ac.kr
[2] Department of Multimedia Engineering, Dongguk University, Seoul, South Korea
{jeonghoon, sung}@dongguk.edu

Abstract. In recent years, the development of unmanned aerial vehicles (UAVs) has increased significantly and they are currently used in various fields and applications. In some applications, multiple UAVs need to be cooperated to accomplish tasks, because a single UAV is not sufficient. However, even when multiple UAVs are used, their autonomous control systems are not perfect, which leads to collisions between the UAVs. In this paper, we propose a path planning method for collision avoidance of UAVs, when multiple UAVs are controlled using a ground control system. Furthermore, using this method, the UAVs have less likelihood to be in a close encounter with obstacles, and collisions are avoided.

Keywords: UAV · Path planning · Collision avoidance · Multiple UAVs · Geometric approach

1 Introduction

As applications of unmanned aerial vehicles (UAVs) have been increasing, more fields are trying to use UAVs. A few examples are search and rescue operations, surveillance systems, crop dusting, and performances using multiple UAVs [1–4].

When a single UAV is used, it can take a long time to accomplish tasks, or even fail. This problem can be resolved using multiple UAVs. However, to operate multiple UAVs in a ground control system, it is necessary to apply a collision avoidance technique to the system.

In this paper, we propose a path planning method for collision avoidance of multiple UAVs based on recorded UAV flight data. A geometric arithmetic algorithm is used to plan the flight paths for multiple UAVs, predict possible collisions before operation, and prevent collisions [6]. While traditional geometric arithmetic algorithms use the velocity information of UAVs, the proposed method computes the differences in recorded UAV flight data. Because the method uses the recorded UAV flight data to plan the paths, it can consider the collision issues less than traditional algorithms.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_9

The rest of this paper is organized as follows: In Sect. 2, we propose a path planning method for multiple UAVs. In Sect. 3, the implementation of the proposed method is described, the operations are verified, and the conclusions are provided.

2 Multiple UAV Control for Collision Avoidance

The method for generating collision avoidance paths is described in Fig. 1. In the first stage, i.e., the flight record correction stage, a user records UAV flight data while controlling the UAV with a controller. In the second stage, i.e., the flight record revision stage, a path revision algorithm is used to revise the paths recorded in the first stage. Finally, in the third stage, i.e., the avoidance path planning stage, the system generates the paths in which no collisions between UAVs would occur.

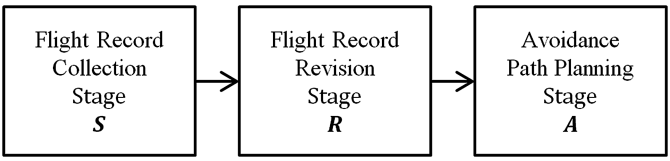


Fig. 1. This figure shows a diagram of the three stages to obtain results from the proposed algorithm. The first stage is flight record collection through the control of a UAV by a user, using a controller, the second stage is the path revision stage, and the last stage is the avoidance path planning stage.

2.1 UAV Flight Data Records

The hourly states of the n^{th} UAV record are defined as the collected state set, S_n, as shown in Eq. (1). The collected state, $S_{n,i}$, indicates the i^{th} collected state of the collected state set, S_n. Then, the collected state, $S_{n,i}$, is described in terms of four components, $x_{n,i}$, $y_{n,i}$, $z_{n,i}$, and $t_{n,i}$, which represent the GPS latitude, the GPS longitude, the absolute altitude, and the flight time from departure, respectively. Finally, the collected states, $S_{n,i}$, are added to the collected state set, $S_{n,i}$, to record.

$$\boldsymbol{S}_n = \left[S_{n,1}, S_{n,2}, \cdots, S_{n,i}, \cdots\right] \tag{1}$$

$$S_{n,i} = \left[x_{n,i}, y_{n,i}, z_{n,i}, t_{n,i}\right] \tag{2}$$

2.2 Path Revision Computations

To revise the collected flight data, the Douglas–Peucker algorithm is applied [5]. This algorithm suggests a path revision method in two-dimensional space ($f{:}p_{n,i} \rightarrow R_{n,j}$) and advances the method to a three-dimensional path revision method. This method is not included in this paper.

The collected state set, S_n, is revised to make the revised state set, R_n, as shown in Eq. (3). The revised state, $R_{n,j}$, is the j^{th} revised state in the revised state set, R_n, and is defined by Eq. (4). The element in the revised state set, R_n, uses the element of the collected state set, S_n, to define the collected state, $S_{n,i}$, which contains the $[x_{n,i}, y_{n,i}, z_{n,i}]$ elements that are outside the error tolerance, as the revised state, $R_{n,j}$.

$$\boldsymbol{R}_n = [R_{n,1}, R_{n,2}, \cdots, R_{n,j}, \cdots] \tag{3}$$

$$R_{n,j} = [x_{n,j}, y_{n,j}, z_{n,j}, t_{n,j}] \tag{4}$$

2.3 Collision Avoidance Path Planning

The avoidance state set, A_n, is defined by Eq. (5). The avoidance state, $A_{n,k}$, is defined as the k^{th} avoidance state, and is shown in Eq. (6). The state-difference set, D_n, is defined by Eq. (7). The state-difference, $D_{n,j}$, is the j^{th} element of the state-difference set, D_n, and is shown in Eq. (8). The avoidance point set, P_n, is defined by Eq. (9). The avoidance point, $P_{n,k}$, is the k^{th} element of the avoidance point set, P_n, and is shown in Eq. (10). Finally, the direction pointing to the avoidance point, $P_{n,j}$, is defined as $U_{n,j} = [x_{n,j}, y_{n,j}, z_{n,j}]$.

$$\boldsymbol{A}_n = [A_{n,1}, A_{n,2}, \cdots, A_{n,k}, \cdots] \tag{5}$$

$$A_{n,k} = [x_{n,k}, y_{n,k}, z_{n,k}, t_{n,k}] \tag{6}$$

$$\boldsymbol{D}_n = [D_{n,1}, D_{n,2}, \cdots, D_{n,j}, \cdots] \tag{7}$$

$$D_{n,j} = [x_{n,j}, y_{n,j}, z_{n,j}, t_{n,j}] \tag{8}$$

$$\boldsymbol{P}_n = [P_{n,1}, P_{n,2}, \cdots, P_{n,k}, \cdots] \tag{9}$$

$$P_{n,k} = [x_{n,k}, y_{n,k}, z_{n,k}] \tag{10}$$

The computational time, $t'_{n,m}$, is defined as the m^{th} element of the computational time set, $T'_n = \{0, 1, \cdots, t_{n,|R|-1}\ of\ R_n - t_{n,1}\ of\ R_n - 1, t_{n,|R|}\ of\ R_n - t_{n,1}\ of\ R_n\}$, where n! = MAX(n). $h(n1, n2, m)$ is $\max_j(t_{n1,j}\ of\ R_{n1} - t_{n,1}\ of\ R_{n1})$ where $t_{n1,j}\ of\ R_{n1} - t_{n1,1}\ of\ R_{n1} < t'_{n2,m}$.

There are nine steps in the generation procedure. First, the values of n1 and n2 that satisfy the condition n1 < n2 need to be determined, and m is set to be 1. If the values of n1 and n2 cannot be determined, then the generation of the collision avoidance path is terminated.

Second, if $A_n \neq \emptyset$, then each element of the avoidance state set, A_n, is added to the revised state set, R_n, in order, and the avoidance state set, A_n, and the state-difference set, D_n, are set to be null sets.

Third, the state-difference set, D_n, is computed using the revised state set, R_n. Each element of the state-difference set, D_n, is calculated as shown in Eq. (11); it is the difference between the adjacent elements of the revised states in the revised state set, R_n.

$$\begin{aligned} x_{n,j} &= x_{n,j+1} - x_{n,j}, y_{n,j} = y_{n,j+1} - y_{n,j} \\ z_{n,j} &= z_{n,j+1} - z_{n,j}, t_{n,j} = t_{n,j+1} - t_{n,j} \end{aligned}, \tag{11}$$

Fourth, the indices j1 and j2 are determined. Set $j1 = h(n1, n1, m)$ and $j2 = h(n2, n1, m)$.

Fifth, the revised states, $R_{n1,j1}$ and $R_{n2,j2}$, are added to the avoidance state sets, A_{n1} and A_{n2}, respectively, using Eq. (12).

$$A_n = A_n \cup \{R_{n,j}\} \tag{12}$$

Sixth, using the geometric arithmetic algorithm and the state-differences, $D_{n1,j1}$ and $D_{n2,j2}$, the time for reaching the collision prediction point, t'', and the directions pointing to the avoidance point, $U_{n1,j1}$ and $U_{n2,j2}$, are computed, where $g: D_{n1,j1}, D_{n2,j2} \rightarrow [t'', U_{n1,j1}, U_{n2,j2}]$ [6]. If $t'' \leq 0$ or $j1 = |R_{n1}|$, the system skips to the ninth step without performing the seventh and the eighth steps, because the collision cannot be predicted under those conditions. Otherwise, the system performs the remaining steps in order.

Seventh, the avoidance point is generated using the directions pointing to the avoidance point, $U_{n1,j1}$ and $U_{n2,j2}$. $U_{n1,j1}$ is always generated, while $U_{n2,j2}$ is not. Therefore, when $U_{n2,j2}$ is not generated, $U_{n1,j1}$ is computed using Eq. (13) to generate the avoidance point, $P_{n1,j1}$.

$$\begin{aligned} x_{n,j} =& x_{n,j} \, of \, R_n + t''_n \times x_{n,j} \, of \, D_n + \left(x_{n,j+1} \, of \, R_n - x_{n,j} \, of \, R_n\right) \times t'' \times x_{n,j} \, of \, U_{n,j}, \\ y_{n,j} =& y_{n,j} \, of \, R_n + t''_n \times y_{n,j} \, of \, D_n + \left(y_{n,j+1} \, of \, R_n - y_{n,j} \, of \, R_n\right) \times t'' \times y_{n,j} \, of \, U_{n,j}, \\ z_{n,j} =& z_{n,j} \, of \, R_n + t''_n \times z_{n,j} \, of \, D_n + \left(z_{n,j+1} \, of \, R_n - z_{n,j} \, of \, R_n\right) \times t'' \times z_{n,j} \, of \, U_{n,j} \end{aligned} \tag{13}$$

Eighth, the generated avoidance point, $P_{n,j}$, is added as the last element of the avoidance state set, A_n. For example, the possible flight time is predicted when the avoidance point, $P_{n1,j1}$, is generated, and added as the last element of the avoidance state, A_n, using Eq. (14).

$$A_n = A_n \cup \{[P_{n,j}, t'_{n,m} + t'']\} \tag{14}$$

Ninth, if $t'_{n,m}$ is not the last element of T'_n, then m is increased by 1 and the procedure is repeated starting from the fourth step. If $t'_{n,m}$ is the last element of T'_n, then the procedure is repeated starting from the first step.

3 Experiments

3.1 Three-Dimensional Experiments Scenario and UAV Flight Data Record

In this study, we conducted experiments to verify the proposed method for the path planning based on the collected UAV flight data and for avoiding collisions between the UAVs. The proposed method was verified using a scenario that we designed. The flight record in this scenario includes external disturbances, such as wind and an inexperienced pilot, and internal disturbances, such as GPS position error. The unstable records are shown in Fig. 2.

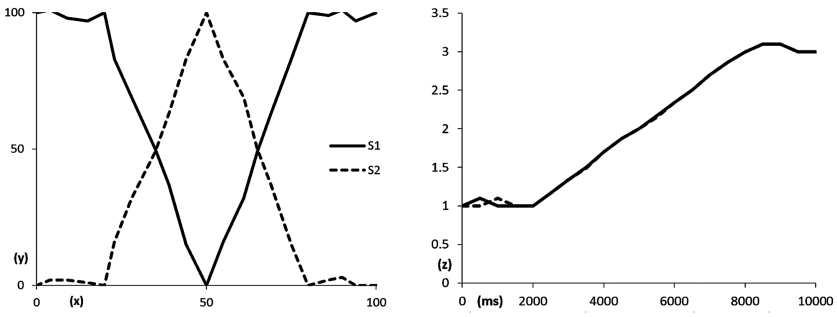


Fig. 2. UAV flight data collected from two different UAVs. The graphs are linear but noisy and unstable.

3.2 Path Revision Computation

In this study, the path was revised in three-dimensional space. The results are shown in Fig. 3. By correcting the collected flight data, the graphs became considerably stable and therefore, it is expected that the UAVs would not stop during autonomous flight.

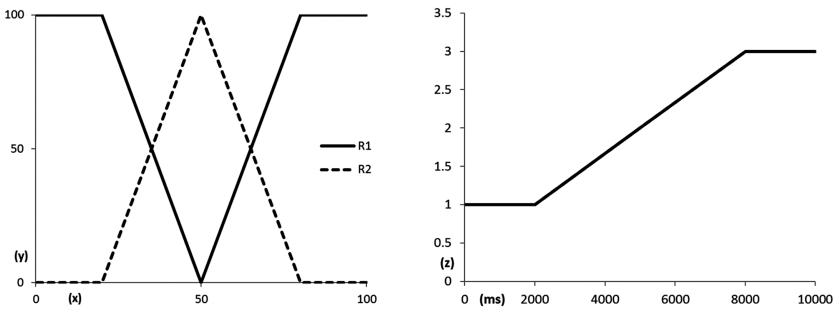


Fig. 3. The corrected path of the collected flight data, S. Dogulas–Peucker algorithm was applied in three-dimensional space.

3.3 Experiments for the Generation of the Collision Avoidance Path

In this study, a three-dimensional collision avoidance algorithm was applied [6]. The path generated through collision avoidance is shown in Fig. 4.

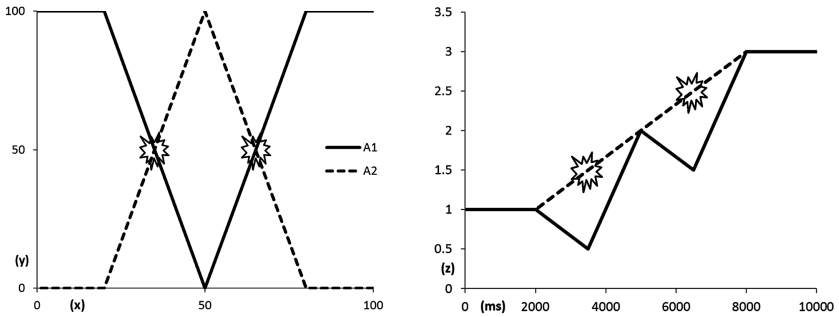


Fig. 4. A generated path, A, by predicting collisions at different time intervals, and generating and adding the avoidance points.

In the experiments, two collision points were predicted and these were reflected in the system such that the UAV could change its path along the Z-axis to avoid collision. Finally, the two collision points, (65.00504, 49.98345, 0.50015, 3.5) and (34.98355, 50.02151, 1.500367, 6.5), were added to the avoidance state set, A_1.

3.4 Conclusion

In this paper, we propose a method to collect and revise UAV flight data through a user with a UAV controller, and path adjusting and planning techniques for collision avoidance of UAVs. When multiple UAVs are controlled by a ground control system using the proposed method, the UAVs have less likelihood to be in a close encounter with obstacles and other UAVs. Therefore, the system would have less burden to consider the collision issues than traditional algorithms.

Acknowledgments. This work was supported by the Korea Foundation for the Advancement of Science & Creativity (KOFAC), and funded by the Korean Government (MOE).

References

1. Fahlstrom, P.G., Gleason, T.J.: Introduction to UAV Systems, 4th edn. Wiley, Chichester (2012)
2. Bellingham, J.S., Tillerson, M., Alighanbari, M., How, J.P.: Cooperative path planning for multiple UAVs in dynamic and uncertain environments. In: 41st IEEE Conference on Decision and Control, Las Vegas, Nevada USA, 10–13 December, vol. 3, pp. 2816–2822 (2002)

3. Cacace, J., Finzi, A., Lippiello, V.: Multimodal interaction with co-located drones for search and rescue. In: 2nd Workshop on Artificial Intelligence and Robotics, Ferrara, Italy, 22 September, vol. 1544, pp. 54–67 (2015)
4. Sung, Y., Kwak, J., Yang, D., Park, Y.: Ground station design for the control of multi heterogeneous UAVs. In: Korea Multimedia Society Spring Conference, 28–30 May, vol. 18, no. 1, pp. 829–830. Andong Culture and Art Center, Andong (2015)
5. Douglas, D.H., Peucker, T.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Can. Cartographer **10**(2), 112–122 (1973)
6. Park, J., Oh, H., Tahk, M.: UAV collision avoidance based on geometric approach. In: The Society of Instrument and Control Engineers (SICE) Annual Conference, 20–22 August 2008, pp. 2122–2126. The University of Electro-Communications, Tokyo (2008)

3D UAV Flying Path Optimization Method Based on the Douglas-Peucker Algorithm

Guichang Sim[1], Jaehwa Chung[2], and Yunsick Sung[3(✉)]

[1] Faculty of Computer Engineering, Keimyung University, Daegu, South Korea
guichang@kmu.ac.kr
[2] Department of Computer Science,
Korea National Open University, Seoul, South Korea
jaehwachung@knou.ac.kr
[3] Department of Multimedia Engineering,
Dongguk University, Seoul, South Korea
sung@dongguk.edu

Abstract. Unmanned Aerial Vehicles (UAVs) have been utilized in various applications in many fields in recent years. The paths the pilots flew can be measured and collected to be utilized to create routes for autonomous flight. However, there is a problem in that GPS errors result in the path being irregularly represented. The measured path can be optimized by using the Douglas-Peucker algorithm. Our research led to the proposal of a method to optimize this path by applying the Douglas-Peucker algorithm, which has been shown to be suitable for a two-dimensional path, in three-dimensional space. Optimization of the 3D path by the proposed method was possible by deleting unnecessary points from the three-dimensional space. Thus, the flight paths that were measured and collected can be utilized to define the autonomous flight path.

Keywords: UAV · Drone · Path planning · Douglas-Peucker · 3D path

1 Introduction

Unmanned Aerial Vehicles (UAVs) [1] have found application in various fields, such as in filming [2] and monitoring [3], and their use is expanding. Especially, during broadcasted sporting events, UAVs can be seen utilized for filming, in which case the UAV can be flown by defining the flight path in advance. Regarding the flight path, to set it intuitively, it is possible to use the path the pilot flew. However, the intuitive approach is problematic in that GPS errors and external influences such as wind would cause the flight path to be irregularly represented. Therefore, there is a need for studies to optimize the measured path.

A method to smooth the path followed by a UAV has been reported [4]. However, even when the UAV flew at right angles, with the smoothing method, the path was always expressed in curves, thereby increasing the flight path error. The Douglas-Peucker algorithm presents a way to optimize the two-dimensional path [5]. This algorithm achieves optimization by reducing the number of points representing a

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_10

complex curve. However, since it has only been used to process a two-dimensional path, further study is required to apply it to a three-dimensional path.

Our research aimed to use a UAV to film a soccer game while following the ball autonomously, and led to the proposal of a three-dimensional flight path optimization method based on the Douglas-Peucker algorithm.

This paper is organized as follows. Section 2 explains how to optimize the 3D path. Section 3 describes the experiments for optimizing the 3D path and presents the experimental results. Section 4 concludes on 3D path optimization.

2 3D Path Optimization Process

The flight path of a UAV filming a football game includes the absolute altitude as well as the GPS-based location. This paper introduces a three-dimensional path optimization method applied to the 3-D path (defined in terms of the GPS coordinates and the absolute altitude) with the Douglas-Peucker algorithm.

2.1 3D Douglas-Peucker Algorithm

A method is required to express the flight paths in three dimensions. thus, a method that can represent the 3D points based on the GPS coordinates and the absolute altitude is needed. Because the GPS coordinates and absolute altitude are measured in different units, this paper represents both GPS and absolute altitude in meters. The three-dimensional Route $R_{n1,n3}$ is a set of 3D positions collected from *n1*-th to *n3*-th. The Point P_n is a three-dimensional position with x_n, y_n, and z_n coordination as Eq. (1). For example, the GPS position of the n-th position is obtained by finding the distance from the meridian to the longitude values and then convert it to x_n, and then finding the distance of the latitude from the equator, then convert it to y_n. The absolute altitude is converted to z_n.

$$R_{n1,n3} = [P_{n1}, \cdots, P_{n3}] \; where \; n1 \neq n3 \tag{1}$$

$$P_n = [x_n, y_n, z_n] \tag{2}$$

The straight line segment that connects the Point P_{n1} and Point P_{n3} is defined as $L_{n1,n3} = \{P_{n1}, P_{n3}\}$ where n1 $\neq$ n3. The distance $d_{n1,n2,n3}$ between any intermediate position of Route $R_{n1,n3}$, Point $P_{n2}(n1 < n2 < n3)$ and line segment $L_{n1,n3}$ is the shortest distance between the line segment $L_{n1,n3}$ and point P_{n2}. In this paper, distance $d_{n1,n2,n3}$ is calculated based on the property that the size of the cross product of any two vectors is equal to the area of the parallelogram consisting of the two vectors. For example, define the vector from the *n1*-th position to *n2*-th position vector as $V_{n1,n2}$, and define the vector from the *n1*-th position to *n3*-th position as the $V_{n1,n3}$. Obtain the area by calculating the size of the vector that are the cross product of two vectors, $V_{n1,n2}$ and $V_{n1,n3}$, where $V_{n1,n3}$ represents the base of the parallelogram. Divide the area by the size of $V_{n1,n3}$, obtain the shortest distance, $D_{n1,n2,n3}$ between the line segment $L_{n1,n3}$ and point P_{n2}.

If the maximum distance, $d_{arg_{n2}MAX(d_{n1,n2,n3})}$ exceeds the offset tolerance ε, based on the maximum distance point $P_{arg_{n2}MAX(d_{n1,n2,n3})}$, divide the Route $R_{n1,n3}$ to generate new 3D routes, Route $R_{n1,n2}$ and Route $R_{n2,n3}$. Apply the 3D Douglas-Peucker algorithm recursively to the 3D routes, Route $R_{n1,n2}$ and Route $R_{n2,n3}$, respectively, for optimization.

2.2 3D Douglas-Peucker Algorithm

The 3D Douglas-Peucker algorithm is processed as follows. Route $R_{n1,n3}$ and offset tolerance ε are entered. If Distance $d_{arg_{n2}MAX(d_{n1,n2,n3})}$ is more than offset tolerance ε, apply algorithm recursively. If Distance $d_{arg_{n2}MAX(d_{n1,n2,n3})}$ does not exceed the offset tolerance ε, then terminate the algorithm.

```
FUNCTION 3D-Douglas-Peucker with Rn1,n3, ε
BEGIN
  IF n3>n1+1 THEN
  BEGIN
     SET n2←arg_nMAX(d_n) where n∈Rn1,n3
     IF d_arg_n2MAX(d_n1,n2,n3) with Pn2, Ln1,n3 > ε THEN
     BEGIN
        CALL Simplify with Rn1,n2, ε
        CALL Simplify with Rn2,n3, ε
     END.
  END.
END.
```

3 Experiments

In this section, we designed a 3-D flight path and offset tolerance to gather flight data by assuming a situation in which the UAV is recording a soccer game being played on a soccer field. Subsequently, the experimental results obtained by applying the collected data based on the design with the proposed algorithm are introduced.

3.1 Design of Experiments

In the experiment, a Parrot AR.Drone 2.0 [6] was used. The GPS sensor of the AR. Drone 2.0 has an accuracy of ±2 m. For the experiment, the offset tolerance was set up to 2.5 m, with consideration for the effect of the wind and the GPS error.

Assuming a situation in which a football game is being filmed, the UAV traversed half of the stadium, along a path such as that represented in Fig. 1. The numbers in Fig. 1 represent the absolute altitude at which the UAV can fly. The UAV takes off from the position marked by Start and lands at the position marked End.

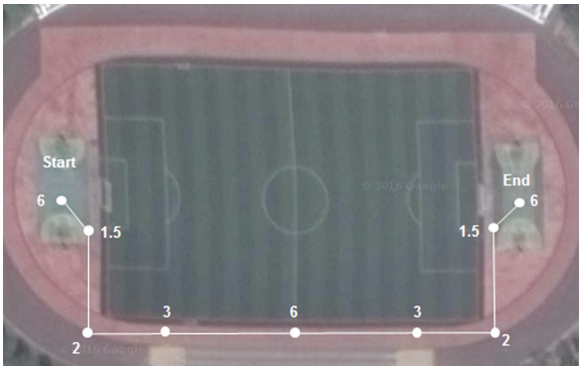
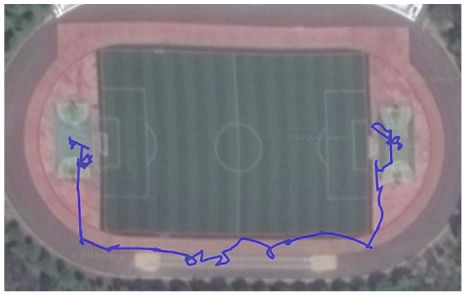
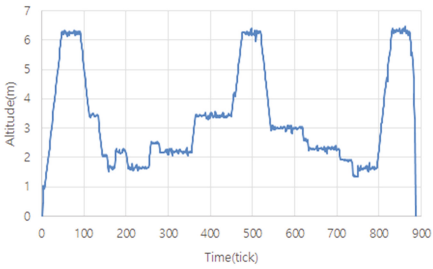


Fig. 1. Flight path specified for filming the soccer game. Start and End indicate the start and end positions of the path. The lines represent the established UAV flight path and the numbers represent the absolute altitude of the UAV.

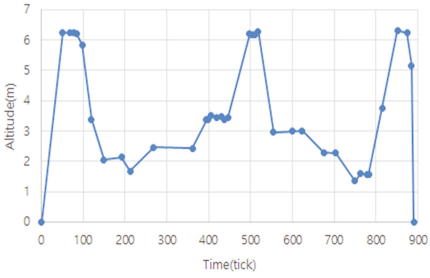


(a) Collected GPS



(b) Collected Absolute Altitude (AA)

(c) GPS applied with the proposed algorithm



(d) AA with the proposed algorithm

Fig. 2. Illustration of the optimized path using the collected UAV flight path and algorithm. (a), and (b) each represents the absolute altitude and GPS collected before the optimization. (c), and (d) represent the absolute altitude and GPS optimized by the Douglas-Peucker algorithm, respectively.

3.2 Test Result

The proposed method is verified in the following experiment. First, it collects the 3D data points along the path of the UAV flight. Figure 2 shows a comparison of the 3D path optimization utilized with the Douglas-Peucker algorithm and the collected UAV 3D path. Figure 2(a) shows the 3D flight path collected by the UAV, and Fig. 2(b) shows the absolute altitude of the collected UAV data.

Second, when the data collection is complete, calculate the optimized 3D path with the proposed method. Figure 2(c) shows the optimized 3D path using the Douglas-Peucker algorithm, and Fig. 2(d) represents the optimized absolute altitude by using the Douglas-Peucker algorithm. A total of 881 collected locations were optimized to 38 locations by the proposed algorithm.

Third, the distance travelled between the collected 3D path and optimized 3D path is found and compared. The flight distance of the collected flight path is about 489.77 m and the flight distance of the optimized 3D flight path is about 370.95 m. Therefore, it was possible to reduce it by about 62.9%.

4 Conclusion

This paper proposes a 3D path optimization algorithm, which is based on the Douglas-Peucker algorithm, for UAV path collection. The experiments verified that the path of the collected UAV is optimized by applying the proposed method. Optimization of the 3D route led to a reduction in the error. By shortening the distance of the 3D path, the UAV was able to fly the route more quickly.

Acknowledgments. This work was supported by the Korea Foundation for the Advancement of Science & Creativity (KOFAC), and funded by the Korean Government (MOE).

References

1. Fahlstrom, P.G., Gleason, T.J.: Introduction to UAV Systems, 4th edn. Wiley, Chichester (2012)
2. Yahyanejad, S., Wischounig-Strucl, D., Quaritsch, M., Rinner, B.: Incremental mosaicking of images from autonomous, small-scale UAVs. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, 29 August–1 September 2010, pp. 329–336 (2010)
3. Kim, D., Hong, S.K.: Target pointing and circling of a region of interest with quadcopter. Int. J. Appl. Eng. Res. **11**(2), 1082–1088 (2016)
4. Yang, K., Sukkarich, S.: 3D smooth path planning for a UAV in cluttered natural environments. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, 22–26 September 2008, pp. 794–800 (2008)
5. Douglas, D.H., Peucker, T.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Can. Cartographer **10**(2), 112–122 (1973)
6. AR.Drone 2.0 official site. http://www.parrot.com/usa/products/ardrone-2/

Encrypted Network Traffic Analysis Method via Secure Socket Layer Handshake Control

Jihoon Yoon, Kangsik Shin, and Yoojae Won(✉)

Department of Computer Science Engineering, Chung-Nam National University, Daehak-ro. 99, Daejeon 34134, South Korea
{yjh6569, sks4315, yjwon}@cnu.ac.kr

Abstract. As the amount of encrypted network traffic on enterprise networks increases steadily, the problem of malicious acts encrypted to bypass security devices has emerged. Previous studies analyzed the encrypted network traffic by changing the network traffic or communication flow between the encrypted communications to analyze such encrypted malicious behavior. However, there are limitations to the existing methods because they require additional prior-data or additional network configurations in order to analyze the encrypted network traffic. In this paper, we propose a system to decrypt secure socket layer network traffic to analyze the encrypted network traffic in the enterprise network environment. The proposed system can be used to analyze encrypted network traffic in order to detect malicious activity and corporate information leaks.

Keywords: SSL traffic decryption · Encrypted traffic analysis

1 Introduction

Currently, the share of encrypted network traffic in the corporate network is growing. According to a survey conducted on corporate security personnel, 87% of the enterprises had 25% of their entire network traffic encrypted, and 97% responded that they have plans to expand the proportion of the encrypted network traffic within the next two years [1].

Although the current network traffic encryption method is used as a means for network security, the security experts are wary of the possibility that network traffic encryption can be used as a means to bypass the detection of malicious cyber-attacks. According to survey data on cyber-attacks, more than 50% of them are expected to be achieved through the secure socket layer (SSL) network traffic by 2017 [2]. Cyber-criminal organizations and hackers use the encrypted network traffic on internal networks for intrusion, malicious code distribution, and command & control (C&C) network traffic transmission. These malicious acts can bypass existing security solutions. In addition, when general sites that provide encrypted communications are infected with malware, the intrusion is undetectable by existing security solutions because malicious codes are delivered through the normal encrypted network traffic.

In this paper, we propose a system that can analyze the encrypted traffic through SSL traffic decryption in the enterprise network environment. This paper is organized as follows. Section 2 introduces the existing research related to encrypted network

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_11

traffic analysis, and Sect. 3 explains the SSL communication security method. In Sect. 4, the structure and operation of the SSL traffic analysis system proposed in this paper are described. Finally, a conclusion is drawn and future research plans are explained in Sect. 5.

2 Related Research

The available methods for analyzing encrypted network traffic include the client's User-Agent method, classification of risk of the encrypted communication by using the Cipher Suites information provided by the applications [3], and using secret sharing schemes in the VPN environment to analyze encrypted traffic [4].

The methods involving the classification of the risks of the encrypted communication through the Cipher Suites and User-Agent information use the server-to-client communication traffic. The User-Agent information is collected through the HTTP header during the HTTP communication process, and Cipher Suites information is collected by the Client Hello message from the SSL Handshake process in HTTPS communication. User-Agent and Cipher Suites information is managed in advance in the form of a pair by using the pre-collected information in bulk for classification in secure encrypted communication and weak encryption communication. If new HTTPS communication is found, the collected pre-information's User-Agent and Cipher-Suites information is compared to the encrypted communication to determine the communication's degree of risk. The analysis speed of this method is high because it uses the traffic delivered prior to the encrypted communication. However, it requires the collection of sufficient pre-data in order to analyze the encrypted traffic analysis.

The encrypted network traffic analysis method, which utilizes the secret sharing scheme in the VPN environment, uses the Shamir secret sharing scheme [5]. It divides the message into n messages according to a predetermined scheme and transmits them to an equal number of Proxy servers. The Proxy server receives the message, divides it according to a predetermined probability, and transmits it to the actual destination and detection system. The actual destination receives the message by restoring the divided message to its original form, and the detection system recovers the message and analyzes the encrypted network traffic. The methods, such as those mentioned in [5], can directly receive the content, making the analysis highly accurate. However, in order to analyze the encrypted network traffic, several Proxy servers and a common secret sharing scheme between the clients and servers is necessary.

The method proposed in this paper analyzes the encrypted network traffic by obtaining the session key through the Handshake message transmitted between the client and server, and it can analyze the encrypted traffic without pre-data or additional environmental configurations used in previous studies.

3 Secure Socket Layer (SSL) Communication

The SSL is a security protocol that provides authentication and encrypted communications between the server and the client on the transport layer. SSL communication generates a session between the server and the client and establishes the encrypted

communication over the secure channel generated. It also performs the SSL Handshake at the session generating step in order to exchange the session key necessary for the encrypted communication.

In the SSL Handshake process, the client and server exchange the random value information through the Hello Message and the server delivers the certificate in addition to the client. After the exchange through the Hello Message, the client encrypts the Secret value with the public key extracted from the server certificate and transmits the message to the server. The server acquires the Secret value by decrypting the encrypted message with the private key and then generates the session key by using a random value shared in advance, including the Secret value.

The sharing method in the SSL Handshake process, which includes the public key of the server certificate, makes it difficult for a third party to obtain the session key. With traditional security equipment, it is impossible to analyze the contents of the encrypted network traffic because it cannot obtain the session key used in SSL communication.

4 SSL Traffic Analyzing System

4.1 System Architecture

The SSL traffic analysis system proposed in this paper is configured with a network interface, ARP Spoofing module, SSL traffic decryption module, and SSL session information DB. The operation of the analytical system proceeds in the following order.

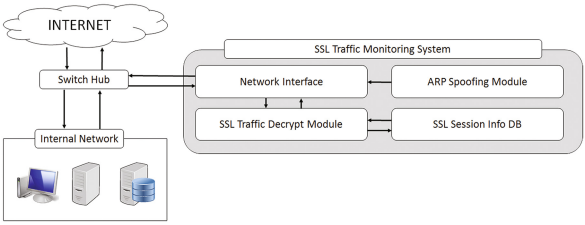


Fig. 1. SSL traffic analysis system architecture

The ARP Spoofing module controls the network traffic flow by sending ARP Spoofing messages over the network interface to the internal network connected to the switch hub [6]. The ARP Spoofing message converts the physical address of the gateway's IP address into a physical address of the analysis system. The host receiving the ARP Spoofing message transmits the traffic through the gateway to the analysis system. The network interface of the analysis system delivers a packet corresponding to the SSL traffic decryption module. The SSL traffic decryption module decrypts the received encrypted traffic and analyzes its contents. The session information necessary to decrypt the SSL traffic is stored in the SSL session information DB and then managed. Figure 1 shows the proposed system architecture.

4.2 Obtaining the SSL Session Key

The system proposed in this paper obtains the session key using an SSL man-in-the-middle [7] in order to decrypt the SSL traffic. Figure 2 shows the process of obtaining a session key from the SSL traffic analysis system.

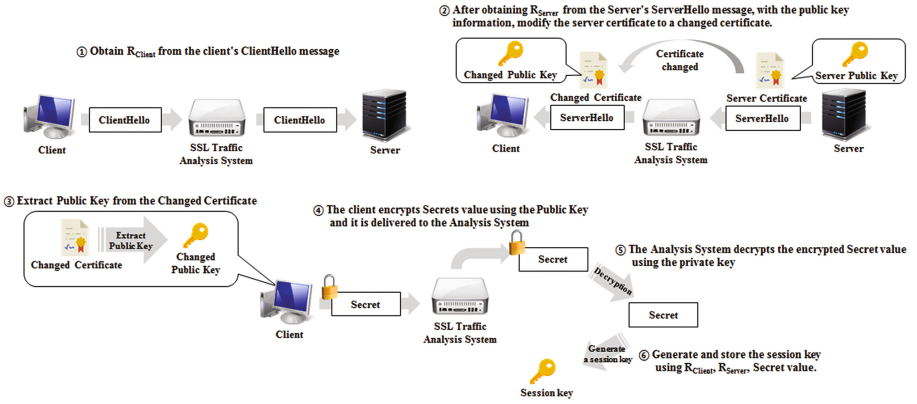


Fig. 2. SSL session key acquisition process

In the SSL Handshake process, the proposed system obtains the random values R_{client}, $R_{serve,}$ and the encryption-algorithm information contained in the Client Hello and Server Hello message. The information of the server certificate's public key included in the Server Hello is changed to the randomly generated public/private key pair and sent to the client. The client extracts a public key of the changed server certificate, encrypts the Secret value using the extracted public key, and transmits to the analysis system. This process provides the Secret values required in the session key generation by using a public/private key pair generated by the analysis system instead of the server private/public key pair, which cannot be obtained during the SSL Handshake process. The analysis system that received the encrypted Secret value obtains the Secret value by decrypting it with a private key from the private/public key pair previously generated. After the Secret value acquisition, the Secret value is re-encrypted with the existing server public key to enable a normal key exchange and passed to the server. The analysis system generates a session key using R_{client}, R_{server}, and the Secret Value obtained in the SSL Handshake process, and the generated session key and encryption algorithm information is stored in a DB and used for decryption of encrypted traffic.

4.3 Server Certificate Changes

The analysis system follows the procedure shown in Fig. 3, creates the certificate, and transmits it to the client in order to pass through the validation phase of a client's server certificate according to the server certificate changes. When the Server Hello is received

from the server, a copy of the certificate is generated during the extraction of server certificate information from the message. The public key information of the certificate copy generated is updated with the public key that is generated by the detection system. We generate a certificate accredited by the private Certificate Authority (CA) by adding the signature of the private (CA) certificate for the changed certificate.

The private CA certificate is generated when the analysis system is installed for the first time, and it is distributed to all the hosts within the enterprise. The distributed private CA certificate is registered with the host's trusted certificate authority, and it should pass the server certificate verifying the steps of the changed certificate in the analysis system.

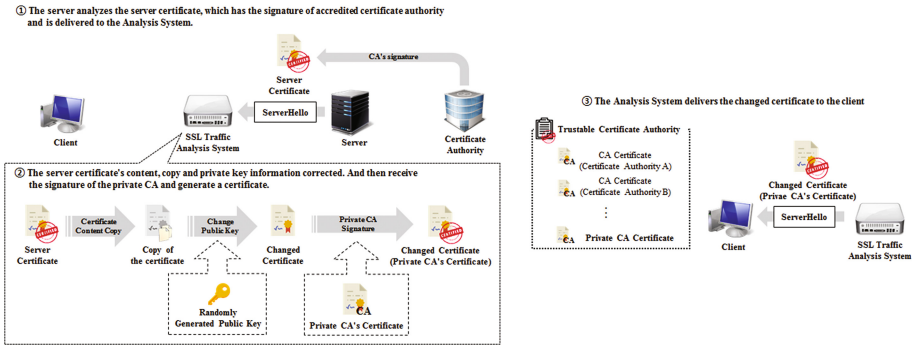


Fig. 3. Certificate generation and delivery process

4.4 Decrypting SSL Traffic

When the encrypted packet enters the analysis system, the corresponding session's session key and encryption algorithm is searched for in the SSL Session Info DB in order to decrypt SSL traffic. If the pertaining session's information does not exist in the DB, it should be considered as an abnormal connection that bypassed the analysis system, and the packet should be discarded. When the corresponding session information exists in the DB, the encrypted packet is decrypted using the stored session key and an encryption algorithm. The decrypted packet may be used to determine whether some action is malicious using the existing malicious behavior detection techniques. The leakage of personal information of a specific pattern can be detected using a regular pattern matching technique. The decrypted packets may adopt the existing malicious behavior detection techniques to determine the presence or absence of malicious action, or apply the regular expression pattern matching technique to detect the leakage of personal information.

5 Conclusion

The detection of malicious behavior using encrypted network traffic, which has recently been on the rise, requires a system that can analyze the encrypted SSL network traffic by decrypting traffic, such as the one proposed in this paper. With this analytical system, we can analyze the encrypted network traffic without the additional data or environment configuration required for existing encrypted network traffic analyses. The proposed analysis techniques for encrypted network traffic can monitor the leakage of internal information via encrypted traffic as well as detect malicious behavior. In addition, it is possible to analyze the encrypted traffic without degrading the overall security level of the analysis system. In future, a detection function for internal information leakage and encrypted malignant behavior will be implemented by using the proposed analysis system. The system will then be applied to the enterprise network environment to demonstrate the performance of the analysis system.

Acknowledgments. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the University Information Technology Research Center support program (IITP-2016-R2718-16-0003) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. ESG: Network Encryption and its Impact on Enterprise Security (2015)
2. Gartner: Security Leaders Must Address Threats from Rising SSL Network traffic (2013)
3. Husák, M., Čermák, M., Jirsík, T., Čeleda, P.: HTTPS network traffic analysis and client identification using passive SSL/TLS fingerprinting, EURASIP J. Inf. Secur. **2016**(6), 1–14 (2016)
4. Goh, V.T., Zimmermann, J., Looi, M.: Detecting attacks in encrypted networks using secret-sharing schemes. Int. J. Cryptol. Res. **2**(1), 89–99 (2010)
5. Shamir, A.: How to share a secret. Commun. ACM **22**, 612–613 (1979)
6. Whalen, S.: An introduction to arp spoofing, Node99 [Online Document] (2001)
7. Chomsiri, T.: HTTPS hacking protection. In: 21st International Conference on AINAW 2007, vol. 1. IEEE (2007)

Property Analysis of SMS Spam Using Text Mining

Manki Baek, Youngkyung Lee, and Yoojae Won(✉)

Computer Science Engineering, Chungnam National University,
Daejeon, Republic of Korea
{bmg8551,yklee,yjwon}@cnu.ac.kr

Abstract. A considerable amount of spam that occur each year can cause the financial damage as well as mental harm to the recipient. This is a serious problem in society. In this paper, we analyze properties of SMS spam in mobile phones to establish a method for effectively blocking SMS spam. As a result, SMS spam can be seen that the surge in the amount shipped during a specific time period. Also, we could find the frequently included word on spam and we could identify spammer that sent smishing messages frequently by comparing several spammers.

Keywords: SMS spam · Text mining

1 Introduction

In modern society, mobile phones are widely used by people due to the development of wireless technologies and the subsequent dependence of numerous people on mobile communication services. According to Ref. [1], the number of mobile communication service users is about 54 million in South Korea as of December 2015. This number is almost the same as the population of South Korea, implying that almost everyone in Korea uses mobile communication services. Because there are so many mobile communication service users, spammers can easily send their spam through the short messaging service (SMS) to many people.

Meanwhile, users receive unwanted messages, which might cause psychological trauma because mobile phones alert users of SMS messages as soon as they are transmitted, unlike e-mails. Furthermore, smishing messages can lead to financial losses [2]. Nevertheless, spammers consistently send spams for their profit. According to Ref. [3], the total number of spam messages collected by a spam trap system was 26.2 million in the second half of 2015. Considering that this number is derived exclusively from the spam trap system, we can infer that there could be much more spam that was not caught by the spam trap system. Thus, SMS spam should be considered as a very serious issue.

Therefore, we definitely need an appropriate solution to filter or block spam. If we can analyze the various properties of spam, the results can be used for policy data to respond to this spam and appropriate solutions can be effectively proposed and designed to block it.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_12

In this paper, we analyze the properties of SMS spam containing Korean characters and share the analysis results with society to develop spam-blocking solutions. In Sect. 2, we explain the types of spam data used in this paper and the analysis method employed. In Sect. 3, we present analysis results of the spam data. Finally, in Sect. 4, we present the conclusion and future work.

2 Dataset and Analysis

2.1 SMS Spam Dataset

The spam data used in this study was collected from October 2015 to April 2016 and includes 70,600 messages. Moreover, the spam data consists of the received timestamp, transmitting mobile number, and advertisement contents. Each item is explained in Table 1.

Table 1. Properties of SMS spam data

Property	Description
Receive timestamp	Time when a receiver received a spam
Transmit mobile number	Mobile number of a person who sent a spam (spammer)
Advertisement contents	Contents of a spam

Among these properties, the transmitting mobile number is personal information; therefore, the mobile number was encrypted using a hash function. As a result, the spammers were distinguished by the encrypted values, instead of their actual numbers.

2.2 Analysis Method

Combination of the Properties. In the spam data, some information, such as a received timestamp, transmitting mobile number, and advertisement contents, is available as mentioned previously. Using this information, we can combine one property with another to understand the correlations between them and use the correlations to understand any other properties that might appear in the spams. For instance, we can perform an analysis to find which spammers send spams frequently or at what time the rate of spams sent is the highest.

Spam Content Analysis. As one of the properties of the spam data, the advertisement contents are in text format, which makes them difficult to be analyzed. Therefore, the data should be refined to a format that can be easily analyzed [4]. Text mining is one means to accomplish this goal. This technique first divides each sentence of the content into morphemes to extract meaningful words and uses these words for analysis. There are several methods that use words to analyze the contents: a method to calculate the frequency of each word, a method to find correlations between words, and a method to cluster data from words [5].

In this study, to extract words from the advertisement contents in text format, KoNLP, an analysis package of the R language and the Korean morpheme, was used. To prevent any meaningless words, such as special characters, from being selected as meaningful words during word extraction, we performed preprocessing and postprocessing. Furthermore, after words were extracted, we performed an analysis based on the frequency of each word.

3 Property Analysis of SMS Spam

In this section, the properties of the SMS spam are analyzed using the data and analysis method described in the previous section. To do this, three types of analyses are defined and described in Table 2.

Table 2. Types of SMS spam analyses

Type	Description
1	Number of transmitted spams
2	Main features of the contents of the spam
3	Main features of the spammer

3.1 Number of Transmitted Spam Messages

Because the collected spam data means that the spam is sent, we can find the total number of transmitted spam messages by counting the number of messages in the spam dataset. Moreover, we can find the features of the spam by counting the number of spam messages that meet a certain property.

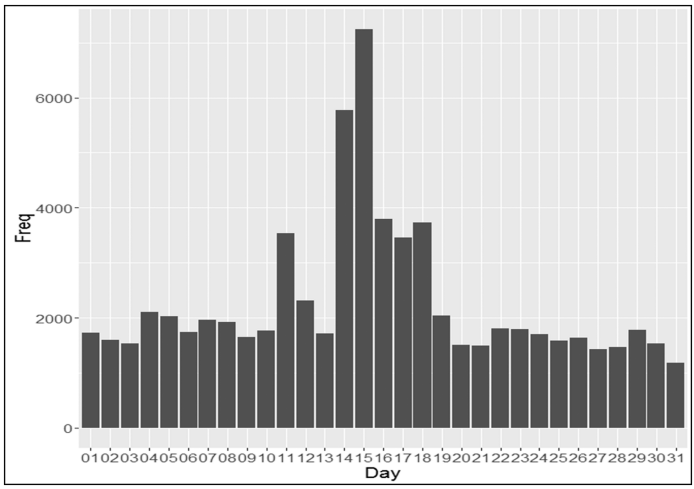


Fig. 1. Number of daily transmitted spam messages

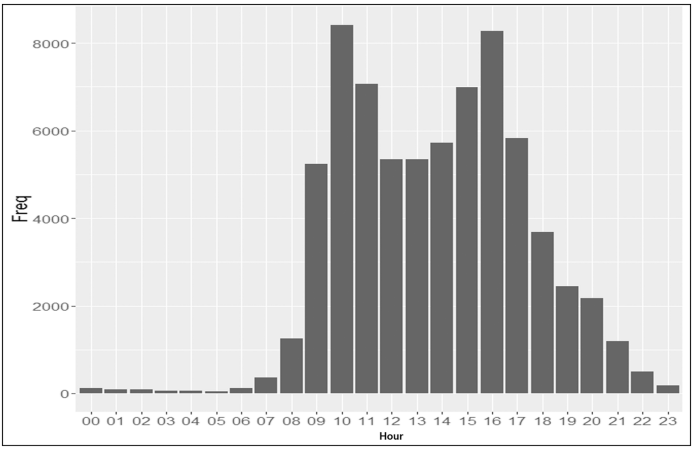


Fig. 2. Number of hourly transmitted spam messages

Because the received timestamp data includes temporal information such as the year, month, day, and hour, we can classify the number of transmitted spam messages according to the received time. Figures 1 and 2 shows the number of transmitted spams at different dates and times. Due to the limitation of the duration of data collection, the yearly, quarterly, and monthly data were omitted for comparison. Therefore, we only performed the analysis with daily and hourly graphs, which were not omitted during data collection.

From the graph of the daily transmitted spam in Fig. 1, we can note that most days in a month showed a similar number of spams, with a sudden increase on the 14th and 15th. Furthermore, from the graph of the hourly transmitted spam in Fig. 2, we can note that spam was rarely sent in the early morning and the number of spam messages rapidly increase later in the morning and showed highest peaks at 10 am and 4 pm.

3.2 Features of Spam Content

The original text content in spam cannot be used for analysis. Therefore, to enable an analysis, the contents should be refined and converted to a different format. Using the techniques described in previous sections, we can refine the spam contents and measure the frequency of word data to find the main message in the contents. Also, shipped spam in specific date includes the issue and status of that time.

When the contents are analyzed based on the frequency of words, a short-term analysis of the spam might not be a meaningful analysis; thus, we conducted the analysis on a quarterly basis. Table 3 shows the frequencies of words in descending order. For convenience, only three words are shown.

Table 3. Word frequency on a quarterly basis

Quarter	Word	Frequency
1	Free	4907
	Candidate	3054
	Congressmen	2234
2	Information	2091
	Election	1710
	Candidate number	1426
3	•	•
	•	•
	•	•
4	Impersonation	14,791
	Cash	14,983
	Warning	15003

In Table 3, the frequencies of the words related to the election such as candidate, congressmen, election, candidate number, and information, were very high in the first and second quarters. In South Korea, the 20th national election was held on April 13, 2016 and the first and second quarters were the periods of the election campaigns. In the case of the fourth quarter, three words showed very similar frequencies and this was mainly because these three words were used together in a sentence. The analysis was not performed for the third quarter due to a lack of data.

3.3 Features of Spammers

By analyzing the contents of the spam sent by each spammer, we can determine the feature of the spammer. More specifically, we can determine if the spam is a smishing or not. Even if a spammer sends numerous spams, it does not necessarily mean that they are smishings, because the number of the smishings sent by the spammer might be very small. On the other hand, even if another spammer sends a small number of spams, that spammer might have sent a relatively larger number of the smishings than the former spammer. Therefore, we should be able to determine the number of smishings sent by each spammer to find malicious spammers.

To check if a spam is a smishing, we should search for any keywords connecting to a certain URL in the spam contents. If a keyword is found, then the spam should be classified as a smishing. Figure 3 shows the number of transmitted smishings by each spammer. For convenience, we only presented the top 20 spammers. As shown in Fig. 3, the encrypted mobile number 484670 sent the largest amount of smishings. Furthermore, because the number of smishings sent by this spammer is much higher than other spammers, we can guess that most spam sent by this spammer would be smishings.

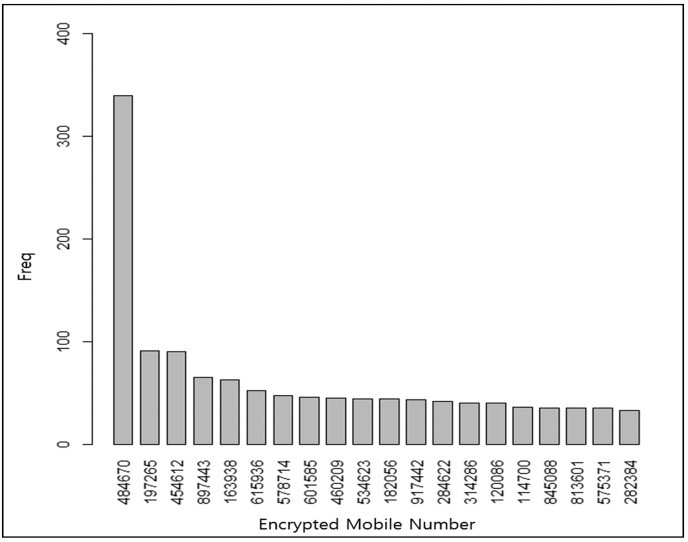


Fig. 3. Number of transmitted smishings by each spammer

4 Conclusion

In this study, we tried to find meaningful features from SMS spam data containing Korean characters by using combinations of SMS spam properties and by standardizing the text data in the spam. As a result, we could make some observations. First, more SMS spams were sent in a certain time period than the other times during the day. Second, by analyzing words in the spam, the frequency of each word could be calculated to find the most frequently used words and these words could be used to shed light on an issue or event occurring around the time the spam was sent. Third, we could distinguish between the spammers who send smishings and those who do not. We think these features could be used to develop a policy and effective technical solutions to block spam.

With regard to future work, we will try to transform the properties of spam to that of standardized data. Furthermore, we will also try to use standardized data to increase the reliability of analysis performed to distinguish between spam and non-spam and classify the spam.

Acknowledgements. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the University Information Technology Research Center support program (IITP-2016-R2718-16-0003) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Ministry of Science, ICT and Future Planning: The current status of wireless communication service users as of May 2016, 30 May 2016. http://www.msip.go.kr/
2. Park, I.W., Park, D.W.: A study on the analysis and security measures for smishing hacking attacks. J. Next Gen. Inf. Technol. **5**(2), 85–91 (2014)
3. KISA: Status of the spam distribution in the second half of 2015, March 2016. http://www.spamcop.or.kr/
4. Sung, H.H.: Text Mining Techniques for Korean Web Documents Classification and Clustering, vol. 12, p. 10 (2015)
5. Lee J.A.: Study on developing a patent citation prediction model using data mining techniques, vol. 12 (2011)

An Automatic Patch Management System with Improved Security

JunHee Kim, MinSeok Sohn, and YooJae Won(✉)

Computer Science Engineering, Chungnam National University,
Daejeon, Republic of Korea
{jhkim89,yjwon}@cnu.ac.kr, mss007daum@hanmail.net

Abstract. As the number of patches in a patch management system increases due to software updates and security issues arise in the existing patch management system, a more efficient patch management system with reinforced security is required. Additionally, existing patch management systems must be improved, as they perform patch collection inefficiently and their patch integrity verification schemes are simple. In this paper, we propose an automatic patch management system with improved security, enhanced patch collection efficiency, and reinforced verification of patch integrity that automatically collects patches through patch sites.

Keywords: Patch management system · Patch collection · Integrity verification

1 Introduction

In general, cyber-attacks can be divided into zero-day attacks and those that use a known vulnerability based on a patch announcement. While it is impossible to prepare for zero-day attacks, attacks that use a known vulnerability can be sufficiently prepared for using patches. To prevent damage caused by such attacks, many security companies have developed and are operating patch management systems.

However, existing patch management systems face problems such as inefficient patch collection and insecure verification of patch integrity. Current patch management systems expend significant computation and time for patch collection due to their inefficiency and can fail to resolve vulnerabilities due to the insecurity of their patch integrity verification.

In this paper, we propose an automatic patch management system that improves upon current systems, reducing inefficiency in patch collection and instability in patch integrity verification. The proposed system includes a plan for automatic patch collection that monitors new patches through vendor websites and automatically downloads new patches, improving patch collection efficiency. In addition, patches are applied in a virtual environment, resolving problems related to insecurity and verifying patch integrity based on the information generated by the clients.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_13

2 Existing Patch Management Systems

2.1 Structure and Components of Existing Patch Management Systems

The structure of current patch management systems as detailed in the literature is illustrated in Fig. 1 and their components and functions are described in Table 1.

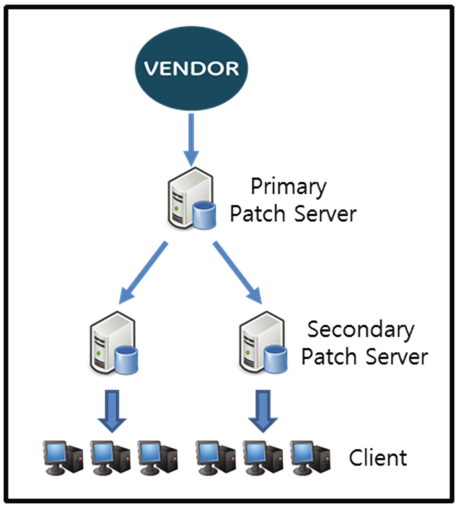


Fig. 1. Structure of patch management systems.

Table 1. Components of patch management systems.

Component	Detailed function
Information collection	Collection of patches and information; cataloguing of patches
Verification	Verification of patch stability and integrity
Scheduling	Management of policies and reservations
Certification	Certification of users and system
Distribution	Encryption and group distribution
Installation	Footprint, report, and roll-back

The existing patch management system enables efficient patch distribution by utilizing the hierarchical server structure [1, 2].

The components of existing patch management systems include information collection, verification, scheduling, certification, distribution, and installation. In information collection phase, patch files and information are collected from vendors and catalogued; in the verification phase, the integrity and safety of the collected patches are examined; in the scheduling phase, patch distribution policies are managed; during certification, certification between the patch servers and clients is carried out; in distribution, the patches are distributed in groups to the clients through encrypted

communication based on established policy; lastly, in the installation phase, appropriate actions are taken for specific situations that arise during or after patch installation—based on information about the system to be patched—and a patch installation report is transmitted to the patch server [3, 4].

Such patch management systems have two significant limitations, which are detailed in Sect. 2.2: patch collection is carried out inefficiently and patch integrity verification methods are insecure.

2.2 Limitations of Existing Patch Management System

Inefficiency in Patch Collection. Existing patch management systems inefficiently collect patches from vendors, accessing patch files by extracting all of the links from a vendor's website. Patch collection accuracy and performance may deteriorate as inadequate links are frequently accessed [5]. Advancements in information technology and the software market have led to steady growth in industry reliance on IT, while also diversifying users' requirements. For this reason, the number of software products that require patch management is increasing, rendering the existing inefficient patch collection method increasingly unsuitable for the management of numerous patches and immediate reaction to the announcements of new patches. Furthermore, it increases the computational and time costs expended for patch management. Accordingly, a method of efficiently collecting patches from diverse vendors is required.

Patch Integrity Verification Insecurity. Existing patch management systems verify patch integrity when the administrator has collected a patch from a vendor and it is distributed to clients. Methods of integrity verification include cyclical redundancy checks (CRCs) and patch file hash value comparison [2, 6]. However, these methods cannot properly verify patch integrity when a hacker has forged or falsified the patch file and information from the patch server. This problem can expose clients to diverse security threats and has already had severe real-world consequences. In 2013, Korea's financial management network was paralyzed due to insecure patch integrity verification and code signing certificates were seized through similar means in 2016. To improve the security of patch management systems, the patch integrity verification process should be reinforced.

3 Automatic Patch Management and Distribution System with Improved Security

3.1 Structure of the Automatic Patch Management and Distribution System

The structure of the proposed automatic patch management system, and its security improvements, is shown in Fig. 2. Most of the structure is the same as that of existing patch management systems; the main difference is that there is a Patch Collection and Test Server between the vendor and the primary patch distribution server. This server

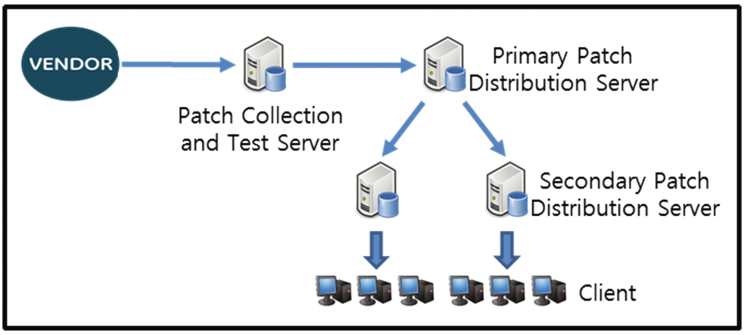


Fig. 2. Structure of the automatic patch management system.

has two primary roles: First, patch collection, which has been previously been inefficient, is automatically carried out. Second, patch integrity information is generated by applying the patches in a virtual environment. This information is used in the process of verifying the patch integrity when applying the patch in clients.

3.2 Components of the Automatic Patch Management System with Improved Security

The components of the automatic patch management system with improved security proposed in this paper are outlined in Table 2. Of these components, scheduling, certification, and distribution are the same as those of existing patch management systems. The information collection, test, and installation and verification processes have been modified or added, in contrast with existing patch collection systems.

Table 2. Components of the automatic patch management system with improved security.

Component	Detailed function
Information collection	Automatic collection and cataloguing of patches and information
Test	Generation of patch integrity information and verification of patch security
Scheduling	Management of policies and reservations
Certification	Certification of users and system
Distribution	Encryption and group distribution
Installation and verification	Patch application, patch integrity verification, and report

Information Collection. Figure 3 shows the process of information collection. The information collection phase is carried out by the Patch Collection and Test Server, which automatically collects patches and patch information from websites using a Web crawler. Most vendors operate an official website and announce new patches through

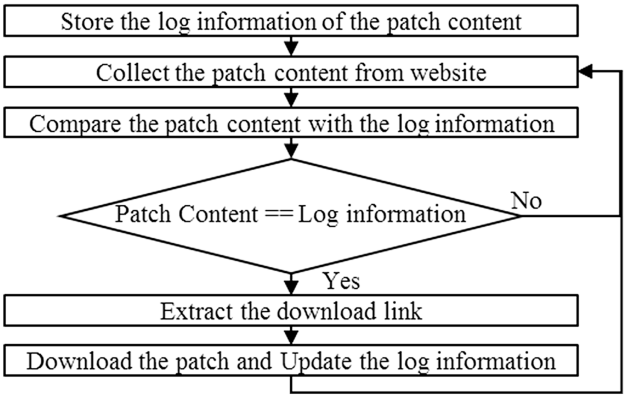


Fig. 3. Information collection process

specific content on that website before providing a patch. The Patch Collection and Test Server stores and manages the log information of this content, with which new patches can be identified, and periodically monitors vendor websites. At this time, patches and patch information are automatically collected by recognizing the new patch announcement based on this information and extracting the download link for the relevant patch.

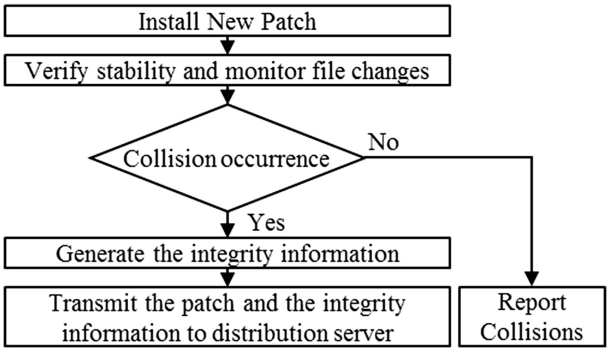


Fig. 4. Test process

Test. Figure 4 shows the process of Test. Testing is the phase in which patch stability is verified and integrity information about the patch is generated to improve the security of the proposed system. The test phase is carried out on the Patch Collection and Test Server. Diverse virtual environments are built in the Patch Collection and Test Server, tailored to the clients it manages. In the test phase, the patches collected in the information collection phase are applied in an appropriate virtual environment, stability verification is carried out, and patch integrity information is generated. Stability verification is the process of checking whether any collision occurs between the patch and

the relevant system. Integrity information is stored as the changes in the relevant system before and after applying the patch. The change details to be stored include the name, path name, size, version, and hash value of the file. The integrity information thus generated is distributed to clients together with the patch at the distribution phase and utilized for verification of patch integrity.

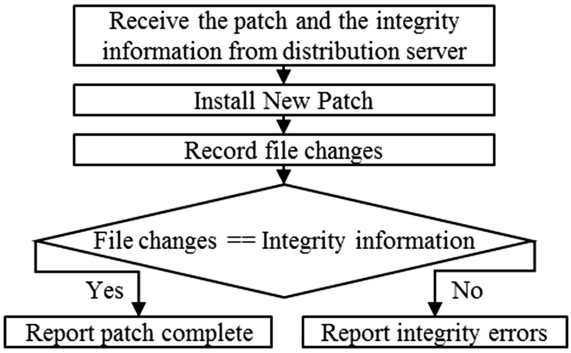


Fig. 5. Installation and verification process

Installation and Verification. Figure 5 shows the process of installation and verification. The final phase installs and verifies the patch and transmits the result to the Patch Distribution Server. Clients receive a patch and integrity information from the Patch Distribution Server when a new patch is announced. The clients apply the patch that was received and record the changes made to the system. Clients then determine whether the patch has been properly applied by comparing the changes and integrity information and transmits the result to the Patch Distribution Server.

4 Conclusion

In this paper, the limitations of existing patch management systems were analyzed and, to resolve these problems, an automatic patch management system with improved security was proposed. The existing method of inefficiently collecting patches was improved through a method that automatically collects patches using a Web crawler to improve efficiency. In addition, to improve the security of the patch management system, patch integrity is guaranteed by utilizing file changes resulting from patch application.

Software patches from more vendors can be managed by utilizing the proposed system, which is efficient, improves security, and can guarantee patch integrity.

Future work on this system will resolve the practical problems found during the implementation and testing of the patch management system proposed in this paper.

Acknowledgments. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the University Information Technology Research Center support program (IITP-2016-R2718-16-0003) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Kim, Y.J., Lee, S.W., Son, T.S., Moon, J.S., Suh, J.T., Yoon, J.B., Park, E.K.: Design of a hierarchical patch distribution system framework considering scalability. In: Spring Symposium of Korean Institute of Information Scientists and Engineers (2004)
2. Suh, J.T., Yoon, J.B., Choi, D.S., Park, E.K., Park, C.S.: Design and implementation of a patch management system for improvement in system security. Convergence Secur. J. Korea Convergence Secur. Assoc. **4**(2), 43–51 (2004)
3. Lee, I.Y., Lee, S.Y., Moon, J.S., Lim, J.I.: A study on efficient components of patch management system. In: Winter Symposium of the Korean Institute of Broadcast and Media Engineers (2008)
4. Son, T.S., Kim, J.W., Park, I.K., Moon, J.S., Suh, J.T., Lim, E.K., Lee, C.W.: Design of a Safe Patch Distribution Structure, vol. 29, no. 2, pp. 559–561. Collection of Autumn Academic Dissertations of Korean Institute of Information Scientists and Engineers (2002)
5. Min, D.W., Son, T.S., Suh, J.T., Koo, W.B., Jang, J.A., Moon, J.S.: A Plan to Automatically Organize a DB for Automatic Distribution of Security Patches, vol. 31, no. 1(A), pp. 367–369. Collection of Academic Dissertations of Korean Institute of Information Scientists and Engineers (2004)
6. Lee, S.W., Kim, Y.J., Son, T.S., Moon, J.S., Suh, J.T., Lee E.Y., Lee, D.H.: Design of a Framework for Generalized Security Patch Distribution and Management System, vol. 31, no. 2, pp. 502–504. Collection of Autumn Academic Dissertations of Korean Institute of Information Scientists and Engineers (2004)

An Empirical Study on the Relationship Between User Interface Design Attributes in Smartphone Applications and Intention to Use

Wonjin Jung[1,2] and HyungRok Yim[1,2(✉)]

[1] The School of Business and Economics, Dankook University, 152, Jook-Jun-Ro, Soo-Ji-Goo, Young-In, Kyung-Ki-Do 16890, Korea
jintari@hanmail.net
[2] The School of Business, Hanyang University, 222, Wangsimli-Ro, Sung-Dong-Goo, Seoul 04763, Korea
hryim@hanyanag.ac.kr

Abstract. Smartphones, which debuted in the form of personal digital assistants (PDAs) in the late 1990s, have evolved continuously. However, hardware features and the external environment restrict their use, making it difficult to ensure high interactivity. The structure and usage of mobile applications are also becoming increasingly complex and it is often found to be difficult to understand the user interface (UI). These user environments and conditions inhibit the smooth interaction of the user with the application. This is expected to negatively affect the user's intention to use the applications eventually. However, past studies on information systems have not shown much interest in the impact of smartphone UI designs on the attitudes and behaviors of users. Thus, this study attempted to empirically explore the impact that UI application designs have on the behavior intentions of the users utilizing the application. This study specifically looked at the following aspects of UI design: simplicity and consistency. The data was collected through a survey and structural equation modeling (SEM) was employed for the analysis. The results showed that these attributes have a significant effect on the interaction as well as a positive impact on the intention to use the application.

Keywords: Smartphone · Application · Interface · Design · Intention

1 Introduction

UI is a part of a system through which users interact with the system [1]. Moran [3] argued that a UI consists of diverse physical and conceptual design components to support the smooth interaction of the user with the system. UI designs vary in a variety of ways depending on the placement and use of the components; different UI designs affect the interactions with the system in different ways.

Several researchers noted that UI design of a system contains the attributes of simplicity and consistency. According to the associated studies, it has been identified that simplicity has a significant impact on the usability, usefulness, and aesthetics of a

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_14

system including interactions with the system [4]. It has been reported that consistency also has a significant impact on interactions with the system [2]. These prior studies have been conducted on PC-based information systems and webpages.

Any study about smartphone UI designs is bound to emphasize the simplicity of the design. Smartphone applications are significantly affected by the physical and environmental conditions of the smartphones because the applications run on smartphones. As the small screens of smartphones cannot contain much information, UI of smartphone applications should restrict the use of the information, content, and other various design components as much as possible and realize minimalism to maximize user convenience. When too much information and unnecessary design components occupy the narrow space of the application's UI, readability will be reduced and interactivity is bound to decrease. In addition, it would be difficult to accurately touch the design components with one's fingers. Therefore, UI designs of smartphone applications require simplicity.

However, in reality, it is difficult to predict the impact of this simplicity in the UI of the smartphone application on the interaction of users. This is because the type and usage of smartphone applications are indeed diverse and user environments of smartphone applications differ from those running on PC-based information system and webpages. It is necessary to examine the effects of the simplicity of smartphone application UIs on interactions, which have not been specifically reported in this context. Thus, the study will attempt to identify these impacts through the hypothesis below.

Hypothesis 1: The simplicity in the UI of smartphone applications positively affects the interactions of users with the applications.

In addition, many researchers in e-learning fields have stated that consistency in the UIs of a system affects learning. In other words, many researchers have clarified that the consistency in the UI design of a system enables users to learn easily through the system and, furthermore, to improve learning and academic performance [5]. Other researchers have claimed that education software needs to be equipped with UI design consistency to ensure the interactivity and success of the system. To support the argument, researchers provided the explanation that the consistency in the UI design enables users to formulate the knowledge of and experience in the use of the software quickly and accurately [6]. Based on these prior studies, it can be predicted that ensuring consistency in the UI design of the smartphone applications will also significantly affect the users' interaction with the smartphone applications; the study will attempt to verify this through the hypothesis below.

Hypothesis 2: The consistency in the UI of smartphone applications positively affects the interactions of users with the applications.

The Technology Acceptance Model (TAM) was proposed on the basis of TRA in IS fields; the model has been widely used in understanding the individual's attitudes associated with the acceptance of information technologies [8]. In other words, TAM explains that the perceived usefulness and ease of use formed on the basis of the experience in an information technology determine the individual's attitudes towards the information technology; and the individual's attitudes affect the individual's behavioral intentions, that is, the acceptance of the information technology.

Looking at studies on the personal beliefs, attitudes and behavioral intentions of users in mobile fields, the impacts of the perceived usefulness and ease of use on the behavioral intentions to use mobile data services were identified. Similarly, the impacts of the perceived usefulness and ease of use of a mobile payment system on the intentions for using the system were also identified. Finally, the smartphone's usefulness and ease of use perceived by doctors and nurses affect the intentions to use the smartphones was explained.

Users who interact with the smartphone applications form attitudes towards the applications on the basis of the experiences and these attitudes are predicted to affect the behavioral intentions of the users based on the TRA and TAM [7]. However, various applications are installed and utilized on the smartphones and these various applications require unique interactions. There have not been many reports as to the impacts of the various forms of the interactions on the user's attitudes toward the applications and intentions for using the applications. Therefore, this study will attempt to identify the impacts through the following hypotheses:

Hypothesis 3: The interactions with smartphone applications positively affect the users' intentions to use the applications.

2 Research Methodology, Data Analysis, and Results

This study explored the impacts that simplicity, consistency and metaphors in UI design have on the attitudes and behavioral intentions of the users. The study achieved this by studying the interactions between the user and the application in the smartphone environment. In this study, these interactions were considered a mediating variable. A survey was conducted to collect data and a total of 258 university students and practitioners volunteered to participate in the survey. One hundred and seventy two students (66.7% of the participants) answered the questions. The participants majored in economics, business administration, computer science, and graphic design at three universities in Korea. The gender ratio of the participants was 53.9% male to 46.1% female. The majority of the participants (78.7%) were in their twenties. The application type that the participants used most frequently just before the survey was social networking and communications-related applications (61.7%).

This study employed Structural Equation Modeling (SEM) to analyze the proposed research model and used SPSS statistics and AMOS ver. 18 as the statistical software. The path coefficients were examined to check the causal effects between the variables. To do so, the significance of the relationships between the variables in the proposed model was analyzed. As predicted in the hypotheses, simplicity had a significant influence on the interaction ($\beta = .396$, $p < .0001$), consistency also had a significant influence on the interaction ($\beta = .331$, $p < .0001$). In addition, the interaction had a positive impact on users' intentions to use the application ($\beta = 412$, $p < .0001$). Thus, all of the hypotheses were supported. Table 1 shows the results of the test of the structural model.

Table 1. Hypothesis test

	Paths	Coeff.	Stand. coeff.	P	Results
H1	Simplicity → Interaction	.259	.396	.0001	Accept
H2	Consistency → Interaction	.264	.331	.0001	Accept
H3	Interaction → Intention to Use	.337	.412	.0001	Accept

3 Discussion and Conclusion

In short summary, this study explored the attributes of UI design for smartphone applications; these attributes include simplicity and consistency. Not only did the study explore these attributes but it also empirically examined the effects that the UI attributes had on users' intention to use the applications through the mediating variables of interaction. The results showed significant direct effects of the UI attributes on the interaction between users and applications. In addition, the interaction also had a positive impact on the users' intentions to use the applications. The results of this study addressed the importance of the UI attributes for smartphone applications in understanding and explaining the factors that influence the interaction with applications and the intention to use them. In particular, this study emphasized how important it is for smartphone applications to apply recognizable simplicity and consistency in UI design. Practitioners can use the findings to improve, not only the interaction between users and applications, but also the intention to use the applications when developing smartphone applications.

References

1. Bostrom, R.P., Olfman, L., Sein, M.K.: The importance of learning style in end-user training. MIS Q. **3**(1), 101–119 (1990)
2. Chu, C., Chan, B.K.: Evolution of web site design: implications for medical education on the internet. Comput. Biol. Med. **28**, 470–472 (1998)
3. Moran, T.: An applied psychology of the use. ACM Comput. Surv. **13**, 1–12 (1981)
4. Moshagen, M., Thielsch, M.T.: Facets of visual aesthetics. Int. J. Hum.-Comput. Interact. **68**, 689–709 (2010)
5. Lin, H.X., Choong, Y., Salvendy, G.: A proposed index of usability: a method for comparing the relative usability of different software systems. Behav. Inf. Technol. **16**(4), 267–278 (1997)
6. Polson, P.: The consequences of consistent and inconsistent user interfaces. In: Guindon, R. (ed.) Cognitive Science and its Applications for Human-Computer Interaction, pp. 59–108. Lawrence Erlbaum, Hillsdale (1988)
7. Schmidt, K.E., Liu, Y.L., Sridharan, S.: Webpage aesthetics, performance, and usability: design variables and their effects. Ergonomics **52**, 641–643 (2009)
8. Szajna, B.: Empirical evaluation of the revised technology acceptance model. Manag. Sci. **42**(1), 85–92 (1996)

Thumb Biometric Using Scale Invariant Feature Transform

Naeun Lim, Daejune Ko, Kun Ha Suh, and Eui Chul Lee(✉)

Department of Computer Science, Sangmyung University, Hongjimun 2-Gil 20, Jongno-Gu, Seoul 03016, Republic of Korea
eclee@smu.ac.kr

Abstract. Recently, biometrics technology has been receiving attention as means of personal authentication in smartphone environment. Fingerprint recognition is generally contained in newest smartphones and other biometric methods such as iris recognition are receiving attention. However, these methods have a problem of being not applicable to existing smartphones because additional devices such as infrared cameras or sensors should be included. To solve this problem, in the present paper, a new biometric method using features on the rear of the thumb is proposed. The similarity between enrolled thumb images and input thumb images is measured through the SIFT (Scale Invariant Feature Transform) method. Through feasibility tests, it could be identified that the proposed method could recognize the thumb with an accuracy level of approximately 99.94%.

Keywords: Biometric · Hand biometric · Thumb recognition · SIFT

1 Introduction

Biometrics is a method of verifying user identity through pattern recognition techniques using unique features in human body data obtained through sensors, which has advantages over existing user authentication methods such as no problem of forgetting or losing and lower risks of fake or forgery [1]. The kinds of biometrics include fingerprint recognition, face recognition, iris recognition, vein recognition, and voice recognition [2–4]. Recently, as Fintech, which is a compound word of financial and technology, has been in the limelight, the number of those that use simple payment systems through personal authentication using fingerprints on mobile devices that contain fingerprint recognition sensors has been increasing [5].

However, such simple payment systems have problems as fingerprint recognition cannot be used on smartphones that do not contain a fingerprint recognition sensor and recognition errors frequently occur due to moisture or injuries on the surface of the thumb. Although face recognition methods can be considered because all smartphones contain a camera, recognition rates can be reduced drastically due to deviations in face information resulting from glasses wearing or differences in hair color, lighting, or facial expressions [6]. Although iris recognition has an advantage of high accuracy, it has a problem of requiring infrared lighting and expensive camera equipment with zoom control [4, 7].

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_15

Therefore, to solve abovementioned problems in biometric systems, in the present paper, a new method is proposed that can obtain thumb rear images through smartphone cameras without any additional separate sensor and conduct personal authentication through image processing.

2 Proposed Method

The proposed method is operated through feature matching using the processing and shift of thumb rear images taken through general visible light cameras. Figure 1 is a flow diagram of the proposed method.

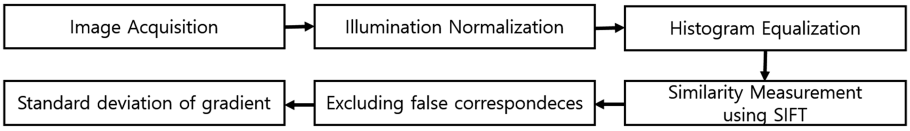


Fig. 1. Flow diagram of proposed method

2.1 Image Acquisition

Actual environments where smartphones are used and quite diverse also the distance between the camera and the thumb cannot be adjusted to be constant. Therefore, the background, thumb size, and thumb position in acquired images cannot but vary all the time. Red solid line in Fig. 2 shows to solve this problems, a guideline for rough positioning of the thumb is provided on the camera preview screen.

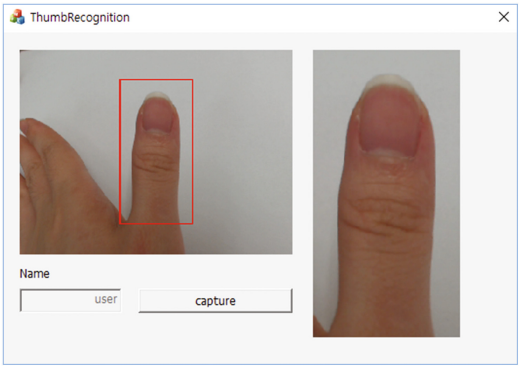


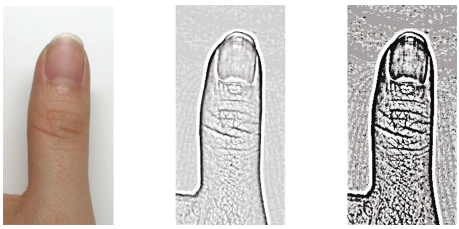
Fig. 2. Examples of captured thumb image and guideline for aligning thumb (red solid line)

2.2 Illumination Normalization

Since acquired thumb images have uneven surfaces and the surrounding lighting environment varies all the time, the components of illumination should be adjusted to be uniform through an illumination normalization process. A Retinex algorithm based illumination normalization method has been used in the existing face recognition and thumb vein recognition areas too [4, 8]. That method is applied to the thumb area to conduct thumb image illumination normalization. In the present study, considering the resolution of images, 5×5 was applied as an average filter size for detection of illumination components.

2.3 Histogram Equalization

The thumb images after illumination normalization generally have low visibility because most pixels appear around 128 based on the 8 bits gray scale with low contrast. To detect robust features through SIFT, images with high contrast and clear edges should be used. To this end, a process to apply histogram equalization to illumination-normalized images is undergone to enhance the images so that features such as wrinkles in the thumb are clearly visible. An example of enhanced images is as shown in Fig. 3.

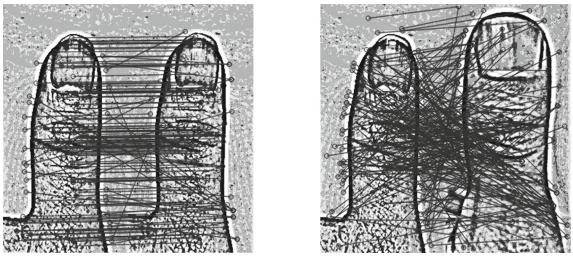


(a) (b) (c)

Fig. 3. Example of thumb image enhancement using histogram equlization. (a) Original captured image. (b) Illumination normalized image. (c) Enhanced image using histogram equlization.

2.4 Similarity Measurement Using SIFT

SIFT (Scale Invariant Feature Transform) method is used to measure the similarity between enrolled images and input images [9]. The SIFT algorithm is a method of detecting corresponding pairs between two images at high speeds regardless of affine geometric variations. Using the SIFT is quite appropriate because geometric variations such as translation, scaling, and rotation exist in thumbs in acquired images. The results of obtaining SIFT correspondences in cases of genuine and imposter are shown in Fig. 4.

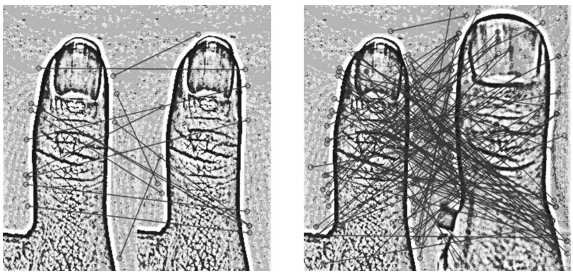


(a) (b)

Fig. 4. Results of obtaining SIFT corresponding pairs between two thumb images. (a) Genuine matching. (b) Imposter matching.

2.5 Excluding False Correspondences

Since images were obtained according to the guideline, images with the results of matching using the SIFT that go out of a certain range, that is, 1/3 of the horizontal size of the image and 1/4 of the vertical size of the image are excluded. Figure 5 shows matching results that are excluded.

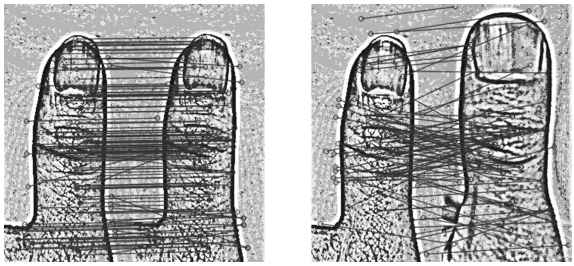


(a) (b)

Fig. 5. Examples of excluded false correspondeces. (a) Genuine matching. (b) Imposter matching.

2.6 Standard Deviation of Gradient

The result after excluding matching results that went out of the range is as shown in Fig. 6. It can be seen that when two images have been obtained from the same person, the gradients in the matching result are almost constant. Therefore, whether two images being matched have been obtained from the same person or from different persons can be identified using the standard deviation of the gradients in the matching result.



(a) (b)

Fig. 6. Result of SIFT correspondences by excluding false ones. (a) Genuine matching. (b) Imposter matching.

3 Experimental Results

To measure the accuracy of recognition, the standard deviation of gradients in matching between two thumb images in case of genuine or imposter were obtained. In the experiment, 100 photos of thumbs (thumbs of 10 persons, 10 photos per person) with specifications as shown in Table 1 were used.

Table 1. Database specification of thumb image

Type	Format	Resolution	File size	Number of images
Thumb image	BMP, RGB24	172 × 340	175,494 Bytes	100

Probability density functions for the measured standard deviation values can be obtained to figure out FAR (False Accept Rate) and FRR (False Reject Rate). Figure 7 shows the probability density functions for the results of thumb matching as a graph. It can be seen that the recognition can be successful at an accuracy level of 99.94% with a FAR of 0.02% and a FRR of 0.1%, respectively.

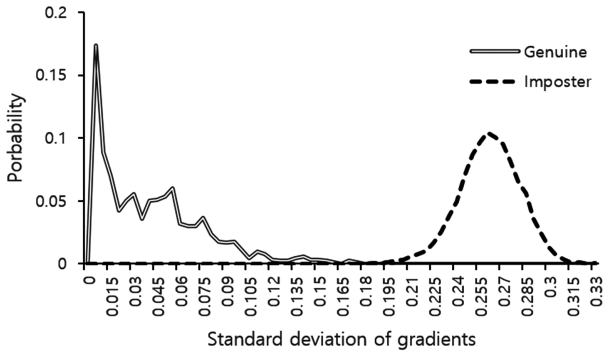


Fig. 7. Probability distributions of genuine and imposter matching

4 Conclusion

In the present paper, a thumb recognition method using SIFT was proposed. The proposed method can be applied to all smart devices containing a camera without any additional sensor and shows an accuracy level of 99.94%. In detail, FAR of the proposed method was 0% in case of 0.4% FRR. Therefore, our proposed method can be applied into the Fintech authentication purposes with high security level. Since our previous work using correlation based thumb recognition method has limitations such as low accuracy caused by geometric variation, in future works, we will enhance the matching performance of thumb recognition by fusing two scores such as the number of SIFT correspondences and correlation coefficients.

Acknowledgements. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (grant number NRF-2016R1C1B2014345). Also, this research was supported by the Ministry of Science, ICT and Future Planning (MSIP), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2016-H8501-16-1014) supervised by the Institute for Information & Communications Technology Promotion (IITP).

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. **14**, 20–44 (2004)
2. Jain, A.K., Nandakumar, K., Nagr, A.: Biometric template security. EURASIP J. Adv. Sig. Process. **2008**, 1–17 (2008)
3. Daugman, J.: How iris recognition works. IEEE Trans. Circuits Syst. Video Technol. **14**, 21–30 (2004)
4. Lee, E.C., Lee, H.C., Park, K.R.: Thumb vein recognition using minutia based alignment and local binary pattern based feature extraction. Int. J. Imaging Syst. Technol. **19**, 179–186 (2009)
5. Lee, S.H., Lee, D.W.: FinTech-conversions of finance industry based on ICT. Korea Convergence Soc. **6**, 97–102 (2015)
6. Huang, G.B., Ramesh, M., Berg, T., Learned-Miler, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical report 1, 07–49, University of Massachusetts, Amherst (2007)
7. Wang, L., Leedham, G., Cho, D.S.Y.: Minutiae feature analysis for infrared hand vein pattern biometrics. Pattern Recogn. **41**, 920–929 (2008)
8. Tan, X., Triggs, W.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process. **19**, 1635–1650 (2010)
9. Bicego, M., Lagorio, A., Grosoo, E., Tistarelli, M.: On the use of SIFT features for face authentication. 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006). IEEE (2006)

Image Classification Using Color and Spatial Frequency in Terms of Human Emotion

Min Woo Park[1], Daejune Ko[1], Hyeonsang Hwang[1], Jiyeon Moon[2], and Eui Chul Lee[1(✉)]

[1] Department of Computer Science, Sangmyung University, Seoul, Republic of Korea
nogood79dle@gmail.com, eclee@smu.ac.kr
[2] Department of Sports ICT Convergence, Sangmyung University, Seoul, Republic of Korea

Abstract. Image classification is helpful for searching and image retrieval in terms of corresponding to the preference of users. However previous works did not consider human emotion but perform the retrieval by using keywords or objects in image. In the field of color psychology, the color has been proven that an impact on the human emotion. Also, visual complexity such as spatial frequency affects to human emotion. In this paper, a new image classification method is proposed for analyzing the relationship between image components such as color and spatial frequency and human emotion. We collected totally 391 images which contained the three different kinds of scene categories such as natural scene, campus scene, and human made scene images from the public image database. Consequently, we confirmed that image can be reasonably classified by using the color and spatial frequency in terms of human emotion.

Keywords: Image classification · Color · Spatial frequency · Human emotion

1 Introduction

Since the numerous image contents have been widely distributed according to the explosive growth of the Web nowadays, image search and classification has become more popular and important technology as well as growing challenging issues. Image classification is helpful techniques in several ways as follows. Firstly, image classification will be useful method as a convenient user interface for searching the image database in terms of corresponding to the preference of users [1]. Moreover, image clustering can be used for improving the performance and speed of content-based image retrieval (CBIR) [2].

In previous works, many researches have performed for image classification by using color as well as texture characteristics which can be easily extracted from images by using the several clustering algorithms such as *k*-means, spectral clustering, minimum distances, and decision rules [3]. Color components or texture can be used for identifying regions of interest or objects in images.

However, because the many previous image classification methods have performed by using the only one low-level visual feature, researches were rarely presented by

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_16

using the two kinds of visual information. Also, many researches above mentioned were not considered in terms of human emotion, the clustering results cannot be satisfied for people. In other words, these applications were not appropriate for affective computing application. Therefore, standard should be existed for the image classification by considering the human emotion.

According to Russell emotion model, human emotion is defined as a two-dimensional emotion model for designing the various emotion such as "Pleasant – Unpleasant" axis (X) and "Aroused – Not Aroused" axis (Y) [4]. Based on Russell emotions model, the several studies have been researched for the relationship between human emotions and color by mapping the various emotional state and the color. Generally, the visualized information is existed for indicating the correlation between color and human emotion so called "Geneva Emotion Wheel" [5].

In the field of color psychology, color components were mainly used in marketing fields for towards an influence on the attitudes and decisions of consumers [6]. That is, color components can be influenced for human emotions. Also, as the spatial frequency is higher in images, the higher spatial frequency (boundary information) can be caused "arousal" emotional factors [7].

Based on the theories of relation between image components and human emotion, we proposed image classification method as 2-dimensional emotion model by using color and spatial frequency in terms of human emotion.

2 Proposed Method

The overall procedure of the proposed method presented in this paper is shown in Fig. 1.

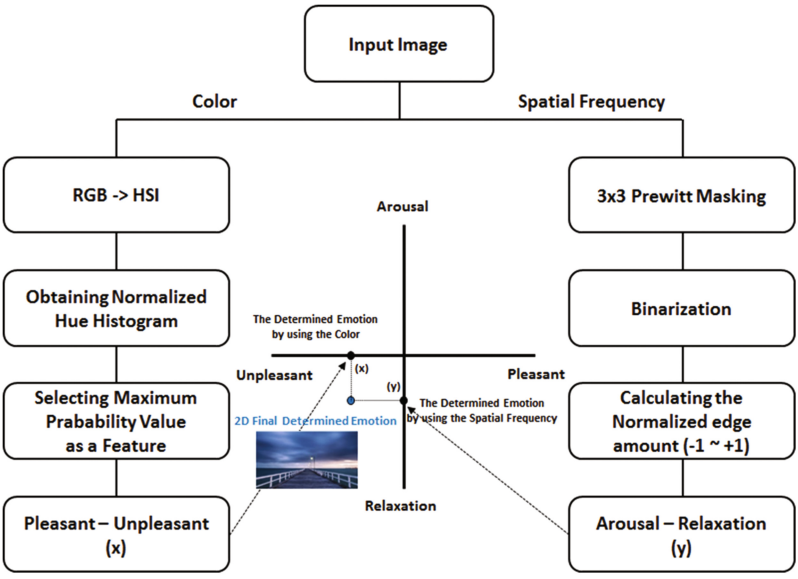


Fig. 1. The overall procedure of the proposed method.

2.1 Color

Pixel intensity of color image can be explained as Hue (H), Saturation (S), and Intensity (I) based on HSI color model [8]. In this paper, color conversion is performed as the first step for extracting color components from RGB to HSI of the original image [9]. The reason for converting a color model from RGB to HSI is as followed. Firstly, because the components imply both the meaning of color and brightness in RGB color model, only using color component is difficult to analysis [8]. Whereas, because the HSI color model can be represented as separately spectrum of color (H), saturation or purity of colors (S), and brightness or luminance value of the color (I), the limitation of such a RGB model can be solved [9]. Consequently, using the HSI model is suitable for analyzing the color effectively by converting RGB into hue. In here, human emotion is mapped into specific colors by dividing 2 dimensions such as "Pleasant – Unpleasant" (X-axis) and "Arousal – Relaxation" (Y-axis) [5]. Secondly, the process of obtaining normalized hue histogram is performed as the between −1 and 1 scales. In here, hue value is only used for image classification in colors. Generally, hue value is calculated from 0° to 360°. In this paper, the value of color (Val_c) is calculated as followed in Eq. (1). Because the mapping between various colors and two 2D emotion model includes the meaning of saturation and brightness factors, only hue value is used for mapping onto the X-axis. In Eq. (1), W and H is represented as width and height in images, respectively. Also, *histo* is represented as number of the hue value which existing in images. In here, because the angular offset between the hue of HSI model and the Geneva Emotion Wheel is about 80°, the parameter of cosine function is modified by adding 80° in order to compensate the offset. As a result, the calculated results was used as the element in charge of the X-axis as the value of range from −1 to 1 scale.

$$Valc = \frac{\sum_{hue=0}^{359} \cos\ (hue + 80)^{\circ} \times histo[hue]}{W \times H} \tag{1}$$

In this paper, we estimated 6 colors such as Red (R), Green (G), Blue (B), Cyan (C), Magenta (M), and Yellow (Y) based on color psychology model. After that, maximum frequency value can be selected as a feature in X-axis as "Pleasant – Unpleasant". Consequently, colors are mapped into the X-axis of unpleasant toward pleasure in order of B, M, C, R, G, and Y.

2.2 Spatial Frequency

Edge components were considered as the important factor in image processing because the boundary characteristics of an object as well as distinctive boundaries between different pixels or textures. Similar as edge component, spatial frequency was also represented as the characteristics of crisp changing of pixel intensity in space. In here, spatial frequency were analysed by converting from color image to gray image.

According to the different spatial frequency, the difference of information can be represented about the appearance of a visual stimulus. In proposed method, a fixed 3 × 3 Prewitt mask were used for estimating the spatial frequency. Because the results

of detected edge through a mask is dependent on resolution of the image, a fixed 3 × 3 mask is applied to various the image resolution. Then, resolution of the image which can be represented the best clear edge was determined such as shape of the objects. Consequently, because the 320 × 180 resolution of the image is the best matched, input images are used as reduce to the corresponding size as shown in Fig. 2.

Fig. 2. Results 3 × 3 Prewitt masking and binarization in cases of various spatial resolution of same image.

Next, to clearly separate edge and background regions, the edge image is binarized. After that, the value of edge (Val_E) is calculated as followed in Eq. (2). Then, the calculated result are represented as range from 0 to 1 scale. In this paper, because the performing normalized the spatial frequency into range from −1 to 1 scales, the values are re-scaled from 0–1 to −1–1 based on 0.4. In here, the clipping process is performed into 0.8 when the value is more than 0.8.

$$ValE = \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} Count(x,y)[Edge(x,y) \geq 0.4]}{W \times H} \tag{2}$$

Finally, the value of Y-axis can be determined by using the obtaining value of the spatial frequency as "Arousal – Relaxation".

3 Experimental Results

In this step, we collected totally 391 images which contained the 3 different kinds of scene categories such as natural scene, campus scene, and human made scene images from Natural Image Database, Natural Scene Statistics in Vision Science, Center for Perceptual Systems, University of Texas at Austin [10]. Among the 3 different kinds of scene categories, the images includes the following contents. Firstly, dataset #1 of Natural Scene Collection were contained the 1,204 images that the nature sense containing no man-made objects or people from Austin, Texas, USA and surrounding areas. In here, totally 1,204 images were divided into 9 of sets, respectively, we selected the only "set 9" database. Secondly, dataset #2 of Campus Scene Collection were contained the 90 images that the university campus scenes containing cars, buildings, and people. Lastly, dataset #3 of Human Made Scene Collection were contained the 197 images that the indoor and outdoor scenes containing human made objects.

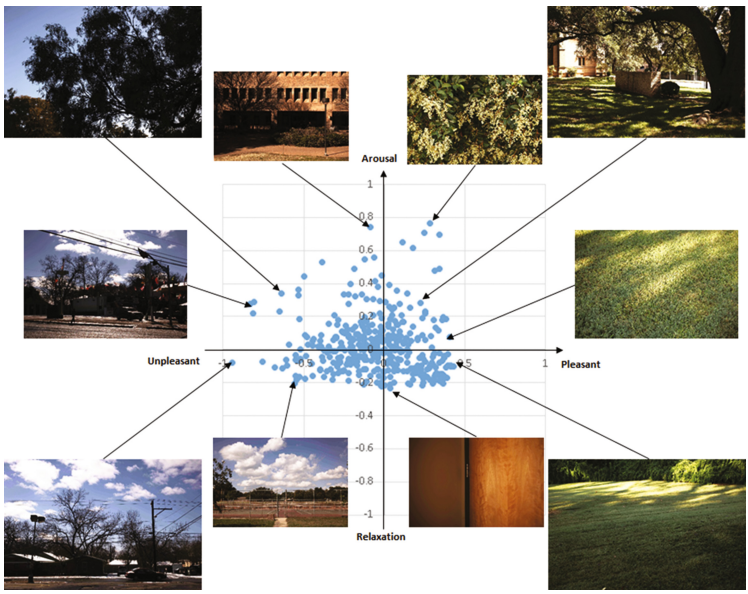


Fig. 3. The plotting results on total of 10 images which positioned at the extremes.

The plotting results are shown by choosing the 12 images of the two extremes both "Pleasant – Unpleasant" as X-axis and "Arousal – Relaxation" as Y-axis, as shown in Fig. 3.

In terms of color, totally 391 images of the natural scene were not almost distributed in variety of color domain, almost distributed in specific color domain such as blue, red and green. Based on Geneva Emotion Wheel Model, images of the series of blue color are distributed in almost left side as shown in Fig. 3. On the other hands, images of the series of green and red color are distributed in around the center.

In terms of spatial frequency, images which contained high frequency are distributed on a little upper side based on the center of (0, 0). On the other hands, images which contained low frequency are distributed on a little lower side based on the center of (0, 0).

Consequently, we confirmed that that image can be reasonably classified by using color and spatial frequency in terms of human emotion.

4 Conclusions

In this paper, we proposed image classification method as 2D multimodal fusion method by using image components such as color and spatial frequency in terms of human emotion. Because the both color and spatial frequency were turned out that can be influenced for human emotions in many researches, the proposed 2D multimodal fusion method can be suitable for affective computing application. For the experiments,

we collected totally 391 images which contained the 3 different kinds of scene categories such as natural scene, campus scene, and human made scene images from the public image database. Consequently, we confirmed that image can be affectively classified by using the color and spatial frequency.

In future works, we will perform for finding the appropriate mask size into each different images instead of using only 3 × 3 Prewitt mask in terms of image classification. Moreover, additional research will be needed for the reaction of people through subjective evaluation.

Acknowledgements. This work was supported by the ICT R&D program of MSIP/IITP [R0126-15-1045, the development of technology for social life logging based on analysis of social emotion and intelligence of convergence content]. Also, this research was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (2016M3A9E1915855).

References

1. Zheng, X., Cai, D., He, X., Ma, W.Y., Lin, X.: Locality preserving clustering for image database. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 885–891. ACM (2004)
2. Jimenez-Rodriguez, L.O., Arzuage-Cruz, E., Velez-Reyes, M.: Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data. IEEE Trans. Geosci. Remote Sens. **45**, 469–483 (2007)
3. Haralick, R.M., Shanmugam, K.: Textual features for image classification. IEEE Trans. Syst. Man Cybern. **6**, 610–621 (1973)
4. Park, M.W., Kim, C.J., Whang, M., Lee, E.C.: Individual emotion classification between happiness and sadness by analyzing photoplethysmography and skin temperature. In: 2013 Fourth World Congress on Software Engineering (WCSE), pp. 190–194 (2013)
5. Sacharin, V., Schlegel, K., Scherer, K.R.: Geneva Emotion Wheel Rating Study. Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences. Aalborg University, Aalborg (2012)
6. Singh, S.: Impact of color on marketing. Manage. Decis. **44**, 783–789 (2016)
7. Sylvain, D., Karim, N., Klaus, S., Didier, G.: Spatial frequencies or emotional effects? a systematic measure of spatial frequencies for IAPS pictures by a discrete wavelet analysis. J. Neurosci. Methods **165**, 144–150 (2007)
8. Park, M.W., Im, J., Kwon, J., Whang, M., Lee, E.C.: Correlation between heart rate and image components. In: Advances in Computer Science and Ubiquitous Computing, pp. 201–207. Springer, Singapore (2015)
9. Ghimire, D., Lee, J.: Color image enhancement in HSV space using nonlinear transfer function and neighborhood dependent approach with preserving details. In: 2010 Fourth Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp. 422–426 (2010)
10. Natural Scene Statistics in Vision Science, Center for Perceptual Systems, University of Texas at Austin. http://natural-scenes.cps.utexas.edu/db.shtml

Human Robot Interaction Method by Using Hand Gesture Recognition

Jaehyun Han[1], Woohyuk Jang[2], Dojoon Jung[3], and Eui Chul Lee[1(✉)]

[1] Department of Computer Science, Sangmyung University, Hongjimun 2-gil 20, Jongno-gu, Seoul 03016, Republic of Korea
eclee@smu.ac.kr
[2] Department of Sports ICT Convergence, Sangmyung University, Hongjimun 2-gil 20, Jongno-gu, Seoul 03016, Republic of Korea
[3] Laboratory of Disaster, National Disaster Management Research Institute, Jongga-ro 365, Jung-gu, Ulsan 44538, Republic of Korea

Abstract. Recently, human robot interaction technique is limelight in which controlling robot by using natural user interface such as using body gesture, eye tracking or human voice recognition. In this paper, hand gesture based robot controlling method is proposed. In the present study, a new natural user interface model is hierachically designed. The hand gesture is recognized based on Kinect V2. The recognized gesture information is send to the robot controlling robot through Bluetooth 4.0 interface. The developed robot controlling interface can be adopted into the field of disaster area in order to restore or surveillance the area of human inavailable. Experimental result showed that the proposed method can be effectively used for controlling robot.

Keywords: Kinect V2 · Human robot interaction · Natural user interface · Gesture recognition

1 Introduction

Natural User Interface (NUI) means input protocol to computer by using user's natural intrinsic method such as behavior, gaze, and voice. Because NUI has advantages such as intuitive and easily learned, the related researches have been widely performed. For example, gesture based smartphone interface and game interface using gesture can be considered [1, 2]. To implement gesture based NUI, raw data including high quality and good information about human body is preferred such as capturing both RGB and depth images using Kinect. Kinect V2 can capture such images from RGB camera and IR sensor [7]. By analyzing RGB and depth map image, maximum six persons can be concurrently analyzed in terms of their 3-dimensional joint coordinates. Such high performance sensor can easily make gesture or posture based NUI to implement [3, 4].

With the growth of robot industry and combining the NUI, Human-Robot Interaction (HRI) technique has been widely researched [5, 6]. Recently, HRI sublates traditional interface methods using key or button press based method. Instead of that, the above mentioned NUI techniques are frequently considered as easy and effective robot controlling method [8]. Especially, humanoid robot can be effectively controlled

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_17

by using human body based NUI such as gesture or posture recognition because of their same structure.

In this paper, a new robot controlling scheme is designed and implemented. For that, several hand gestures are proposed and systematically combined. In the present paper, five gestures are used such as forward, stop, backward, left turn, and right turn. The gestures are recognized by using Kinect V2, then the recognized gesture information is send to robot control module through Bluetooth wireless connection.

2 Proposed Method

In this section, the proposed method and design are described for implementing hand gesture based HRI method. Conceptual environment of the proposed method is shown in Fig. 1. A distance between a user and Kinect V2 is about 2 m in our method. However, because the depth range of recognition is maximum 4.5 m, the proposed method can be available within the range.

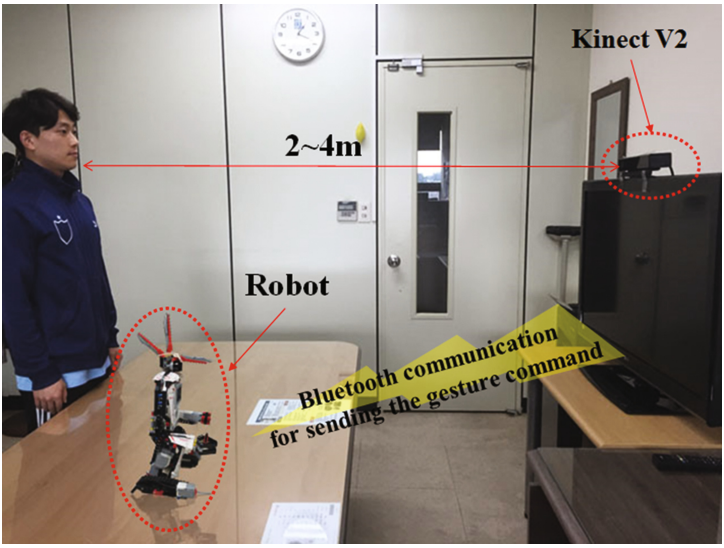


Fig. 1. System environment of the proposed method

According to the recognized gesture which is already defined one, robot can be controlled based on the Bluetooth received gesture data. In our method, LEGO Mindstorm EV3 model [9] is used for HRI implementation and feasibility test because the model is appropriate in terms of humanoid and Bluetooth communication. The procedure of proposed method is shown in Fig. 2.

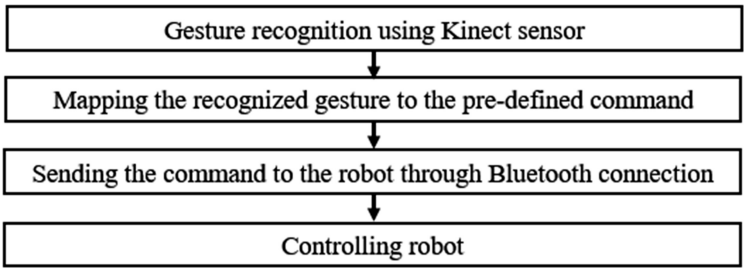


Fig. 2. Flow chart of the proposed method

2.1 Definition of Hand Gestures

In this section, hand gesture based NUI hierarchy is defined for controlling robot through five kinds of command such as forward, backward, stop, left turn, and right turn. Figure 3 shows the defined hierarchical hand gesture model which explains the mapping relation between each gesture and its corresponding command for controlling robot.

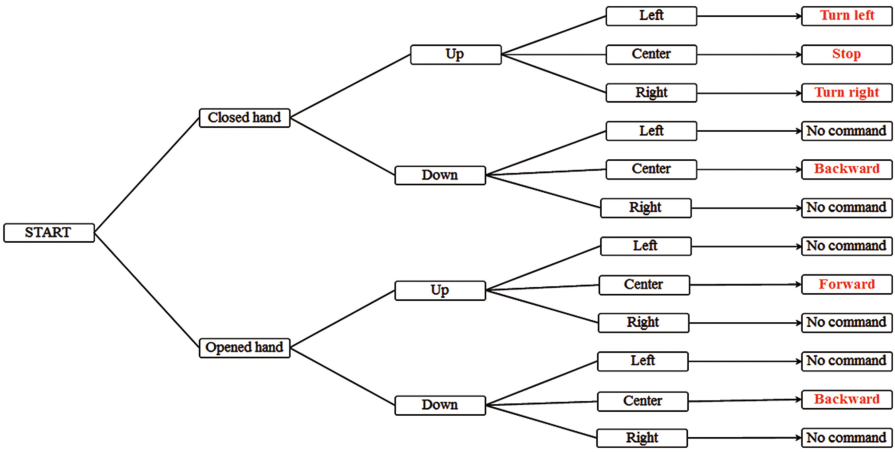


Fig. 3. Hierarchical model of the used hand gestures in our proposed robot controlling method

Any other gestures excepting for the defined ones are invalid in terms of controlling robot. Also, both left and right hand are usable for recognizing the defined gestures. In here, only one hand is analyzed for recognizing gesture.

Forward Command

"Forward" command is the one for moving robot of going straight. For that, user raises arm with open hand against the Kinect sensor as shown in Fig. 4(a). Even though the raised arm does not meet perfectly with the orthogonal vector of floor, the angular offset between 0°–±20° from the orthogonal vector regarded as the raised one. In here, closed hand or out of the mentioned angular offset is regarded as invalid gesture.

Backward Command

"Backward" command is the one for moving robot of going the back direction. To perform the movement, user should drop arm with close or open hand against the Kinect sensor as shown in Fig. 4(b). As similar way with the mentioned in "Forward Command", little angular offset between 0°–±20° is allowed for recognizing the "Backward" command. In here, although both open and close hand do not care for the command, out of the mentioned angular offset is regarded as invalid gesture.

Stop Command

"Stop" command is defined for no movement of robot. To generate this command, user should take same gesture with the "Forward" one with close hand as shown in Fig. 4 (c). As similar way with the mentioned gestures, open hand or angular offset between 0°–±20° is regarded as invalid gesture.

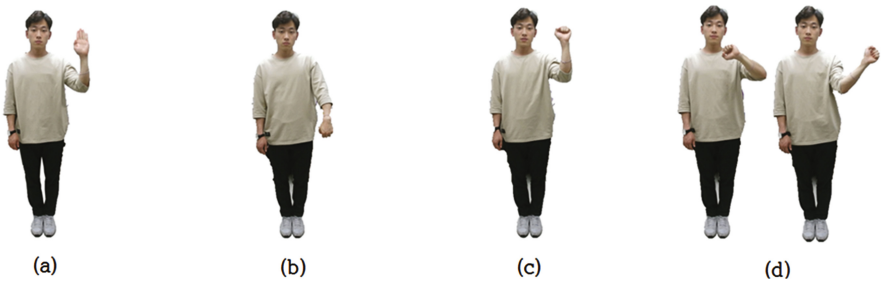


Fig. 4. Defined gesture commands for controlling robot. (a) Forward. (b) Backward. (c) Stop. (d) Left: Turn Left, Right: Turn Right.

Commands for Turning Left and Right

"Turn Left" and "Turn Right" commands are defined for respectively rotating robot to left and right at fixed position. For that, user tilts arm to left or right side from the mentioned "Stop" gesture, as shown in Fig. 4(d). Empirically, because arm tilting to inside (left image of Fig. 4(d)) is easier than the one to outside (right image of Fig. 4 (d)), the earlier generally make more tilting angle than the later. Based on the body structural reason, the valid range of arm tilting to inside is configured between 30° and 75° from the gesture of "Stop". Also, the valid range of arm tilting to outside is configured between 30° to 55°. In here, because turn commands should be performed from the "Stop" command, close hand is necessary. That is, open hand is regarded as invalid gesture even if the raised arm is positioned within valid angular range.

2.2 Robot Controller

In this section, specification of robot module is described. In our method, the used robot is "EV3" model released from LEGO Mindstorms [9]. In this model, various

machine parts such as wheel, gear, motor, infrared sensor, touch sensor, and so on are included in order to implement automatic driving robot. In our study, the robot is assembled as humanoid type among recommended five types. The pre-defined command based on the recognized gesture is sending from the Bluetooth dongle to the Bluetooth receiver at robot controlling module.

According to the received command, robot performs the corresponding movement such as moving forward or backward, turning left or right.

3 Experimental Results

To measure the accuracy of recognizing five kinds of gestures, we performed the following experiments. To confirm the error caused by body figure variation, six subjects whose height are 190 cm, 183 cm, 175 cm, 171 cm, 160 cm, and 157 cm were participated. The subjects were categorized into three groups such as tall, medium, and short based on the height. Each subject takes 50 times per each gesture then the accuracy of correct recognition is calculated. Experimental environment was coinciding with Fig. 1. Accuracy figure for each gesture is shown in Table 1.

Table 1. Recognition accuracy of five pre-defined robot controlling gestures in terms of subject height.

Height	Execution times	Go straight		Back		Stop		Turn left		Turn right	
		T	F	T	F	T	F	T	F	T	F
Higher than 180 cm	1	50	0	48	2	50	0	49	1	50	0
	2	50	0	50	0	50	0	49	1	49	1
	Recognition rate (%)	100%		98%		100%		98%		99%	
Higher than 170 cm, but lower than 180 cm	1	50	0	50	0	50	0	50	0	50	0
	2	50	0	49	1	50	0	50	0	49	1
	Recognition rate (%)	100%		99%		100%		100%		99%	
Lower than 170 cm	1	50	0	49	1	50	0	48	2	50	0
	2	50	0	50	0	50	0	49	1	50	0
	Recognition rate (%)	100%		99%		100%		97%		100%	

Based on the result of measuring recognition accuracy, our proposed method could accurately recognize the designed gestures for robot controlling. Also, we confirmed that there was no individual variation of recognition accuracy in terms of body figure variation such as height.

4 Conclusion

In this paper, we proposed human robot interacting method by using hand gesture based natural user interface. Five kinds of hand gestures are hierarchically designed and mapped into commands for robot controlling by considering cross effects between different gestures. Based on the recognized gestures by Kinect V2, the send command through Bluetooth make robot to move. Experimental results showed that the proposed method recognized five kinds of gestures at the accuracy of about 99%.

In future works, we will apply the NUI based HRI concept into robot interface field for disaster area. Our proposed method can be effectively applied into disaster area where is dangerous to person. In here, robot based remote monitoring or simple rescue or recovery may possible based on our proposed method.

Acknowledgements. This research was supported by the Ministry of Science, ICT and Future Planning (MSIP), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2016-H8501-16-1014) supervised by the Institute for Information & Communications Technology Promotion (IITP). Also, this work was supported by National Disaster Management Research Institute.

References

1. Cho, S., Byun, H., Lee, H.K., Cha, J.: Arm gesture recognition for shooting games based on Kinect sensor. J. KISS Softw. Appl. **39**, 796–805 (2012)
2. Siddiqui, M., Medioni, G.: Human pose estimation from a single view point, real-time range sensor. In: Workshop on Computer Vision for Computer Games at Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2010)
3. Kim, B.C., Kim, H., Ko, D., Lee, E.C.: Hand mouse manipulation by supporting moving sensitivity modification and click event. Asia Pac. J. Multimedia Serv. Converg. Art Humanit. Sociol. **5**, 19–26 (2015)
4. Oikonomidis, L., Kyriazis, N., Argyros, A.A.: Efficient model-based 3D tracking of hand articulations using Kinect. In: British Machine Vision Conference (2011)
5. Jung, J., Yang, C., Lee, T., Lee, M.: Live-block interacting with computers. J. HCI Soc. Korea 209–214 (2005)
6. Ahn, H., Song, J.S.: A ubiquitous interface system for mobile robot control in indoor environment. J. Inst. Control Robot. Syst. **12**, 66–71 (2006)
7. https://developer.microsoft.com/ko-kr/windows/kinect/hardware. Accessed 15 Oct 2016
8. Park, S.H., Shim, J.H., Hong, K.T., Park, J.Y., Kim, S.H., Song, H.J.: An Obstacle Avoidance Control of a Mobile Robot using a Kinect Sensor (2011)
9. http://mindstorms.lego.com/. Accessed 15 Oct 2016

A Beacon-Based User Direction Estimating Method in Indoor Environments

Jeonghoon Kwak and Yunsick Sung(✉)

Department of Multimedia Engineering, Dongguk University,
Seoul, South Korea
{jeonghoon, sung}@dongguk.edu

Abstract. Applications using parts of a user's body as an interface are increasing in number. In an indoor environment the location of a user can be measured using a beacon, and the direction of movement can be estimated from their location as measured by the beacon; however, a limitation exists when estimating the direction in which the user is facing. This paper proposes a method for estimating the direction of the user's body using two beacons. It verifies experimentally that the proposed two-beacon method can correctly estimate the direction of a user's body, with the error in the user direction reduced by 74.13% when compared to an existing method.

Keywords: NUI · Natural user interface · Nature user experience · Beacon · Location · Direction

1 Introduction

Research into the use of interfaces based on user motion is increasing [1–3]. A head-mounted display (HMD) such as Oculus [4] generates a user-based virtual character on a screen, on which the head direction matches that of the user; however, if the character is controlled using only the HMD, the virtual character is able to achieve movements - such as full 360° head rotation - that are impossible in real life. For an effective interface, the direction of the user's body is required in addition to the direction of the user's head.

Previous studies estimating the user's location in an indoor environment have used beacons to determine the user direction [5–7]. It is possible to predict the direction of user movement by utilizing the location of the beacon, however, it cannot be determined if the user is facing forwards or backwards while walking. Therefore, a more effective method for estimating the direction of a user's body using beacon location is required.

This paper proposes a method for estimating the direction in which a user's body is facing, by utilizing the locations of two beacons. Access points (APs) and beacons are set in place, and the locations of two beacons attached to the user are calculated using the APs. The proposed method allows for realistic movements in virtual reality systems that utilize the direction of the user's body.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_18

The remainder of the paper is organized as follows: In Sect. 2, the user direction estimation method is presented; Sect. 3 validates the method of user direction estimation; finally, Sect. 4 summarizes the proposed method.

2 User Direction Estimation Framework in Indoor Environments

This paper proposes a framework for estimation of user direction by using the user position calculated using beacons. Figure 1 shows the process used to recognize the direction in which the user's body is facing. In the proposed method the beacon signal is measured, and then the direction of the user is calculated using the distance to the beacons as measured by the APs; the Applications provide the calculated angle.

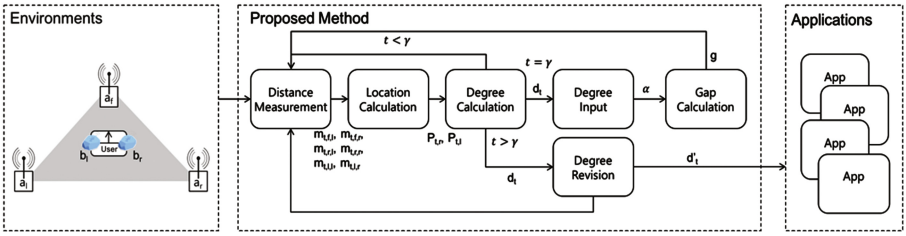


Fig. 1. Process of calculating user direction with two beacons attached to the user.

Figure 2 shows the environmental settings used to recognize the direction in which the user's body is facing. A beacon attached to the right side of user's body is set as beacon b_r, and a beacon attached to the left side is set as beacon b_l. The beacon locations are calculated based on the beacon distance measured at each AP, which is defined as follows: The location of AP a_f is defined by $[x_f, y_f]$, the location of AP a_r is defined $[x_r, y_r]$, and the location of AP a_l is defined by $[x_l, y_l]$. At time t, the location of beacon b_l is point $P_{t,l} = [x_{t,l}, y_{t,l}]$, and the location of beacon b_r is point $P_{t,r} = [x_{t,r}, y_{t,r}]$.

In the distance measurement step, the six distances from AP a_f, AP a_l, and AP a_r to beacon b_l and beacon b_r are measured. The six distances measured are at the time t; the distance measured from AP a_f to beacon b_l is $m_{t,f,l}$, and to beacon b_r is $m_{t,f,r}$; the distance measured from AP a_l to beacon b_l is $m_{t,l,l}$, and to beacon b_r is $m_{t,l,r}$; and the distance measured from AP a_r to beacon b_l is $m_{t,r,l}$, and to beacon b_r is $m_{t,r,r}$.

In the location calculation step, the location is estimated based on the distances from AP a_f, AP a_l, and AP a_r to beacon b_l and beacon b_r [7]. The location of AP a_f, AP a_l, and AP a_r, and the locations of the two points point $P_{t,l}$ and point $P_{t,r}$ are set using a trilateration method; this uses the location of AP a_f, AP a_l, and AP a_r, with beacon b_l of distances $m_{t,f,l}$, $m_{t,l,l}$, and $m_{t,r,l}$, and beacon b_r of distances $m_{t,f,r}$, $m_{t,l,r}$, and $m_{t,r,r}$ as

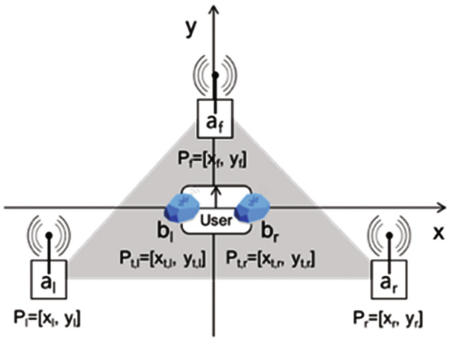


Fig. 2. Two beacons are attached to the right and left sides of the user, respectively. The beacon attached to the right side is defined as beacon b_r, and the beacon attached to the left side is defined as beacon b_l. APs are placed in each environment to measure the beacon locations. The AP on the user's right side is defined as AP a_r, the AP on the user's left side is defined as AP a_l, and the AP in front of the user is defined as AP a_f.

shown in Eq. (1). The point $P_{t,l}$ is calculated by function f_l, and the point $P_{t,r}$ is calculated by function f_r.

$$\begin{aligned} P_{t,l} &= f_l\left(a_f, m_{t,f,l}, a_l, m_{t,l,l}, a_r, m_{t,r,l}\right) \\ P_{t,r} &= f_r\left(a_f, m_{t,f,r}, a_l, m_{t,l,r}, a_r, m_{t,r,r}\right) \end{aligned} \tag{1}$$

During the degree calculation step, the direction of the user is calculated using point $P_{t,l}$ and point $P_{t,r}$, which are the locations of the two beacons. The 2D vector $\overrightarrow{v_t}$, is calculated based on point $P_{t,l}$ and point $P_{t,r}$, and is defined as shown in Eq. (2). $v_{t,x}$ is calculated as $x_{t,l} - x_{t,r}$, and $v_{t,y}$ is calculated as $y_{t,l} - y_{t,r}$. After a calculation using point $P_{t,l}$ and point $P_{t,r}$, vector $\overrightarrow{v_t}$ is converted into a unit vector through a normalization process. The direction of the user's body, which is calculated using the locations of the two beacons, is defined as degree d_t at time t. As shown in Eq. (3), vector $\overrightarrow{v_t}$ is used to calculate the angle rotated by the user based on the x-axis, to set degree d_t. If time t is smaller than the γ, it moves to the distance measurement step; if time t is equal to γ, it moves to the degree input step; if time t is greater than γ, it moves to the degree revision step.

$$\overrightarrow{v_t} = \left[v_{t,x}, v_{t,y}\right] \tag{2}$$

$$d_t = \begin{cases} \textit{if } \operatorname{atan2}\left(v_{t,y}, v_{t,x}\right) \times \frac{180}{\pi} - 90 < -180 \\ \qquad \operatorname{atan2}\left(v_{t,y}, v_{t,x}\right) \times \frac{180}{\pi} + 270 \\ \textit{else} \\ \qquad \operatorname{atan2}\left(v_{t,y}, v_{t,x}\right) \times \frac{180}{\pi} - 90 \end{cases} \tag{3}$$

The degree input step defines the angle viewed by the user as α, and sets α. The gap calculation step usees the collected degree d_t to calculate the difference in user direction; gap g is defined as the difference in the direction in which the user is looking, and is calculated by averaging the difference between the α and degree d_t, as shown in Eq. (4).

$$g = \frac{1}{\gamma} \times \sum_{t=1}^{\gamma} (d_t - \alpha) \tag{4}$$

During the direction revision step, the angle is revised using angle d_t and gap g. If angle d_t is added to represent −180° to 180° as shown in Eq. (5) and it is larger than 180°, 360° is decreased. Angle d'_t is ouptut as shown in Fig. 3.

$$d'_t = \begin{cases} if\ d_t - g > 180 \\ \quad d_t - g - 360 \\ else \\ \quad d_t - g \end{cases} \tag{5}$$

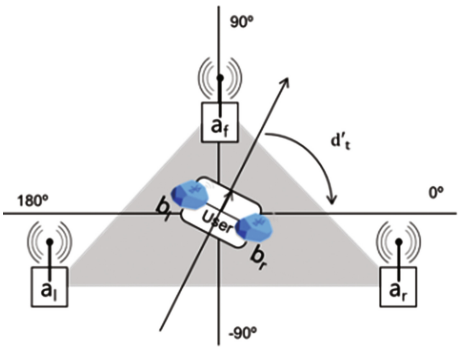


Fig. 3. User direction as calculated by the proposed method. User direction is provided by degree d'_t.

3 Experiments

To verify the proposed method, three APs and a user were placed as shown in Fig. 4. At start-up, the user direction was set to −90°, α was set by −90. The gap g was set to −93.43.

To verify the proposed method, the user changed direction four times, as shown in Fig. 5. Figure 6 shows the position of the two beacons measured when the user direction changed.

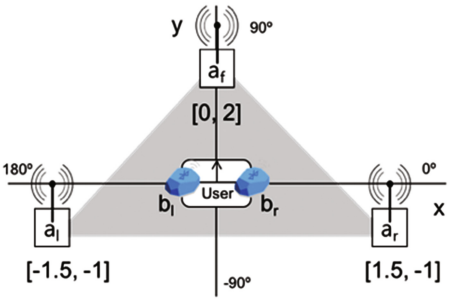
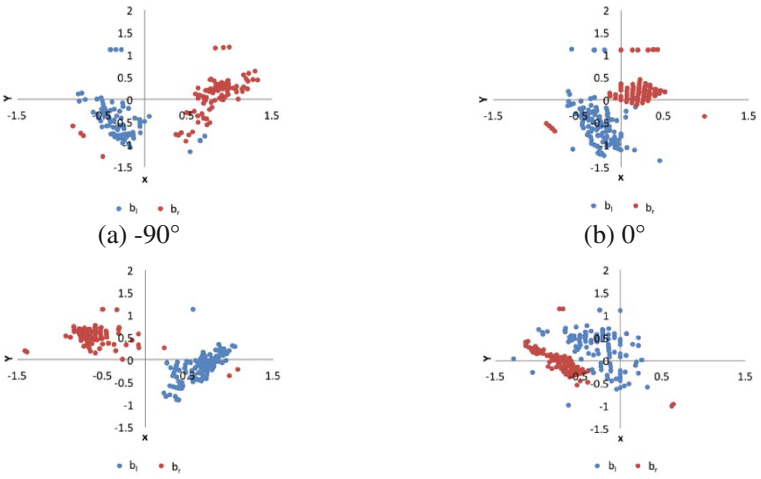


Fig. 4. Location of AP a_l was [−1.5, −1], the location of AP a_r was [1.5, −1], and the location of AP a_f was [0,2]. The heights of the three APs and two beacons were set to 0.7 m.

(a) -90° (b) 0° (c) 90° (d) -180° or 180°

Fig. 5. User direction was changed in the order of −90°, 0°, 90°, and then −180°.



(a) -90° (b) 0°

(c) 90° (d) -180° or 180°

Fig. 6. Locations of beacon b_l and beacon b_r with changing user direction.

The results of calculating the direction using the positions of the two beacons alone, and the results using the proposed method, are shown in Fig. 7. When using only the location of the two beacons, the cumulative error was 92179.77°; when using the proposed method the cumulative error was 68335.14°, giving a reduction of 74.13%.

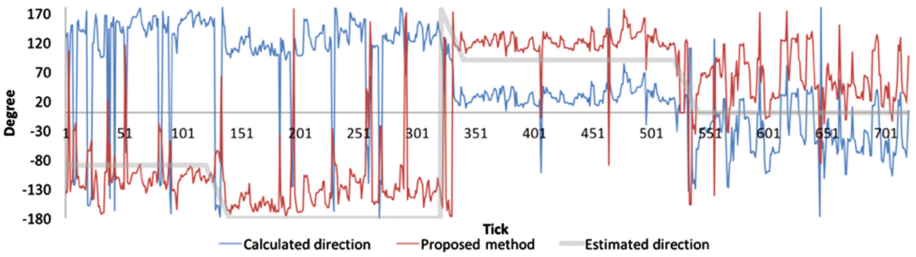


Fig. 7. Degrees calculated per the user direction. User direction calculated by the position of two beacons is shown in blue, and user direction calculated by the proposed method is shown in red; the actual user direction is shown in gray.

4 Conclusions

This paper proposed a method for estimating the direction of a user's body by utilizing two different beacons. The beacon locations are obtained using the distance between the beacons and several APs, and the direction in which the user is heading toward is derived using the location-based vector of the two beacons. In experiments, measurement of user direction was attempted using four different angles; when compared to an existing method the proposed method reduced the cumulated error by 74.13%.

Acknowledgments. "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-H8501-16-1014) supervised by the IITP (Institute for Information & communications Technology Promotion)".

References

1. Kim, J.W., Kim, P.Y., Sung, Y.: Planet editing method using leap motions. In: Juang, J. (ed.) Proceedings of the 3rd International Conference on Intelligent Technologies and Engineering Systems (ICITES2014). Lecture Notes in Electrical Engineering, vol. 345, pp. 225–230. Springer, Cham (2016)
2. Zhang, L., Sturm, J., Cremers, D., Lee, D.: Real-time human motion tracking using multiple depth cameras. In: The 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2389–2395 (2012)
3. Lee, S., Sung, Y., Park, J.H.: Motion estimation framework and authoring tools based on MYOs and Bayesian probability. Multimedia Tools Appl. 1–20 (2016)
4. Oculus Information. https://www.oculus.com/
5. Lee, H.C., Lee, D.M.: The 3-dimensional localization system based on beacon expansion and coordinate-space disassembly. J. Korea Inf. Commun. Soc. **38B**(1), 80–86 (2013)
6. Kwak, J., Sung, Y.: Indoor location-based natural user interface for ubiquitous computing environment. In: Park, J.J. (Jong Hyuk), Yi, G., Jeong, Y.S., Shen, H. (eds.) Advances in Parallel and Distributed Computing and Ubiquitous Services. Lecture Notes in Electrical Engineering, vol. 368, pp. 131–136. Springer, Singapore (2016)
7. Lee, H.C., Lee, D.M.: A study on localization system using 3D triangulation algorithm based on dynamic allocation of beacon node. J. Korea Inf. Commun. Soc. **36**(4), 378–385 (2011)

User Selection Based Backpropagation for Siamese Neural Networks in Visual Filters

Hanju Park and Sukho Lee(✉)

Department of Software Engineering, Dongseo University,
47 Jurye-ro, Sasang-gu, Busan 47011, Korea
petrasuk@gmail.com

Abstract. In this paper, we propose a user selection based backpropagation method for siamese networks which we will use as visual filters in mobile contents. The loss function used to train the network is affected by the user's interaction which also affects the update of the weights in the network such that the visual filter takes subjective similarities into account. As a specific application of the proposed algorithm, we expect that the visual filter can be applied for mobile services which provides the user with 3D visual products appearing above the phone. The products appearing in sequence to the user are those which have similar appearances to that selected by the user where the subjective similarity is also taken into account.

Keywords: Visual filter · Mobile contents · Siamese neural network · Backpropagation

1 Introduction

In this paper, we propose a user interactive training method for the Siamese neural network which can be used for commercial purposes. The target application of the proposed algorithm using the Siamese network is a mobile application with 3D display, where the user can pick up the desired product via a touch in the air above the display of the mobile phone. After the user has selected the product, other products having visual similarities appear also in 3D above the phone, and the user can further select the next product which he thinks is closer to the product in mind.

However, visual similarity is a subjective issue, which is not easy to describe. For example, even though two products have similar sizes and forms, the user may regard them as being more different than other combinations of two products. Therefore, we think that the subjective value of the user should also be taken into account in the training stage of the Siamese neural network. For this purpose, we propose a user interactive backpropagation method which updates the weights in the Siamese neural network by the user selection of the product he thinks to be closer to the product appearing in a similarity test. In Sect. 2, we explain the related works to the proposed method, then, in Sect. 3, we propose our user selection based backpropagation method for the Siamese neural network.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_19

2 Related Works

In this section, we explain the related works of the proposed algorithm. This gives also an insight into the concept of the proposed backpropagation method.

2.1 3D-Display for Mobile Phones

The target of the proposed algorithm is a 3D augmented mobile shopping content which is played above the display of a smartphone wearing a lenticular optical filter case. The lenticular optical filter consists of many lenslets which makes the underlying elemental images to appear above the display. The underlying technology making it possible is called integral imaging [1, 2]. This concept is called the mobile hologram reality space and is now under development by us. Figure 1 shows a conceptual picture of this system. Here, we want the product to appear in 3D above the display, and let the user pick-up in consequence other products which he thinks are closer in appearance to the product he has in mind.

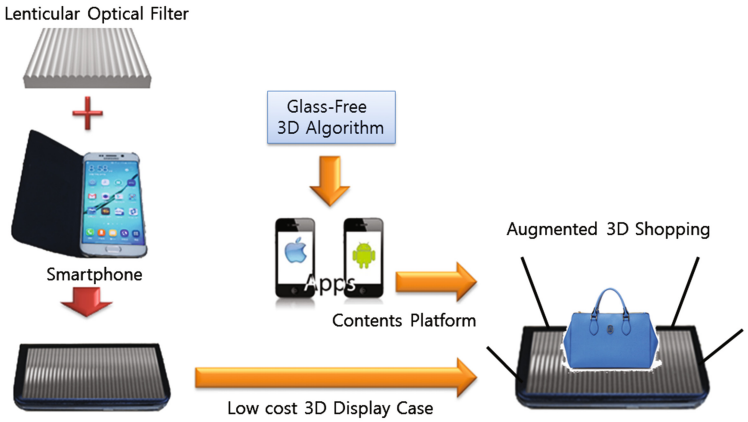


Fig. 1. Concept of augmented 3D shopping using a low cost 3D Display which consists of lenslet arrays to display the underlying elemental images as a 3D object above the display.

2.2 Siamese Neural Networks

Siamese neural network refers to the neural network architecture which contains sub-networks of the same structure. Parameters like the weights can be shared across the sub-networks or not. Originally, the Siamese neural network has been proposed to identify if two objects appearing in slightly different forms are identical or not, for example, to verify if two fingerprints are from the same person or not [3]. Another application is two verify if a frame captured by two different CCTVs contain the same person [4]. Nowadays, Siamese neural networks find their use as visual filters which offer products which are visually similar to that the user selects. A successful adoption of the Siamese neural network is the visual search application of the Shoes.com

company which uses an Siamese neural network-powered visual filter developed by sentient technology [5]. Figure 2 shows how the visual filter of Shoes.com works. First, arbitrary shoe products appear on the screen. Then, the user tabs at a product that is most like to that he is looking for. Based on the selection, the software suggests other visually similar products on the screen. By keeping on to refining the display of the products, the consumer can come closer to the product he is keeping in mind.

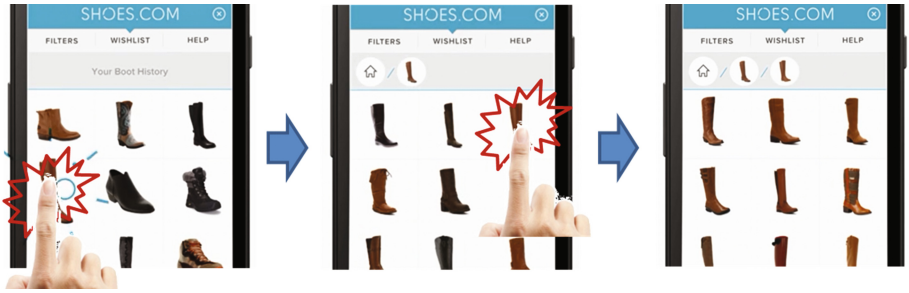


Fig. 2. Example of the working of the visual filter of Shoes.com. As the consumer keeps on clicking, the display is updated with more and more visually similar products, so that the consumer comes closer to the product he has in mind.

Figure 3 shows how the Siamese neural network is used in the visual filter. After the user clicks on a product, the image of it and those of the candidate products in the database are put through the Siamese neural network together to produce feature vectors at the output of the convolutional neural network (CNN). A metric, which computes the similarity of the output vectors is applied, and the product which produces a small value of this metric is considered to be visually similar to the user selected product. The products which have lowest metric values are displayed on the screen for further search.

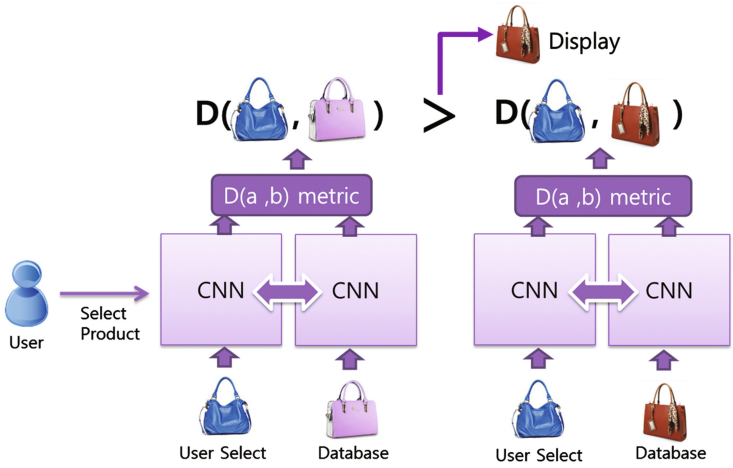


Fig. 3. Showing the use of the Siamese neural network in a visual filter

Normally, the weight parameters of the CNNs in the Siamese neural network are set to give a minimum metric value to identical objects. However, in the visual filter there exist no labels for which products are visually similar. Therefore, normally, the CNNs in the Siamese neural network are trained by unsupervised methods, such as variational auto-encoders. This results in feature vectors that are robust against small deformations. However, there exist products that consumers regard as being closer to some products even though their outer appearance is farther than other products. Therefore, some subjective criteria should affect the weights in the networks too. In the next section, we propose an algorithm that does this kind of work.

3 User Interactive Backpropagation for Siamese Network

The CNNs in the Siamese neural network are trained first by an auto-encoder, which initializes the weights in the CNNs. We used the ConvNetJS [6] open source library to implement the auto-encoder. The layers for the auto-encoder have the following parameters: Layer-1 has 2500 nodes to take a 50 × 50 image with 1 channel as an input. Layer-2, layer-3, layer-5, and layer-6 have 100, 50, 50, and 100 nodes, respectively. Layer-4, which will be the output layer for the CNN after it has been trained by the auto-encoder, has 10 nodes, which means that every product will be represented by a vector of length 10. The activation function is the tanh function. For experimental case, we used bags as the products to test our algorithm.

Figure 4 shows the diagram of the proposed user selection based backpropagation method for the visual filter. This backpropagation updates the weight values in the CNN which have been initialized by the auto-encoder.

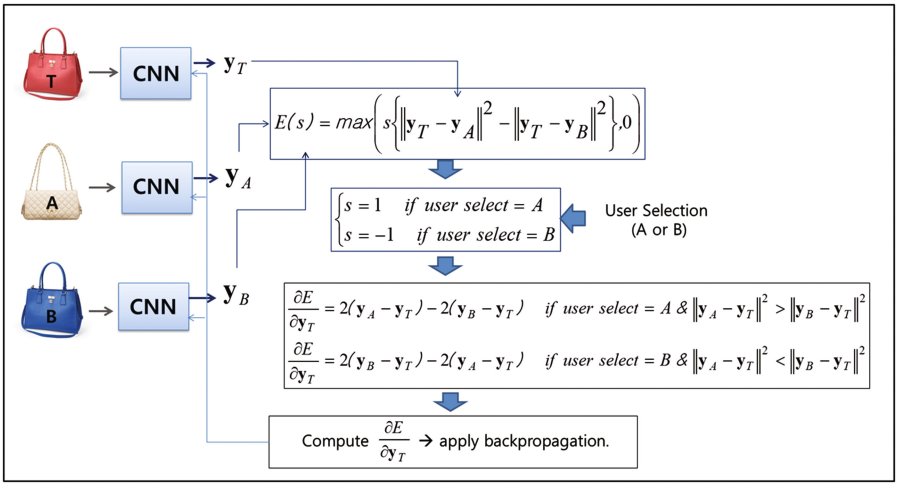


Fig. 4. Example of the working of the visual filter of Shoes.com.

The software randomly displays three bag images on the screen: bag-A, bag-B, and bag-T. Then, the user has to select between bag-A and bag-B which looks more similar to bag-T. The energy function (cost function) which is minimized by the backpropagation, has the following form:

$$E(s) = max\Big(s\Big\{\|\mathbf{y}_T - \mathbf{y}_A\|^2 - \|\mathbf{y}_T - \mathbf{y}_B\|^2\Big\}, 0\Big) \tag{1}$$

Here, $\mathbf{y}_A, \mathbf{y}_B, \mathbf{y}_T$ are the output feature vectors corresponding to the bags bag-A, bag-B, and bag-T, respectively. $max(a, b)$ is a function which takes the value a if $a \geq b$ and b if $a < b$. s is a value which takes on the value 1 if the user selects bag-A, and -1 if the user selects bag-B, i.e.,

$$\begin{cases} s = 1 & if\ user\ select = bag\ A \\ s = 0 & if\ user\ select = bag\ B \end{cases}$$

The partial differential of the energy with respect to the output of the CNN becomes then,

$$\begin{aligned} \tfrac{\partial E}{\partial \mathbf{y}_T} &= 2(\mathbf{y}_A - \mathbf{y}_T) - 2(\mathbf{y}_B - \mathbf{y}_T) \quad if\ user\ select = A\ \&\ \|\mathbf{y}_A - \mathbf{y}_T\|^2 > \|\mathbf{y}_B - \mathbf{y}_T\|^2 \\ \tfrac{\partial E}{\partial \mathbf{y}_T} &= 2(\mathbf{y}_B - \mathbf{y}_T) - 2(\mathbf{y}_A - \mathbf{y}_T) \quad if\ user\ select = B\ \&\ \|\mathbf{y}_A - \mathbf{y}_T\|^2 < \|\mathbf{y}_B - \mathbf{y}_T\|^2. \end{aligned} \tag{2}$$

This can be inserted in the ordinary backpropagation to solve for the weights in the CNN. The reason that the condition for $\|\mathbf{y}_T - \mathbf{y}_A\|^2$ to be larger than $\|\mathbf{y}_T - \mathbf{y}_B\|^2$ in the first equation in (2) is to prevent the case that the CNN is updated such that $\mathbf{y}_T$ becomes closer to $\mathbf{y}_A$ when it is already closer to $\mathbf{y}_A$ than $\mathbf{y}_B$. This is in contrast with the following energy, which one might think instead of (1),

$$E(s) = s\|\mathbf{y}_T - \mathbf{y}_A\|^2 + (1 - s)\|\mathbf{y}_T - \mathbf{y}_B\|^2 \tag{3}$$

with

$$\begin{cases} s = 1 & if\ user\ select = A \\ s = 0 & if\ user\ select = B \end{cases}$$

for which the gradients of E with respect to $\mathbf{y}_A$ and $\mathbf{y}_B$ become,

$$\begin{aligned} \tfrac{\partial E}{\partial \mathbf{y}_T} &= 2(\mathbf{y}_A - \mathbf{y}_T) \quad if\ user\ select = A \\ \tfrac{\partial E}{\partial \mathbf{y}_T} &= 2(\mathbf{y}_B - \mathbf{y}_T) \quad if\ user\ select = B \end{aligned} . \tag{4}$$

In this case, there is no constraint on the weights to be updated in the direction of $\mathbf{y}_A - \mathbf{y}_T$ (for the first equation in (4)), which can result in an over-modified CNN.

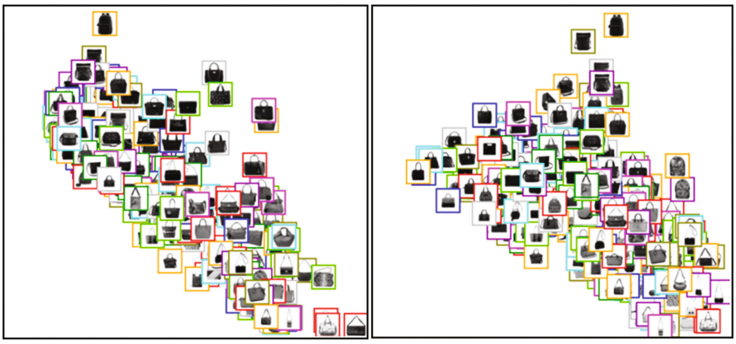


Fig. 5. Distribution of products shown in 2D after applying auto-encoder (left) and auto-encoder +proposed backpropagation.

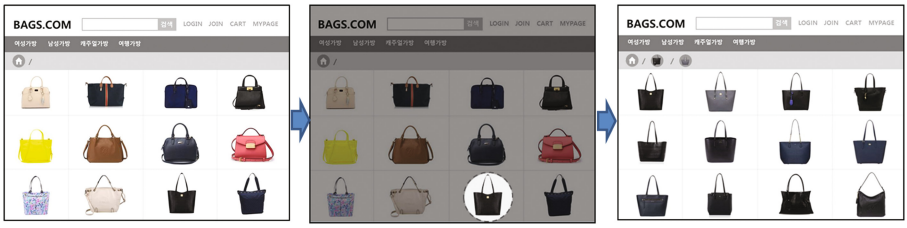


Fig. 6. Working of the visual filter trained by the proposed method.

Figure 5 shows the distances of the output vectors of the variational auto-encoder and those after applying our method as a refinement. The output vectors are projected on a 2D plane with the corresponding bags used as icon images. The left image shows the relative distribution of the output vectors after it has been trained by the auto-encoder, and the right one shows the distribution after applying the proposed method. As can be seen visually similar bags tend to converge together. Figure 6 shows an example of the appearance of products after the user has clicked the highlighted product. As can be seen, some visually similar products seem to become closer after our method is applied, but it is rather subjective and depends on the taste of the user. Therefore, a quantitative comparison is not possible.

Acknowledgments. This work was partly supported by the Industrial-Academic Cooperative R&D Program funded by the Small and Medium Business Administration (SMBA, Korea) [C0395100] and “Dongseo Frontier Project” Research Fund of 2015.

References

1. Stern, A., Javidi, B.: Three-dimensional image sensing, visualization, and processing using integral imaging. Proc. IEEE **94**, 591–607 (2006)
2. Park, J.-H., Hong, K., Lee, B.: Recent progress in three-dimensional information processing based on integral imaging. Appl. Opt. **48**, H77–H94 (2009)
3. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a Siamese time delay neural network. Int. J. Pattern Recogn. Artif. Intell. **7**(4), 669–688 (1993)
4. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: CVPR (2014)
5. Shoes.com homepage. http://www.shoes.com
6. ConvNetJs. www.ncbi.nlm.nih.gov http://cs.stanford.edu/people/karpathy/convnetjs/

Methodological Route to Designing Optimized Bedroom Environment for Active-Aging

Sung Jun Park(✉)

Faculty of Interior and Environmental Design, Architectural Engineering,
Keimyung University, Daegu, South Korea
sjpark@kmu.ac.kr

Abstract. This paper stresses the importance of monitoring indoor environment quality (IEQ) in a bedroom environment for the pre-elderly, by comparing data on sleep behavior with the corresponding IEQ variables. Although concerns regarding the influence of IEQ on residents' health are increasing, solutions for optimizing bedroom environments have not yet been determined. In particular, for the elderly and pre-elderly, optimization of the indoor environment may relieve the symptoms of chronic disease and degradation of their physical abilities. This study presents a method for obtaining basic data to create an optimized bedroom environment for the aging society.

Keywords: IEQ · Sleep behaviors · Bedroom environment · Pre-elderly

1 Introduction

The bedroom environment is the space in which sleep occurs for the majority of home occupants, and this environment has functions and features designed to satisfy the associated human physiological needs. The majority of residents spend 30% or more of the 24-h day in the bedroom, because of their behavioral needs regarding sleep and rest. Further, elderly home occupants often have physical, psychological, or social difficulties, along with various chronic diseases. Thus, when planning a bedroom environment for an elderly occupant, the suitability of that space must receive particular consideration. The quality of the environment is determined based on variables related to the indoor environment quality (IEQ), such as the noise, light, temperature-humidity, and indoor air quality (IAQ). In a previous study [1], the present authors examined sleep behavior in the bedroom using Actigraph monitoring devices (Actigraph WGT 3X-BT, USA) and self-diagnosis. Hence, we determined the scale of tossing and turning during sleep, the sleep efficiency (using an applied sleep algorithm [2]), and the subjective sleep satisfaction. In that study, however, the sleep efficiency and sleep satisfaction results for the participants exhibited variations, raising the possibility that variables related to the IEQ may cause tossing and turning during sleep. Therefore, the purpose of the present study is to find a basic direction for creating an optimized sleep environment, through comparative analysis to determine the relationship between sleep behaviors and the bedroom IEQ.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_20

2 Chronic Disease and Indoor-Environment Pollutants

The elderly have low immunity to indoor-environment pollutants. Further, 76% of the participants examined in a previous study were found to have multiple chronic diseases due to physical function degradation. In particular, the influence of indoor-environment pollutants is a very significant concern for elderly residents afflicted by chronic illnesses, because these individuals tend to spend more time indoors than any other age group. Therefore, sensors to measure indoor-environment pollutants are required in order to configure an optimized sleep environment from an integrated viewpoint. Figure 1 illustrates the correlation between the main chronic diseases and indoor-environment pollutants and indicates the items that must be monitored by the sensors to facilitate IEQ measurement [3].

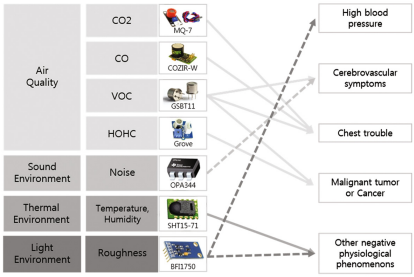


Fig. 1. Correlation between chronic diseases and indoor-environment pollutants

3 Method

The system for monitoring indoor-environment pollutants used in this study comprised sensors, which monitored the items listed on the left of Fig. 1. Figure 2 shows the system configuration for the IEQ-related sensors, with the system architecture being divided into three parts. The first part is comprised of sensors measuring the air, temperature-humidity, sound, and light variations in the monitored environment. The second component contains a processor that retrieves the data collected by the sensors, and the third part is a wireless network controlling data transmission. To achieve low power consumption for the processor and a wireless network, an ATmega328 microcontroller (Atmel, USA) and Xbee ZB series 2 module antennas (Xbee, USA) [3], respectively, were used here.

The sleep behaviors of the targets examined in this study were analyzed using an Actigraph triaxial acceleration sensor (Actigraph WGT 3X-BT, USA), as shown in Fig. 3. The Actigraph is designed to be worn on the ankles, wrists, and waist, to monitor physiological behavior during both sleep and wakefulness. The collected data were analyzed using the Actilife 6 software package. The study participants were a pre-elderly couple living in an apartment facing south. The IEQ-monitoring sensors

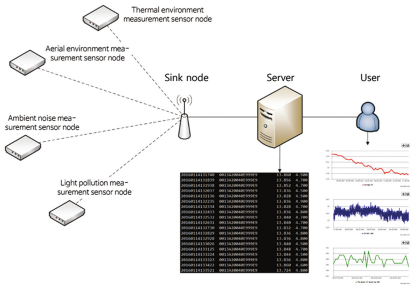


Fig. 2. Sensor system architecture

Fig. 3. Actigraph WGT3X-BT

were positioned beside the bed, and the participants wore the Actigraph devices during bedtime hours for a 4-d period [4].

4 Results

4.1 Bedroom Indoor Environment Quality

Table 1 presents the maximum and minimum values of the data collected over the 4-d period [4]. Compared to the acceptable indoor environment limits by Ministry of Environment (Korea), the NO_2, HCHO, and noise levels were low. On the other hand, the humidity level was significantly higher than the indoor environment limit. The maximum temperature and CO_2 values are above the acceptable levels for an indoor environment. On the third of the four days, the highest noise, humidity, and CO_2 values were obtained.

Table 1. Bedroom indoor environment quality

Measurement/limits	Maximum value				Minimum value			
	Day 1	Day 2	Day 3	Day 4	Day 1	Day 2	Day 3	Day 4
NO_2 (E−11/ppm)/ 0.05–0.3 ppm (h)	1.91E−11	1.91E−11	1.90E−11	1.09E−11	1.70E−12	1.08E−12	1.42E−12	5.73E−13
CO (ppm)/ 10–25 ppm (h)	0.00997551		0.00997551		0.00997551		0.00997551	
HCHO (ppm)/ 100 μg/m^3 (half)	0.029	0.029	0.03	0.03	0	0	0.004	0.002
Noise (dB)/30 dB	3.029	2.794	2.84	2.73	0.316	0.155	0.13	0.12
Temperature (°C)/ 21.2–26.5 °C	28.71	28.68	27.46	27.3	27.76	25.82	26.67	26.11
Humidity (%)/28–40%	55.11	61.51	64.02	54.19	50.26	44.65	55.19	44.37
CO_2 (ppm)/ 1,000 ppm (h)	1610	1180	2530	2320	30	130	1130	1060

4.2 Sleep Behavior in Bedroom

In general, the female subject tossed and turned to a greater extent than the male. The analyzed total sleeping time, time to waking after sleep onset, number of awakenings, and average awakening length were determined from the Actigraph data using sleep algorithms [2]. Overall, the values for the female are high, as shown in Table 2, which presents the results obtained from the data analysis [4]. For the male, the average number of awakenings per day was 9.5, while the average length of each awakening was 2.75 min. The maximum, minimum, and average values of the sleep efficiency were 95.56%, 90.57%, and 93.29%, respectively. On the other hand, for the female, the average number of awakenings per day and the average length of each awakening were 15.25 and 2.3 min, respectively. For the female, the maximum, minimum, and average sleep efficiencies were 97.5%, 87.3%, and 91.05%, respectively. On the third of the four days, both the male and female exhibited the lowest sleep efficiencies.

Table 2. Sleep efficiency and related behavior

Classification	Efficiency (%)		Total minutes in bed (min)		Time to waking after sleep onset (min)		Number of awakenings (no.)		Average awakening length (min)	
	M	F	M	F	M	F	M	F	M	F
Day 1	94.75	88.72	400	390	21	44	8	20	3	2
Day 2	95.56	90.71	360	420	16	39	7	20	2	2
Day 3	90.57	87.3	350	370	33	47	12	12	3	4
Day 4	92.31	97.5	390	400	30	10	11	9	3	1

4.3 Actigraph Data Regarding Sleep Behavior

Figure 4 shows the Actigraph data for the female on the second day, showing the variations in the X-, Y-, and Z-axis data. Figure 5 shows the variation in the temperature and humidity on the second day. Comparison of the variations in the trends in Figs. 4 and 5 for the periods of 3:30–4:00 a.m. and 7:00–8:30 a.m. indicates that the temperature and humidity varied sharply and that the X, Y, Z-axes exhibit significantly higher activity compared to the other time zones.

Figure 6 shows the variations in the X-, Y-, and Z-axis data for the male on the third day, while Fig. 7 shows the variations in the temperature, humidity, and CO_2 on the same day. Comparison of the variations illustrated in Figs. 6 and 7 for the period of 1:30–3:00 a.m. suggests a correlation. That is, the temperature, humidity, and CO_2 values fluctuated dramatically, while higher values were obtained for the X-, Y-, and Z-axis trends compared to the other time zones. Note that, on the third day, the highest CO_2 value was recorded, at 2,320 ppm. Overall, sharp changes in the variables related to the bedroom IEQ influenced the X-, Y-, and Z-axis values of the targets.

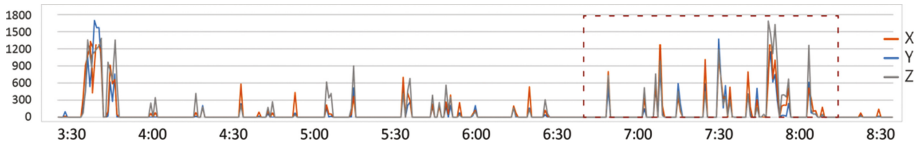


Fig. 4. Variations in X-, Y-, and Z-axis trends for female on second day.

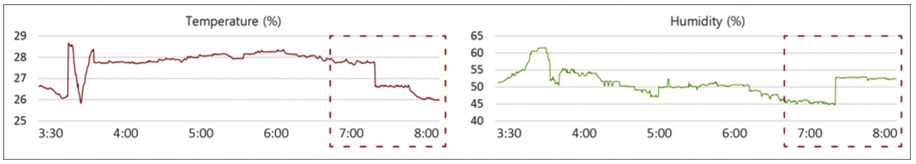


Fig. 5. Temperature and humidity variations for female on second day.

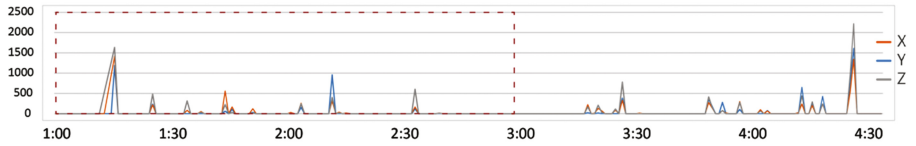


Fig. 6. Variations in X-, Y-, and Z-axis trends for male on third day.

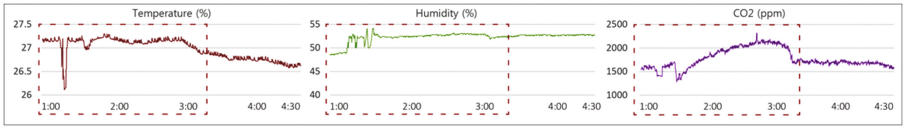


Fig. 7. Temperature, humidity, and CO_2 variations for male on third day.

5 Conclusion

This study explores the influence of variables related to indoor environment quality (IEQ) in the bedroom, which have been determined in previous studies [1, 3, 4], on the sleep behaviors of pre-elderly participants. These data are collected with the aim of creating an optimized bedroom environment. Note that this study focuses on the variations in the IEQ values as determined by IEQ-measuring sensors rather than accurate measurement. The variations in the IEQ values are compared with variations in the subjects' sleep behaviors, which are determined using an Actigraph wearable-device monitoring system.

Bedroom environment planning for the elderly and pre-elderly requires particular attention, because such individuals may suffer from difficulties affecting their physical abilities and health. Thus, this study provides basic data facilitating optimization of the

bedroom environment, with the aim of developing a remodeling plan involving the application of advanced technology.

Overall, the female participant exhibited higher sensitivity to changes in the indoor environment compared to the male, and presented lower sleep efficiency. This result demonstrates the need to develop technology to monitor IEQ integrally and to properly control the IEQ using smart home applications. Such technological advancement is crucial for the development of housing to satisfy the needs of the actively aging society. Note that this study was conducted as a preliminary investigation. Future research based on the findings of the present investigation should seek to provide improved suggestions for IEQ optimization through analysis of a greater number of study participants and a supplemented IEQ sensor system.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, (NRF-2015R1C1A1A02037539).

References

1. Lee, E.J., Park, S.J.: The analysis of sleep behavior for the pre-elderly by using actigraph. In: 2016 Proceedings of KIID Spring Conference, pp. 215–218. Korean Institute of Interior Design, Korea (2016)
2. Cole, R.J., Kripke, D.F., et al.: Automatic sleep/wake identification form wrist activity. Sleep **12**(5), 461–469 (1992). American Sleep Disorders Association and Sleep Research Society
3. Choi, H.K., Park, S.J.: The system configuration of indoor environment sensors for the elderly diseases. In: 2016 Proceedings of Spring Conference, pp. 225–229. Korean Housing Association, Korea (2016)
4. Lee, E.J., Park, S.J.: Bedroom environments and sleeping patterns of pre-elderly. In: 2016 Proceedings of Autumn Conference, pp. 1188–1189. Architectural Institute of Korea, Korea (2016)

Grayscale and Halftone Gel Lithography as Promising Techniques for Swelling-Induced Deformation of Smart Polymer Hydrogel Films

Myunghwan Byun(✉)

Department of Advanced Materials Engineering,
Keimyung University, Daegu, South Korea
myunghbyun@kmu.ac.kr

Abstract. Differential swelling of spatially patterned gel sheets offers an indirect avenue to rational understanding how things in nature grows as time elapses. In the present study, we demonstrate swelling-driven deformation of spatially designed gel sheets sensitively relying on film thickness and pattern dimension. Through chemical copolymerization of poly (N-isopropyl acrylamide) (PNIPAm) with pendent benzophenone UV-crosslinkers, photo-crosslinkable hydrogels were prepared. Various kinds of spatially designed features with different equilibrium degrees of swelling are created by grayscale and halftone gel lithography techniques, wherein time-sequential UV exposures with a number of photomasks, thus selectively embedding densely-crosslinked features into a lightly-crosslinked area. Deformation of the photo-patterned gel sheets by energy competition between the regions strongly depends on a characteristic dimension and sheet thickness, providing fruitful information on the contrast in modulus between the regions.

Keywords: Differential swelling · Smart polymer hydrogels · Grayscale and halftone gel lithography

1 Introduction

Smart polymer hydrogels have attract considerable interest in light of their ability undergo significant volumetric deformation in direct response to externally applied stimuli or environmental changes as well as their biomimetic characteristic. Stimuli-responsive structures of smart polymer hydrogels can be dynamically induced by non-uniform growth, which can be recognized as a basic rule for growth of natural thing. It is generally acknowledged that elastic thin films with laterally differential degrees of swelling or growth rate have strong tendency to spend the elastic energy by choosing the shapes of wrinkles or buckles, which are as a close function of the film thickness [1–4]. Therefore, there is still a big challenge to overcome to rationalize why

This research was supported by the Keimyung University Research Grant of 2016.

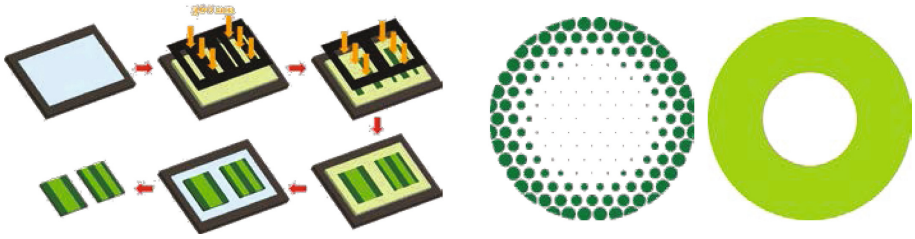
J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_21

non-uniform swelling and why the designing of transforming surfaces are both of critical importance in smart polymer hydrogels.

In the present study, we explore swelling-mediated deformation of thin hydrogel films strongly depending on both film thickness and feature dimension. By spatially designing and embedding the photo-crosslinked areas comprising high- and low-swelling rate, we place emphasis on understanding how in-plane stress arising in between two distinctive areas in terms of crosslinking density is relaxed to affect the swelling-driven deformation of the patterned hydrogel films. Our approach can provide a simple and robust pathway to understanding of the principles of differential growth, which is very similar with a basic rule for natural things.

2 Materials and Methods

UV-crosslinkable hydrogels were chemically synthesized through a free-radical copolymerization process. For this chemical synthesis, N-isopropylacrylamide (NIPAm; 1.5 g), acrylic acid (AAc; 57.1 μL), acrylamidobenzophenone (BPAm; 266.5 mg), fluorescein isothiocyanate (FITC; 20.2 mg), and azobisisobutyronitrile (AIBN; 2.5 mg) as a free radical initiator were used. Compositions of the final material, poly(NIPAm-co-BPAm-co-AAc-co-FITC), were 87.3, 7.0, 5.5, and 0.2 mol%, respectively. Figure 1 fully describes both grayscale and halftone gel lithography processes.



(a) Grayscale Lithography (b) Halftone Gel Lithography

Fig. 1. Schematic for grayscale and halftone gel lithography techniques to get the photo-patterned hydrogel films (a) grayscale lithography using two alternative photomasks for embedding highly crosslinked regions (dark green) into a lightly crosslinked region (light green) or vice versa. (b) halftone gel lithography by varying sizes of highly crosslinked dots embedded into a lightly crosslinked area.

Thickness of smart hydrogel films was fixed at approximately 12 μm. Figure 2 shows areal swelling of uniformly crosslinked hydrogel disks exposed by different UV doses. Areal swelling ratios of the UV crosslinked hydrogel disks decrease nonlinearly as a function of UV dose. Two distinctive sets of swelling ratios (Ω), 3.6 and 2.2, 4.8 and 2.1, were applied for fabricating the samples (i.e., lower values are the swelling ratios of the highly-crosslinked areas and higher values are the swelling ratios of the lightly-crosslinked areas).

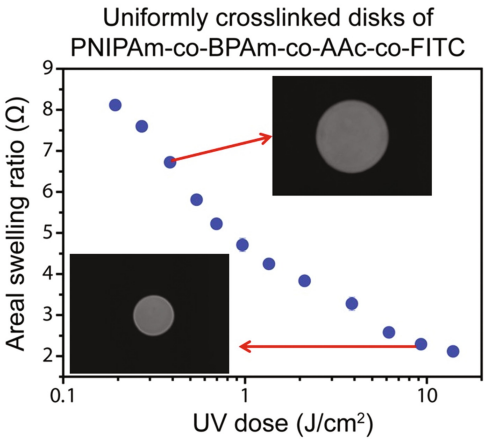


Fig. 2. Areal swelling ratio of the crosslinked hydrogel films. Areal swelling plot showing nonlinear relationship between areal swelling ratio and UV dose.

3 Results and Discussion

Single hinges were manufactured by the sequential UV exposure with alternative two photomasks. Figure 3 is a set of fluorescent microscopy images showing the strong dependence of bending angles on the sample dimension including width and length. Bending angles of the single hinges also depend on thickness of the sample.

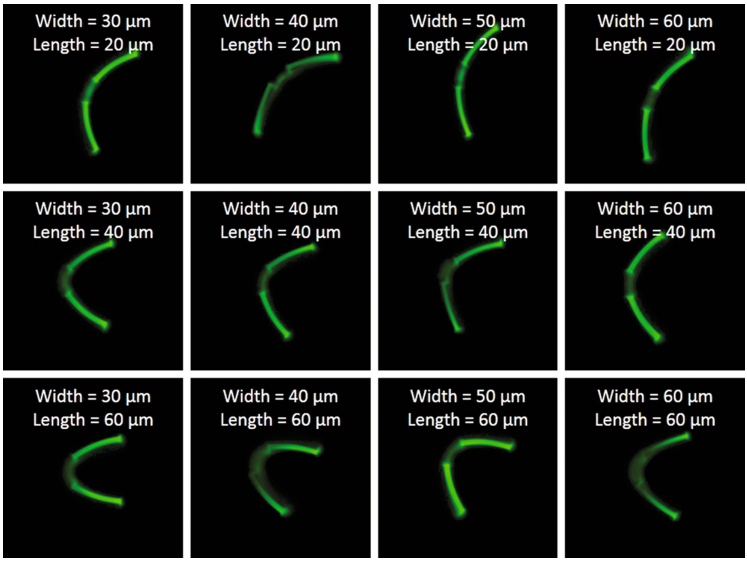


Fig. 3. Bending deformation of the single hinges. A series of the fluorescent microscopy images of the bent single hinges.

(a) Plane view (b) Flank view

Fig. 4. Confocal microscopy images of the cylindrical front (a) plane view of the cylindrical front. (b) flank view of the cylindrical front. The scale bars are all 500 μm.

Turning our attention to halftone gel lithography, transformation of two dimensional (2D) hydrogel disk shape to three dimensional (3D) ring shape was observed through the confocal microscopy. In order to get this transformation, we first set the outer/inner radii corresponding to 195 μm and 90 μm and dot spacing corresponding to 30 μm. Two distinctive swelling ratios for the highly- and lightly crosslinked areas, $\Omega_{low} = 2.1$ and $\Omega_{high} = 8.1$, were applied for fabricating the sample. (Fig. 4).

4 Summary

In summary, we successfully explored grayscale and halftone gel lithography techniques as simple and robust tools for obtaining swelling-driven deformation of the temperature-responsive hydrogel films with spatial patterned crosslinked areas. These strategies possibly offer versatile platforms for understanding the mechanics regarding shape transition of smart hydrogel materials systems.

Acknowledgement. We gratefully acknowledge support the Keimyung University Research Grant of 2016.

References

1. Byun, M., Santangelo, C.D., Hayward, R.C.: Swelling-driven rolling and anisotropic expansion of stripped gel sheets. Soft Matter **9**, 8264–8273 (2013)
2. Kim, J., Hanna, J., Byun, M., Santangelo, C.D., Hayward, R.C.: Designing responsive buckled surfaces by halftone gel lithography. Science **335**, 1201–1205 (2012)
3. Na, J.-h., Bende, N.P., Bae, J., Santangelo, C.D., Hayward, R.C.: Grayscale gel lithography for programmed buckling of non-euclidean hydrogel plates. Soft Matter **12**, 4985–4990 (2016)
4. Silverberg, J.L., Na, J.-h., Evans, A.A., Liu, B., Hull, T.C., Santangelo, C.D., Lang, R.J., Hayward, R.C., Cohen, I.: Origami structures with a critical transition to bistability arising from hidden degrees of freedom. Nat. Mater. **14**, 389–393 (2015)

Methodology for Improving Usability Model of Multiple Devices

Jeyoun Dong[1] and Myunghwan Byun[2(✉)]

[1] Daegu-Gyeongbuk Research Center, Electronics and Telecommunications Research Institute (ETRI), Daegu, South Korea
jeyoundong@gmail.com
[2] Department of Advanced Material Engineering, Keimyung University, Daegu, South Korea
myunghbyun@kmu.ac.kr

Abstract. Recently usage of multiple devices including mobile devices, TVs, OTTs, game consoles, have been extremely increased, allowing multi tasks to users. In addition, the rapid increase of connectivity between the devices has occurred. However, the issues of limited connectivity, battery (power) consumption rates, different remote controls and limited input modalities give rise to uncomfortable experience for users. The more usage of devices increase, the better users require usability. The present study describes which factors and attributes are most relevant to improve usability. We place emphasis on designing methodologies of pilot study to enhance usability thorough data collection and analysis. As a result, we define usability model of multiple devices by those factors and attributes.

Keywords: Usability model · Mobile devices

1 Introduction

Mobile technological development has enabled to use for diverse purpose by people on the move. Moreover, need for the connected usage between the devices as well as mobile device has rapidly increased. However, some of the issues including limited connectivity, battery (power) consumption rates, one or more remote controls and limited input modalities provide the device users with inconvenience.

To date, there have been numerous studies focusing on mobile usability. They have indicated that cognitive overload can be important aspect of usability [1] as well as existing attributes. It seems that mobile devices have trouble with sensitive effect of cognitive overload due to multiple task settings and limitation of size. In addition, the existing usability models do not consider this aspect of usability.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2015R1C1A1A01053111).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_22

The present study is sailing for finding possible factors and attributes that directly affect towards improving connected usability between the multiple devices as well as mobile device. Our pilot study is about how to collect datasets and analyze them to come up with the best solution, thus enhancing usability. We establish a correlation between tasks and context of use and identifying the relevant factors for improving usability.

The remainder of this paper comprises as follows: Sect. 2 depicts our new usability model and Sect. 3 states design of our pilot study to identify and clarify relevant factors and attributes for multiple devices. Finally, Sect. 4 concludes this paper and suggests future work.

2 Suggestion of New Usability Model

This chapter depicts the methods to improve usability. In existing models of usability, ISO (International Organization for Standardization) [2] standard outlines three measurable attributes (Effectiveness, Efficiency, and Satisfaction) and Nielsen [3] identified five attributes of usability (Efficiency, Satisfaction, Learnability, Memorability, and Errors). The models mentioned above have focused on traditional systems rather than current devices. It encounters new usability challenges that are difficult to model using traditional models of usability. Therefore, our usability model includes new issues as well as attributes and factors of all existing models. Figure 1 shows a new usability model that includes a number of the issues to be carefully considered of the advent of multiple devices.

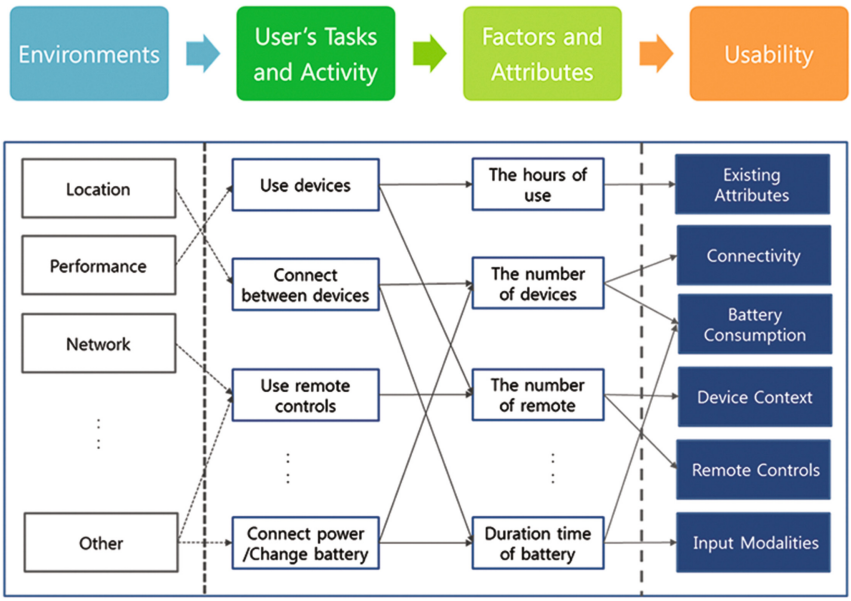


Fig. 1. Conceptual overview for our new usability model

In order for the users to have better connected usability between the mobile devices, self-powered energy harvesting systems are of critical importance in light of ease/portable charging efficiency, dispensing with need for externally applied power sources. Therefore, in the initial design stage of new usability model, self-powered energy harvesting systems should be considered in a harmonious way.

3 Design of Pilot Study

In this section, we design small pilot study about usage behavior of multiple devices to validate and update our usability model. The goal of this study is to know how to use multiple device and which factors and attributes is most important to use devices.

These tasks of pilot study are as follows;

1. The number of devices connecting with mobile device
2. The number of connection failure between the devices
3. The kind of devices which occurs connection failure
4. Continuous time of battery in mobile device
5. The kind of tasks
6. The number of remote controls

For our study, we aim to achieve the following things; 1. The device number that the users want to connect one another, 2. The number of failure in connection between the

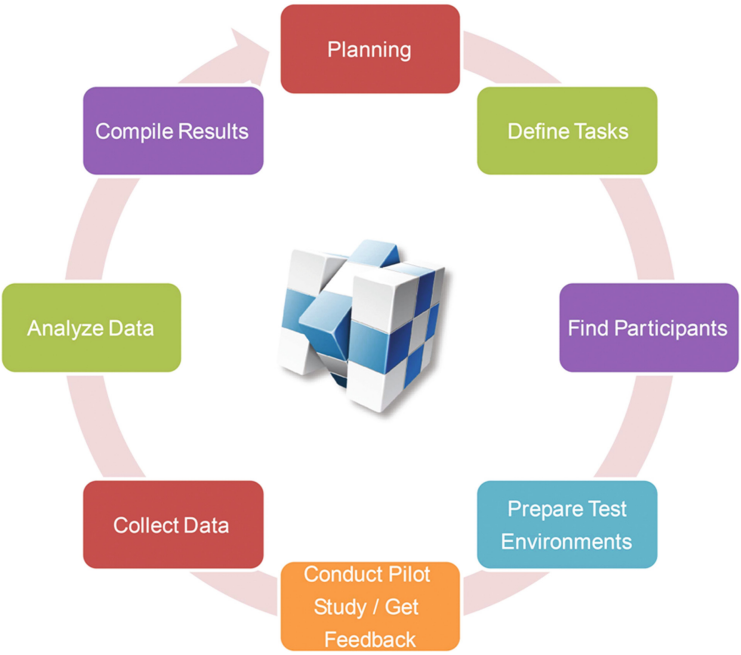


Fig. 2. Process flow chart schematically showing our pilot study

devices, 3. The kind of the device with connection failure, 4. Usage time of the device battery. Based on data and analysis obtained from the corresponding experiments, we place intensive focus on rationalizing what kind of factors benefits connection of multiple devices. Subsequently, the factor that we apply for our proposed new usability model would be useful for supplement or/and overcoming demerits of the existing models (Fig. 2).

4 Conclusions

This paper suggests a new usability model to overcome the issues of existing usability models. For this model, we design pilot study and describe which factors and attributes are most relevant to improve usability. As our future work, we plan to conduct a pilot study and collect data. Based on the results of pilot study, we validate and we improve our new usability model.

Acknowledgements. We gratefully acknowledge support from Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2015R1C1A1A01053111).

References

1. Harrison, R.: Derek flood and david duce “Usability of mobile application: literature review and rationale for a new usability model”. J. Interact. Sci. **1**, 1 (2013)
2. Nielsen, J.: Usability engineering. Elsevier, Amsterdam (1994)
3. ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, International Standards Organization (1998)

Development of SMART Base Isolation Using Linear Motion Guide

Chunho Chang(✉) and Sangyoung Shin

Department of Civil Engineering, Keimyung University, Daegu, South Korea
{chunho,llll140}@kmu.ac.kr

Abstract. This study addresses the design of nonlinear oil-spring damper that render an effective re-centering mechanism for a linear motion (LM) guide base isolation system. A nonlinear stiffening behavior of the damper offers added advantage of re-centering mechanism. The proposed LM guide with damper system works in a similar fashion as that of the LRB isolation system supplemented with re-centering mechanism for small to high level of shaking. For high intensity shaking, the proposed system minimizes the peak horizontal displacement in addition to keeping the residual displacement close to zero. To demonstrate the concept of the proposed base isolation system, a numerical study is conducted with a steel moment-resisting frame when subjected to ground motions of varying hazard levels. It has been found from this study that the proposed LM guide base isolation system is effective in limiting the peak bearing displacement and making the residual bearing displacement negligible for varying hazard levels.

Keywords: SMART base isolation · Linear motion guide

1 Introduction

The seismic design philosophy for infra-structures has undergone a series of modifications based on the performance of constructed facilities under various earthquakes. In the past, the seismic design of a infra-structure was governed by a minimum level of lateral strength requirement. This design concept was modified later by introducing ductility in the design procedures. Earthquakes such as the 1994 Northridge and 1995 Kobe further demonstrated that buildings and bridges designed as per contemporary codes may undergo severe damages and the economic loss arising due to damage of structures and/or loss of functionality may be beyond the acceptable level. Therefore, it was reasoned that the ductility based criteria may alone not be sufficient to guaranty the desired performance of a structure during earthquakes. As an improvement, the performance-based design philosophy has developed. In performance-based design of a building and bridges, a pre-specified level of performance (performance objectives with an acceptable damage level) is envisioned when subjected to an event of specific hazard level FEMA-356 [1], FEMA-349 [2]. As per FEMA-356 [1], (i) for a minor intensity shaking, normally, a negligible structural damages without hindering the functionality of the system is acceptable and the desired performance level is known as 'Immediate Occupancy'; (ii) in case of moderate shaking, the damage includes major structural

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_23

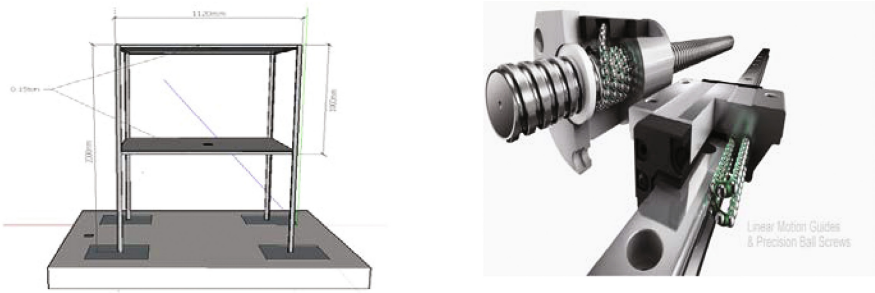
damage without collapse, minimal falling hazard and the performance level is known as 'Life safety' [1]; (iii) for a strong shaking the extent of damage can go up to a severe structural damage with probable falling hazard and the performance is the 'collapse prevention'. For a infra-structure, these performance objectives can be achieved by proper design that may include application of smart devices such as isolation bearings.

This research evaluate the seismic performance of smart base isolation system. This paper is organized as follows: Sect. 2 introduce analysis procedures. Finally, Sect. 3 concludes this paper.

2 Analysis of Isolated Structures

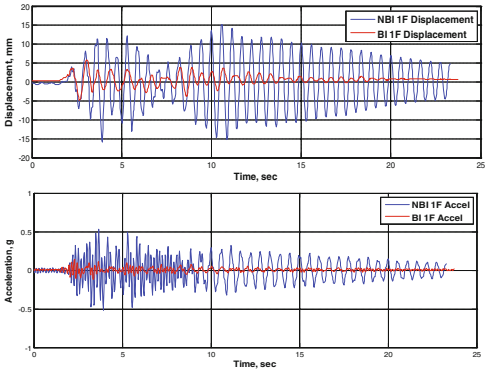
A two-story one-bay base steel moment-resisting frame (SMRF) using linear motion guide is considered for the analysis. Figure 1(a) and (b) respectively, show the plan and the elevation of the building considered in the study. The building consists of Korea standard steel sections with a uniform mass distribution and a non-uniform distribution of lateral stiffness over the height. The overall floor plan dimension of the building is shown in Fig. 1(a). The element stiffness (beams and columns) is considered in such a way that the fundamental period of the building can represent that of the existing steel buildings of similar dimensions in Korea. The fundamental period of the building is 0.515 s. Additionally, except the beam-to-column connections at the top floor, the steel sections of structural members are selected in such a way that the sum of the plastic section modulus of the columns framing into each beam-column joint was higher than the sum of plastic section modulus of the beams framing into the same joint. The building model is developed in Midas software for the selected frame. The beams and columns are assigned with various angle.

For numerical simulation of the dynamic response of the hybrid base isolation system that is introduced in this paper, a MIDAS finite element method analysis is used. This approach is advantageous in that a Kelvin model of the LM Guide and damper. Experimental and simulated responses for a controlled hybrid base isolation system using oil-spring damper and LM guide are compared in Fig. 2. Relative base

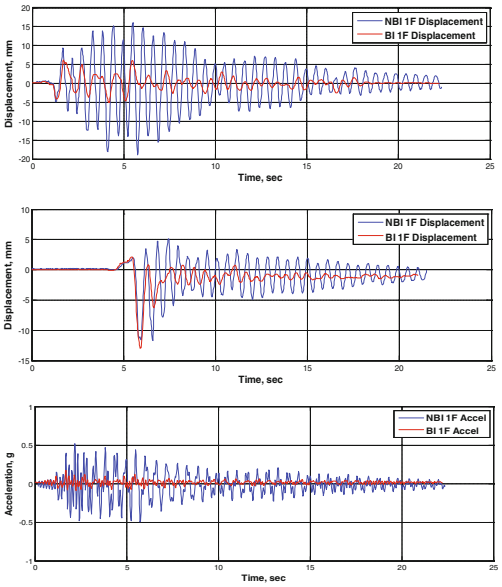


(a) Configuration (b) Linear Motion Guide

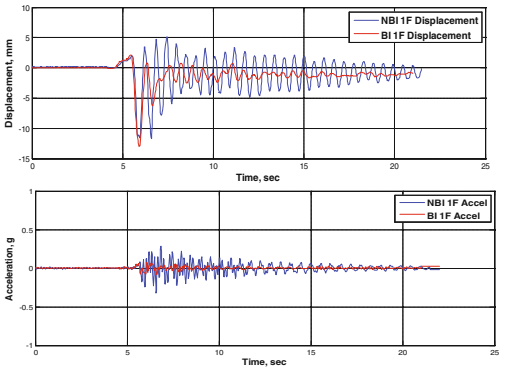
Fig. 1. The configuration and finite element modeling of the bridge in Daegu



(a) Morgan Hill 1984 1F



(b) North Palm Spring, 1986 1F



(c) Kyungjoo 2016

Fig. 2. Acceleration response result

acceleration of the isolated structure are plotted for the Morgan Hill, North Palm Spring and Kyngjoo earthquakes with PGA levels of 0.5, 0.51 and 0.18 g. Note the very good agreement between results predicted in the simulation and those obtained in the experiment. This indicates that the suggested models can accurately represent behavior of the oil-spring damper and LM guide of the hybrid base isolation system. It can be seen that the discrepancy between the predicted and experimental displacements increases with the larger PGA levels.

3 Conclusions

In this paper, smart hybrid base isolation has been proposed as an effective re-centering device for new type base isolation system. The oil-spring damper shows a stiffening nonlinear behavior with displacement while providing a better lateral buckling resistance. The proposed system has the advantages of hybrid base isolation (with weak restoring mechanism) for small to medium level of shaking and minimizes the peak bearing displacement (PBD) for high intensity shaking. In order to show the effectiveness of the proposed isolation system, a numerical study is conducted with a 2-story one-bay steel moment-resisting frame along with a series of US and Korea ground motions representing different hazard levels. Finally, new smart hybrid base isolation for obtaining seismic design is developed in this work.

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-2014R1A2A1A11052469).

References

1. FEMA-356. Prestandard and commentary for the seismic rehabilitation of buildings. Technical report. Federal Emergency Management Agency (2000)
2. FEMA-349. Action plan for performance based seismic design. Technical report. Federal Emergency Management Agency (2000)

Analysis on Work Zone Characteristics in South Korean Expressways Using Text Mining Technique

Oh Hoon Kwon[1], Je-Jin Park[2], and Shin Hyoung Park[1(✉)]

[1] Department of Transportation Engineering, Keimyung University, Daegu, South Korea
{ohoonkwon, shpark}@kmu.ac.kr
[2] Transportation LabKorea Expressway Corporation Research Institute, Hwaseong-si, Gyeonggi-do, South Korea
jjpark@ex.co.kr

Abstract. The study proposes the association between work zone types and work zone lane closure types used in the construction of South Korean expressways through the use of the text mining and association analysis techniques. Of the factors that greatly affect the capacity of work zones, *work zone lane closure types* and *night work* were analyzed. Words were extracted from the narrated work types using the text mining technique, and the associations among the words, *lane closure types*, and *night work* variables were identified. The analysis revealed a significant difference between the work types associated with main lane closure and shoulder lane closure. It is expected that this study can be effectively applied to establishing transportation management plans (TMPs) for work zones.

Keywords: Work zone type · Work zone lane closure · Text mining · Association analysis

1 Introduction

Expressways are constantly subjected to various works to repair pavements damaged by deterioration, install a wide range of safety facilities, and carry out traffic flow improvement projects. In South Korea, in particular, a constant stream of repair projects is expected to take place to deal with the issue of deteriorating expressways. Such works on expressways greatly disrupt traffic flow, causing congestions and economic losses.

The relevant laws provide that each work should be performed based on transportation management plans (TMPs) to minimize the damage caused by the work zones. In the United States, the Federal Highway Administration (FHWA) [1] provides guidelines for TMPs, and each state publishes its own version of the guidelines based on FHWA guidelines. As for South Korea, while the Korea Expressway Corporation publishes its own TMP manual for expressway works, the country's TMP scheme is still in need of vast improvements.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_24

TMP guidelines normally divide TMPs into (a) Basic TMPs, (b) Intermediate TMPs (or Minor TMPs), and (c) Major TMPs in terms of work type, work intensity, work duration, length of work zone, and lane closure type. As the effect on traffic flow in the work zone increases, TMP moves from Basic TMP to Major TMP. Therefore, it is crucial to determine which TMP to apply. This determination should be done by information on work zone type, intensity, work zone length, work duration, and lance closure type.

This study analyzed the association between work zone types and lane closure types using big data on expressway work zones. Words were extracted from the narrated work types in the work zone data collected by Korea Expressway Corporation using the text mining technique, and the association rules among the words, *lane closure types*, and *night work* variables were identified.

2 Literature Review

The FHWA of the United States manages its strategies for management of impact from work zones in several aspects: Temporary Traffic Control, Public Information, and Transportation Operations. Basic TMPs are applied to work zones with minimal disruption to the traveling public and adjacent businesses and community; Intermediate TMPs are applied to work zones that are anticipated to have more than minimal disruption but have not been identified as significant projects; and Major TMPs are applied to large-scale projects with significant impacts such as multiple lane closures or total closure of an important corridor in an urban area.

California Department of Transportation (DOT) [2], Colorado DOT [3], and Oregon DOT [4] also publish TMP guidelines based on the FHWA guidelines. These latter guidelines also divide TMPs into three categories by work type, work duration, and lane closure type. South Korea's Ministry of Land, Transport and Maritime Affairs [5], on the other hand, publishes Korea Highway Capacity Manual (KHCM), which divides work types into bridge repair, road surface repair, facility repair, and lane painting. Each type is applied with different capacity values. The findings of this study are expected to be applicable to setting the criteria for determining TMP categories to apply by looking into which work types are associated with which lane closure types.

3 Data Description and Analysis Methods

The analysis was performed on the data for short-term work zones (less than 12 h) which were collected by the Korea Expressway Corporation. Unrealistic values were excluded as data record errors, and 20,000 work zones from the remaining 43,807 work zones were randomly selected for the analysis. The variables used for the analysis was *night work*, *shoulder lane closure*, *main lane closure*, and *work contents*, as shown in Table 1. Night work is defined as works that began at or after 18:00.

To derive work types from the *work contents*, key Korean words were extracted using the R KoNLP package [6] and tm package [7], which were then applied to the text mining analysis. The Apriori algorithm [8] was employed for association analysis between work types and lane closure types using R arules package [9].

Table 1. Summary of work zone data.

Variable type	Variable name	Value and statistics
Night work	*night work*	0 (day work): 17,788 (88.9%) 1 (night work): 2,212 (11.1%)
Type of lane closure	*shoulder lane closure*	0 (lane open): 6,295 (31.5%) 1 (lane closure): 13,705 (68.5%)
	main lane closure	0 (all of lanes open): 7,528 (37.6%) 1 (any of lanes closure): 12,472 (62.4%)
Type of work zone	*work contents*	Narrated work contents

Text mining. Text mining is a technique designed to identify valuable and meaningful information from irregular text data. In comparison with simple data search, the text mining provides findings by extracting significant information from a vast pool of information using complicated statistical algorithms, identifying its association with other information, and discerning the categories within the texts.

Association analysis. Association analysis is to discover interesting relations between variables in large databases. Apriori algorithm, which is most widely used in association analysis, finds association rules from frequent item-sets using indicators such as *support*, *confidence*, and *lift*. *Support* is an indication of how frequently the item-set appears in the data set, $support(A \rightarrow B) = P(A \cap B)$. *Confidence* is an indication of how often the rule has been found to be true, $confidence(A \rightarrow B) = P(B \mid A)$. *Lift* is a ratio of confidence to the probability that the consequent (B) occurs in the overall data set, $lift(A \rightarrow B) = P(B \mid A)/P(B)$. If the *lift* is larger than 1, it means that the rule is potentially useful for predicting the consequent in future data sets.

4 Findings

Of the most frequently identified words from the *work contents* of 20,000 work zones, the words related with work types are listed in Table 2. The most frequent words were those related with recurring everyday tasks, such as cleaning and maintenance,

Table 2. Top frequent items (words) related to type of work zone.

Frequent items	Frequency	Proportion
Cleaning	1,343	6.7%
Shoulder lane	1,325	6.6%
Maintenance	1,034	5.2%
Routine	883	4.4%
Pavement	704	3.5%
Painting	613	3.1%
Median strip	561	2.8%
Lane line	547	2.7%
Tunnel	538	2.7%
Light bulb	499	2.5%

followed by the words related to facilities requiring routine maintenance such as tunnels, median strips, pavements, and lightbulbs.

Table 3 shows the top association rules that are ranked by *lift*. Work types associated with main lane closure were works related with landscaping, pavement, night work, and median strip. Of those words, landscaping had a *confidence* value of 1, always causing main lane closure. Night work and main lane closure showed *support* values of around 0.1 (10% of all works), which indicates that main lane closures in night works frequently occurred. The high *confidence* value (*confidence*({night work} => {main lane closure}) = 0.95) means that most night works led to main lane closure. In addition, pavement work is 3.5 times more likely to involve night works as compared with proportion of night works in the overall work zones (*lift*({pavement} => {night work}) = 3.52). Works associated with shoulder lane closure were frequently recurrent

Table 3. Top ranked association rules by *lift* for types of work zone and lane closure.

Category	Association rules	*support*	*confidence*	*lift*
Main lane closure	{landscaping} => {main lane closure}	0.010	1.000	1.604
	{night work, pavement} => {main lane closure}	0.014	0.985	1.580
	{concrete, pavement} => {main lane closure}	0.010	0.967	1.550
	{median strip} => {main lane closure}	0.027	0.955	1.532
	{night work} => {main lane closure}	0.105	0.953	1.528
Shoulder lane closure	{maintenance, routine} => {shoulder lane closure}	0.032	1.000	1.459
	{landscaping} => {shoulder lane closure}	0.010	1.000	1.459
	{collecting refuse} => {shoulder lane closure}	0.022	0.991	1.446
	{cleaning} => {shoulder lane closure}	0.064	0.960	1.401
	{weeding} => {shoulder lane closure}	0.014	0.955	1.393
Night work	{pavement} => {night work}	0.014	0.389	3.519
Bridge	{bridge} => {shoulder lane closure}	0.018	0.756	1.104
	{bridge} => {main lane closure}	0.013	0.537	0.861
Tunnel	{tunnel} => {main lane closure}	0.023	0.842	1.350
	{tunnel} => {shoulder lane closure}	0.011	0.407	0.594
Road surface	{road surface} => {main lane closure}	0.018	0.800	1.283
	{road surface} => {shoulder lane closure}	0.015	0.670	0.977
Lane line painting	{painting, lane line} => {main lane closure}	0.017	0.866	1.388
	{painting, lane line} => {shoulder lane closure}	0.014	0.706	1.031

everyday works such as routine maintenance, landscaping, cleaning, and weeding. Of those works, landscaping was found to always cause shoulder lane closure as well as main lane closure.

KHCM [5] divided work types into bridge repair, road surface repair, facility repair, and lane painting. With regard to the categorization, this study looked into their associations with lane closure types. The findings showed that bridge-related works had a *lift* value of 1 or higher in terms of the rule with shoulder lane closure (*lift* ({bridge} => {shoulder lane closure}) = 1.10). On the other hand, tunnel-related works showed a *lift* value of 1 or higher in terms of the rule with main lane closure (*lift* ({tunnel} => {main lane closure}) = 1.35). In addition, road surface and lane line painting showed higher *lift* and *confidence* values with a main lane closure than with a shoulder lane closure (*lift*({road surface} => {main lane closure}) = 1.28, *lift*({painting, lane line} => {main lane closure} = 1.39). Thus, when setting TMPs for tunnel repair, road surface repair, and lane line painting, the impacts of main lane closure need to be taken into more consideration.

5 Discussion and Conclusions

This study analyzed the association between work types and lane close types using a vast amount of work zone data. Lane closure was divided into two types: main lane closure and shoulder lane closure. As for work types, related words were extracted using the text mining technique. The findings showed that recurrent everyday works tend to cause more shoulder lane closures, while tunnel repair, night work, pavement, median strip, road surface, and lane line painting were more associated with main lane closure. This study analyzed the characteristics of work zones from large-scale text data using the data mining technique. This technique is expected to be applicable to establishing TMPs.

This study is limited by the failure to include the number of closed lanes, although it represents a significant factor affecting the traffic flow. In addition, as the data were compiled from records of working-level staff, they may include misentries by human errors, or data based on different data entry criteria. These shortcomings will have to be addressed in future research.

Acknowledgement. This research was supported by a grant (16RDRP-B115529-01) from Regional Development Research Program (RDRP) funded by Ministry of Land, Infrastructure and Transport of Korean Government.

References

1. Federal Highway Administration: Developing and Implementing Transportation Management Plans for Work Zones. U.S. Department of Transportation (2005)
2. California Department of Transportation: Transportation Management Plan Guidelines. Division of Traffic Operations, Office of Traffic Management, California Department of Transportation (2015)

3. Colorado Department of Transportation: Work Zone and Mobility Rule Procedures Document. Colorado Department of Transportation (2014)
4. Oregon Department of Transportation: Transportation Management Plan Project Level Guidance Manual. Oregon Department of Transportation (2013)
5. Ministry of Land, Transport and Maritime Affairs: Korea Highway Capacity Manual. Ministry of Land, Transport and Maritime Affairs (2013)
6. Jeon, H., Kim, T.: KoNLP: Korean NLP package. R package version 0.80.0 (2016)
7. Feinerer, I., Hornik, K.: tm: Text Mining package. R package version 0.6–2 (2015)
8. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
9. Hahsler, M., Buchta, C., Gruen, B., Hornik, K., Borgelt, C.: arules: Mining Association Rules and Frequent Itemsets. R package version 1.5–0 (2016)

Seismic Performance Evaluation of a Prestressed I-Type Girder Bridge in Daegu for ICT Based Disaster Management in Daegu Metropolitan City

Chunho Chang[1(✉)], Sung Jig Kim[2], and Shin Hyoung Park[3]

[1] Department of Civil Engineering, Keimyung University, Daegu, South Korea
chunho@kmu.ac.kr
[2] Department of Architectural Engineering, Keimyung University, Daegu, South Korea
sungjig@kmu.ac.kr
[3] Department of Transportation Engineering, Keimyung University, Daegu, South Korea
shpark@kmu.ac.kr

Abstract. One of the most typical bridges in the Province of Daegu (Korea) is prestressed I-type girder bridges. Due to the lack of seismic detailing, this bridge may be vulnerable to earthquake events. In an effort to prevent an interruption of the transportation network, which could be catastrophic for Daegu, a comprehensive research program was conducted whose aims were to assess the seismic vulnerability of typical bridge classes. The understanding of the impact of various modeling parameters on structural-component responses is the first step in a forward-vulnerability assessment of bridges in Daegu. This paper presents part of a comparative study using detailed finite element nonlinear models that was conducted to assess the longitudinal and transverse responses of multi-span continuous prestressed I-type girder bridges in their as-built configurations using code-based loading of Korea. Deterministic responses in terms of column curvature demand, abutment footing deformations are provided for the longitudinal and transverse directions of bridge models.

Keywords: Seismic performance · ICT based disaster management

1 Introduction

In Korea, the roadway system is very critical for the transportation of people and goods and for the nation's economy. Hence, its interruption after a seismic event could be catastrophic. Furthermore, recent earthquakes have shown that bridges [1–3] are the most vulnerable element in a roadway network system. About 64% of the bridges in Daegu were built more than 20 years ago and, consequently, they were not designed according to important seismic regulations and detailing adopted in the 1992, and later. Moreover, the region's seismic-hazard level was increased in the last edition of the Roadway Code of Korea with the use of 10% probability of exceedance in 50 years.

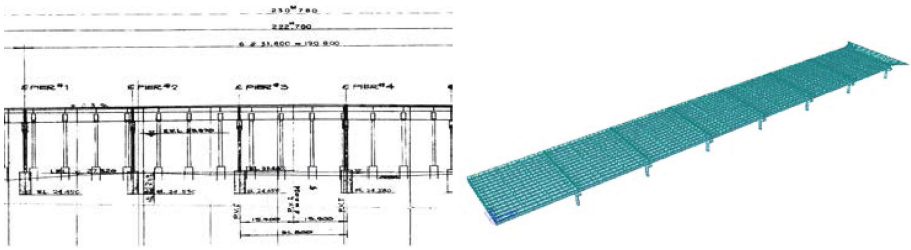
J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_25

This underlines the need to develop reliable procedures for the evaluation of the structural capacity of bridges [4] in Daegu, before and after an earthquake, and to evaluate effective techniques for retrofitting deficient bridges [4]. In recent years, the use of statistical tools for estimating the extent of probable bridge damage and loss assessment under a range of hazard levels has increased around the world using empirical data or analytical approaches.

This research evaluates the seismic performance of the most typical bridge in Daegu. This paper is organized as follows: Sect. 2 introduce the bridge specification and analysis procedures. Finally, Sect. 3 concludes this paper.

2 Seismic Performance Evaluation

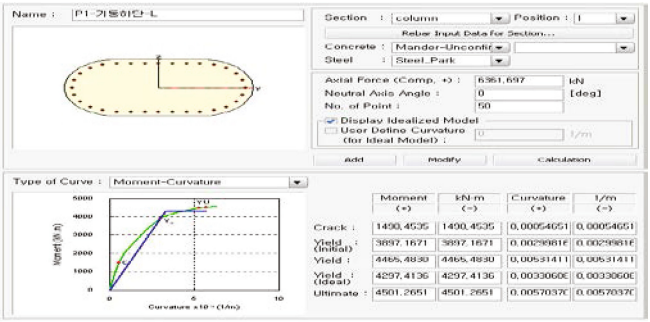
Based on data from the inventory of Daegu's multi-span prestressed I-typed continuous bridges, typical model representing as-built bridge configurations is evaluated. This bridge models are composed of elements that may exhibit highly nonlinear behavior, such as columns, abutments and elastomeric bearings. The portfolio of prestressed I-type bridges comprises a large number of structures 20 years old or more. This bridge was designed with a lack of seismic detailing implemented in recent design codes. For each bridge class, eight geometric configurations are defined to be representative of real-bridge configuration and finite element modeling are presented in Fig. 1.



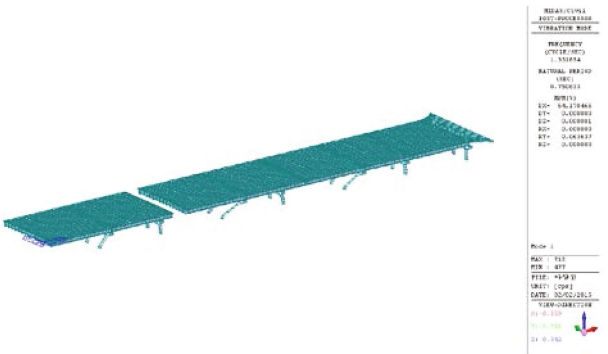
(a) Configuration (b) Finite Element Modeling

Fig. 1. The configuration and finite element modeling of the bridge in Daegu

The moment-curvature result of this bridge (longitudinal direction) as in Fig. 2 Based on this result, we can calculate the section strength and demand-displacement-ductility of the pier. We also can get the important mode-shape of the bridge. In Fig. 2(b), you can see the mode shape along longitudinal direction (t = 0.7508 s). Based on the Korea code, the seismic performance evaluation of the bridge is as follow in Table 1. As you can see, some section capacity is not enough to resist against earthquake. For example, several important aspects such as displacement ductility, shoe force, concrete section force in pier 2 are not good. Now we need seismic retrofit based on this result, and also give a good information to make alternative transportation pass-root in emergency.



(a) moment-curvature relationship



(b)mode shape

Fig. 2. Analysis result (a) moment-curvature relationship (b) mode shape

Table 1. The seismic performance of the bridge in Daegu

Classify				Demand	Capacity	Result	
Fixed pier evaluation (R = longitudinal: 1.138, lateral top = 0.359, lateral bottom = 0.386)							
Displacement ductility			Longitudinal	1.030	1.138	0.906	N.G
Bending moment (kN·m)	Longitudinal		Top of column	7,827	2,812	2.784	O.K
	Transverse		Bot. of column	7,294	2,812	2.594	O.K
Evaluation of shoe at pier							
Pier (P1)	Shoe (kN)		Longitudinal	3,000	2,408	1.246	O.K
			Transverse	3,000	3,468	0.865	N.G
	Anchor bolt	Steel failure	Longitudinal	1,237	80	15.403	O.K
			Transverse	1,237	116	10.695	O.K
		Con. failure	Longitudinal	318	80	3.960	O.K
			Transverse	352	231	1.522	O.K
		Flyout failure	Longitudinal	13,293	2,408	5.520	O.K
			Transverse	13,293	3,468	3.833	O.K

(*continued*)

Table 1. (*continued*)

Classify				Demand	Capacity	Result	
Pier (P2)	Shoe (kN)		Longitudinal	–	–	–	
			Transverse	3,000	6,365	0.865	N.G
	Anchor bolt	Steel failure	Longitudinal	–	–	–	
			Transverse	445	212	2.098	O.K
		Con. failure	Longitudinal	–	–	–	
			Transverse	294	424	0.694	N.G
		Flyout failure	Longitudinal	–	–	–	
			Transverse	13,293	6,365	2.088	O.K

3 Conclusions

The severe earthquake brought an intense impact on the seismic resistant design and technical development of bridge engineering. From the bridge disaster reconnaissance, some valuable lessons were learned to review and improve the current seismic design code. This paper presented a comparative seismic performance evaluation multi-span continuous prestressed I-typed concrete bridge, which are typical in Daegu. This bridge was represented by 3-D analytical models for as-built configurations subjected to a suite of the 10% probability of exceedance in 50 years hazard level. It is necessary to proceed with both the improvement of seismic evaluation skill and retrofit of old bridges. In addition, many verified seismic evaluation techniques should be applied to repair or strengthen old bridges.

Acknowledgement. This research was supported by a grant (16RDRP-B115529-01) from Regional Development Research Program (RDRP) funded by Ministry of Land, Infrastructure and Transport of Korean Government.

References

1. Roy, N., Paultre, P., Proulx, J.: Performance-based seismic retrofit of a bridge bent: design and experimental validation. Can. J. Civil. Eng. **37**(3), 367–379 (2010)
2. Basöz, N.I., Kiremidjian, A.S., King, S.A., Law, K.H.: Statistical analysis of bridge damage data from the 1994 Northridge, CA, earthquake. Earthq. Spectra **15**(1), 25–54 (1999)
3. Shinozuka, M., Feng, M.Q., Kim, H., Uzawa, T., Ueda, T.: Statistical analysis of fragility curves. Report, Multidisciplinary Center for Earthquake Engineering Research, MCEER-03-0002 (2003)
4. Hwang, H.M., Jernigan, J.B., Lin, Y.W.: Evaluation of seismic damage to Memphis bridges and highway systems. ASCE J. Bridge Eng. **5**(4), 322–330 (2000)

Implementation of Sitting Posture Monitoring System with Kinect

Dong-Jun Shin[1], Min-Sang Kim[1], Wook Song[1], Se Dong Min[2], and Min Hong[1(✉)]

[1] Department of Computer Software Engineering, Soonchunhyang Univeristy, Asan-si, Chungcheongnam-do, Korea
{daons, ben399399, wook2735, mhong}@sch.ac.kr
[2] Department of Medical IT Engineering, Soonchunhyang Univeristy, Asan-si, Chungcheongnam-do, Korea
sedongmin@sch.ac.kr

Abstract. Most modern office workers are seated on the chair for working with the computer during most working time. Generally, the correct sitting posture is critical for human life because wrong sitting posture can lead the back or neck disorder related sickness. The majority of previous researches have been utilized various sensors with special chairs to measure the sitting posture. However, these special chairs can be uncomfortable for users due to unfamiliar sitting environment. Therefore, we proposed the sitting posture monitoring system using Kinect which provides IR depth camera to record user's sitting posture data and alert wrong posture to user. The proposed system provided 62% increased correct sitting posture during 35 min for 5 participants. We believe that the proposed system can help to improve the behavior of sitting posture of user without any uncomfortable adhesive sensors.

Keywords: Sitting posture · Kinect · Monitoring system

1 Introduction

Correct sitting posture is a one of crucial factors for measuring the quality of human life. Recent modern office workers have been seated on the chair for working with the computer for most work time in everyday life. Thus sitting posture is critically related to individual health and incorrect sitting posture can cause the back or neck disorder related sickness. However, most people can not recognize the importance of sitting posture. In this paper, we developed the sitting posture monitoring system to correct user's sitting posture using Kinect.

1.1 Effects of Sitting Posture

According to the study of Carney et al. [1], the body posture can increase the male hormone value and can gain self-confidence. In addition, this study introduces that human and other animals express the power through open and expansive postures, and

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_26

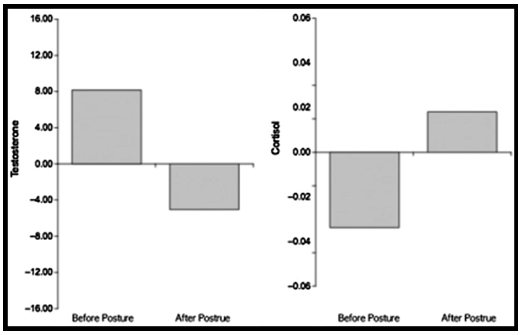


Fig. 1. Mean changes in the dominance hormone testosterone and the stress hormone cortisol following high-power and low-power postures.

they express passive attitude through closed and contractive postures. Thus it means that the posture affects the confidence of human or animals.

Figure 1 shows the value of dominance hormone testosterone and the stress hormone cortisol which has been changed by posture. When the posture is changed correctly, testosterone level was increased about 20% and cortisol level was reduced about 25% which are shown in Fig. 1. The result of this research shows the important influence of the correct posture of human body for health. On the other hand, wrong posture of human body for daily life can lead bad effects for health.

Figure 2 shows the severe deformation of human body such as the forward head posture by incorrect body posture. Most these patients have difficulty to realign incorrect posture because they have unrecognized bad habit for body posture during daily life. In this paper, we proposed the simple sitting posture monitoring system which can provide the notice and daily report for user's sitting posture to rectify incorrect sitting posture. We believe that the proposed sitting monitoring system can improve user's sitting posture effectively.

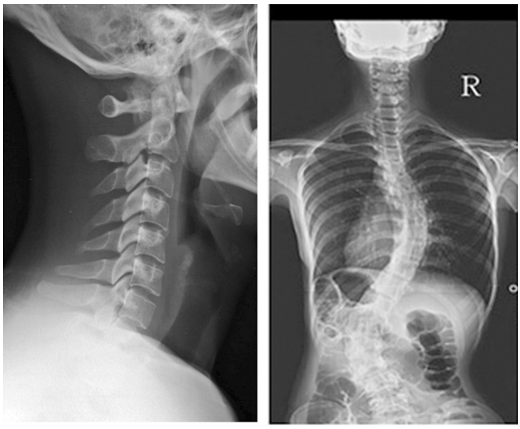


Fig. 2. Deformation of neck bone and backbone by sitting posture

2 System Environments for Proposed Method

The proposed system utilizes Microsoft Kinect V2 to detect user's body posture conveniently. Kinect is a motion sensing input devices by Microsoft for Xbox 360 and Xbox One video game consoles or Windows PC. Kinect provides IR depth camera to figure out the depth information of human body.

Since Kinect V2 provides information of x, y, and z positions of 25 joints which are shown in Table 1, we can readily detect key body joints for sitting posture. In most previous researches for sitting posture monitoring system have been utilized the pressure sensors to detect sitting posture, but it makes users feel uncomfortable due to special sensors on chair. In addition, these approaches only detect the pressures on the heap and can't figure out the detailed sitting posture visually. The proposed sitting posture monitoring system doesn't require any uncomfortable ahherable sensors for posture detection using image processing technology with Kinect V2.

Table 1. Provided joint information from Kinect V2

Joints	Description
AnkleLeft, AnkleRight	Left and right ankle
ElbowLeft, ElbowRight	Left and right elbow
FootLeft, FootRight	Left and right foot
HandLeft, HandRight	Left and right hand
HandTipLeft, HandTipRight	Tip of the left and right hand
Head	Head
HipLeft, HipRight	Left and right Hip
KneeLeft, KneeRight	Left and right knee
Neck	Neck
ShoulderLeft, ShoulderRight	Left and right shoulder
SpineBase, SpinerMid	Base and middle of the spine
SpineShoulder	Spine at the shoulder
ThumbLeft, ThumbRight	Left and right thumbs
WristLeft, WristRight	Left and right wrist

Figure 3 shows the major joints of body such as head, shoulder, forearms. Caneiro et al. [2] introduced that the habit of sitting posture can mainly affect the health of neck and waist. According to this study, the proper sitting posture can be determined by whether the neck and waist are straight or not. Therefore, the proposed sitting posture monitoring system detects 5 key body joints such as head, mid, shoulder, neck, and base and all 3D positions of information are recorded in the monitoring server using PHP and MySQL.

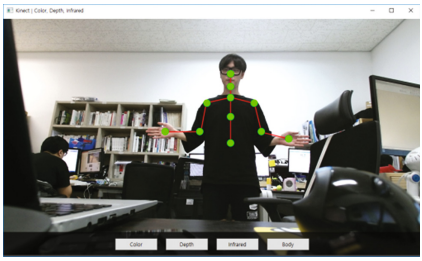


Fig. 3. Detected key joints of body using Kinect V2

3 Definition of Correct Sitting Posture

According to the definition by KABPA [3], the proper posture should minimize the stress and tension, efficiently utilize body at the most, and represents a mental state. Firstly, before determining the sitting posture is proper or not, we need to define which state is correct posture. Generally, the correct sitting posture is defined as Fig. 4.

Fig. 4. The definition of proper sitting posture by Sagewood wellness center [4]

The top of PC screen should be at eye level and user's shoulder should be relaxed. The forearms comfortably are paralleled to the floor. The chair should have a backrest that supports the lower back and user's legs should be bent in a 90 to 110° angle. Lastly, feet should be flatted on the floor. These sitting postures are proper sitting posture on the chair during office work [4]. Accordingly, the proposed monitoring system detects key joints from head, neck, shoulder, chest, and belly in order to properly detect the correct sitting posture. To determine the proper sitting posture with 5 key joints, the proposed monitoring system measures the average of angles between key joints according to reference posture in Fig. 4 from 5 participants. The 4 angles

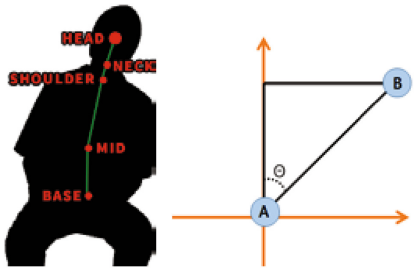


Fig. 5. Detected 5 key joints from Kinect and measurement of angle between two joints

between head and neck, neck and shoulder, shoulder and mid, and mid and base which are shown in Fig. 5 are measured for each participant. Table 2 shows the average values of these angles and we applied these average angles as reference value to determine the correct posture.

We set up the threshold within 15° from the experimental references which are shown in Table 2 as proper angle between key joints. In Fig. 5, the line that is linked between mid and shoulder is somewhat tilted to the right. The proposed method calculates the angle between two key joints in Fig. 5 with Eq. (1). For the two key joints such as A and B, we can make a triangle with the two points which is shown in Fig. 5. To determine the correct sitting posture based on the angle θ, we calculated the angle θ using following formula with the inverse trigonometric function. When this angle is out of threshold, the proposed system defines that this link is in the incorrect posture. The angles for Z-axis are also calculated as same manner.

$$\theta = \operatorname{atan} X * 180/\pi \tag{1}$$

Table 2. Average of angles between two key joints

Joints	Angle of X-axis	Angle of Z-axis
Head – Neck	82.45	137.64
Neck – Shoulder	89.94	120.84
Shoulder- Mid	90.02	125.72
Mid - Base	89.68	118.86

Figure 6 shows the snapshots of results for proposed posture monitoring system. When the user sits incorrect posture which is one of angle between two key joints over predefined threshold value, the monitoring system gives the warning window message. We believe that the proposed system can help users to realign the sitting posture during the work time without uncomfortable sensors. In addition, sitting posture related data which is average value for every 30 s is stored in the server using PHP and MySQL.

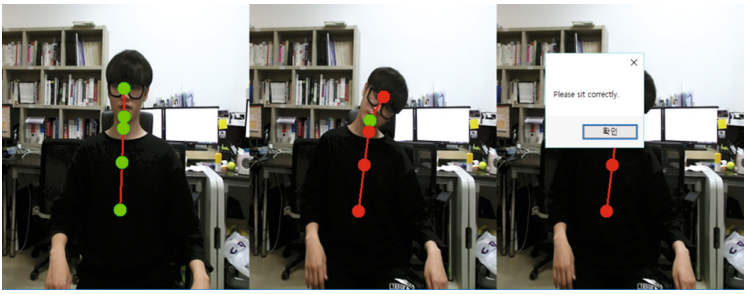


Fig. 6. Snapshot of proposed sitting posture monitoring system

4 Monitoring Results

The result of tested user's sitting posture with monitoring system is increased 62% after experimental test of 35 min. However, the result was somewhat different depending on the body type of person, so the realigned values showed also different. Since the wrong sitting posture can lead the illness like forward head posture, these illnesses can be cured to realign wrong posture. We believe that monitoring system can realign wrong posture easy (Fig. 7).

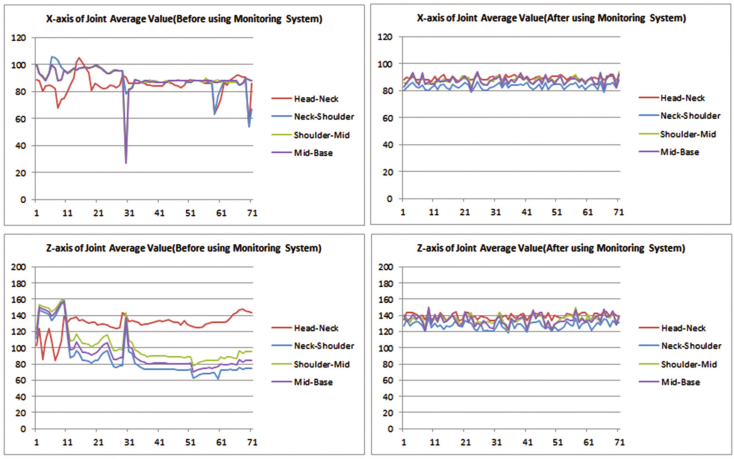


Fig. 7. Compare before using monitoring system and after using monitoring system.

5 Conclusions

We proposed sitting posture monitoring system to realign modern people's sitting posture. The proposed system is the monitoring system can't determine user's body type and determine posture with average value of some people's posture data, so it may

not be exactly correct. In the future, we will improve the monitoring system and will make that record of user's data by user's self, so it can realign a better sitting posture fit on user's body type. Also, the recognition rate is decreased in Kinect on sitting posture but we believe that monitoring system can determine sitting posture effectively using different image processing technology in the future.

Acknowledgment. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-H8601-16-1009) supervised by the IITP (Institute for Information & communications Technology Promotion). This research was supported by the Soonchunhyang University Research Fund.

References

1. Carney, D.R., Cuddy, A.J.C., Yap, A.J.: Power posing brief nonverbal displays affect neuroendocrine levels and risk tolerance. Psychol. Sci. **21**(10), 1363–1368 (2010)
2. Caneiro, J.P., O'Sullivan, P., Burnett, A., Barach, A., O'Neil, D., Tveit, O., Olafsdottir, K.: The influence of different sitting postures on head/neck posture and muscle. Activity **15**(1), 54–60 (2010)
3. Korea Accurate Body Posture Association
4. Sagewood wellness Center

Solving the Subgraph Isomorphism Problem Using Harmony Search

Hoyeong Yun[1], Yongjoon Joe[1], Byung-Ok Jung[1], Hyoguen Bang[2], and Dongmyung Shin[1(✉)]

[1] LSWare Inc., Seoul, Korea
{allen, eugene, luca, roland}@lsware.com
[2] Sungkyunkwan University, Seoul, Korea
revoice@skku.edu

Abstract. The active usage of open source software contributes many areas. However, there are many problems like ignoring license or intellectual properties infringement which can lead litigation. In this paper, we try to find original open source software by using similarity of source code. Source code similarity analyze resembles plagiarism detection problem, and using program dependence graph can be handled as a subgraph isomorphism problem which is one of NP-complete. In this paper, we apply harmony search, one of a metaheuristic algorithm, to solve the problem efficiently.

Keywords: Open source · Subgraph isomorphism problem · Harmony search

1 Introduction

Open source software is computer software which discloses its source. Open source software is allowed to use, copy, modify and distribute freely. Against proprietary software side, the open source society is a representative example of sharing economy. The activity of open source software contributes many areas of ICT like embedded, mobile, cloud computing, big data, etc.

The background of the free use of open source is based on the obligation of open source license. The free of use, copy, modification and distribution is provided under the condition of licenses.

However, there are many litigations because of ignoring or misunderstand of the licenses. Even many methods like SPDX format are suggested to solve the license and intellectual property problem; the developer and users have exposed the risk of software intellectual property infringement because of its weak activity.

When an initial developer distributes open software, other can modify and redistribute it again and again. However, in this progress, there could be license deletion, corruption or collision. Therefore, this study tracks an original open source software reasoning from a similarity of each source code group to provide developer information about the original open source software of the distribution which a user have.

An analyzing similarity of software source codes resembles plagiarism detection. Methods of detecting plagiarism could be classified four categories crudely like Table 1.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_27

Table 1. Software plagiarism methods

Category	Speed	Reliability
String based	Fast	Low
Token based	↑	↓
AST based		
Program dependence based	Slow	High

String based method is the simplest method which handles source code as a general string. This method could detect fast without any modification. However, when only variables or identifiers are modified, it easily misses for detecting them.

Token based method which is applied broadly excludes non-effect factors like a comment or space characters and generate token sequence from the preprocessed code. This method has the advantage to coding style, identifier, and spacing variation. However, it still has no robustness against code injection or reordering function calls.

Abstract Syntax Tree (AST) based method generates AST from source code. Because AST based method has structure information which string or token based does not, it is robust against reordering function or merging. However, it is still fragile against statement rearrangement.

Program dependence graph based method analyzes a control flow and data dependencies of software as a graph. A method representing original source code and target source code and analyzing is a case of graph isomorphism problem. This study chooses program dependence graph focusing on responsibility.

This paper describes research background about software plagiarism detection and graph isomorphism problem at Sect. 2. Section 3 introduces the algorithm for solving graph isomorphism problem and design and result about the experiment. Finally, Sect. 4 concludes the experiment and further tasks.

2 Preliminaries

2.1 Literature Reviews

In these days, the growing of importance about intellectual property makes people nothing to studies for plagiarism detection about academic reports, media articles, and software source code. In this study, we focus on previous works about software plagiarism.

Ji et al. [1] provide plagiarism detection method. Their study which is based on bytecode because of security problem about source code similarity shows a strong relation between source code and bytecode. Their method analyzes similarity based on token sequence. Kim et al. [2] introduced AST based method to evaluate the similarity

between two C program. Moreover, they also suggest an algorithm which clustering many source code groups by similarity to reduce comparing iteration (frequency). Kim [3] studies PDG based method against Javascript.

Koschke et al. [4] detect plagiarism by AST against C program. Liu et al. [5] introduce the process that utilizes PDG to translate it as a graph isomorphism problem with source code preprocessing.

SoVAC(Software Verification and Analysis Center), developed by Pusan National University [6], translate token sequences to DNA and compares DNAs by local alignment method. JPlag and MOSS, SIM, and YAP3 [7–9] also applies token-based method.

2.2 Graph Isomorphism Problem

Graph theory describes and analyzes a set of nodes and the relation among the nodes. Many of fields like physics, chemistry, communication science, computer technology, electronic and civil engineering, architecture, genetics, psychology, sociology, economics, and linguistics apply graph theory [10].

Graph isomorphism problem is determining the identity between to graph. Figure 1 shows two different shapes of the graph. However, the nodes of left graph *{1, 2, 3, 4}* correspond to the nodes of right graph *{D, B, A, C}*. In this case, the two graphs are isomorphic.

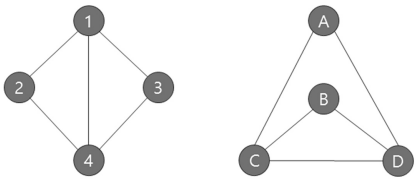


Fig. 1. Examples of graph isomorphism

Let the group of nodes (vertexes) of graph G and H as $V(G)$ and $V(H)$. And let the group of edge for each graph as $E(G)$ and $E(H)$. When the two graph is isomorphic, these satisfy the following conditions, and they are represented as $G \simeq H$.

(i) $|V(G)| = |V(H)|$, and $|E(G)| = |E(H)|$
(ii) Every connect components are same
(iii) Every degree sequences are same

In this study, to find more similar software among the open source software, subgraph isomorphism problem is more appropriate. Subgraph isomorphism problem says graph G is a subgraph of graph H when graph G and H satisfies $V(G) \subset V(H)$ and $E(G) \subset E(H)$.

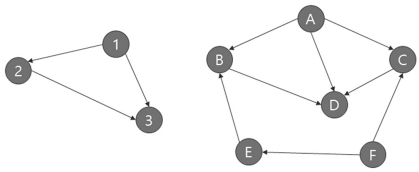


Fig. 2. Examples of subgraph isomorphism

According to Fig. 2, the nodes group of left graph *{1, 2, 3}* corresponds to the subgroup *{A, B, D}* about the nodes group of right graph *{A, B, C, D, E, F}*. In this case, the left graph is a subgraph of the right graph.

Cheminformatics applies Subgraph isomorphism problem to compare chemical constitution, and pattern analyzing of a database, protein-protein interaction networks, and social network also applies the problem.

3 Experiments

Subgraph isomorphic problem is a typical NP-Complete problem, so there is no polynomial time algorithm for it. Therefore, when the size of a problem grows, time complexity also grows as exponentially. An exhaustive search method is not tractable.

There are two approaches for this kind of problems; relaxation and approximation. Because the former approach is restricted for relaxation conditions, this study uses latter one.

3.1 Harmony Search

In this study, we use metaheuristic algorithm to get the approximate value. Meta-heuristic algorithm imitates natural phenomenon to get global optimality instead of local optimality. There are many meta-heuristic algorithms like a genetic algorithm, simulated annealing, Tabu search, ant colony optimization, particle swarm optimization, etc. Metaheuristic algorithm can be designed to almost every problem, so subgraph isomorphism problem also can utilize it. There are previous studies which are applied a genetic algorithm and simulated annealing [11–13]. In this study, we apply harmony search which imitates a process of music players finding harmony.

Harmony search, a population-based algorithm, has five operators, but we use only three in this study. Memory considering refers previous note; Pitch adjustment twists previous note, and Randomization refers random note. Table 2 shows pseudo-code of Harmony search.

Table 2. Pseudo code of the harmony search

Pseudo Code : Harmony Search

```
begin
    Objective function f(x), x = (x_1, x_2, ..., x_d)^T
    Generate initial harmonics (real number arrays)
    Define pitch adjusting rate ( P_PAR ) pitch limits and bandwidth
    Define harmony memory accepting rate ( P_HMCR )
    while (not_termination)
        Generate new harmonics by accepting best harmonics
        Adjust pitch to get new harmonics
        If (rand() < P_HMCR)
            if (rand() < P_PAR)
                Adjust the pitch randomly within limits
            else
                Choose an existing harmonic randomly
            end if
        else
            Generate new harmonics via randomization
        end if
        Accept the new harmonics if better
    end while
    Find the current best solutions
end
```

3.2 Experiment Design

3.2.1 Representation

When $V(G) \leq V(H)$, we represent G and H as an array. Each index of an array means a node of graph G, and each value of an array means a node of graph H. An array can represent how graph G correspond to graph H.

3.2.2 Fitness Function

A barometer of subgraph isomorphism is defined by a multi-objective function which concerns the difference of edges between two graphs. When $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ and $|V_1| \leq |V_2|$, $G_S = (V_S, E_S)$ means that graph G_1 is subgraph of graph G_2. At this time, the first objective function f_1, the difference of edges, is defined as (1)

$$f_1 = \sum_{e \in E_1} I(e, E_s) + \sum_{e \in E_s} I(e, E_1)$$
$$where\, I(e, E) = \begin{cases} 0, & \text{if } e \in E \\ 1, & otherwise \end{cases} \tag{1}$$

The second graph, corresponding graph, is defined like (2). $e_{in\text{-}out}$ means that incoming and outcoming edge.

$$f_2 = \sum_{v \in E_1} S(v)$$
$$S(v) = \begin{cases} 0, & \text{if } e_{\text{in}-\text{out}}(G_2(v)) \geq e_{\text{in}-\text{out}}(G_1(v)) \\ 1, & \textit{otherwise} \end{cases} \tag{2}$$

The proportion of objective function f_1 and f_2 are 50% each; $f(w_1, w_2) = w_1 f_1 + w_2 f_2$ when w_1 and w_2 are 0.5 [14].

3.3 Experimental Setup

We generate graph data randomly with a density of edge and number of nodes in a graph. We base on graphs with 150 nodes and generate graphs with 10, 15, 50, 75 and 100 nodes which reduced from them.

These graphs are subgraphs for each based graphs. Densities of graphs are 0.2, 0.1, 0.05, and 0.02. We experiment based on 20 sets of data.

Experiments are repeated 30 times for statically significance. We set HMCR is 0.95 and PAR 0.7 which are parameters of harmony search. The bandwidth of pitch adjustment is set to ±20% about the node number of each graph; HMS starts from 10 to 100; iteration is set from 3000 to 30000. These parameters are determined by the size of graphs.

We implement the algorithm by C#, and the algorithm is executed on 3.3 GHz CPU and 8 GB RAM computer.

3.4 Experiment Result

Table 3 shows the results under 20 different conditions about five graph sizes and four densities of edges. When the number of nodes is 10, the algorithm could get global optimal value regardless of the density of edges. Also, it is same when the number of nodes is 25, and the densities of edges are 0.02 and 0.05.

However, when the number of the nodes is above 25, and the density is above 0.1, the algorithm could not reach the global optimal value with the conditions. The difference between global optimal and the output grows with the number of nodes. Also, the result shows a relation between a graph with a density of edges and a gap with optimal value. When we expand candidate in proportion to the number of nodes and changed iteration times, it shows a little improvement. However, when the size of a graph is big, the algorithm cannot escape from local optimal. This result shows the weak point of the algorithm.

Table 3. Results of subgraph isomorphism problem

Node	Density	Best	Average
10	0.02	**0**	**0**
	0.05	**0**	0.067
	0.1	**0**	0.167
	0.2	**0**	1.15
25	0.02	**0**	0.633
	0.05	**0**	0.8
	0.1	0.5	2.1
	0.2	2.5	2.95
50	0.02	3	4.783
	0.05	4.5	6.017
	0.1	6	7.6
	0.2	7	8.783
75	0.02	12	13.767
	0.05	12	13.933
	0.1	12.5	13.8
	0.2	15.5	16.6
100	0.02	18.5	20
	0.05	24.5	20.817
	0.1	25.5	28.033
	0.2	27	29.767

4 Conclusions

In this study, we find the approximate solution of subgraph isomorphic problem to solve the similarity problem about open source software by program dependence graph, because subgraph isomorphic problem is NP-Complete, so there is no polynomial time exact algorithm. We experiment graphs with 150 nodes and its subgraphs and could find the optimal solution when the subgraph is small. However, the number of nodes and the density of edges are.

As further tasks, we execute our algorithm with real open source software source code. Because of syntactical differences of each programming language, the graph translation preprocess is indispensable. Also, improvement of the algorithm will be implemented.

Acknowledgments. This research project was supported by the Ministry of Culture, Sports, and Tourism (MCST) and by the Korea Copyright Commission in 2016.

References

1. Ji, J.H., Woo, G., Cho, H.G.: A plagiarism detection technique for Java program using Bytecode analysis. J. Korean Inst. Inf. Sci. Eng. **35**(7), 442–451 (2008)
2. Kim, Y.C., Hwang, S.C., Choi, J.Y.: A program similarity evaluation algorithm. J. Korean Soc. Inf. **6**(1), 51–64 (2005)
3. Kim, S.H.: Plagiarism detection using dependence graph analysis specialized for Javascript. Master Thesis, Korea Advanced Institute of Science Technology (2011)
4. Koschke, R., Falke, R., Frenzel, P.: Clone detection using abstract syntax suffix trees. In: Proceedings of the 13th Working Conference on Reverse Engineering, pp. 253–262. IEEE (2006)
5. Liu, C., Chen, C., Han, J., Yu, P.S.: GPLAG: detection of Software plagiarism by program dependence graph analysis. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining, pp. 872–881 (2006)
6. Kim, Y.E., Cheon, J.S., Byun, S.W., Woo, G.: A parallel performance comparison of Haskell using a plagiarism detection method. In: Korea Computer Congress, pp. 1724–1726 (2016)
7. Prechelt, L., Malpohl, G., Philippsen, M.: Finding plagiarism among a set of programs with JPlag. J. Univ. Comput. Sci. **8**(11), 1016–1038 (2002)
8. Gitchell, D., Tran, N.: Sim: a utility for detecting similarity in computer programs. ACM SIGCSE Bull. **31**(1), 266–270 (1999). ACM
9. Wise, M.J.: YAP3: improved detection of similarities in computer program and other texts. ACM SIGCSE Bull. **28**(1), 130–134 (1996)
10. Suresh Singh, G.: Graph Theory. PHI Learning, New Delhi (2010)
11. Choi, J.E., Yoon, Y.R., Moon, B.R.: An efficient genetic algorithm for subgraph isomorphism. In: Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, pp. 361–368 (2012)
12. Farahani, M.M., Chaharsoughi, S.K.: A genetic and iterative local search algorithm for solving subgraph isomorphism problem. In: International Conference on Industrial Engineering and Operations Management (IEOM), pp. 1–6. IEEE (2015)
13. Li, Z., Chen, B., Che, D.: Solving the subgraph isomorphism problem using simulated annealing and evolutionary algorithms. In: Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), pp. 293–299 (2016)
14. Choi, J., Yoon, Y., Moon, B.R.: An efficient genetic algorithm for subgraph isomorphism. In: Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation (2012)

Persuading Recommendations Using Customized Content Curation

Keonsoo Lee[1] and Yunyoung Nam[2(✉)]

[1] Medical Information Communication Technology, Soonchunhyang University, Asan, Republic of Korea
keonsoo@sch.ac.kr
[2] Department of Computer Science and Engineering, Soonchunhyang University, Asan, Republic of Korea
ynam@sch.ac.kr

Abstract. A recommendation system finds the most suitable item for customers. However, the most suitable item is sometimes refused. The most significant reason for this failure of recommendation is that we do not want what is good for us. Therefore, to make the recommendation for acceptance, persuasion should be provided. In this paper, a method of curating information for persuasion is proposed. Cognitive bias are used to determine the layout of information for customers to accept recommendations. With the proposed method, options which are beneficial but not preferable are provided without offending the customers.

Keywords: Recommendation · Cognitive bias · Content curation

1 Introduction

Contemporary production systems are ruled by the long-tail principle [1]. Even though, the long-tail principle is more lavish than the Pareto principle in an economical view, highly developed technology realizes the long-tail principle and the public chooses it as a winner. Since then, recommendation systems are highlighted [2]. As the number of options for selection increases, the difficulty of finding the right option becomes harder. Recommendation systems take the role of deputy for evaluating each option, comparing each option's pros and cons, and ranking them. With the help of recommendation systems, customers are able to be saved from the anguish of choice [3]. However, customers are trapped in accustomedness. For a customer whose preference is registered, novel or unfamiliar items are not recommended. For items in the category of tastes, it can be allowed. However, for items in the category of essentials, options which is against the registered preference should be provided and persuaded to be selected [4, 5].

For a diabetic who likes to eat pizzas, recommending hamburgers may be correct but is not right. Sometimes, what is good for us is not what we want. In such situation, right recommendations are easily refused and get discontents. Therefore, a way of persuading is needed. Good recommendation system should propose not only good options but also persuade the options to be selected. In this paper, we propose a way of persuading options using cognitive bias of customers.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_28

2 Background

2.1 Recommend System

The objective of recommendation systems is to find a suitable items for system users [2]. When the request of users is represented explicitly, the result for recommendation can clearly evaluated. However, in most recommendations, the request is ambiguous. Therefore, filtering methods which is a pre-pruning process to reduce the size of options for the recommendation, are researched. Content-based filtering and collaborative filtering are most widely employed filtering methods [4, 5]. As shown in Fig. 1, content-based filtering method finds items which are similar to the item which is selected beforehand. For a user who likes item-A, item-B which is similar to item-A is recommended to the user. Collaborative filtering method recommends items which is popular in a group whose members are similar to the user. The basic idea of these two filtering methods is that a preference can be inducted from the history of customers. Once the preference is made, items which are not similar to the preference are filtered out. For collaborative filtering, novel items may be recommended as long as the items are popular.

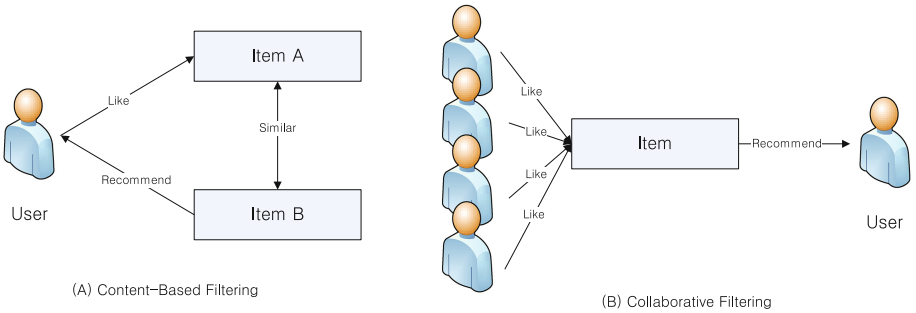


Fig. 1. Two ways of filtering for recommendation

2.2 Cognitive Bias

There are two kinds of thinking which is slow and fast [6]. In order to make a correct decision, we need to collect all the possible facts and infer logically. Even if this process makes better results, it consumes more time. This is why it is called as slow thinking. For problems which require quick responses, decisions need to be made as fast as possible. In such case, stereotypes become useful. Cognitive bias is a way of how human beings use stereotypes by removing unnecessary information and making decisions rapidly [7]. Table 1 shows examples of cognitive biases which are famous.

Table 1. Well known biases and their descriptions.

Bias	Description
Anchoring	The tendency to rely too heavily on some pieces of information to make a decision
Availability cascade	The tendency to believe what is shown repeatedly
Bandwagon effect	The tendency to do what many other people do
Decoy effect	The tendency to change the preference when new option which is similar but worse, is presented
Information bias	The tendency to believe that any information, even if it is not necessary, will be better
Gambler's fallacy	The tendency to think that the probabilities will be altered by the previous events

3 Proposed Method

As mentioned in Sect. 2, good recommendations are refused sometimes. One of the most significant reasons is that the discrepancy between what is necessary and what is desired. Humans have a psychological impulsion of self-destruction [8]. In this paper, we propose a method of persuading recommendations so as for good suggestions not to be refused. The persuasion is made using customized content curation. Content curation is a way of display information [9]. As shown in Table 2, customers have various cognitive biases and each customer has his/her major cognitive bias. Therefore, recommendations which are provided in proper way will be more easily accepted by the customers. The curation is executed according to the each customer's cognitive bias.

Table 2. Templates of curation layout for each bias.

Bias	Template
Anchoring	A set of information that defines the preference of the target user
Availability cascade	A series of short message to make the recommendation accustomed to the target user
Bandwagon effect	A set of reviews that shows many other people think the item is good enough
Decoy effect	A specification of competitive item that shows the proposed item is better than other items
Information bias	A set of information that shows the domain of item is important and the messenger of the recommendation is reliable
Gambler's fallacy	A number of applicants and winning tickets, which shows the chance of winning, are to buy the item right now

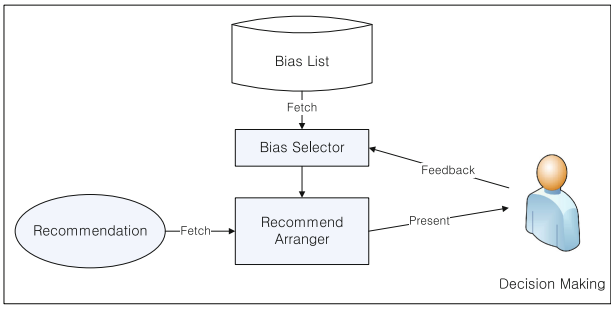


Fig. 2. A model of customized recommending system using customized content curation

Figure 2 shows general architecture of the proposed method. *Bias Selector* (BS) module that selects the customer's major cognitive bias. And *Recommend Arranger* (RA) displays the recommendations to be suit for the customer's cognitive bias. For BS, it collects the purchasing history of the customer and finds the relation between the action of purchasing and given information. For example, the time interval from searching and buying shows how impulsive s/he is. If s/he takes lots of times to make a decision, the suggested option which is recommended is provided with its pros and cons comparing with other options. If s/he has lots of friends and spends lot of time in SNS chatting with others, we can assume that bandwagon effect is available. Recommended option is shown with how many others have selected it. If s/he has a tendency for availability cascade, the option which needs to be selected should be exposed as many as possible even in unrelated recommending requests. Then when the moment of selection, s/he will choose the familiar option as intended. BS has the responsibility of finding the suitable cognitive bias for each customer, and RA has the responsibility of content curation. Table 2 shows templates for each bias's layout.

As the templates in Table 2 are based on the heuristics, even for the same bias, the resulted curation can be different for customers. For bandwagon effect, if the target customer likes celebrities, it is important to show that the celebrities also take the same option. If the target customer is loyal to his/her flock, to show that the flock's preference is matched with the option is more effective.

4 Conclusion

The objective of recommendation systems is to reduce the size of options so as to make quick and right decisions and make a happy experience in selections. However, existing filtering methods provide convenience by evading challenges and protecting with accustomed mannerism. Providing similar options may be safe from being failed. However, such recommendation system is valueless. In order to make more valuable recommendations, it should persuade users to be better not just to keep the existing comfortable life. For this objective, the proposed method extracts the customer's cognitive bias from his/her history and suggests options which is not only correct but

also righteous by customized information arrangement. In order to realize the proposed method and evaluate in real situation, the way of customization which defines how the bias based content curation is applied for each customer will be researched as future work.

Acknowledgments. This work was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-H8601-16-1009) supervised by the IITP (Institute for Information & communications Technology Promotion) and was also supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (NRF-2015M3A9D7067219).

References

1. Brynjolfsson, E., Yu, H., Simester, D.: Goodbye Pareto principle, hello long tail: the effect of search costs on the concentration of product sales. Manag. Sci. **57**(8), 1373–1386 (2011)
2. Perugini, S., Gonçalves, M.A., Fox, E.A.: Recommender systems research: a connection-centric survey. J. Intell. Inf. Syst. **23**(2), 107–143 (2004)
3. Schwartz, B., Kliban, K.: The Paradox of Choice: Why More Is Less. Brilliance Audio, Grand Haven (2014)
4. Van Meteren, R., Van Someren, M.: Using content-based filtering for recommendation. In: Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop (2000)
5. Ekstrand, M.D., Riedl, J.T., Konstan, J.A.: Collaborative filtering recommender systems. Found. Trends Hum. Comput. Interact. **4**(2), 81–173 (2011)
6. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York (2013)
7. Mathews, A., Mogg, K., Kentish, J., Eysenck, M.: Effect of psychological treatment on cognitive bias in generalized anxiety disorder. Behav. Res. Ther. **33**, 293–303 (1995)
8. Holzberg, J.D., Cahen, E.R., Wilk, E.K.: Suicide: a psychological study of self-destruction. J. Projective Tech. **15**(3), 339–354 (1951)
9. Whittaker, S.: Personal information management: from information consumption to curation. Annu. Rev. Inf. Sci. Technol. **45**(1), 1–62 (2011)

Improving the Quality of an R-Tree Using the Map-Reduce Framework

Viet-Ngu Huynh Cong[1], Kang-Woo Lee[2], In-Hak Joo[2], Oh-Heum Kwon[1], and Ha-Joo Song[1(✉)]

[1] Department of IT Convergence and Application Engineering, Pukyong National University, Busan, Republic of Korea
hcvngu@pukyong.ac.kr, {ohkwn,hajoosong}@pknu.ac.kr
[2] Electronics and Telecommunications Research Institute, Daejon, Republic of Korea
{kwlee,ihjoo}@etri.re.kr

Abstract. An R-tree is an index structure that enables fast access to multi-dimensional data. Constructing an R-tree for a given data set yields a more efficient R-tree structure than incrementally building one as data are inserted. However it usually takes a lot of time to construct an R-tree for a huge volume of data. In this paper, we propose a parallel R-Tree construction scheme based on a Hadoop framework. The proposed scheme divides the data into partitions, builds local R-trees in parallel, and merges them to construct an R-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: R-tree · Hadoop · MapReduce · Bulk-loading

1 Introduction

Nowadays, aside from the fact large amounts of traditional data are still increasing significantly, there is an explosion in the amount of spatial data that is being produced from many devices such as satellites or smart phones. In order to handle this amount of spatial data efficiently, the R-tree is considered as an optimal index mechanism that will help retrieve data quickly according to its spatial locations. An R-tree [2] is a data structure in which each node contains a certain number of index entries, each of which consists of a Minimum Bounding Rectangle (MBR) and the pointer to an object or its child node if it is not a leaf node. Both objects and non-leaf nodes are always represented by MBR. A layout of MBRs of objects is shown in Figs. 1 and 2 show the R-tree structure with four nodes.

An R-tree, is built by inserting new items iteratively as they arrive [2]. But with this method, when an object is inserted into an R-tree node, in some cases of the node splitting operation, it requires the locking of the R-tree, therefore the concurrent insertions are prevented. In fact, in some applications where all the items is available as with the Geometry Information System (GIS) problems, if an R-tree is built using the packing technique in a parallel way, it's much faster than the traditional method as

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_29

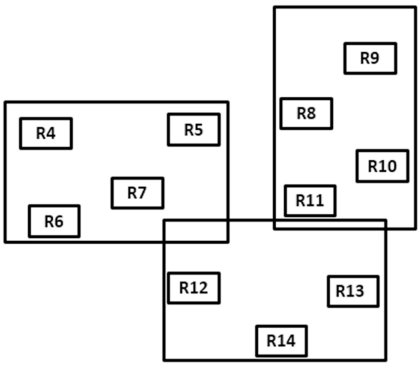


Fig. 1. A layout of MBRs

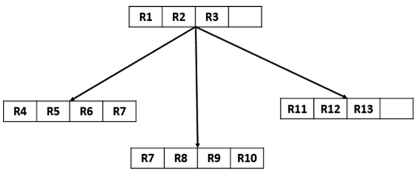


Fig. 2. The R-tree structure with four nodes

mentioned previously. Since its release in April 2005, Hadoop [1, 3] was adopted as an optimal solution for scalable processing of huge datasets in many applications, e.g., machine learning, image processing, web crawling or text processing, and so on. Hadoop employs MapReduce [4], a simplified programming paradigm for distributed processing, to build an efficient large-scale data processing framework.

An R-tree is an index structure that enables fast accesses to multi-dimensional data. Constructing an R-tree for a given data set yields an efficient R-tree structure that builds itself incrementally as data are inserted [9–12]. However, it usually takes a lot of time to construct an R-tree for a huge volume of data. In this paper, we propose a parallel R-Tree construction scheme based on a Hadoop framework. The proposed scheme divides the data into partitions, builds local R-trees in parallel, and merges them to construct an R-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 Works

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. 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. Furthermore, MapReduce consists of

	Input	Output
Map	<k1,v1>	list(k2,v2)
Reduce	<k2,list(v2)>	list(k3,v3)

Fig. 3. Input and output in MapReduce

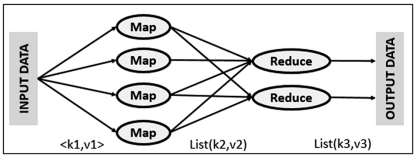


Fig. 4. MapReduce framework

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. 3), 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 k2 and produces a list of final output key/value pair (k3, v3). The data flow of the MapReduce framework is illustrated in Fig. 4.

2.2 Parallel R-Tree Construction

As we discussed in the first chapter, in some applications where all items are available, a good packing method would build be able to quickly build an R-tree with maximum node occupancy and high node quality with Minimum Bounding Rectangles MBRs of minimum area, with some margin and overlap between them. In general, there are two main goals that are considered during the construction of an R-tree. The first goal is how to minimize the area of the MBRs of the non-leaf nodes that are not covered by MBRs at the leaves these enclose, which is also called the "Dead Area." This goal can improve search performance since decisions on which paths have to be traversed can be taken on higher levels. The second goal is how to minimize the overlap between MBRs, this goal is also designed to decrease the number of paths to be traversed.

There are some packing techniques for R-trees that have been proposed before such as Nearest-X, Hilbert sort, Sort-Tile-Recursive (STR) or Z-order curve. One of them, STR algorithm [5] is considered as one of techniques that brings out the simplicity of implementation yet a good query performance, even it outperforms with other techniques in some cases. For parallel R-tree construction, in [6], author have proposed a method for parallel R-tree construction on Hadoop through MapReduce framework, it is performed in a bottom-up fashion and using Z-order curve to partition a huge data volumes.

In [7–9], to build R-tree structure in CPU environment, the author have proposed new two strategies to improve the STR (Sort-Tile-Recursive) algorithm, and they have been proven to achieve a better search performance than original STR, Nearest-X, and Hilbert sort. From these strategies, motivated our own parallel R-tree construction on Hadoop through MapReduce framework, and use one of those strategies as the idea to partition a huge data volumes in our scheme, it's much simpler than using Z-order curve, named for our partition method is ISTRMR (STR on MapReduce).

3 Parallel R-Tree Construction Using Hadoop

Before we describe how to build R-tree indexing on a MapReduce framework, there are some notations that we will use in the rest of the paper which are as follows: M is maximum node occupancy, r is the number of total spatial objects, and we use the coordinate of the center point of objects to represent objects location. For simplicity, in this paper, we only concentrate on two-dimensional objects.

When building an R-tree, the quality of the resulting R-tree is our main consideration. As discussed in the previous section, for a good R-tree, it should be built by grouping spatially neighboring MBRs, so that the "Dead Area" and overlap between these MBRs can be decreased. Our parallel R-tree construction is performed in a bottom-up fashion and has three phases, and two of them are implemented in the Hadoop environment using the MapReduce model:

- Partitioning phase – using ISTRMR to determine the longest coordinate and partition boundary of big data set.
- R-tree construction – small R-trees are built concurrently.
- R-tree consolidation – merge small R-trees into the final R-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 object, it contains <o.id, o.P> where o.id is the object's unique identifier and o.P is the location of an object which is represented by a list of coordinates.

Our scheme consists of three phases executed in sequence, as shown in Fig. 5. First, we find out the "longest" coordinate that has the two most distant centers of the rectangles in the coordinate, then the number of partition and the partitions boundary will be determined. Next, data is partitioned into the corresponding partition which creates small R-Trees. Finally, the small R-Trees are merged into the final R-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.

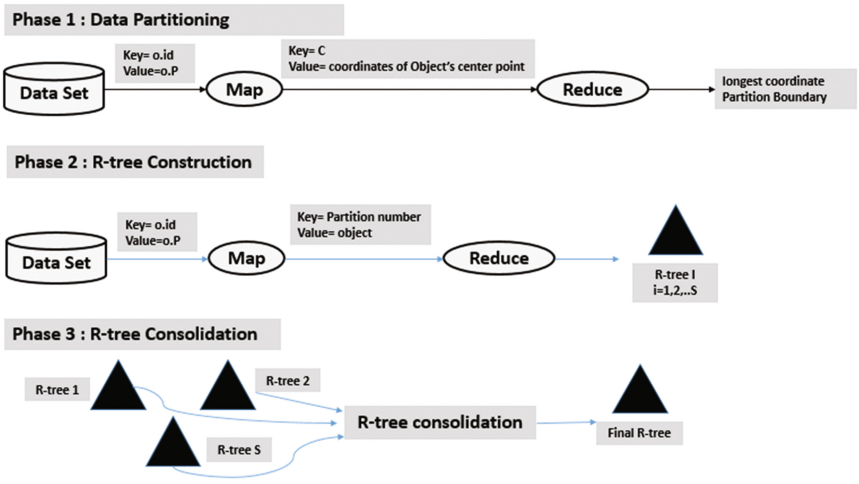


Fig. 5. Phases involved in building an R-tree in MapReduce

3.1 Data Partitioning

In this phase, we assume all objects lie in the plane, each object' location is represented by a center point with its coordinate. To find out the longest coordinate that has the two most distant centers of the hyper- rectangle in the coordinate, 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 calculates the coordinates of center point of each object. Then a single Reducer, firstly, it will calculate the distance between those objects and determine the longest coordinate as shown in Fig. 6. After that, it determines a new list K of R–1 partition boundary that split the longest coordinate of sample into R nearly equal-size partition.

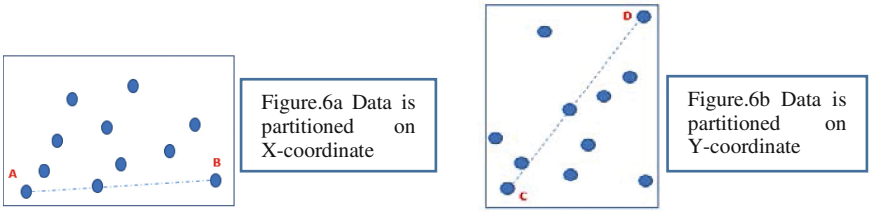


Fig. 6. Determine the longest coordinate

The specific MapReduce key/value input pairs are presented in Fig. 7. Mappers read the default ratio of data from input file and calculates the coordinates of center point of the objects. The intermediate key is a constant that helps to send all the Mappers' outputs to a single Reducer. Then Reducer receives all center point with theirs coordinates from Mappers, firstly find out the longest coordinate by calculate the two most distant centers of the hyper- rectangle in the coordinate, then determine the list splitting point K base on the longest coordinate as shown in Fig. 8.

Function	Input	Output(Key,Value)
Map	(o.id, o.P)	(C,center point)
Reduce	(C, list(center point))	K

Fig. 7. Map and reduce inputs/outputs for data partitioning

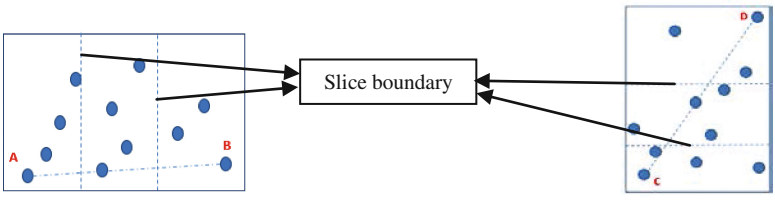


Fig. 8. Example of data grouping into partition with maximum objects equal 4

3.2 R-Tree Construction

In this phase, an individual R-tree is built concurrently. Mappers partition the input file into R groups, then every partition is executed by a different Reducer, each Reducer, R-tree is built independently. The output of every Reducer is a root node of their constructed R-Trees, as shown in Fig. 9.

Function	Input	Output(Key,Value)
Map	(o.id, o.P)	(partition Number,object)
Reduce	(partition Number, list(objects))	Tree, root

Fig. 9. MapReduce inputs/outputs for R-tree construction

3.3 R-Tree Consolidation

In this phase, we are going to combine the R individual R-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.

4 Conclusion and Future Work

In this paper, we proposed a scheme for parallel R-tree construction with ISTRMR packing techniques on Hadoop environment using the MapReduce model. With our scheme, we hope to contribute to the improvement of the quality of R-tree construction and reduce the construction time. With our proposed scheme, it can be used in many kinds of applications in the field of databases in general and in particular, spatial data.

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 spatial 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.

Acknowledgement. This research, 'Geospatial Big Data Management, Analysis and Service Platform Technology Development' was supported by the MOLIT (The Ministry of Land, Infrastructure and Transport), Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for Infrastructure Technology Advancement) (16NSIP-B081011-03).

References

1. Polato, I., Ré, R., Goldman, A., Kon, F.: A comprehensive view of Hadoop research—a systematic literature review. J. Netw. Comput. Appl. **46**, 1–25 (2014). Elsevier
2. Guttman, A.: R-trees-A dynamic index structure for spatial searching. In: ACM SIGMOD, pp. 47–57 (1984)
3. Apache Hadoop. http://hadoop.apache.org
4. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Comm. ACM **51**(1), 107–113 (2008). 50th Anniversary Issue
5. Leutenegger, S.T., Edgington, J.M., Lopez, M.A.: STR: a simple and efficient algorithm for R-tree packing. In: 13th International Conference on Data Engineering, pp. 497–506 (1997)
6. Cary, A., Sun, Z., Hristidis, V., Rishe, N.: Experiences on processing spatial data with MapReduce. In: 2009 SSDBM, vol. 5566. LNCS, pp. 302–319. Springer, Berlin, Heidelberg (2009)
7. Eldawy, A., Mokbel, M.F.: SpatialHadoop: a MapReduce framework for spatial data. In: 2015 IEEE 31st International Conference (ICDE), pp. 1352–1363 (2015)
8. Giao, B.C., Anh, D.T.: Improving sort-tile-recusive algorithm for R-tree packing in indexing time series. In: 2015 IEEE International Conference (RIVF), pp. 117–122 (2015)
9. Huh, J.H., Seo, K.: RUDP design and implementation using OPNET simulation. In: CUTE 2014, vol. 330. LNEE, pp. 913–919. Springer (2015)
10. Park, J., et al.: Design of the real-time mobile push system for implementation of the shipboard smart working. In: CUTE 2015, vol. 373. LNEE, pp. 541–548. Springer (2015)
11. Huh, J.H., Seo, K.: Design and test bed experiments of server operation system using virtualization technology. HCIS **6**(1), 1–21 (2016). Springer
12. Huh, J.H., Je, S.M., Seo, K.: Design and configuration of avoidance technique for worst situation in Zigbee communications using OPNET. In: ICISA, vol. 376. LNEE, pp. 331–336. Springer (2016)

An Energy-Efficient and Reliable Routing Protocol for Cognitive Radio Sensor Networks

Zamree Che-aron(✉)

Faculty of Communication Sciences, Prince of Songkla University (PSU),
Hat Yai 94000, Pattani, Thailand
zamree.c@psu.ac.th

Abstract. With rapid advances in wireless communications and networking technologies, the demand for high speed, wireless and ubiquitous connectivity continues to increase in order to cope with new services and applications. Increasing demand for spectral resources has introduced the problem of spectrum scarcity. Moreover, the static spectrum allocation policy for wireless communications can cause the issue of spectrum underutilization. Cognitive Radio (CR) offers promising solution for spectrum shortage and underutilization problem by means of dynamic spectrum management. Routing in Cognitive Radio Sensor Networks (CRSNs) is a very challenging task due to energy constraints, opportunistic spectrum access, dynamic topology changes as well as intermittent connectivity caused by activities of Primary Users (PUs). This paper proposes the Energy-efficient and Reliable Cognitive Ad-hoc Routing Protocol (ERCARP) with an aim to provide reliable transmission path and prolong network lifetime in CRSNs. The protocol takes account of packet loss probability, link latency and residual energy for path establishment. Furthermore, by utilizing the joint path and spectrum diversity in routing, the multi-path multi-channel routes are given for fast route recovery. The protocol performance is compared with that of the Dual Diversity Cognitive Ad-hoc Routing Protocol (D2CARP) through simulations using NS-2 simulator. The simulation results obviously prove that ERCARP outperforms D2CARP in terms of packet loss, energy efficiency, end-to-end delay and jitter.

Keywords: Reliable routing · Energy efficiency · Cognitive radio · Sensor network · Fast route recovery

1 Introduction

Cognitive Radio (CR) [1–3] is a candidate for the next generation of wireless communication system which applies the Dynamic Spectrum Access (DSA) scheme to improve spectrum utilization efficiency and communication quality. CR has the ability to know the unutilized spectrum in a licensed and unlicensed spectrum band. In Cognitive Radio Networks (CRNs) [4], unlicensed users (or Secondary Users (SUs)) are able to sense their environment and are allowed to opportunistically access the temporally unused licensed spectrum bands without harmful interference to licensed users (or Primary Users (PUs)).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_30

In addition, recently, significant growth in the applications of Wireless Sensor Networks (WSNs) [5, 6] has been made such as air pollution monitoring, wildlife tracking, heal monitoring, intruder tracking, etc.

With recent advances in CR technology, it is possible to apply the DSA model in WSNs, called Cognitive Radio Sensor Networks (CRSNs) [7, 8], to mitigate overcrowded unlicensed spectrum bands. CRSN is a specialized ad hoc network of distributed wireless sensors that are equipped with CR capabilities. The components of CRSN architecture are shown in Fig. 1.

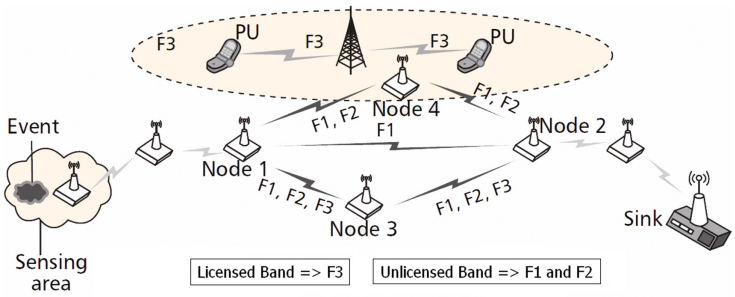


Fig. 1. Cognitive radio sensor network.

With unique characteristics of CRSNs, data routing in CRSNs faces various significant challenges that require in-depth studies [7, 9]. The main challenge is to deal with the dynamic spectrum availability. In CRSNs, the available spectrum bands are time and location varying because of dynamic PU activities. Another challenge is how to determine the optimal path for data transmission which provides high network performance. Furthermore, in CRSNs, a link failure frequently occurs caused by not only the node mobility but also the appearance of PU activity. The data transmission of SUs may be interrupted immediately after a PU activity is detected.

In this paper, we propose the Energy-efficient and Reliable Cognitive Ad-hoc Routing Protocol (ERCARP) with an aim to provide reliable transmission path and prolong network lifetime in CRSNs. In ERCARP protocol, the novel routing metric that takes account of packet loss probability, link latency and residual energy is introduced and implemented. The protocol establishes a transmission path with small delay, low packet loss and high residual energy. Moreover, to offer fast route recovery, the protocol applies the joint path and spectrum diversity in routing to provide multi-path multi-channel routes so that the source node is able to dynamically switch to different paths or channels when PU activity is detected during data transmission. The performance evaluations are conducted through simulations using NS-2 simulator [10]. The simulation results obviously show that ERCARP protocol outperforms the Dual Diversity Cognitive Ad-hoc Routing Protocol (D2CARP) [11] in terms of the number of dropped packets, energy efficiency, end-to-end delay and jitter.

The remainder of the paper is organized as follows. In Sect. 2, we present the related work. The overview of ERCARP protocol is provided in Sect. 3. The protocol's routing metric is described in Sect. 4. Section 5 explains the protocol operations including route discovery and route maintenance process. In Sect. 6, the simulation configuration is described. We show the simulation results and discuss the protocol performance evaluation in Sect. 7. Finally, we state the concluding remarks in Sect. 8.

2 Related Work

Several routing protocols have been proposed in the literature in order to address the different concerns for CR ad hoc networks which can be applied in CRSNs. In [12], Chowdhury and Felice has proposed a spectrum aware routing protocol for cognitive scenarios based on geographic routing paradigm, i.e. each SU can determine the location of other nodes, to elude regions of PU activity during the route formation. In [13], Beltagy et al. has presented a multipath routing protocol with a purpose to improve the reliability of transmission paths. The "Route Closeness" metric has been introduced to create transmission routes based on non-closeness to each other. The main goal of this routing design is to provide less vulnerability to the impact of PU activity as an active mobile PU is unable to interrupt all the selected paths in the same time if they are not close to each other. The authors of [14] has proposed the on-demand Spectrum and Energy aware Routing (SER) protocol to support the energy-constraint multi-hop CR ad hoc networks, where each SU has limited energy supply, with an objective to give high throughput and robust routing as well as prolong the network lifetime. The protocol provides the channel access scheduling and balances the network load by dividing the data traffic over different channels and timeslots according to the channel-timeslot assignment. The article in [15] has introduced the Cognitive Ad-hoc On-demand Distance Vector (CAODV) routing protocol, which applies individually path and spectrum diversity, with an aim to support dynamic CR ad hoc networks. In [11], the Dual Diversity Cognitive Ad-hoc Routing Protocol (D2CARP) has been proposed by sharing some common functionalities with the CAODV. The protocol exploits the joint path and spectrum diversity in routing to reduce the impact of performance degradation experienced by SUs due to PU activities. Nevertheless, both CAODV and D2CARP consider the number of hops as the routing metric in order to select the transmission path with minimum hop count. Although the main advantage of this metric is its simplicity, the quality of wireless links and the energy efficiency of the network are not taken into account, resulting in the establishment of non-optimal transmission paths which can significantly cause poor protocol performance. Additionally, scalable routing protocols for CRSNs have been proposed in [16, 17]. Even though, the protocols take account of the network scalability, energy efficiency, resource constraint and spectrum efficiency, the issue of efficient route recovery has not provided.

3 Overview of ERCARP Protocol

The ERCARP protocol is a reactive routing protocol which triggers the process of route discovery when a source node requires data delivery. The protocol shares some common features with D2CARP protocol [11]. Each node in the network is unnecessary to know the complete path from source to destination for data transmission. To circumvent the routing loop problem [18], the sequence number is utilized to indicate the freshness of routing information. Furthermore, the protocol applies the joint path and spectrum diversity in routing to provide multi-path multi-channel routes for rapid route recovery. The protocol's routing metric takes account of energy efficiency, packet loss probability and link latency. A transmission path with small delay, low packet loss and high residual energy is established in order to provide reliable communications and prolong network lifetime in CRSNs. The protocol also enables sensor nodes to go into sleep mode when they do not have any packet to transmit or receive in order to conserve their battery resources. The main routing packets include RREQ (Route REQuest), RREP (Route REPly), RERR (Route ERRor), ETX (Expected Transmission Count) probe, RTT (Round-Trip Time) probe and RTT acknowledgement packet.

4 Routing Metric

In CRSNs, a SU selects an appropriate available channel among various channels for communication with each other based on a routing metric. The ERCARP protocol relies on the combined routing metric that is composed of three essential components: packet loss probability, link latency and residual energy. The protocol selects a path with small delay, low packet loss and high residual energy for data transmission. The link cost, *C(l)*, is calculated by using the following equation:

$$C(l) = RE(l) * \left(\frac{1}{ETX(l)}\right) * \left(\frac{2}{RTT(l)}\right) \tag{1}$$

where $RE(l)$ is the residual energy of a node on the link l. The $RTT(l)$ is the round-trip time of the link l, denoting the interval between the sending of an RTT probe packet and the receiving of the corresponding RTT acknowledgement packet over the link l. The $ETX(l)$ is the Expected Transmission Count of the link l, representing the expected number of retransmissions required to successfully transmit an ETX probe packet over the link l. The $ETX(l)$ can be measured as:

$$ETX(l) = \frac{1}{[1 - P_f(l)][1 - P_r(l)]} = \frac{1}{[d_f(l) * d_r(l)]} \tag{2}$$

where $P_f(l)$ and $P_r(l)$ are the probability of packet loss in the forward and reverse direction of the link l respectively. The $d_f(l)$ (forward delivery ratio) denotes the

probability that an ETX probe packet is successfully transmitted to the neighbor over the link l during the time window (i.e. *ETX_PROBE_WINDOW*). The $d_r(l)$ (reverse delivery ratio) represents the probability that an ETX probe packet sent from the neighbor is successfully received through the link l during the time window (i.e. *ETX_PROBE_WINDOW*). The $d_f(l)$ and $d_r(l)$ can be calculated as:

$$d_f(l) = \frac{n_f(l)}{\text{ETX\_PROBE\_WINDOW}} \tag{3}$$

$$d_r(l) = \frac{n_r(l)}{\text{ETX\_PROBE\_WINDOW}} \tag{4}$$

where $n_f(l)$ is the number of ETX probe packets successfully sent to the neighbor through the link l. While $n_r(l)$ denotes the number of ETX probe packets successfully received from the neighbor via the link l.

In ERCARP protocol, a path from source to destination with highest path cost, $C(p)$, is always selected for data delivery. With maximum value of $C(p)$ for the path, the value of ETX and RTT is minimized and the RE value is maximized. The maximum value of RE ensures that the selected path has high residual energy, i.e. energy-efficient path. On the other hand, the smaller ETX and RTT value give the data transmission with low packet loss and delay respectively. The $C(p)$ is calculated by using the following equation:

$$C(p) = \sum_{l \in p} C(l) \tag{5}$$

5 Protocol Operations

5.1 Route Discovery

When a source node requires a path towards a destination node for data communication, it broadcasts a Transmission Route REQuest (T-RREQ) packet to its all neighbors through all its available channels (i.e. not occupied by a PU).

If a T-RREQ packet reaches the destination node or an intermediate node which has a valid route towards the destination, it generates a Transmission Route REPly (T-RREP) and sends it back to the previous node through the same channel that the T-RREQ packet has been transmitted. The destination node will not discard the further T-RREQ packets received from the same node but on different channels.

In the route reply phase, an intermediate node which receives the first T-RREP packet creates a forward route pointing to the packet sender through the same channel that the packet has been received and then forward the copies of the packet over all its valid reverse routes with different vacant channels (i.e. not used by a PU) towards the source of the T-RREQ packet.

When a T-RREP packet arrives at the source node, a path from source to destination is created and the node can begin sending data packets. Also, the source node will not discard the additional T-RREP packets received from the same sender but on different channels.

5.2 Route Maintenance and Recovery

As data packets flow from source to destination, each node over the transmission path updates the lifetime (i.e. expiration time) of its forward and reverse routes in order to maintain the link connectivity. However, if a route's lifetime is expired, the routing entry for that route will be invalid.

In the ERCARP protocol, each SU sets a timer for every specific time to sense a PU activity. During data transmission, if a SU detects a PU activity, it is unable to transmit a data packet through the channel which overlaps the PU's transmission frequency, thus resulting in a link failure. For that case, with the fast and efficient route recovery mechanism, the node responds to the failure by immediately selecting another available channel or path for data delivery without the need to trigger a new route discovery process.

In addition, a link failure, which is detected by link-layer feedback, can result from not only PU activity but also node mobility, node fault, link degradation, etc. During data delivery, if a node detects a link breakage and no alternative available channel or path is found, a new route discovery process will be triggered.

6 Simulation Configuration

The efficiency of the ERCARP protocol is evaluated by using NS-2 simulator [10] with an extension to support the CR environments. The simulation area of 1000×1000 m^2 in which 100 movable SUs are located is specified. There are 10 PUs randomly placed in the simulation area. The PU activities are modeled according to the ON/OFF process with exponential distribution with parameter λ of 250, referred to as PU activity parameter [12]. Each SU sets a timer of every 5 s to sense a PU activity. There are 5 non-overlapping channels given for multi-channel data communications. The transmission range of SUs and PUs is set to 150 m. Additionally, to calculate the updated link cost, an ETX and RTT probe packet are periodically broadcasted to the neighboring nodes every 1 and 0.5 s respectively. The ETX time window is set to 10 s. The simulation parameters are summarized in Table 1.

In our network model, we assume that each SU is equipped with multiple wireless interfaces. Each interface can operate only on one of non-overlapping channels. The locations of PUs are assumed to be unknown to SUs.

Table 1. Simulation parameters.

Parameter name	Value
Simulation area	1000 × 1000 m^2
Simulation time	500 s
Number of SUs	100
Number of PUs	10
PU activity parameter (λ)	250
Number of channels	5
Source node	Node 0
Destination node	Node 99
Traffic type	CBR
Data packet size	512 bytes
MAC layer	IEEE 802.11
Transport layer	UDP
SU transmission range	150 m
PU transmission range	150 m
Radio propagation	Two-Ray Ground Reflection
PU activity checking interval	Every 5 s
RTT probe interval	Every 0.5 s
ETX probe interval	Every 1 s
ETX probe window	10 s

7 Performance Evaluation

The protocol performance is evaluated through simulations (with varying data rate) based on the performance metrics including percentage of packet loss, average end-to-end delay, average jitter and minimum node lifetime. The NS2 Visual Trace Analyzer [19] is used to analyze the simulation results. To validate the performance improvement of ERCARP protocol, the simulation results are compared with that of the D2CARP protocol [11] under identical scenarios.

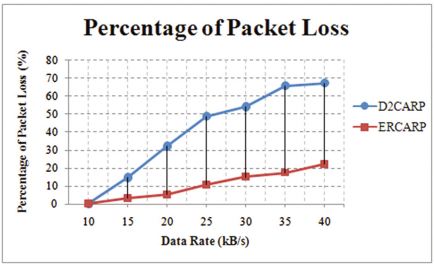


Fig. 2. Simulation results of percentage of packet loss.

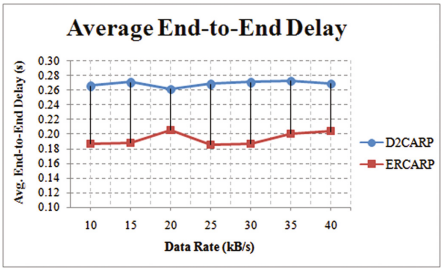


Fig. 3. Simulation results of average end-to-end delay.

Figure 2 shows the results of percentage of packet loss against the increased data rate. For D2CARP protocol, it is obvious that the percentage of packet loss grows dramatically when increasing the data rate. In contrast, the packet loss result for ERCARP protocol increases slowly in the same situations. In ERCARP protocol, the impact of packet loss is taken into account as a routing metric for path selection. Therefore, the ERCARP protocol provides significantly lower number of dropped data packets than the D2CARP protocol.

In Fig. 3, the simulation results of both protocols are evaluated in terms of the average end-to-end delay by increasing the data rate. Although the figure shows the fluctuation in the results of average end-to-end delay as the data rate increases, it is observed that the ERCARP protocol outperforms the D2CARP protocol in all cases. As expected, the ERCARP protocol utilizes the routing metric which considers the link delay for path decision, i.e. a path with lower latency is chosen for data transmission. As a result, the D2CARP protocol provides higher end-to-end delay result.

Figure 4 exhibits the comparative results of average jitter as the data rate increases. From the graph, when the data rate is increased, the ERCARP protocol provides lower average jitter as compared with that of the D2CARP protocol. The reason behind that is because, with more effective routing metric, the ERCARP protocol provides more reliable transmission paths. In addition, by utilizing the multi-path multi-channel routes, the ERCARP protocol keeps the data delivery process running continuously even in presence of path failures caused by PU activities. Therefore, the ERCARP protocol gives lower communication interruptions which produce smaller average jitter result.

In Fig. 5, the minimum node lifetime result of ERCARP protocol is compared with that of D2CARP protocol under the different data rate. For ERCARP protocol, a transmission path that has high residual energy is selected. From the simulation results, as compared to the D2CARP protocol, the ERCARP protocol can prolong the lifetime of individual SUs and overall network. As expected, with new routing metric in ERCARP protocol, the data packets are forwarded to SUs which have higher residual nodal energy; thus providing longer minimum node lifetime.

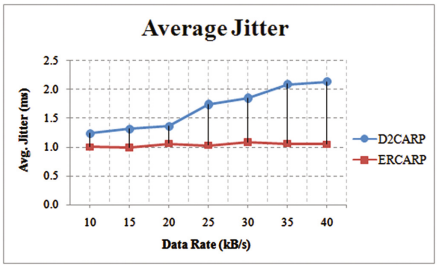


Fig. 4. Simulation results of average jitter.

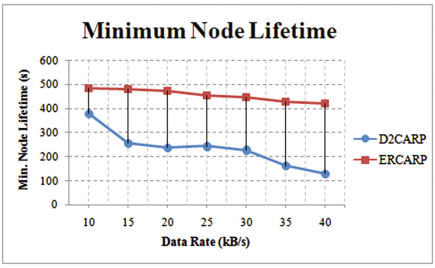


Fig. 5. Simulation results of minimum node lifetime.

8 Conclusion

In this article, we have proposed the ERCARP protocol which is an energy-efficient and reliable routing protocol for CRSNs. The novel routing metric that takes account of packet loss probability, link latency and residual energy is applied in the protocol. It ensures the selected transmission path has small delay, low packet loss and high residual energy. Furthermore, to offer the efficient route recovery mechanism, multi-path multi-channel routes are provided to cope with path failures caused by a PU activity. To evaluate the protocol effectiveness, the performance comparison between the proposed protocol and the D2CARP protocol has been conducted. As compared with D2CARP protocol, the simulation results obviously prove that ERCARP provides lower percentage of packet loss, reduces average end-to-end delay, decreases average jitter, and achieves longer minimum node lifetime in CRSNs.

References

1. Amanna, A., Reed, J.H.: Survey of cognitive radio architectures. In: Proceedings of the IEEE SoutheastCon Conference, pp. 292–297 (2010)
2. Haykin, S.: Cognitive radio: brain-empowered wireless communications. IEEE J. Sel. Areas Commun. **23**(2), 201–220 (2005)
3. Marinho, J., Monteiro, E.: Cognitive radio: survey on communication protocols, spectrum decision issues, and future research directions. Wireless Netw. **18**(2), 147–164 (2012)
4. Akyildiz, I.F., Lee, W.-Y., Vuran, M.C., Mohanty, S.: Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey. Comput. Netw. **50**(13), 2127–2159 (2006)
5. Puccinelli, D., Haenggi, M.: Wireless sensor networks: applications and challenges of ubiquitous sensing. IEEE Circ. Syst. Mag. **5**(3), 19–31 (2005)
6. Akyildiz, I.F., Weilian, S., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Commun. Mag. **40**(8), 102–114 (2002)
7. Akan, O.B., Karli, O., Ergul, O.: Cognitive radio sensor networks. IEEE Netw. **23**(4), 34–40 (2009)
8. Cavalcanti, D., Das, S., Jianfeng, W., Challapali, K.: Cognitive radio based wireless sensor networks. In: Proceedings of the 17th International Conference on Computer Communications and Networks, pp. 1–6 (2008)
9. Yau, K.L.A., Komisarczuk, P., Teal, P.D.: Cognitive radio-based wireless sensor networks: conceptual design and open issues. In: Proceedings of the IEEE 34th Conference on Local Computer Networks, pp. 955–962 (2009)
10. Issariyakul, T., Hossain, E.: Introduction to Network Simulator NS2. Springer, New York (2012)
11. Rahman, M.A., Caleffi, M., Paura, L.: Joint path and spectrum diversity in cognitive radio ad-hoc networks. EURASIP J. Wirel. Commun. Netw. **2012**(1), 1–9 (2012)
12. Chowdhury, K.R., Felice, M.D.: SEARCH: a routing protocol for mobile cognitive radio ad-hoc networks. Comput. Commun. **32**(18), 1983–1997 (2009)
13. Beltagy, I., Youssef, M., El-Derini, M.: A new routing metric and protocol for multipath routing in cognitive networks. In: Proceedings of the IEEE Wireless Communications and Networking Conference, pp. 974–979 (2011)

14. Kamruzzaman, S.M., Eunhee, K., Dong-Geun, J.: Spectrum and energy aware routing protocol for cognitive radio ad hoc networks. In: Proceedings of the IEEE International Conference on Communications, pp. 1–5 (2011)
15. Cacciapuoti, A.S., Calcagno, C., Caleffi, M., Paura, L.: CAODV: routing in mobile ad-hoc cognitive radio networks. In: Proceedings of the 3rd IEEE/IFIP Wireless Days Conference, pp. 1–5 (2010)
16. Spachos, P., Hantzinakos, D.: Scalable dynamic routing protocol for cognitive radio sensor networks. IEEE Sens. J. **14**(7), 2257–2266 (2014)
17. Shah, G.A., Akan, O.B.: Spectrum-aware cluster-based routing for cognitive radio sensor networks. In: Proceedings of the IEEE International Conference on Communications, pp. 2885–2889 (2013)
18. Perkins, C.E., Belding-Royer, E.M., Das, S.R.: Ad hoc on-demand distance vector (AODV) routing. In: RFC 3561, Internet Engineering Task Force (IETF) (2003)
19. Rocha, F.: NS2 visual trace analyzer (2012). http://nsvisualtraceanalyzer.wordpress.com/. Accessed 1 July 2016

A Novel on Automatic K Value for Efficiency Improvement of K-means Clustering

Se-Hoon Jung[1], Kyoung-Jong Kim[2], Eun-Cheon Lim[3], and Chun-Bo Sim[4(✉)]

[1] Department of Multimedia Engineering, GwangYang SW Convergence Institute, Sunchon National University, Suncheon, Korea
iaml710@hanmail.net
[2] Research and Development Team, GwangYang SW Convergence Institute, Suncheon, Korea
kkj0201@nate.com
[3] Harvard Medical School, Boston, USA
euncheon_lim@hms.harvard.edu
[4] School of Information Communication and Multimedia Engineering, Sunchon National University, Suncheon, Korea
cbsim@sunchon.ac.kr

Abstract. The development of H/W and S/W has shortened the repetition cycle of new data generation and produced various categories of data. Machine learning, in particular, attracts explosive interest as it categorizes and analyzes data through artificial intelligence and contests against man. Once generated, data have their importance highlighted in terms of utilization. It is critical to analyze the data from the past and cluster new data for the utilization of data. The present study thus investigated an algorithm of determining the initial number of clusters automatically, which is part of problems with the K-means algorithm used in data clustering. The study also proposed an approach of optimizing the number of clusters through principal component analysis, a pre-processing process, with the input data for clustering. Its performance evaluation results show the accuracy rate of 87.6% or so.

Keywords: K-means · K value · Clustering · PCA

1 Introduction

Today's modern society is entering a new age of smart revolution beyond information revolution. Smart revolution will enable the implementation of man- and object-based ways of thinking through the creative utilization of information and communication technologies and data. One of the core elements to lead the society toward the smart revolution is Big Data, whose production and expansion has been gaining momentum. Big Data include unstructured data on Facebook and Twitter and enjoy high-speed expansion. Also developing are researches on machine learning to predict the future based on the produced data. Machine learning is similar to Big Data in that it collects and analyzes data that are being produced anew to predict the future, but it is different

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_31

from Big Data analysis in that it can collect Big Data and learn with the system itself. Being a sector of artificial intelligence, machine learning has been in spotlight as a core technology of Big Data. There are two types of data analysis in Big Data analysis and machine learning: supervised learning makes an inference of a function based on learning data that have been entered and trains algorithms with the input data including labels. Its representative algorithms are regression analysis and classification. The other type is unsupervised learning makes an inference of relations with input data with the absence of labels and training data for the input data. Its representative algorithms are clustering and neural networks. Clustering is a statistical method of measuring similarity between input data and grouping similar data together. There are two types of clustering: hierarchical and non-hierarchical clustering. The former is a statistical method of putting together similar clusters sequentially based on the premise that all the input data form separate individual clusters, eventually allowing all the data to establish a single cluster. Its utilization peaks when the number of data to be analyzed is relatively small. When the number of input data grows, hierarchical clustering drops in its utilization and requires a lot of time and memory to calculate distance between data, thus facing limitations with application. Non-hierarchical clustering includes data that are the closest to the center of set clustering one by one with the number of clusters set in advance. Its good example is K-means clustering. Non-hierarchical clustering needs the number of clusters that has been set to calculate results, but it is a rare case to measure proper numbers of clusters in advance in actual data analysis. The optimal number of clusters is determined by changing the number of clusters and considering the results in most cases. Previous studies employed Rule of Thumb, R-Squared, Elbow, and Information Criterion Approach to measure the optimal number of clusters [1–7]. R-Squared and Elbow, in particular, confirm the optimal number of clusters by increasing them sequentially. This study thus set out to propose an optimization method for the number of clusters with principal component analysis-based high-dimensional data resolutions. The range of data for clustering encompasses from numerical data to categorical data. The range of data is distinguished according to the matrix. High-dimensional data process the reduction of dimensions through principal component analysis. The present study proposed an approach of obtaining the median values of high-dimensional data based on the determinant of dispersion matrix in each group of input data through the reduced label data and determining them with the number of clusters.

2 Related Work

2.1 K-means

K-means algorithm is a clustering technique to classify input data into k clusters based on unsupervised learning similar to supervised learning. Unlike supervised learning, which updates weight vectors every time a vector is entered, the K-means algorithm updates weight vectors simultaneously after all the input vectors are entered. The criteria of clustering classification are the distance between clusters, dissimilarity among clusters, and minimization of the same cost functions. Similarity between data

objects increases within the same clusters. Similarity to data objects in other clusters decreases. The algorithm performs clustering by setting the centroid of each cluster and the sum of squares between data objects and distance as cost functions and minimizing the cost function values to repeat cluster classification of each data object.

$$\text{Error} = \beta \sum_{i=0}^{k} (\textit{Intra CD}) - \gamma \sum_{i=0}^{k} (\textit{ICD}) \tag{1}$$

Intra-cluster distance (IntraCD) is the addition of distance to all the input vectors allocated to the concerned cluster from the centroid of each cluster. Inter-cluster distance (ICD) is the distance of weight vector between two clusters. As seen in Formula (1), errors are calculated by adding the sum of IntraCD of all the clusters and subtracting the sum of ICD for all the cluster pairs. β and γ are weighted values.

2.2 PCA

PCA is a technique of unsupervised learning to reduce information loss of multi-dimensional input vectors through analysis and to return them to lower-dimensional vectors. It is one of the multivariate data processing techniques presented in a couple of principal component values. When there is a vector of n dimension, eigenvector is obtained through average vector and variance-covariance matrix from the application of Formulas (2) and (3). Then eigenvector is arranged according to the size of the corresponding proper value to add a new matrix. The new matrix is applied as a transformation matrix to convert Vector x into Vector y as seen in Formula (4). Then new variables in Row y have non-correlation and are arranged in the order of monotone decreasing variance to reduce the dimensions with the big principal components of high variance value.

$$m_x = \frac{1}{M} \sum_{k=1}^{M} x_k. \tag{2}$$

$$C_x = \frac{1}{M} \sum_{k=1}^{M} x_k x_k^T - m_k m_k^T. \tag{3}$$

$$\text{y} = \text{nMatrix}(\text{x} - m_x). \tag{4}$$

Cluster Count Optimization Method of K-means

A method of clustering not based on a probability model usually uses a criterion to measure how similar or dissimilar prediction values are and performs clustering based on non-similarity (or distance) rather than similarity. K-means clustering minimizes the sum of distance squares within a cluster for the optimized number of clusters K. For its optimization, one needs to measure the sum of distance squares among the clusters for the input data and the number of clusters K and find the optimum separation of the smallest value. It is extremely important to determine the number of clusters K for

optimal clustering. The present study thus enabled the application of K to discrete and continuous variables to be entered. The multi-dimensional data of data vector were put to linear changes based on the dispersion matrix to analyze the number of clusters in the clustering of input data. The minimum values not in correlations with each other were treated as data objects and extracted as principal components. Figure 1 shows the clustering extraction algorithm of input data with the application of PCA.

```
Start K-means with PCA
 Centroid initial
 Centroid [] = K principal components = n objects * P variables
begin
  for k do
   k ≠ k(pre)
   k(pre) <- k
   compute object distances to k centroids(Euclidean distance)
     Assign each object to the cluster to which it is the closest
      for x do
       data point k in X do
        If new cluster then
          compute a new centroid for each cluster
        end
      end
     If convergence then
      compute less than the selected convergence criterion
     end
end
end
```

Fig. 1. The clustering extraction algorithm of input data with the application of PCA.

For the principal component analysis of entire input data, the primary principal component axis obtained through eigenvalue decomposition was projected onto the input data with the new axial values. Distance between each data was measured based on the axis where changes to the projected values became maximum. Data were extracted with the median values as the center of cluster based on the sum of measured distance squares. If the dispersion of projected values is assumed as the predicted center (z) of input data cluster, it will be expressed as in Formula 5.

$$\sigma^2 = \frac{1}{n}\sum_i (z \cdot w)^2 - (\frac{1}{n}\sum_i (z \cdot w))^2 \tag{5}$$

3 Experiment and Evaluation

In an effort to assess the proposed K-means optimization for the number of clusters, the investigator extracted approximately 1,000 articles from an Internet news group(sports) for a month and selected dates for analysis through dispersion matrix. Figure 2 shows

the clustering of dispersion matrix according to the chosen dates. Related articles and dates were analyzed through dispersion matrix. The analysis results show that the number of article words was high on Sundays. Those results were then used to evaluate the performance of the proposed approach, which applied the PCA-based method to produce the number of clusters.

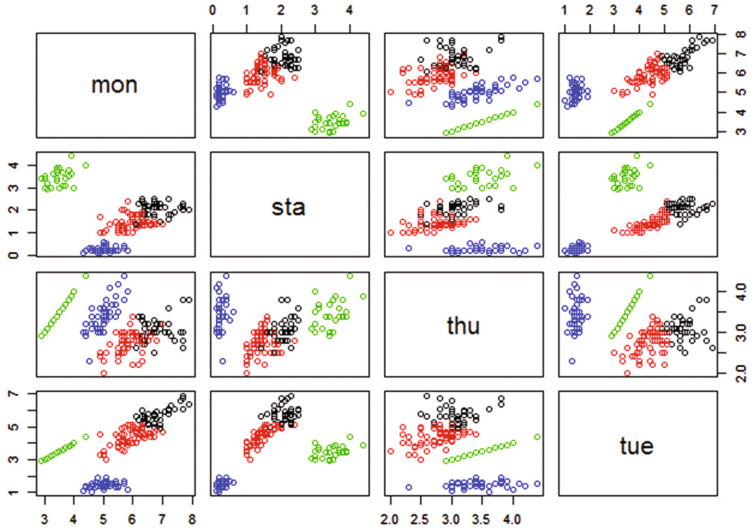


Fig. 2. The clustering of dispersion matrix (day)

The 500 sports articles issued on Sunday were clustered according to the sports events. There were total ten sessions of clustering based on PCA. Figure 3(a) presents the outcomes of clustering that determined the number of clusters based on PCA. Figure 3(b) presents the outcomes of clustering based on Elbow.

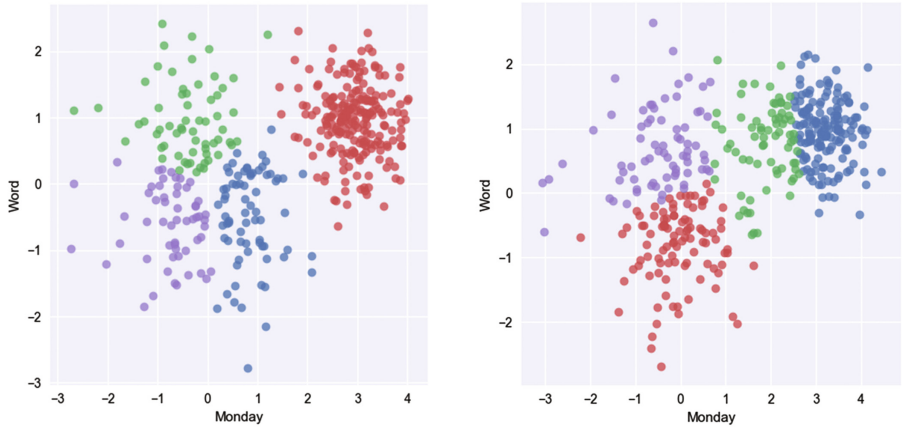


Fig. 3. The clustering result ((a) K-means with PCA, (b) K-means with elbow method))

The performance evaluation results show that the number of clusters based on PCA was K = 4 in all of ten sessions. When the 500 articles were examined in the PCA-applied clusters, however, it was found that about 2–15 articles were in a different cluster from the one they were supposed to belong to. The accuracy rate of clustering was approximately 87.6% after ten sessions.

4 Conclusion

The present study chose principal component analysis over the old methods of determining the number of clusters including Rule of Thumb, R-Squared, Elbow, and Information Criterion Approach and proposed an approach of optimizing the number of clusters through principal component analysis. The approach involved projecting the primary principal component axis, which was obtained by applying eigenvalue decomposition to the input data, as the new axial values of input data, measuring distance between each data based on the axis that maximized changes to the projected values, and extracting data with the median values as the center of cluster based on the sum of measured distance squares. The proposed approach was assessed for its performance, and the results show that the accuracy rate of clustering for the approximately 500 newspaper articles was 87.6%. Follow-up study will conduct a performance evaluation by applying continuous variables and supplement the boundary data of K-means algorithm to lower the error rate of clustering.

Acknowledgments. The research was supported by 'Area Software Convergence Commercialization Program', through the Ministry of Science, ICT and Future Planning (S0417161012).

References

1. Madhulatha, T.S.: An overview on clustering methods. J. Intell. Data Anal. **11**(6), 719–725 (2007)
2. Kodinaiya, T.M., Makwana, P.R.: Review on determining number of cluster in K-means clustering. Int. J. Adv. Res. Comput. Sci. Manag. Stud. **1**(6), 90–95 (2013)
3. Zhang, K., Bi, W., Zhang, X., Fu, X., Zhou, K., Zhu, L.: A new kmeans clustering algorithm for point cloud. Int. J. Hybrid Inf. Technol. **8**(9), 157–170 (2015)
4. Anderberg, M.R.: Cluster Analysis for Applications. Academic Press, New York (1973)
5. Khan, S.S., Ahmad, A.: Cluster center initialization algorithm for K-means clustering. Patter Recogn. Lett. **25**, 1293–1302 (2004)
6. Pena, J., Lozano, J., Larranaga, P.: An empirical comparison of four initialization method for the K-means algorithm. Patter Recogn. Lett. **20**, 1027–1040 (1999)
7. Jung, S.H., Kim, J.C., Sim, C.B.: Prediction data processing scheme using an artificial neural network and data clustering for Big Data. Int. J. Electr. Comput. Eng. **6**(1), 330–336 (2016)

Study on Integrity Verification and Compatibility-Conflict Analysis for Safe Patching

Jeongmin An, Sangmoon Jung, and Yoojae Won(✉)

Department of Computer Science Engineering, Chung-Nam National University, Daehak-ro. 99, 34134 Daejoen, South Korea
{jm_an, moon9l, yjwon}@cnu.ac.kr

Abstract. A Patch Management System (PMS) distributes and manages security patches for patch-server agents after collecting the patch files from software vendors. The PMS must account for the integrity and safety of the patch files to prevent huge damage arising from possible security incidents at the agents' environment. As software vendors cannot consider the patch compatibility of all patch-agent environments, the cause of a compatibility conflict must be analyzed when a patch fails. Existing PMSs manually verify the integrity of the patch files in a test environment. This study presents a method to automate patch testing and application, while monitoring the file modification, and reduce the time needed to analyze compatibility conflicts by using the modified file information.

Keywords: Patch management systems · Integrity · Compatibility conflicts · Automated patches

1 Introduction

Attacks that exploit software vulnerabilities are increasing. However, even after patches addressing vulnerabilities are released, it is difficult to apply them to the numerous varying PC configurations. Enterprises use patch management system (PMS), which collect patch files from software vendors and manage agent patches [1].

However, security incidents continue to occur because patch servers are hacked, and forged patches are applied to the agent. On March 20, 2013, three broadcasting stations and three banks in South Korea suffered hard-drive destruction, which paralyzed their computer networks. The hackers of the 3.20 cyber-attack used virus-infected computers to gain remote access to the affected vaccine servers and planted viruses into the vaccine servers' update files. The agents that downloaded the forged files were infected with viruses [2].

In March 2016, an agent PC that managed certificates was infected with malicious code after it downloaded forged files. Its managed certificates were leaked and threatened the integrity of the signed data.

To address the problem, this study proposes a method to automate the test process used by personnel responsible to manually verify the integrity of patch files and also to reduce the time required to analyze the cause of compatibility conflicts that may occur in a patch agent.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_32

2 Existing Patch Management Systems (PMS)

A PMS collects the patch files from software vendors under a server-agent structure and manages the patches of patch agents at enterprises. Figure 1 shows the composition and patching process of a general PMS.

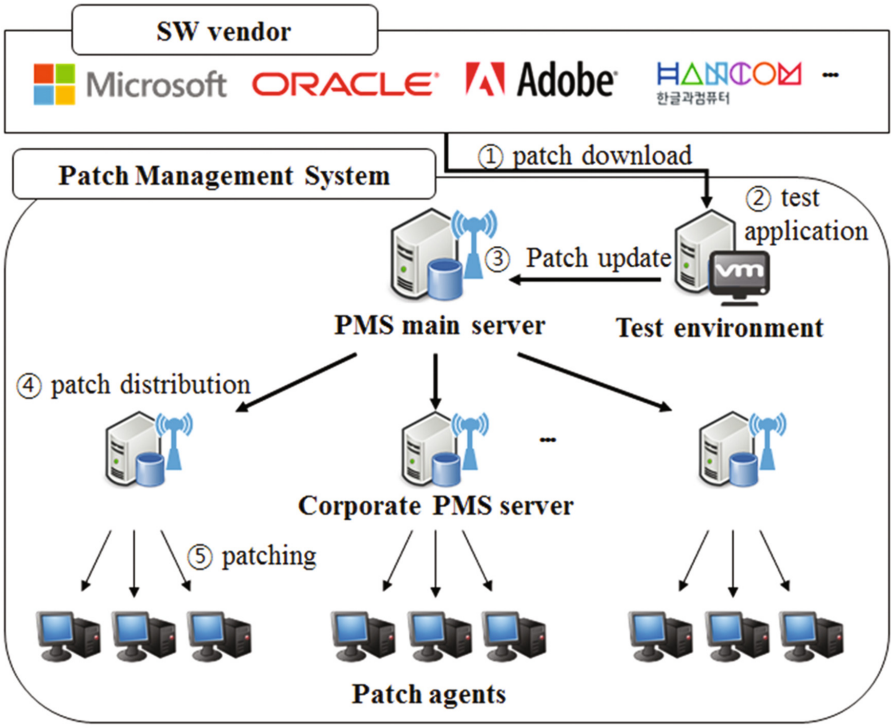


Fig. 1. PMS composition and patching process

After the software vendor distributes the patch file, the PMS emails a notice to the patch personnel and collects the patch file [3]. Upon downloading the patch file, the enterprise must verify the patch's integrity to prevent any security incident from happening to the agents [4]. Any time a new patch is released, patch personnel apply the patch file manually in a test environment to verify its integrity; they electronically sign it when the patching is complete, before distributing it to each enterprise PMS server. However, it is difficult for a software vendor to make a patch compatible with the agent environments of all enterprises. Therefore, automation is needed to accelerate the manual integrity test and the analysis to resolve agent patch-compatibility problems.

3 Automatic Integrity Verification Using Modified File Information

If a new security patch is available, the PMS downloads and installs it on a PC in a test environment. This study proposes a method for generating modified file information through file-change monitoring and providing it for the patch agent's use when the patch file is applied to the test environment.

3.1 Inspecting a Patch File in the Test Environment

The PC in the test environment selects the downloading route for the patch file to monitor the file generation. When the event is observed, the PC saves the file information of the downloaded software folder before it is applied. After saving the existing file information, it applies the patch and monitors the file modification. Figure 2 presents the proposed patch-test process.

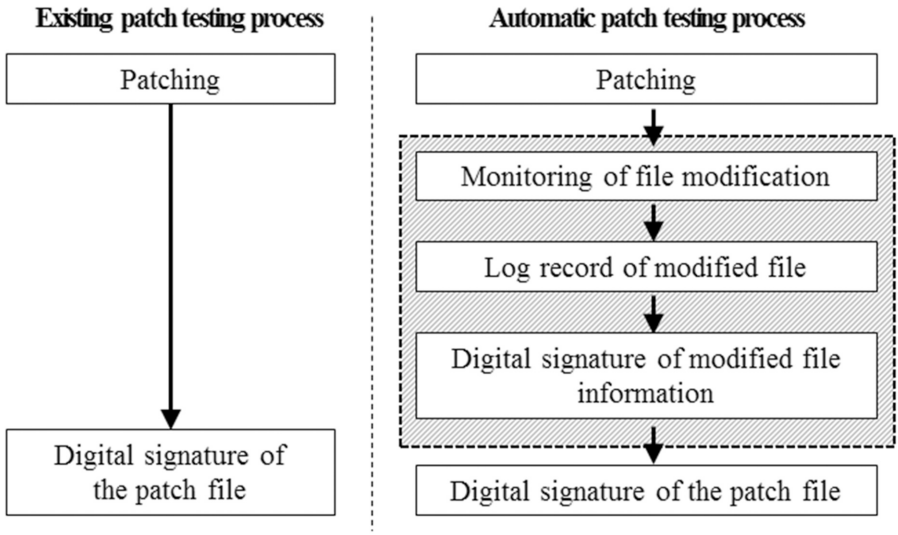


Fig. 2. Automatic patch test process in a test environment

In real time, the computer monitors the events of file generation, modification, and deletion in the patched folder and records the modified file information in order. If the patching is complete, the PMS main server electronically signs the information (shown in Table 1) collected by the file-change monitoring. When the signature is complete, the PC transmits the signed patch file and modified file information to the PMS server of each enterprise. The signature on modified file information uses XML (Extensible Markup Language) digital signature [5, 6].

Table 1. Patch-related information collected in the test environment

Collected information	Description
Patch file name	Used as integrity verification information for the patch file downloaded from the PMS server
Patch file size	
Patch file hash value	
Modified file name	When patching stops because of a compatibility conflict, the moment when the conflict occurs is notified to the patch personnel
Modified file path	
Modified file size	
Modified file hash value	

3.2 Applying a Patch File in the Agent Environment

Having received the signed patch file and modified information, the enterprise PMS server inspects the validity of the digital signature. If it is found to be valid, the PMS server electronically signs the patch file information again before distributing it to the agent. When the agent finds a valid signature on the sent patch file and patch information, it applies the patch, as in Fig. 3. If the signature is invalid, the agent reports it to the enterprise PMS to suspend its distribution.

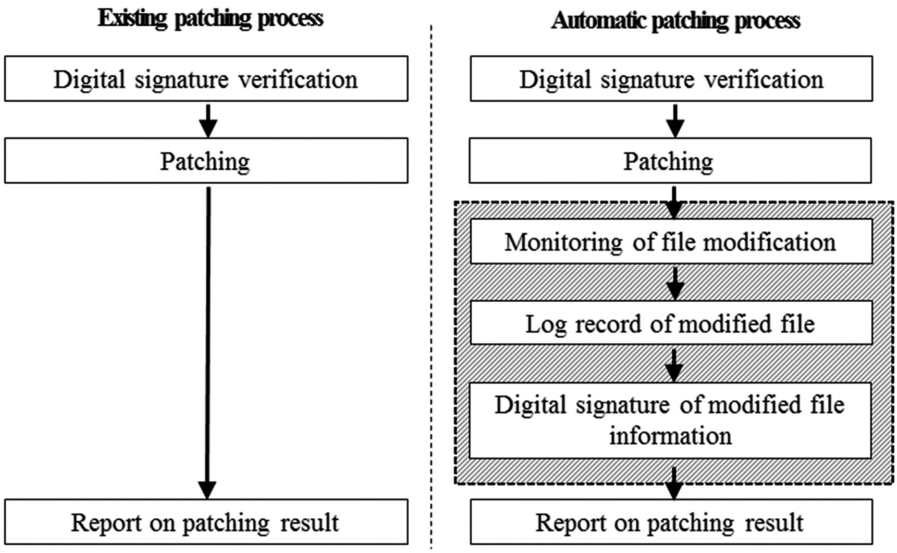


Fig. 3. Patch agent's patching process

After applying the patch, the agent records the modified file information by monitoring the file modification and compares the completed patch information with the transmitted modified information. If the file information modified by the patch matches the modified information sent, the agent reports the patching result to the

enterprise PMS server. If they do not match, it informs the patch server of the altered file so it can suspend the patch distribution and address the problem immediately.

3.3 Analysis of a Software Compatibility Conflict Using the File-Change Log

It is challenging for software vendors to consider all PC environments before distributing a patch. However, a security incident against an enterprise's client PC must be dealt with quickly, as it can inflict grave damage. If patching is suspended because of a compatibility conflict with the agent environment, the patch agent is patched, and the agent's change-log file and the change-log file sent from the enterprise PMS server are signed electronically by the agent and transmitted to the enterprise PMS server. The enterprise PMS server can find the moment of compatibility conflict by comparing the change log file of the agent that experienced the conflict and the change-log file when the patch was correct.

4 Conclusion

Since the alteration of a patch file and subsequent security incidents against an enterprise agent can result in serious damage, efforts to prevent and respond to them quickly are essential. It is also demanding for vendors to test whether a patch is successfully installed on all PC types before they distribute it.

This study enabled the automation of a new-patch test that was manually performed by the patch personnel using efficient allocation of work forces. It also notified the agent of a patch compatibility conflict that stemmed from an enterprise's patch-agent environment, helping reduce the time needed to analyze the cause.

It is expected that signing the patch file and patch information together, as described in this study, will strengthen the integrity verification procedure, save time solving compatibility conflict issues, and eventually allow for more stable patch management than in existing PMSs.

Acknowledgment. This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing and Testing Technology Development).

References

1. Secunia: Secunia vulnerability review 2015, March 2015. https://secunia.com/?action=fetch&filename=secunia_vulnerability_review_2015_pdf.pdf
2. KISA: Major internet incidents of South Korea, October 2013. http://www.kisa.or.kr/uploadfile/201310/201310071957453995.pdf

3. Zhao, D., Furnell, S.M., Al-Ayed, A.: The research on a patch management system for enterprise vulnerability update. In: WASE International Conference on ICIE 2009, vol. 2, pp. 250–253 (2009)
4. NIST.SP.800-40r3: Guide to enterprise patch management technologies, July 2013. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-40r3.pdf
5. W3C: Extensible Markup Language. http://www.w3.org/xml/
6. Eastlake III, D., Reagle, J., Solo, D.: (Extensible Markup Language) XML-signature syntax and processing, No. RFC 3275 (2002). https://www.rfc-editor.org/rfc/pdfrfc/rfc3275.txt.pdf

OFART: OpenFlow-Switch Adaptive Random Testing

Dong-Su Koo[1] and Young B. Park[2(✉)]

[1] Department of Computer Engineering, Dankook University, Juk-Jeon, Republic of Korea
ninedongsu@gmail.com
[2] Department of Computer Science, Dankook University, Juk-Jeon, Republic of Korea
ybpark@dankook.ac.kr

Abstract. In the advent of SDN paradigm, the accumulated verification technologies in the existing software fields are being used to verify the SDN. Data Plane consists of Forwarding Devices and is controlled by Control Plane. If correctness of the Forwarding Device is not verified, it affects to the whole network. However, doing every testing by manually is a huge time-cost consuming act, so it requires an automation. In this paper, it suggests a framework which applies ART (Adaptive Random Testing) technique which considers OpenFlow Switch to be Black Box from the Controller point of view and is easy to do a testing automation.

Keywords: SDN · OpenFlow · Adaptive Random Testing

1 Introduction

SDN paradigm has brought us new challenges. SDN is a networking paradigm which deals with networking direction settings and controls or complicated operating management through software programming. SDN has acquired programmability through its operating ways and structure in which Control Plane and Data Plane are logically separated [1]. The Control Plane controls over Data Plane which consists of Data Forwarding Devices (switch, router, etc.) like Operating System. There is one which represents SDN: OpenFlow and it virtually is located in de-facto standard. The behavior of OpenFlow is that OpenFlow Control Plane, called OpenFlow Controller, updates Flow Table Entries in Data Plane, called OpenFlow Switch, and OpenFlow Switch is processed by flow based on the Flow Table [2]. So, since the role of software has grown huge, we could manage as well as control the network more flexibly and closely. However, the side effects of the software highly made the risk of having more defects. One of the network factors that could not be found from the beginning can affect to the whole network. According to severity of the defects, it would cause a serious result which ranges from packet loss to paralysis of the whole network. For example, if a switch actually run by network could not do a packet processing function, it would have caused a serial accident in the end. Thus, testing a correctness of network function is a very basically and fundamentally important activity.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_33

The testing or SDN-OpenFlow verification research which has been done above are substituting the testing methods or the existing software fields from the researchers' point of views towards SDN. However, most of the researches make SDN approach to Finite State Machine and only do Formal Verification by using authentication tools as in theorem prover, model checking [3–9]; it is actually difficult for testers to implement because of high learning curve. In the paper, we suggest functional testing methods in which OpenFlow Switch regards to Black-Box in a Controller point of view. Controller generates packets for a functional testing and put the packets in a Switch. Getting the results from the switch which contains the packets, compare them to ideal results. A way to generating packet used Adaptive Random Testing technique. In order to lessen time consuming, tedious functional testing characteristics, we suggest OpenFlow Switch testing automation framework which can automatize such as test execution, test generation, test selection, and test result comparison.

2 Background

2.1 Black-Box Testing

Black-box Testing is one of the Software Testing techniques that test software activities about internal software without any knowledge. The black-box testing is widely being used in almost every software [10]. The paper watches Switch as Black-box from the controller's point of view because when Ingress Packet is sent from the controller, black box switch plays a role to do as a packet and throw them back to Egress Packet.

2.2 Adaptive Random Testing

Random Testing can be applied to test generation, since it has the most popular and easiest advantages. However, for the Random Testing, as it tests randomly and selects, it has some performative problems in balancing between error detection and test generation. To fix these issues, Adaptive Random Testing is suggested by Cheng. ART generates test data more effectively and evenly based on Failure Pattern in Input Domain, and there is a variety of ways existed and is being researched. This paper uses FSCS-ART which chooses next test by distance in between a fixed size test candidate set and executed test [11, 12].

2.3 Testing SDN-OpenFlow Switch

SDN-OpenFlow verification and testing research are substituting existing software verification or testing methods from the researcher's point of view towards SDN. OFTEN is regarded as black-box, so it proceeds integrated network testing and detects unrevealed inconsistencies through state-space exploration techniques [13]. Through simple processes, we suggested approaches about integrated network testing, but by

using the Model Checking technique, overhead factors about additional preliminary things are remained. SOFT is checking about Functional Equivalence of different OpenFlow agents to test inter-compatibility. Individually different OpenFlow agents' functional equivalence is checking over whether they return the same results or not through symbolic execution [3].

Automated Test Packet Generation (ATPG) automatizes test packet generation based on topology information of organization and suggest the ways to minimize the number of test packets [14]. For FlowTest, by using Al Planning Tool, it generates Test Plan and suggests the way of testing Data Plane [15]. In case of InSP, it gets information from Packet Template Table and Flow Table and suggests the way of generating packet [16]. J. Yao et al. [4] provided a way and a tool that are able to test Black-Box in SDN Data Plane. Based on OpenFlow switch specification which supports multi table, formal model switch in a specification language, and after it makes Data Graph and Data Paths based on this model, it uses a way to creating test case through those two things. However, there are limits that the flow table status of an actual switch that constantly changes cannot be instantly applied to the switch model. Besides, for OFTest, it suggests a test case about OpenFlow Switch [17].

3 OFART: OpenFlow Switch Adaptive Random Testing

OFART consists of three big parts (1) Test Packet Generation, (2) Test Selection, (3) Test Execution. The flow chart is illustrated in Switch Testing Tool provided by SDN Open Source Framework RYU [18].

3.1 Overall

The whole OFART Framework architecture is illustrated in Fig. 1. At first, test generator creates the number *N* of Test Data and stores them into Test DB. Test Selector randomly chooses one of the Test Data out of *N*. After the Test Data which is chosen by Test Selector is created as a Test Pattern, is sent to Tester. Test Pattern is a test script file which is written in JSON format. Tester makes Test Pattern a packet and gives the packet to the switch which is to be tested. After the packet given to the switch is matched with Flow Table within the switch, output packets taking actions accordingly. The egress packet decides Pass or Fail after comparing Desired Packet with Egress Packet through Test Executor. Decided results are sent to Test Selector. The Test Selector chooses 10 candidates from Test DB, makes Candidate Set, and has distance value from recently executed Data. If compared results are Pass, Test Selector chooses Test Data which is the farthest distance. If it is Fail, it stores the Data in Fail DB, chooses the closest Test Data. The chosen Test Data is sent back to Generator and made as a test pattern. It repeats the acts shown above.

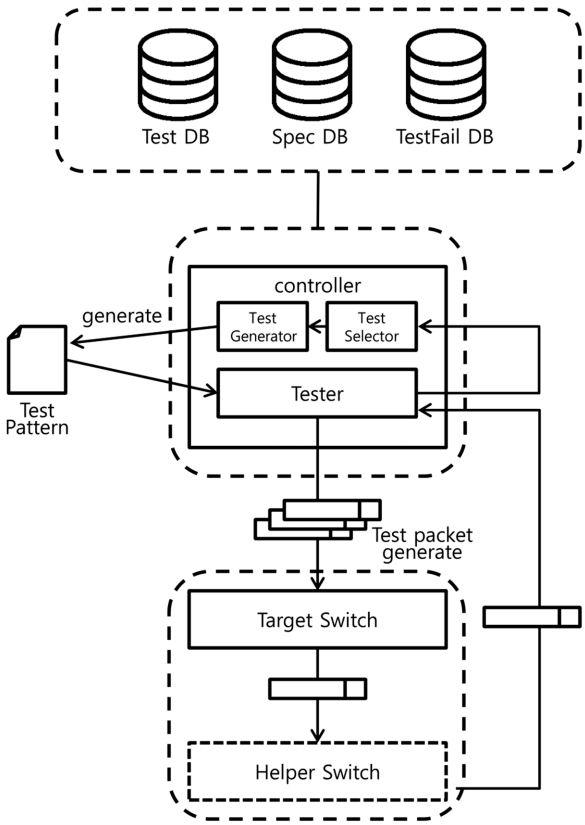


Fig. 1. High-level architecture of OFART framework

3.2 Adaptive Random Test Packet Generation

In this section, we describe of how Test Generator makes Test Pattern to make Test packet. Test Pattern is JSON which is to test OpenFlow Switch in RYU. Test Pattern consists of Prerequisite, Ingress Packet, Egress Packet. Test Generator generates Test Pattern written in JSON format by ART technique. The created Test Pattern is sent to Tester and Tester interprets the Test Pattern and makes Test Packet to test switch. All the processes will be named as 'Adaptive Random Test Packet Generate', and acronym of the term is ARTPG.

3.3 Similarity-Based Test Selection

FSCS-ART picks up test data which will be executed next based on the distance between candidate set and executed test. The next data would be chosen as farthest data from the test data which was executed before. Numerical Input Domain can measure the distance through Euclidian distance, but for OFART, each factors of Test Data have

a variety of shapes and scopes. In order to fix such problems, we defined a distance in between Tests based on similarity of tests. In other words, testing subset has a high similarity, and other functions a low similarity, Thus, tests with high similarity have a short weight: short distance, low similarity have a short weight: long distance.

Selector sends the test which is randomly chosen by Test DB to the Generator. And then, after calculating distances of the chosen test and randomly chosen 10 candidate tests, temporarily remember the farthest test, the closest test. After receiving a test result (Pass/Fail) through Generator and Test, if the result is Pass, sends the farthest test to Generator, or Fail, sends to the closest test to Generator.

4 Evaluation

In order to evaluate proposed method, we measure effectiveness using F-measure that is number of test cases required to detect the first defect. The defect is injected at a 'SET_DL_DST' function of Open vSwitch for F-measure. To compare traditional random testing with proposed method, the test case count and execution time are measured to find a injected defect. Setup time do not included in execution time. Table 1 shows simulation result of 100 times.

Table 1. Test result of random testing and OFART

	Random testing	OFART
Test case count	12	4
Execution time	0.924 s	0.346 s

As shown in Table 1, generated test cases of OFART are reduced 67% than random testing. Also, execution time is reduced 62.5%. As a result, OFART's performance show higher than random testing.

5 Conclusion

This paper applies ART technique in order for SDN Switch to do Functional Testing, suggests Framework which automatizes it. OFART can be used when Regression Testing and Tester are systematically verified in developed functions. Testers which will not involve in the development directly would be able to easily test the SDN Switch without implementing a high-learning curve. By applying ART technique, testing is under processing based on Failure Pattern, so it is more effective than the randomly pick-up way, and through automation, there would appear a reduction effect in the time and cost phases which are consumed by repeated actions of functional testing.

Acknowledgments. This research was supported by The Leading Human Resource Training Program of Regional Neo industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (No. NRF-2016H1D5 A1909989).

References

1. Nunes, B.A.A., Mendonca, M., Nguyen, X.N., Obraczka, K., Turletti, T.: A survey of software-defined networking: past, present, and future of programmable networks. IEEE Commun. Surv. Tutor. **16**(3), 1617–1634 (2014)
2. Open Networking Foundation: Software-defined networking: the new norm for networks. ONF white paper (2012)
3. Kuzniar, M., Peresini, P., Canini, M., Venzano, D., Kostic, D.: A SOFT way for openflow switch interoperability testing. In: 8th International Conference on Emerging Networking Experiments and Technologies, pp. 265–276. ACM (2012)
4. Yao, J., Wang, Z., Yin, X., Shiyz, X., Wu, J.: Formal modeling and systematic black-box testing of SDN data plane. In: 2014 IEEE 22nd International Conference on Network Protocols, pp. 179–190. IEEE (2014)
5. Guha, A., Reitblatt, M., Foster, N.: Machine-verified network controllers. In: ACM SIGPLAN Notices, vol. 48, no. 6, pp. 483–494. ACM (2013)
6. Skowyra, R., Lapets, A., Bestavros, A., Kfoury, A.: A verification platform for sdn-enabled applications. In: 2014 IEEE International Conference on Cloud Engineering (IC2E), pp. 337–342. IEEE (2014)
7. Ball, T., et al.: Vericon: Towards verifying controller programs in software-defined networks. In: ACM SIGPLAN Notices, vol. 49, no. 6, pp. 282–293. ACM (2014)
8. Al-Shaer, E., Al-Haj, S.: FlowChecker: configuration analysis and verification of federated OpenFlow infrastructures. In: 3rd ACM Workshop on Assurable and Usable Security Configuration, pp. 37–44. ACM (2010)
9. Kang, M., Kang, E.Y., Hwang, D.Y., Kim, B.J., Nam, K.H., Shin, M.K., Choi, J.Y.: Formal modeling and verification of SDN-OpenFlow. In: 2013 IEEE Sixth International Conference on Software Testing, Verification and Validation, pp. 481–482. IEEE (2013)
10. Jorgensen, P.C.: Software testing: a craftsman's approach. CRC Press, Hoboken (2016)
11. Chen, T.Y., Leung, H., Mak, I.K.: Adaptive random testing. In: Advances in Computer Science-ASIAN 2004, Higher-Level Decision Making, pp. 320–329. Springer, Heidelberg (2004)
12. Chen, T.Y., Kuo, F.C., Merkel, R.G., Tse, T.H.: Adaptive random testing: the art of test case diversity. J. Syst. Softw. **83**(1), 60–66 (2010)
13. Kuzniar, M., Canini, M., Kostic, D.: OFTEN testing OpenFlow networks. In: 2012 European Workshop on Software Defined Networking, pp. 54–60. IEEE (2012)
14. Zeng, H., Kazemian, P., Varghese, G., McKeown, N.: Automatic test packet generation. In: 8th International Conference on Emerging Networking Experiments and Technologies, pp. 241–252. ACM (2012)
15. Fayaz, S.K., Sekar, V.: Testing stateful and dynamic data planes with FlowTest. In: Third Workshop on Hot Topics in Software Defined Networking, pp. 79–84. ACM (2014)
16. Bifulco, R., Boite, J., Bouet, M., Schneider, F.: Improving SDN with InSPired switches (2016)
17. OFTest. http://www.projectfloodlight.org/oftest
18. RYU. https://osrg.github.io/ryu

Patch Alarm and Collecting System

Jaegeun Oh, Chaeho Cho, and Yoojae Won(✉)

Department of Computer Science Engineering, Chung-Nam National University,
Daehak-ro. 99, Daejeon 34134, South Korea
{dhvkd89, greatopen, yjwon}@cnu.ac.kr

Abstract. Since more systems require management by the patch management system, it is difficult for patch administrators to collect patches using existing methods. In addition, importance of managing Microsoft (MS) patch has been coming to the fore as occurrence frequency of MS related software's vulnerability is increasing. The paper proposed a system that enabled them to automatically check and collect the MS patches and manage them efficiently through automatic patch alarm and collection using WSUS.

Keywords: Patch management system · WSUS · Automation

1 Introduction

As PCs left unpatched for reported vulnerabilities continue to suffer security incidents, the patch management system or PMS has been increasingly adopted [1]. Various systems need patch management and, accordingly, patch administrators consider it cumbersome to visit each vendor's website to check and collect the announced patches as they have done so far.

The number of security vulnerabilities found from Windows also soared and, correspondingly, patch management for the Microsoft (MS) software has become quite important for the patch administrators [2]. Hence, automatic confirmation and collection of the MS patches are emphasized as an essential factor for downsizing the number of patch administrators and provide patches quickly. MS has two ways to offer patches. One is downloading them from its website, which can be automated through web crawling. The other is using Windows Server Update Services or WSUS, a patch management solution of MS. The challenge in web crawling is its high sensitivity to changing contents such as changes of web page size and format [3].

The paper proposed a system where patch administrators used WSUS to find and collect patches automatically, breaking away from the existing method of manually colleting them.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_34

2 Related Studies

2.1 Patch Management System

A patch management system is one where the central management server manages the entire patches for the operation system and software to raise the overall security level of a PC within a network.

If a software vendor announces a new patch, central patch administrators visit the vendor's website to find and collect it. They generate integrity information by applying a new patch to a test environment and encrypt it before distributing it to a patch administrator server of an organization or institute to prevent any forged file from causing security incidents.

As existing patch management systems made it difficult to perform patching manually, studies have been carried out to automate the processes. Under the circumstances, the automatic alarm and collection of MS software patches with the increasing number of vulnerabilities detected and security breached can be an efficient patch management method.

2.2 WSUS (Windows Server Update Services)

Microsoft makes regular security announcements on its website and an announcement has a Knowledge Base (KB) patch by specific software under it. WSUS is a patch management solution to provide and manage this KB patch information and files [4]. It operates with the Windows server and allows the administrator download the patch information and files automatically from the MS download center as well as control the patch distribution and management in person. It also saves patch-related information at a database called SUSDB, from which the patch administrator can collect the patch-related information and obtain the patch files [5].

3 MS Patch Alarm and Collecting System

3.1 Configuration of MS Patch Alarm and Collecting System

The MS patch alarm and collecting system proposed by the paper consists of the patch file server and client agent as shown at Fig. 1.

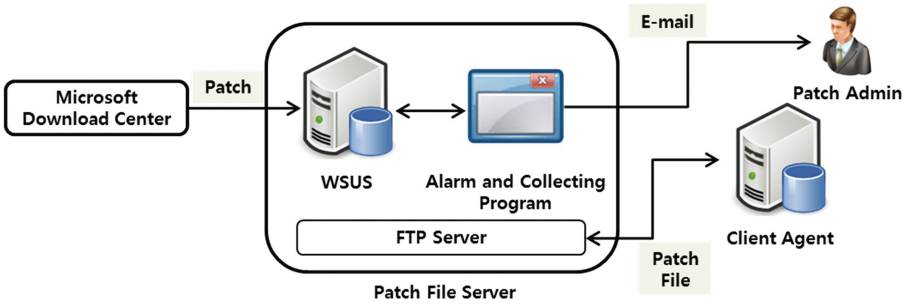


Fig. 1. Configuration of the proposed system

The patch file server has WSUS, an alarm and collecting program, and an FTP (File Transfer Protocol) server. WSUS is linked to the MS download center to automatically download the patches on a regular basis. The patch file is saved at the FTP server and made available to the patch administrator. The alarm and collecting program on a PC installed with WSUS monitors the changing status of SUSDB, collects the modified patch information, and sends an e-mail to inform the patch administrator thereof.

The client agent runs on the patch administrator's PC and downloads the patch file from the FTP server after it receives the notification e-mail from the patch file server. It uses Secure Sockets Layer or SSL for safe patch transmission.

3.2 Information to Be Collected

The patch administrator establishes a distribution policy to maintain the stability of the system to be patched. Under the policy, the administrator identifies a system that needs patching urgently for prior distribution. He should also be knowledgeable of what effects patching would have on the system in advance to prevent system operation error after it is patched. Patching should be done after the system to be patched finishes an important task. To set up the patch management policy, the system of this study collects the information of Table 1 from SUSDB to send it when a new patch is released and notified to the patch administrator via e-mail.

Table 1. Information about patches to be collected

Name	Description
SecurityBulletinID	To identify security notice showing new patches
KB number	To identify patching
KB title KB description	To verify description of new patches, identify effects of their application on the system and establish priority
MoreInfoURL	To confirm detail patching information
Target	To select the system for patching
MsrcSeverity	To determine the order of patch distribution
RebootBehavior	To schedule rebooting after patching
Superseded	To see the new patching information resulting from patching

3.3 Patch Alarm and Collection

SUSDB is monitored for patch alarm and collecting related information. Its patching list is saved based on a field named UpdateID, unique identifier by patching at the tbUpdate table. The tables that have patch information like Table 1 form relations with each other by referring to the tbUpdate table. The information can be accessed via UpdateID. When a new patch is downloaded or modified, the tbUpdate table including patch-related tables is modified and the version of patch modification is managed with a field named RevisionNumber.

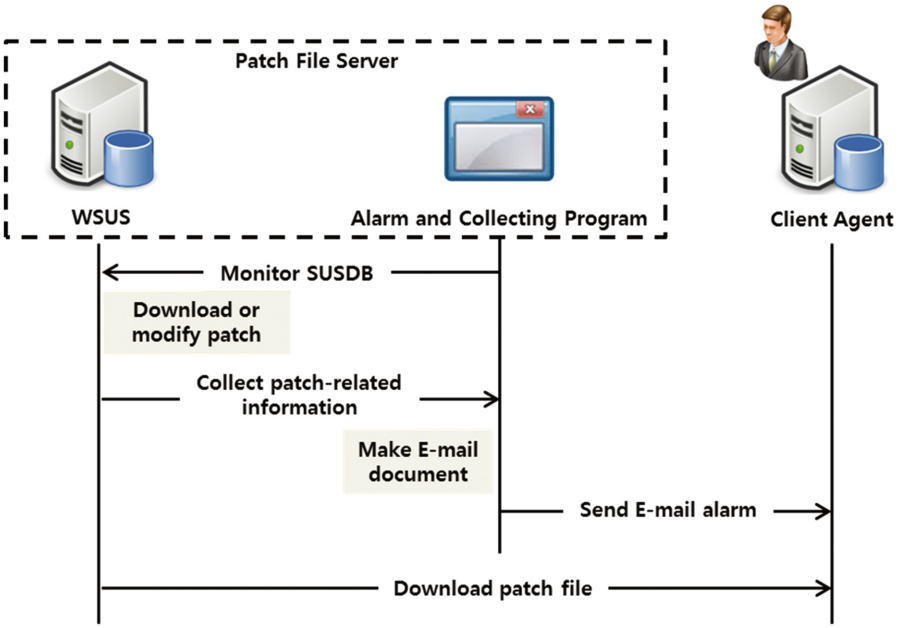


Fig. 2. Patch alarm and collecting process

Figure 2 demonstrates the patch alarm and collection, which is carried out automatically. The alarm and information collection program collects the modified patch information by monitoring the tbUpdate table regularly. It compares the existing tbUpdate table with the one updated by added and modified patches to detect any change in the patching list. It collects the modified patch information in response to MSSQL to be collected at Table 1. Upon completion of the collection, the program produces a collection document and transmits the e-mail alarm to the patch administrator. After he receives the alarm, the client agent downloads the patch file from the FTP server of the patch file server.

3.4 Operation and Verification

The patch file server composed of WSUS (ver. 6.3) and FTP server on Windows Server 2012 R2 (64 bit) applied C# to operate the alarm and collecting program as well as the client agent. For verification, the study installed the patch file server at a test network and client agent on the patch administrator PC to see whether the program receives the alarm e-mail and collects the patch file automatically when it is released.

The test proved that the patch alarm was sent to the e-mail account of the administrator and the patch file was downloaded automatically from the FTP server of the patch file server on the PC. Figures 3 and 4 below illustrate the result.

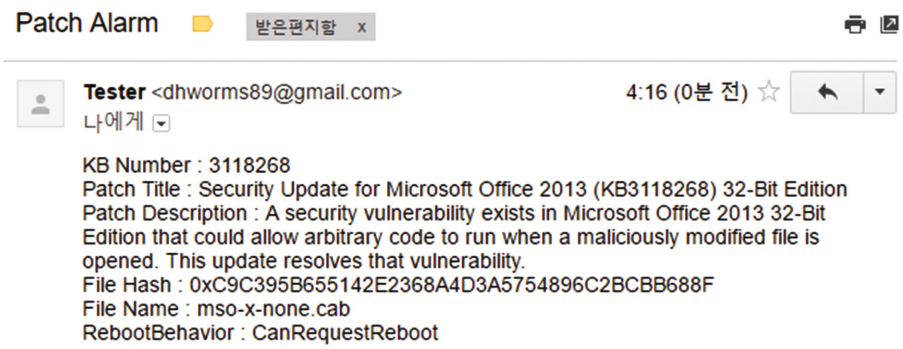


Fig. 3. E-mail transmitted to the patch administrator

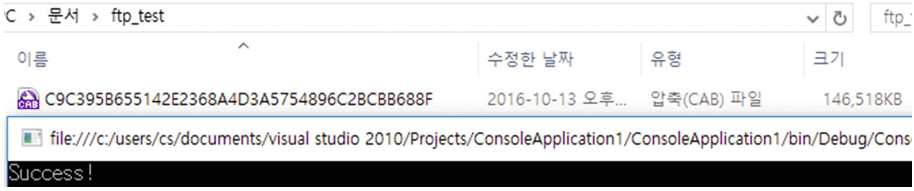


Fig. 4. Patch file downloaded by the client agent

4 Conclusion

Existing downloading methods made patch management difficult because the patch administrator had to find and download new patches by visiting the software vendors' websites in person to collect various operation systems and software patches. Though this process can be automated through web crawling, the crawling policy should be revised anytime the contents change.

The study proposed the system that enabled efficient MS software patch management. This system applied WSUS to overcome the weakness of web crawling and also automated the patch confirmation and collection that had been conducted by the patch administrator manually and this saved human labor. The system enables the patch administrator to check, collect, and distribute the patch quickly and eventually improve the efficiency of the MS patch management.

Acknowledgments. This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing and Testing Technology Development).

References

1. Why is Patch Management the Key to Corporate Security? AhnLab, Korea (2016)
2. Flexera Software: Vulnerability Review 2016 (2016)
3. Pei, J.: Information Retrieval and Web Search-Web Crawling (2010)
4. White, D., Irwin, B.: Microsoft Windows Server Update Services Review (2005)
5. Microsoft Corporation: Windows Server Update Services 3.0 SP2 Operations Guide (2009)

Method of Building a Security Vulnerability Information Collection and Management System for Analyzing the Security Vulnerabilities of IoT Devices

Kisu Kim, Jongsoo Lee, and Wonkang Jung(✉)

Lutes Co. LTD., 14th Floor, 155 Dongnam-ro,
Songpa-gu, Seoul, Republic of Korea
{kisukim, jslee, jwk}@lutes.co.kr

Abstract. This paper presents a method of building a security vulnerability information collection and management system in order to promptly identify which types of security vulnerability are included in IoT devices.

Keywords: IoT device · Security vulnerability · Security risk · Information collection

1 Introduction

According to the 2014 Gartner Report, application services based on IoT devices are increasing at a consistent rate, with the number of IoT devices connected to the Internet forecast to reach 26 billion by 2020. However, although more than 70% of the IoT devices used for such IoT-based application services are exposed to security vulnerabilities, there are problems in terms of information security as it is difficult to identify which IoT devices contain security vulnerabilities due to the diversity of protocols and system S/W used in IoT devices, the utilization of open source S/W, and so on.

The problems of information security that are inherent to such IoT devices can be exploited in various hacking attempts as IoT-based services are enhanced. Also, it might become increasingly difficult to identify which type of hacking attack has been used on which IoT device when an actual hacking attack is being made, and to determine what solution should be used to resolve such situations, ultimately leading to possible social confusion in which a hacking attack may expand to the many systems and information devices that are connected to the network.

This paper presents a method of building a "security vulnerability information collection and management system" with the aim of creating a system that performs periodic or real-time analysis of the presence of security vulnerabilities in various types of IoT devices. The method of building a security vulnerability information collection and information management system proposed herein can be utilized to build a responsive system that promptly identifies and responds to security vulnerabilities in IoT devices which are utilized in various services and to build the relevant control systems.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_35

This paper classifies the security vulnerability information collection and management system for IoT devices into the following functional aspects:

(1) System composition plan: Explains the software composition of the security vulnerability information collection and information management system.
(2) Collected object: Explains which security vulnerability will be collected from which information source.
(3) Collection motion: Explains how information on security vulnerabilities will be collected.
(4) Management of collected details: Explains how to manage the vulnerability information that is collected.

2 Main Points

The IoT market is beginning to take off as evidenced by the recent spate of launches of new products and services related to IoT. There are various environmental conditions in IoT, such as the connection method for IoT devices, networks, and properties of things. Just like existing Internet devices, IoT devices have various weaknesses and are vulnerable to security risks that take advantage of such weaknesses.

The "IoT security vulnerability information collection system" currently under development is a system that collects vulnerability information which is distributed through various routes periodically, analyzes and classifies such information, and processes and manages it so that it can be used efficiently in analyzing the vulnerability of the device.

2.1 System Composition

The "IoT security vulnerability information collection system" is being developed using JAVA language, and consists of a collection module, an analysis module, a classification module, and a collection management module according to its detailed functions.

When collecting data using a single file link, the collection model utilizes the file download function created with Java-based API. When collecting subordinate links and various content files in a complex manner starting with a specific URL, it collects data using Nutch[1], which is distributed by Apache.

The analysis module performs JSON parsing and analysis using the Gson[2] library according to the types of collected data (HTML, XML, JSON, EXCEL, etc.), parses and analyzes EXCEL documents using Apache's POI[3] library, and analyzes information provided in XML and HTML standards.

The classification module re-processes all collected information based on the analyzed data and builds relationships between the analyzed vulnerability data.

[1] Nutch, an open-source search engine created based on Lucene, provides excellent modularization for attaching various plug-ins. It enables Hadoop-based MapReduce and distributed file systems to perform tasks in multiple machines.

[2] Gson is a Java library that converts Java objects to JSON expression.

[3] Library, made by Apache Software Foundation, enables the reading and writing of Microsoft Office file format using pure Java language.

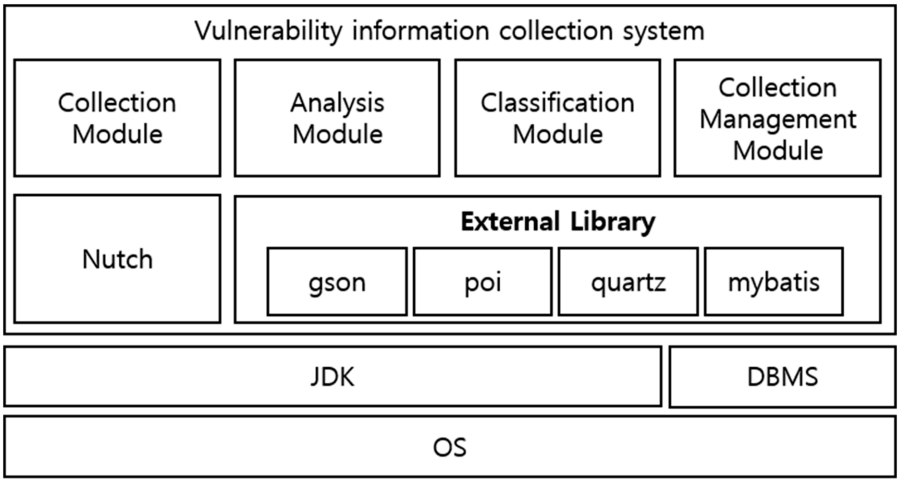


Fig. 1. System structure

The collection management module manages the collection schedule of the collection module, and stores and manages the collected data (Fig. 1).

2.2 Security Vulnerability Information to Be Collected

The national vulnerability database of the National Institute of Standards and Technology (NIST) and the security bulletins of the Microsoft Security TechCenter were selected as the primary objects of security vulnerability information collection by the IoT security vulnerability information collection system. Periodic collections and updates are made by the collection module.

2.2.1 NVD Information

Classification	Details
Download address	https://nvd.nist.gov/download.cfm
Data format	XML file
Scope of vulnerability collection	2002–Current
Collected information	CVE information, CPE information, CWE information, CVSS information

2.2.2 Information on Security Bulletins of the Security TechCenter

Classification	Details
Download address	https://www.microsoft.com/en-us/download/details.aspx?id=36982
Data format	EXCEL file
Scope of vulnerability collection	1998–Current
Collected information	Microsoft Security Bulletin information, CVE information, Information on affected software

2.3 Collection of Security Vulnerability Information

The IoT security vulnerability information collection system collects data from the national vulnerability database (NVD) of NIST and security bulletins from the Microsoft Security TechCenter anywhere from once to multiple times each day. The collection motion involves either accessing each URL which provides vulnerability information and automatically downloading the relevant files, or using Nutch to collect various contents (HTML, PDF, Doc, …) that are subordinate to the URL of the website. NVD information and Microsoft Security Bulletins, which are the current collection objects, are collected using the file download method (Fig. 2).

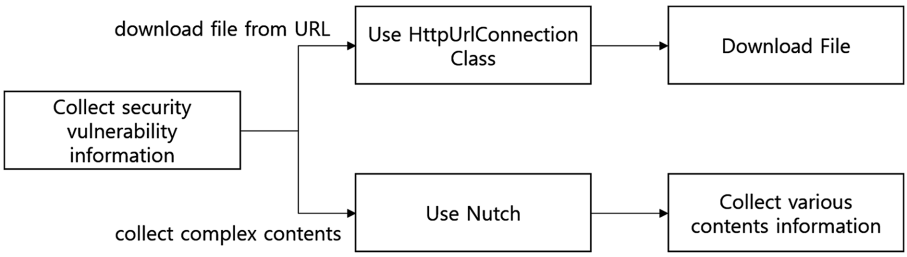


Fig. 2. Security vulnerability information collection motions

2.4 Analysis of Security Vulnerability Information

The IoT security vulnerability information collection system enables the parsing module (supports XML, JSON, EXCEL, HTML formats) to selectively analyze data according to the standards of the collected security vulnerability information data, and selectively uses the relevant parser according to the expression standard of the internal document for the collected vulnerability data in order to perform security vulnerability data analysis (Fig. 3).

The NVD security vulnerability information collected by this system is collected in the XML format. Once collection is complete, a document analysis motion using the XML Parser is performed. For the Microsoft Security Bulletins, a document analysis

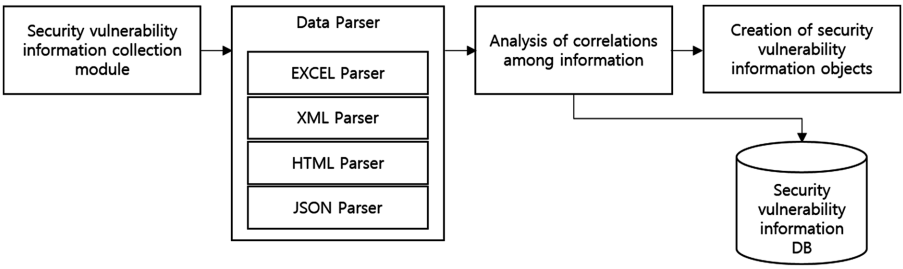


Fig. 3. Analysis of security vulnerability information

motion using the Excel Parser is performed. The results of the analysis are stored in a precomposed data management object (Java Value Object) as well as in the database.

2.5 Classification of Correlations Among Security Vulnerability Information

The IoT security vulnerability information collection system classifies the correlations among security vulnerability information based on the object(s) containing the security vulnerability information data that has been collected and analyzed (Java Value Object). The classification of correlations is conducted by first classifying the correlation between CVE and CPE based on the NVD vulnerability, between CVE and CWE, and then between CVE and CVSS in order to identify all correlations among the subordinate information for each vulnerability (Fig. 4).

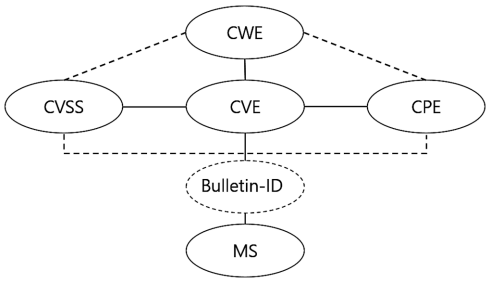


Fig. 4. Correlations among detailed vulnerability information

To classify the correlations among the security vulnerability information of Microsoft and NVD, the Bulletin-ID in CVE information and the Bulletin-ID in MS security vulnerability information are compared with one another in order to identify the correlation for the first time, and then the CVE-ID included in Microsoft security vulnerability information and CVE-ID analyzed using NVD vulnerability information are compared in order to classify the correlations.

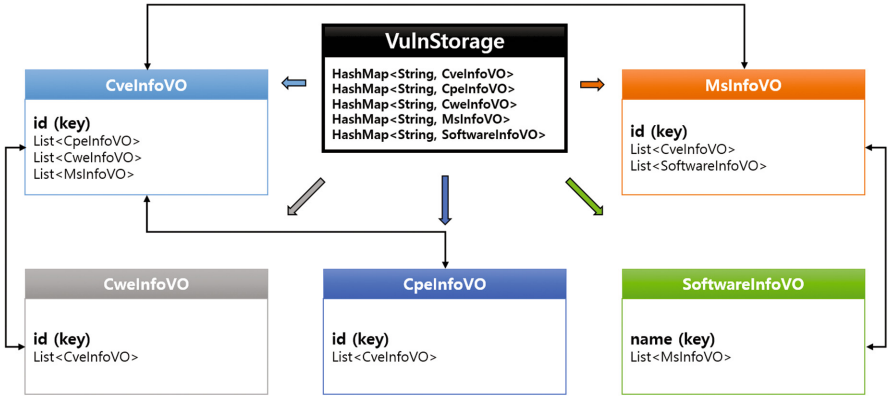


Fig. 5. Data objects of detailed vulnerability information

Vulnerability information thus classified makes it possible to establish and objectify the correlations among different collection objects as follows in order to be stored and managed (Fig. 5).

Based on the objects produced in this way, which store the correlations of security vulnerability, the security vulnerability analysis of IoT device information, a major function of this system, starts to operate.

Acknowledgments. This work was supported by Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing, and Testing Technology Development).

A Cluster Head Selection Method by Restricting Selection Area in Wireless Sensor Networks

Jong Won Lee[1], JiSu Park[2], Heung-Keun Kim[1], and Jin Gon Shon[1(✉)]

[1] Department of Computer Science, Graduate School, Korea National Open University, Seoul, South Korea
{neuronet, hgrikim, jgshon}@knou.ac.kr
[2] National Center of Excellence in Software, Chungnam National University, Daejeon, South Korea
bluejisu@cnu.ac.kr

Abstract. Energy conservation is one of the most important issues in selection of cluster head in wireless sensor networks. Traditional cluster head selection methods reduce cluster communication distance, but there is not enough increase in the network lifetime. In this paper, we propose CHSM-RSA (Cluster Head Selection Method by Restricting Selection Area). CHSM-RSA reduces cluster communication distance by using partitioning and restricting network area where nodes can be selected as cluster head.

Keywords: Cluster head selection · Restricting selection area · Wireless sensor networks

1 Introduction

The Wireless Sensor Network (WSN) consists of hundred and thousand number of small, relatively inexpensive, and low power sensor nodes deployed in broaden area. It can sense data, transmit and receive data, and aggregate data with the help of wireless communication technologies. A sensor node consists of communication unit, power unit, processing unit, and sensing unit. Resources such as battery, bandwidth, and computational capabilities constrain the sensor node [1]. Grouping the network in small clusters is an energy efficient way to prolong network lifetime in WSN. Clustering is a hierarchical structure where the CHs are in a higher level and the member nodes are in a lower level. The CH performs the task of data aggregation and the task of sending the aggregation data to the base station (BS) directly or through cluster heads (CHs) in a self organized method. Member nodes perform the task of sensing, measuring, processing, and transmitting of their data to the CH [2].

Cluster head selection plays an important role in energy efficiency in clustering algorithms. Intra cluster communication distance is decided by the position of cluster head in a cluster. The intra cluster energy consumption depends upon intra cluster communication distance. Therefore clusters with high intra cluster communication

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_36

distance consume more energy than other clusters with low intra cluster communication distance. Whereas inter cluster communication distance is decided by the position of cluster head in network area. The inter cluster energy consumption depends on the distance of CHs to BS. Therefore clusters with high inter cluster communication distance consume more energy than other clusters with low inter cluster communication distance [3]. In this paper, we use quartile, outlier, and node communication radius to restrict CH selection area. Network area is divided into three areas, namely inner area, middle area, and outer area on the basis of quartile, outlier, and node communication radius. All cluster heads are selected within the middle area only. As a result, it contributes to the reduction of the energy consumption in every node and the increase of energy efficiency in WSN.

2 Related Work

2.1 Quartile

Quartiles, a set ranked by distance values, are the three points that divide the data set into four equal groups. The first quartile is defined as the middle number between the smallest number and the median of the data set. The second quartile is the median of the data. The third quartile is the middle value between the median and the highest value of the data set. The interquartile range is the difference between the upper and lower quartiles. We counted the interquartile range according to Eq. 1. Outlier is an observation value that is distant from other values. We also counted outlier according to Eqs. 2 and 3 [4].

$$\text{Interquartile range} = \text{third quartile} - \text{first quartile} \tag{1}$$

$$\text{Outlier} < \text{third quartile} + 1.5 * \text{interquartile range, or m} + 1.5\text{SD} \tag{2}$$

$$\text{Outlier} > \text{first quartile} - 1.5 * \text{interquartile range, or m} - 1.5\text{SD} \tag{3}$$

Where m is mean value, and SD is standard deviation.

2.2 Traditional Cluster Head Selection Methods

Low Energy Adaptive Clustering Hierarchy (LEACH) is one of most well-known clustering algorithms for WSNs. It forms clusters based on probability. It uses the CH nodes as routers in the base-station [5]. But LEACH does not consider the intra cluster distance between the cluster head and the base station. Also it does not consider the inter cluster distance between the cluster head and the member nodes. The uneven distribution of nodes would lead to uneven distribution of cluster heads. And the cluster heads would be far away from the base station. Consequently, the cluster heads spend more energy forwarding the collected data [6]. Energy Efficient Unequal Clustering (EEUC) is an energy-efficient unequal clustering protocol for wireless sensor networks. It has different cluster radius of node according to the distance between node and BS.

Smaller clusters are formed in the region near the BS to reduce the members in cluster. Cluster head can use the saved energy to transfer the data coming from other clusters [7]. A new cluster head selection scheme (NCHSS) can be implemented, regardless of distributed clustering environment. In this scheme, network area is divided into two parts on the basis of distance (d). One is border area and another is inner area. Border area is from border of field to d. And inner area is from center of field to d. All cluster heads are selected within the inner area only. Nodes within border area will work as a cluster member all the time [1, 3].

3 Cluster Head Selection Method by Restricting Selection Area

3.1 Problem Statement

There are two types of communication in network divided into cluster. One is intra cluster communication and another is inter cluster communication. Intra cluster communication distance of a cluster is the total distance from member nodes to cluster head node. On the other hand, inter cluster communication distance is the total distance from cluster head nodes to base station. Cluster communication distances are the sum of intra cluster communication distance and inter cluster communication distance. A cluster head node positioned in outer area of network has longer inter cluster communication distance compared to a cluster head node positioned in middle area. Figure 1 shows a sensor network having 100 nodes over 100×100 m^2 area, and it shows cluster communication distance in outer area and inner area much longer than cluster communication distance in middle area. Figure 1(a) shows the case of CH in outer area, and Fig. 1(b) shows the case of CH in inner area.

(a) CH in outer area (b) CH in inner area

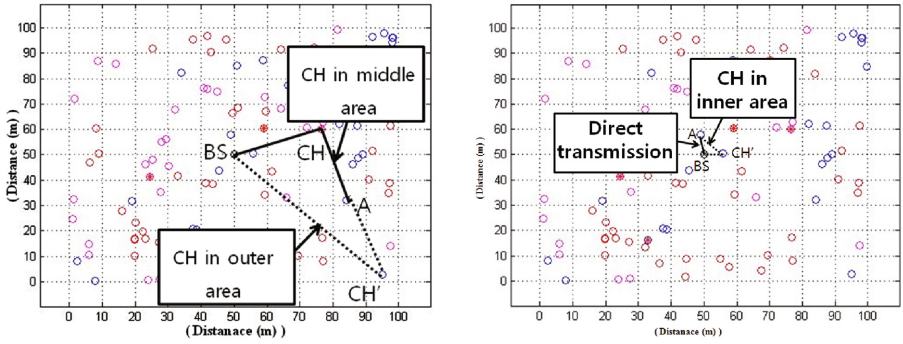


Fig. 1. Cluster communication distance depend upon position of CH

3.2 Cluster Head Selection Method by Restricting Selection Area

In this paper, we propose the energy efficient cluster head selection method. Aim of Cluster Head Selection Method by Restricting Selection Area (CHSM-RSA) is to reduce intra cluster communication distance and inter cluster communication distance. Therefore, CHSM-RSA decreases energy consumption in communication between nodes and BS. CHSM-RSA divides network area into three parts, the inner area, the middle area, and the outer area. Figure 2 shows division of network area.

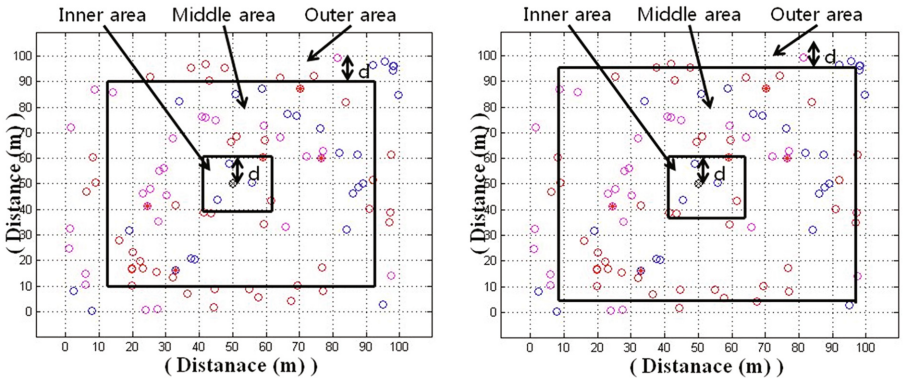


Fig. 2. Division of network area

Let d be the distance for partitioning of field. Outer area is from farthest node to the distance d. Middle area is from outside distance d to inside distance d. And inner area is from inside distance d to BS. The nodes within outer area and inner area will not participate in cluster head selection. Only the nodes within the middle area will participate in cluster head selection. The nodes within outer area are always member nodes in each round, and the nodes within inner area always communicate with BS directly. Cluster head is always selected from the middle area. Therefore there are more chances to be a cluster head. As a result, the cluster communication distance and the energy consumption is reduced. Because we use the value of distance d for partitioning network, deciding the value of distance d is important. Distance d is computed using two kinds of method which are explained as follows. First, we can compute distance d using the node communication radius. In case of Fig. 2(a), node communication radius is predetermined by 10 m for partitioning network. Second, we can compute distance d using quartile and outlier. In case of Fig. 2(b), we counted quartile, interquartile range, and outlier according to Eqs. 1, 2, and 3.

4 Performance Evaluation

4.1 Simulation Environment

In this paper, we compare CHSM-RSA with NCHSS and LEACH by using Matlab 2014a. 100 nodes were randomly distributed in a 100 m × 100 m with BS at center (50, 50). The number of cluster heads were set to 5% in order to coincide the simulation conditions. The environment variables are given in Table 1. T-test was conducted to examine mean comparison between cluster communication distances. SPSS version 18 for Windows was used for statistical analysis.

Table 1. Parameters of simulations.

Parameter	Value	Parameter	Value
Distance d	10 m	Network size	100 * 100
No. of nodes	100	Percent of CH	5%

4.2 Simulation Result

We counted the quartile, the interquartile range, and the outlier according to Eqs. 1, 2, and 3 through Fig. 2(b). The second quartile was 41.17 m, the interquartile range was 21.52 m, and the outlier was above 63.42 m and below 16.99 m. As a result, 6 nodes were in the inner area and 8 nodes were in the outer area. The results are given in Table 2.

Table 2. Nodes within inner and outer areas.

Node ID	Distance from BS	Area	Node ID	Distance from BS	Area
60	65.17	Inner	69	5.87	Outer
26	65.35	Inner	58	7.82	Outer
83	65.37	Inner	100	8.02	Outer
20	66.00	Inner	75	13.25	Outer
86	66.74	Inner	90	13.38	Outer
56	66.74	Inner	6	13.72	Outer
			22	14.10	Outer
			40	16.53	Outer

Simulation result of cluster communication distance is in detail as following. Figure 3 shows comparison of cluster communication distance between CHSM-RSA and LEACH, and CHSM-RSA and NCHSS. First, Average cluster communication distance was 2429.59 ± 32 m on CHSM-RSA and 2546.37 ± 94.49 m on LEACH. There was significant improvement in performance that average cluster communication distance of CHSM-RSA was 116.77 ± 79.71 m shorter than LEACH (<0.01). Second, Average cluster communication distance was 2429.59 ± 32 m on CHSM-RSA and 2437.07 ± 29.56 m on NCHSS. There was no significant improvement in performance. And the average cluster communication distance of CHSM-RSA was 7.50 ± 41.86 m shorter than NCHSS (<0.01).

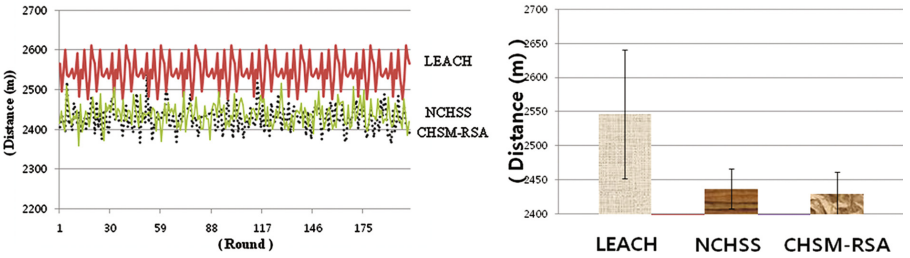


Fig. 3. The sum of distance between in this experiment

The results were given in Table 3.

Table 3. Cluster communication distance.

Parameter	M ± SD	p-value
LEACH - CHSM-RSA	116.77 ± 79.71	.000
NCHSS - CHSM-RSA	7.50 ± 41.86	.076

5 Conclusion

We proposed a cluster head selection method by restricting the selection area. Network area is divided into three area such as inner area, middle area, and outer area according to the distance d. The value of distance d is computed using the node communication radius, quartile and outlier. The nodes within inner area will be selected cluster head. Thereby, the cluster communication distance and energy consumption is decreased as above. According to the performance comparison, CHSM-RSA can reduce cluster communication distance compared to LEACH and NCHSS. Consequently, CHSM-RSA can reduce energy consumption and increase network lifetime evidently.

References

1. Biswas, S., Saha, J., Nag, T., Chowdhury, C., Neogy, S.: A novel cluster head selection algorithm for energy-efficient routing in wireless sensor network. In: 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 588–593 (2016)
2. Jain, N., Trivedi, P.: An adaptive sectoring and cluster head selection based multi-hop routing algorithm for WSN. In: 2012 Nirma University International Conference on Engineering (NUiCONE), pp. 1–6 (2012)
3. Pal, V., Singh, G., Yadav, R.P.: Cluster head selection scheme for data centric wireless sensor networks. In: 2013 IEEE 3rd International Advance Computing Conference (IACC), pp. 330–334 (2013)
4. Wikipedia: Quartile, Wikipedia (2016). https://en.wikipedia.org/wiki/Quartile. Accessed 11 Oct 2016

5. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: IEEE Proceedings of the Hawaii International Conference on System Sciences, January 2000, vol. 2, pp. 1–10 (2000)
6. Gao, H., Cheng, Y.: A relative distance based clustering scheme in wireless sensor networks. In: 2nd International Conference on Intelligent Computation Technology and Automation (ICICTA 2009), pp. 426–429 (2009)
7. Zou, Y., Zhang, H., Jia, X.: Zone-divided and energy-balanced clustering routing protocol for wireless sensor networks. In: 4th IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT), pp. 112–117 (2011)

An Improved Pedestrian Detection System that Utilizes the HOG-UDP Algorithm

Pyeong-Kang Kim, Hyung-Heon Kim, and Tae-Woo Kim(✉)

R&D Planning Team, INNODEP Co., Ltd., Seoul, Korea
david@innodep.com

Abstract. There is a high level of interest in pedestrian detection systems based on worldwide acknowledgement of pedestrian safety, and the need for research in this area is increasing. The HOG based pedestrian detection method proposed by Dalal and Triggs has been recognized as being less sensitive to the clothing and poses of pedestrians and also changes in lighting and is therefore one of the main methods used for pedestrian detection. But because the HOG based method requires a significant amount of computations, it is difficult to implement this method in real-time. Therefore in this research study, to improve the speed and detection rate for pedestrian detection, a pedestrian detection method that was improved upon by reducing dimensions of the particular feature vector extracted using UDP dimension reduction was proposed, and the results of performance evaluation showed that compared to previous HOG, HOG-PCA and HOG-LPP etc. algorithms, the speed and detection rate of the proposed algorithm were confirmed to have been improved.

Keywords: Pedestrian · Pedestrian detection · Intelligent systems · HOG

1 Introduction

There is a high level of interest in pedestrian protection systems through worldwide acknowledgement regarding the safety of pedestrians, and the need for research in this area is increasing. Methods for extracting features to detect pedestrians that are heavily being researched include the Haar wavelet based method, which is a method that utilizes the directionality of slopes, and the LRF (Local Receptive Field) method etc., and these methods can be categorized into to the SVM (Support Vector Machine) models, neural network models, and the AdaBoost algorithm. [2] Out of these methods, the HOG based pedestrian detection method [1] proposed by Dalal and Triggs that has been recognized as being less sensitive to both changes in lighting and also the clothing and poses of people and is the most widely used method. But the HOG based pedestrian detection method has the issue of requiring a significant amount of computations and degraded performance due to feature dimensions. And while there are

This work was supported by the ICT R&D program of MSIP/IITP.
(B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_37

multiple research studies being conducted to resolve these issues, PCA and ICA, which are the main dimension reduction method algorithms that are used, are unable to express location related information that is present in the data because they utilize the characteristics of the entire data, and the LPP method has a drawback in that it does not reflect information related to the overall structure because it only minimizes location related information [3–5]. Therefore, in this research study, to address the issues mentioned above, an improved pedestrian detection algorithm was designed and proposed based on the UDP algorithm, which utilizes a manifold learning based dimension reduction functionality.

2 Designing a Pedestrian Detection Method Using HOG-UDP

2.1 Problems with Previous Research Studies

In the case of SVM, which has a high sorting accuracy, because the space saved by the support vector is determined by the input dimensions, when there are few feature dimensions, the speed can be improved because the sorting time is shortened. And because issues such as curse of dimensionality are addressed when reducing the dimensions, various related research studies are being conducted in this area.

Because the PCA algorithm is the main algorithm used in the area of dimension reduction and expresses level of importance based on the size of data variation, dimensions can be reduced while minimizing differences in the information between the original data using only a few of the higher priority elements. In addition to PCA, the ICA algorithm extracts the independent elements of multi dimensions using statistical characteristics. The LPP algorithm, which is a manifold learning method, maintains the distance relation between the two data points as much as possible while having the objective of finding a low dimensional structure.

However, because the features of the data as a whole are used for the PCA or ICA algorithms, it is not possible to express local information that exists within the information, and in the case of the LPP algorithm, and by locating the data in the vicinity of the circular data space to also be close to the low dimension space that has been linearly transformed, because the objective of this algorithm is to preserve the local data structure as much as possible, there is the issue that the structural information as a whole is not reflected. When this algorithm is used to detect pedestrians, it can affect the background or categorization, and for pedestrians that actually exist in the video feed that are blocked or overlapped, it is not possible to show a strong image and this can result in a decreased rate of detection. Therefore, in this section, to address the two issues mentioned above, as one method of implementing the manifold learning method, using UDP, which improves upon the weaknesses of the LPP, to reduce the dimensions, both the local and overall characteristics are reflected to design a pedestrian algorithm that improves the speed and detection rate.

2.2 Pedestrian Detextion Methods Using the HOG-UDP Algorithm

The pedestrian detection methods using the HOG-UDP algorithm goes through a detection process and a learning process, and the details of the detection process are as follows. First, to improve the detection rate for the video input feed, the feed is pre-processed. Second, the features are extracted using HOG by calculating the gradient of the video, and from this calculated gradient the magnitude and direction are calculated. Further details on extracting details are shown in the paragraph below (Fig. 1).

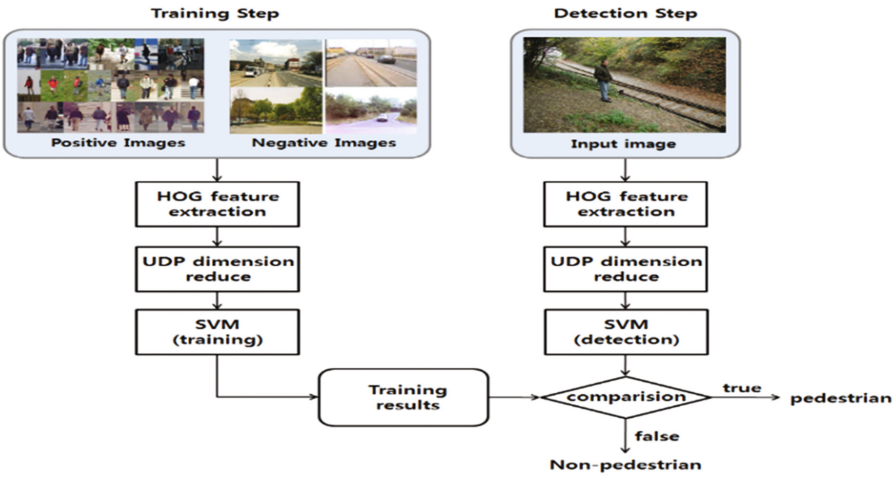


Fig. 1. HOG-UDP

2.3 Extracting HOG Algorithm Features

The first step is to calculate the value of the slope. The most common method is to apply a one-dimensional center point discrete differential mask in both the horizontal and vertical directions as shown in formula (1)

$$D_X = [-1\ 0\ 1], \quad D_Y = [-1\ 0\ 1]^T \tag{1}$$

Then when image I is given, convolution calculations are conducted on both the mask and image of formula (1) in the x and y directions.

The second step is to calculate the magnitude and the direction of the slope for which the I_X and I_Y values calculated in step 1 are used. The third step is related to calculating the cell histogram. The direction of the slope are calculated using the calculations of the slopes derived using the pixel values of each of the divided cells, and these values propagate to each of the direction histogram ranges that are configured as bin numbers. The fourth step requires local normalization, which in turn requires the magnitude of the slopes to be grouped in terms of connected blocks that are connected in bigger and more spacious blocks in order to take into consideration the changes in

the lighting and contrast. Through this process, the features of the HOG algorithm are populated into cell histogram vectors that normalize cells of the block scope. The L2-Hys method was used in this research study. The L2-Hys is calculated using the L2-norm, and the maximum value for υ is 0.2. n is the number of pixels included in the block and ε is a constant that does not significantly affect the results that is used to prevent error. Therefore because the overlap ratio was set as 50% based on a 64×128 pixel image in this research study, there are a total of $7 \times 15 = 105$ blocks, and because each cell has 9 bins, each block has $4 \times 9 = 36$ feature vectors. In this research study a total of $105 \times 4 \times 9 = 3780$ features are deduced.

2.4 Dimension Reduction Using the UDP Algorithm

Once the HOG algorithm features vectors are extracted, dimension reduction must be conducted using the UDP algorithm. The first step for UDP dimension reduction is to set the value of the parameter K, which decides the number of neighbor values once the input data, or in other words the HOG Feature Vector = $[f_1, f_2, \ldots, f_N]$, using the HOG algorithm, is input. If the K value is set to a value too small, a problem where the entire graph cannot be completed connected can occur. On the other hand, because if the K value is too large the chance of using a distance value that does not take into consideration the data structure increases, it is necessary to set an appropriate K value. This depends on the density of the input data collection, and therefore the K value was set to 5 in this research study. The second step is to connect the two data points (x_i, x_j), which are included as neighboring values to K-nearest, as the edge, and also to include the two data points (x_k, x_l), which are not included as neighboring values to K-nearest, as the edge (e_{ij}). The two distances connected for both of the edges are defined as the euclidean distance through formula (2).

$$\left|e_{ij}\right| = \left\|x_i - x_j\right\|^2 \tag{2}$$

The third step is to have the UDP algorithm simultaneously minimize the Local scatter so that the adjacent data in the circular space is also adjacent in the transformation space, while also maximizing the non-Local scatter. To accomplish this, the Local scatter is minimized using formula (3) and the non-Local scatter is maximized through formula (4). The distance matrix d is calculated using formula (1), and matrix B, which is used to analyze the eigen values, is calculated. Formula (5) is used to analyze the eigen values.

$$J_L(\omega) = \frac{1}{2}\sum\nolimits_{i=1}^{M}\sum\nolimits_{j=1}^{M} H_{ij}(y_i - y_j)^2 \tag{3}$$

$$J_N(\omega) = \frac{1}{2}\sum\nolimits_{i=1}^{M}\sum\nolimits_{j=1}^{M} (1 - H_{ij})(y_i - y_j)^2 \tag{4}$$

$$\mathrm{B} = \frac{1}{2}(I_N - 1_N)D^2(I_N - 1_N) \tag{5}$$

In formula (5), D^2 refers to $d_{ij}^{2_{i,j=1\ldots N}}$, and 1_N refers to an N × N matrix where every element has the value $\frac{1}{N}$. The fourth step is to calculate the eigen value matrix and the eigenvector matrix through eigenvalue analysis. Formula (6) is used to calculate the eigenvector matrix.

$$\mathrm{B} = U_B \Lambda_B U_B^T \tag{6}$$

In formula (6), U_B is an eigen vector matrix, and refers to a matrix that has the eigen vectors of matrix B as each of its rows. Λ_B is the eigen value matrix.

The fifth step is to calculate the transformation vector ω for formula (1). The transformation vector ω can be calculated using the eigenvector that corresponds to the largest eigenvalue that satisfies Eq. (7).

$$S_N \omega = \lambda S_L \omega \tag{7}$$

Finally the projection axis of the UDP algorithm is selected according to the eigenvectors $\omega_1, \omega_2, \ldots, \omega_d$ of Eq. (7) that correspond to the largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ \ldots \ \geq \lambda_d$ of d, and once the projection axis is calculated, a new feature vector of m dimensions can be calculated using formula (8).

$$\mathrm{HoG} - UDP_i^k = \sqrt{\lambda_k \omega_k^i} \tag{8}$$

The vector ω_k^i in formula (8) refers to the i th element in ω_k.

2.5 Categorization of SVM

Using the INRIA Person Dataset given by SVM, the feature vectors extracted from the HOG features from 2,146 positive images and 1,218 negative images were processed (learned) after conducting UDP dimension reduction. Based on the processed (learned) data, the final categorization process is conducted.

3 Implementation and Testing

To implement and evaluate the performance, this testing was conducted using Window 7 OS on an Intel Core2 Duo E8400 2.99 Ghz machine with 4096 MB of RAM. The INRIA Person Dataset was the pedestrian dataset to be referenced for implementation and performance evaluation. The INRIA Person Dataset is made up of a total of 5,213 images, which included 64 × 128 pixel dimensions images of pedestrians and other images of various sizes that are not of pedestrians. Figure 2. shows the example screens. To evaluate the pedestrian detection algorithm that utilizes HOG-UDP implemented in this research, Matlab 2012b was used to comparatively evaluate the results of this research with those of previous research.

Fig. 2. Screen embodiment

3.1 Performance Evaluation According to Dimension

Table 1 shows the comparison of performance efficiency according to the dimensions of HOG-UDP. To find the dimension with the highest efficiency, the evaluation was conducted from 10 to 110 dimensions. As can be seen from the results in Table 2, in the case of HOG-UDP for 50 dimensions, the detection rate was shown to be 92.83% and the average detection time was shown to be 13.879 ms, showing that this case had both a superior detection rate and execution time compared to those of other dimensions. Therefore the dimension for the HOG-UDP algorithm used in this research study was set as 50.

Table 1. Comparison of efficiency

Dimension	Detection rate (%)	Duration (ms)
10	72.87	8.457
30	81.13	10.136
50	92.93	13.879
70	93.02	14.115
90	92.71	14.945
110	91.69	15.263

Table 2. Comparison of proposed algorithm and existing algorithm

Category	Proposed algorithm	HOG-LPP	HOG-PCA	HOG
Pedestrian detection rate $(10^{-4}\ FPPW)$	**92.93%**	92.31%	91.58%	88.95%
Average detection duration (ms)	**13.879**	14.054	15.321	20.174

3.2 Performance Evaluation According to the Dimension Reduction Method

Figure 3 is a graph showing previous HOG algorithms and an HOG algorithm that utilizes the main dimension reduction methods in this research, and Table 2 shows a comparison of the detection durations for each of the algorithms. To conduct performance analysis for the HOG-PCA and HOG-LPP algorithms, the HOG-PCA algorithm was set to 400 dimensions and the HOG-LPP algorithm was set to 70 dimensions, which was the dimension setting that showed the highest performance based on previous research. Finally, for reasons mentioned above, the proposed HOG-UDP was set to 50 dimensions.

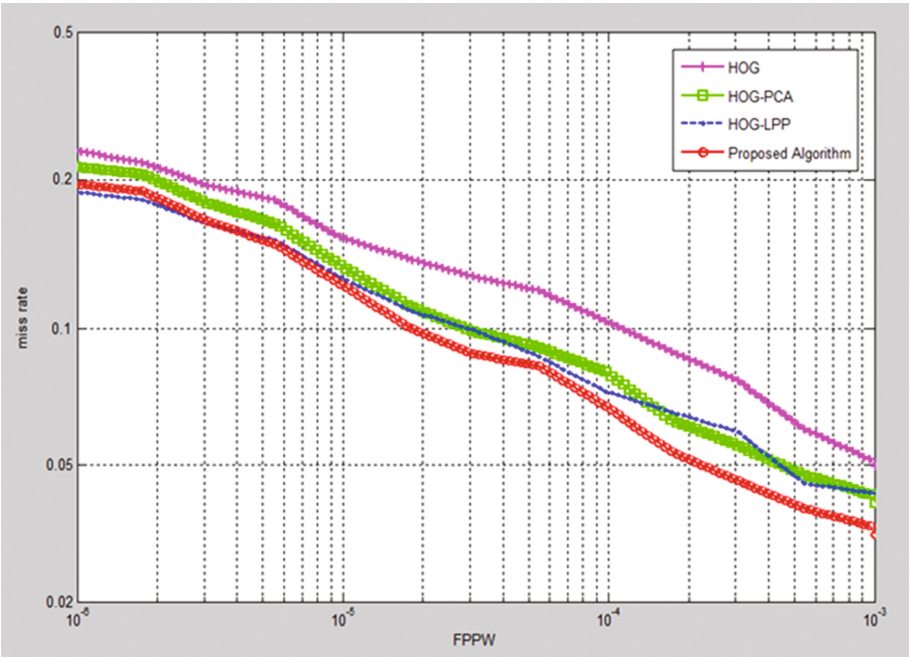


Fig. 3. HOG-UDP performance evaluation

4 Conclusion

This research study designed and implemented an improved pedestrian detection algorithm using a UDP algorithm that included a dimension reduction functionality based on manifold. To design the algorithm proposed in this research study, the first step was to extract features from the HOG. The second step was to reduce the dimensions of the extracted HOG features using UDP. The final step was to design and implement a pedestrian detection algorithm using SVM. Previous pedestrian detection algorithms and the pedestrian detection algorithm proposed in this research study were

comparatively evaluated using the Person Dataset, and the results of the analysis confirmed that compared to previous pedestrian detection algorithms, the algorithm that used dimension reduction based on the UDP algorithm showed a 3.98% increase in terms of detection rate. The results also confirmed that dimension reduction increased the calculation speed by about 31%. Using the UDP dimension reduction algorithm proposed in this research study, it was possible to resolve the previous issues of the HOG algorithm, which include requiring an excessive amount of calculations and a degradation in detection rate due to an unnecessary number of dimensions. It is considered that this increase in performance of pedestrian detection will be beneficial and widely applicable in areas such as the automotive industry and the robotics industry etc.

References

1. Dala, N., Triggs, B.: Histogram of oriented gradints for human detection. In: CVPR (2005)
2. Jeon, H.: A fast human detection method by applying adaptive steps. Master's thesis, Chonbuk National University (2013)
3. Jiang, J., Xiong, H.: Fast pedestrian detection based on HOG-PCA and gentle AdaBoost. Int. J. Secur. Its Appl., 1819–1822 (2012)
4. Wang, C., Zhou, J., Huang, S.: Motion human detection based on mixture of Gaussians and PCA-HOG. Appl. Res. Comput. **6** (2012)
5. Huang, K., Wang, F., Xu, X., Cheng, Y.: Pedestrian detection using hog dimension reducing in video surveillance. In: Pan, J.-S., Snasel, V., Corchado, E.S., Abraham, A., Wang, S.-L. (eds.) Intelligent Data analysis and its Applications. Springer, Cham (2014)

An Approach to Fast Protocol Information Retrieval from IoT Systems

Onur Soyer, Kwan-Young Park, Nomota Hiongun Kim, and Tae-soo Kim(✉)

Mobigen Co. Ltd., 3rd, 5th Fl. 552, Yeoksam-ro, Gangnam-gu, Seoul 06187, Korea
onrsyr@gmail.com, {kypark, nomota, tesse}@mobigen.com

Abstract. The Internet is the global system of interconnected computer networks which use the Internet protocol suite (TCP/IP) to link billions of devices worldwide [1]. These devices are the part of Internet of Things which are used to bring simplicity to our lives like a home surveillance camera we set to our home. It is crucial to secure those connected things around us by discovering things with insecure configuration. To achieve the security measure, we are to set up a framework. Are we sure that we are the only one who can access it? Our approach is to search and gather the specific information of the IoT devices on the internet to provide a data to analyze the vulnerabilities. To search all these devices, we need to check each IP in IPv4. This task takes a long time since handshaking cannot be accomplished so fast. Our approach gives a solution to this problem.

Keywords: IP Scan · Protocol scan · IoT device scan · Vulnerabilities · IPv4 · IPv6 · Security · Vulnerabilities Search

1 Introduction

The Internet is the global system of interconnected computer networks which use the Internet protocol suite (TCP/IP) to link billions of devices worldwide. These devices are part of Internet of Things which are used to bring simplicity to our lives like a home surveillance camera we set to our home. Are we sure that we are the only one who can access it? Our approach is to search and gather the specific information of the IoT devices on the internet to provide a data to analyze the vulnerabilities. To search all these devices, we need to check each IP in IPv4/v6. Then for each device found, a handshake has to be performed to collect the details of the protocol. It takes a long time finish to check each device since handshaking cannot be accomplished so fast. Our approach gives a solution to this problem.

Number of IoT device connected to internet are expected to reach to 21 billion by 2020 [2].

In 2015, research shows that there are 70% of the IoT devices used unencrypted network service that could be led to security issues [3].

Internet is a big cloud where loT devices are connected to each other and can be reachable by anyone including hackers. Lots of IoT devices are built by taking not a

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_38

good care in terms of security. Most of their code contains lots of "copy & paste" code in which the reference code contains lots of vulnerabilities.

This project is for discovering the weaknesses of the IoT systems/devices. Therefore, we need to discover IPs and their protocols and gather their data as much as we can. After enough data is gathered, their vulnerabilities can be found.

It is a hard task to perform quick scan for the whole Internet and gather the details of each protocol found during the search. Our approach to solve this problem is to divide the problem into two parts. An IP search system for a given port and a protocol analyzer for the IP and the port. Since the system has two modules and each module handles lots of processes inside to take the most advantage of all the system resources as much as possible.

1.1 About KISA IoT

Our project is being developed as a part of KISA IoT project. KISA IoT is a 3-year-long project. The project is separated into modules that are being developed by different teams. For the first version of KISA IoT, each module is expected to be completed until end of this year. These modules are IP and Protocol Search system, IoT Security Vulnerability Analyzer System, Information Sharing System, IoT Device Security Test System (Fig. 1).

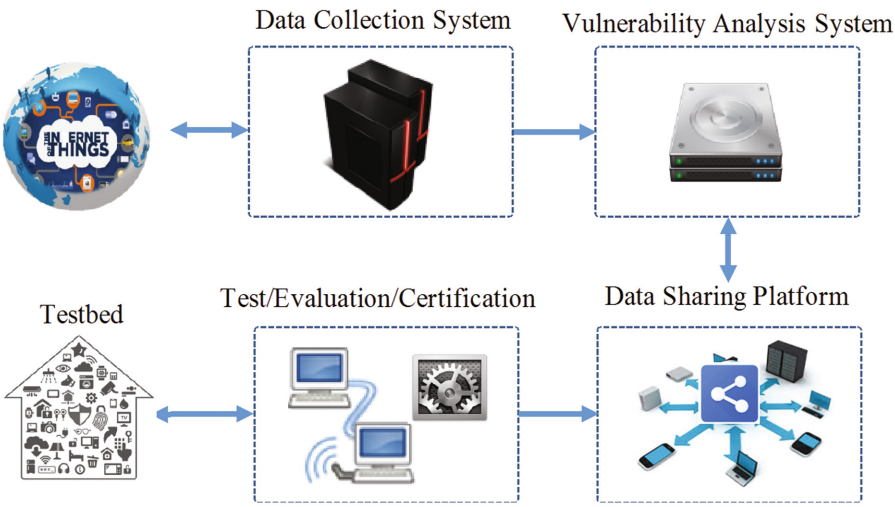


Fig. 1. Architecture of KISA IoT system

Our module is to search the all IP sets in IPv4/v6 and gather the information of the protocols and make them ready to be analyzed for vulnerability testing.

2 Related Works

Censys [4] is a search engine allows computer scientists to ask questions about the devices and networks that compose the Internet. Driven by Internet-wide scanning, Censys lets researchers find specific hosts and create aggregate reports on how devices, websites, and certificates are configured and deployed.

Censys collects data on hosts and websites through daily ZMap and ZGrab scans of the IPv4/v6 address space, in turn maintaining a database of how hosts and websites are configured.

Shodan [5] is a search engine lets their users find specific types of computers (web cams, routers, servers, etc.) connected to the internet using a variety of filters. Some have also described it as a search engine of service banners, which are meta-data the server sends back to the client. This can be information about the server software, what options the service supports, a welcome message or anything else that the client can find out before interacting with the server.

Shodan collects data mostly on web servers (HTTP, port 80), as well as FTP (port 21), SSH (port 22) Telnet (port 23), SNMP (port 161), SIP (port 5060), and Real Time Streaming Protocol (RTSP, port 554). The latter can be used to access webcams and their video stream.

Shodan works by banner grabbing and collecting all information being leaked by insecure devices. The info gets categorized and stored in a large database. The search engine allows users to search this database by type of device, brand name, operating system, location, plus many more.

Like Shodan, Censys also scans the Internet for devices not properly configured to prevent unauthorized access and stores the information in a database that can be broken down categorically. What makes Censys more advanced is its use of two tools, ZMAP and ZGRAB, that scan the entire IPv4 address space everyday to maintain a much more comprehensive database of information [6].

3 IP and Protocol Scanning

3.1 Issues of High Speed Scanning

There are three main issues for IP and Protocol scanning. Network stability and congestion affects the search speed, using system resources effectively and legal issues for IP search as it might be perceived as like an attack.

IP set is too big to process. When testing a protocol of an IP, a request is sent to host and then receive a response from the host. This takes an amount of time to test all the IP set for a given protocol. And there are lots of protocols to test which takes really a long time to complete.

Almost every one of us is connected to internet somehow. In a day there may be one or more pick duration of time of high internet usage. High internet usage means there is a congestion of bandwidth due to the high usage. This may lead some slow-down on network speed.

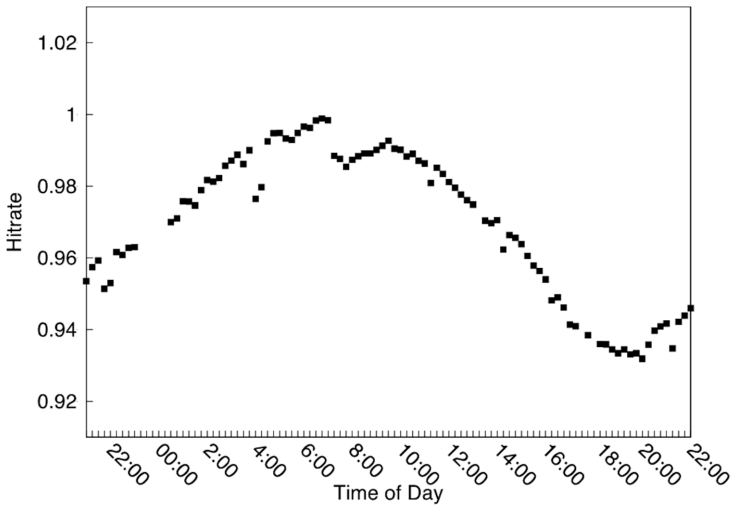


Fig. 2. Fluctuation in hit rate (proving packet's successful response rate) in a day [9]

Typically, the peak internet usage hours are between 7 and 11 pm. During this time frame, users commonly experience slowness while browsing or downloading content. The congestion experienced during the rush hour is similar to transportation rush hour, where demand for resources outweighs capacity [7] (Fig. 2).

Additionally, there are 3.7 billion IP is reserved and used by IoT devices. For each request sent there is a waiting time. Since there is 3.7 billion IP, it would take a really long time to finish handshaking task for each.

3.2 Multiprocessing Approach for High Rate Scanning

The IP/Protocol Scanner system works like a search engine for IoT networks. An IoT network contains lots of devices and systems all over the Internet. We need a system to monitor their status as quick as possible to find the vulnerabilities. That is where this idea came from.

The IP/Protocol Scanner system is mainly composed of 3 modules. IP Search, Protocol Search and Notify Manager. IP Scan module finds IPs for a given port. App Scan module performs tests and gathers information for the IPs found by IP Search module. Notify Manager sends all found data to Data Analysis Server to further data processing (Fig. 3).

IP Scan module finds IPs for a given port. It has a fast algorithm to generate all IPv4 set. For each generated IP, it performs a TCP SYN scan on a specified port at a maximum rate as possible, and finally writes all the found IPs to a file as output.

App Scan module performs tests and gathers information for the IPs found by IP Scan module. It basically waits for IP Scan to finish its task so that it can use the generated IP list file. Then it sends a request for a specified protocol on the specified

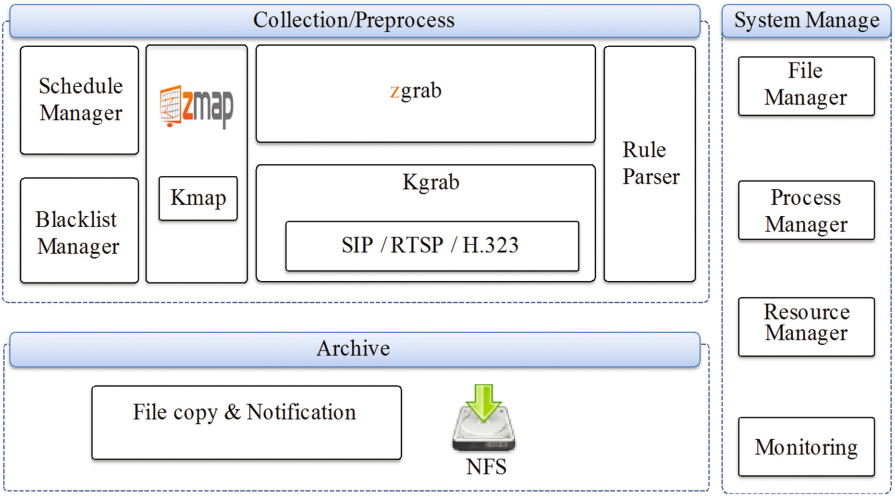


Fig. 3. IP/Protocol Scanner system architecture

port and then collects the response and TCP and IP packet details. Then it writes them to a file in JSON format. Currently it can perform tests on over 14 standard protocols.

App Scan contains a process manager that creates sub-processes to perform protocol tests. We have achieved to handle around 1000 process performing protocol tests.

Notify Manager monitors App Scan's progress and sends all found data to the Data Analysis Server for further data processing.

IP Scan is build on ZMap [8], ZMap is a fast single packet network scanner designed for Internet-wide network surveys. On a typical desktop computer with a gigabit Ethernet connection, ZMap is capable scanning the entire public IPv4 address space in under 45 min. With a 10gigE connection and PF_RING, ZMap can scan the IPv4 address space in under 5 min [9].

App Scan is build on ZGrab [10]. ZGrab is a TLS Banner Grabber and application layer scanner.

ZMap IP generation algorithm uses random permutation approach to generate IP. This algorithm can be improved by dividing IP into 4 pieces and generate each piece by one process.

ZGrab currently supports around 11 protocols. There are much more protocols are actively being used on internet. They need to be inspected too. Additionally, ZGrab does not support multiprocessing. It processes all IPs one by one. To achieve a high throughput, multiprocessing is needed.

We have extended the ZGrab system by adding more protocols to be covered. Some of them are SIP, RTSP and H.323 because these are commonly used protocols among IoT devices.

4 Future Work

We would like to achieve scanning all IPv4 and IPv6 as fast as possible. Hence, we will focus on improving our scan algorithms and multi processing approach. Moreover, there can be a distributed approach that is to perform a search by taking advantage of distributed systems to improve the scan rate.

We currently are focusing on software improvement since we want to use all of the system resources as much as possible. We use cloud system to perform our tests. But in the future, there can be a performance investigation to improve performance as well.

IP Scan module will support PF_RING, raw socket search. Also will a IP blacklist feature will be added too. In addition to that, App Scan module will support more the 15 protocol in near future.

5 Conclusion

Analyzing the security vulnerabilities of IoT devices is a modern day problem due to their high volume. Since there is no way to know of the IP of each IoT device, a full scan of IPv4/v6 is a necessity. This takes a long time to check one by one and collect information of each IoT for each protocol.

Scan speed is important because there may be new IoT devices connected to internet or removed from internet. Certainly there will some other IoT devices updated their software. These updates are needed to be re-checked regularly as quick as possible.

Our multi processing approach provides solution to this problem by scanning whole internet and then analyzing each protocol with multi process to finish scan quickly. Collected data can be easily used for analyzing the vulnerabilities. We hope this approach will provide big help for gathering information easily.

Acknowledgement. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing and Testing Technology Development).

References

1. Wikipedia. https://en.wikipedia.org/wiki/Internet
2. Meulen, R.: Gartner. http://www.gartner.com/newsroom/id/3165317 (2015)
3. Hewlett Packard Enterprise: Internet of things research study (2015). https://www.hpe.com/h20195/V2/GetPDF.aspx/4AA5-4759ENW.pdf
4. Censys. https://censys.io/overview
5. Shodan. https://www.shodan.io/
6. Occupy4Eles: Null-Byte. http://null-byte.wonderhowto.com/forum/move-over-shodan-meet-censys-0168850/

7. Wikipedia. https://en.wikipedia.org/wiki/Internet_Rush_Hour#Bandwidth_Throttling
8. ZMap. https://github.com/zmap/zmap
9. Zakir Durumeric, E.: ZMap: Fast Internet-Wide Scanning and its Security Applications, p. 15 (2013)
10. ZGrab. https://github.com/zmap/zgrab

IoT Vulnerability Information Sharing System

Taeho Seo(✉)

Igloo Security, Inc.,
6 Floor, 429 Teheran-ro, Gangnam-gu, Seoul, Korea
seotaeho@igloosec.com

Abstract. A system that takes collected and analyzed IoT vulnerability information and transform into STIX/TAXII format to share and store, in order to effectively share information among users.

Keywords: IoT · Sharing system

1 Introduction

Internet of Things (IoT) has been chosen by Gartner every year as one of the top 10 strategic technology since 2012, and it has been expected to become a core value-added business of ICT market. As smart devices such as mobiles get further distributed, increased number of smart sensors completed integration and connection between devices which then rapidly led to a greater interest on IoT environment over ICT integration fields. The more IoT gets applied to our lives, the more various economic value and higher efficiency as well as comfort will be created.

IoT consists of various core technologies, such as sensing, communicating/networking, chip, OS/embedded, platform, bigdata processing and web service technology, to provide specific service to market; meanwhile, it is exposed to a highly vulnerable environment in terms of cyber security as each technology's vulnerabilities can be abused to penetrate IoT security.

The complexity of core technologies convolutes security vulnerabilities and their responses for the security managers. Thus, the importance of system that can efficiently share various vulnerability information and analysis result has become higher than ever.

2 IoT Vulnerability Information Sharing System

2.1 Features of Security Vulnerability for IoT Service

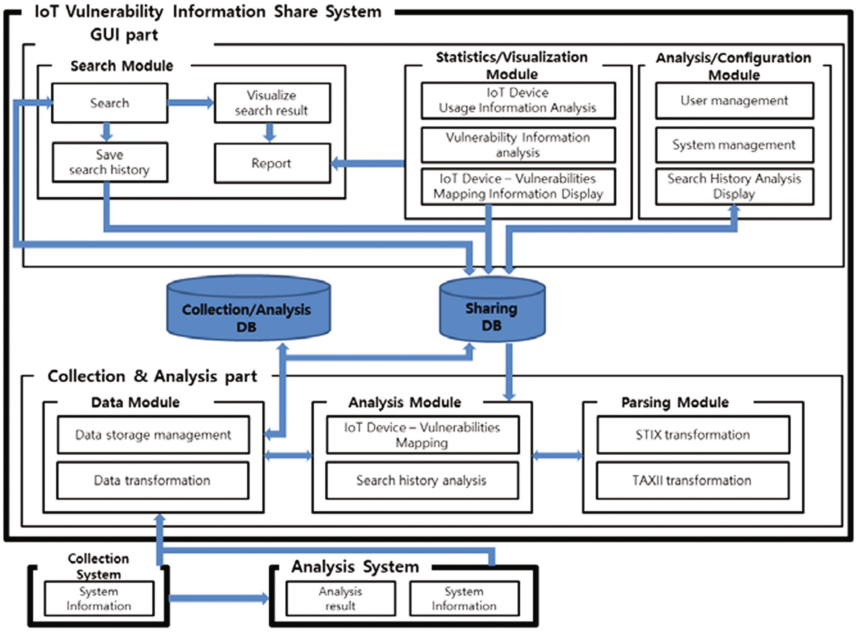
Recently, integration of human, things and service called IoT has become a platform for various products and services for our lives. It is now common to name IoT on things that we used to just call healthcare product/service, wearable devices, control system connected with web service, situation awareness system, etc. This implies that IoT is not simply defined by a single technology and that means if any pre-existing technology is integrated with human, things or services, then it can be call as IoT product or

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_39

service. This trend of integrating human, things and service is basically due to development of various technologies such as the weight lightening, developing sensory technology, software/hardware/open service platform technology, situation awareness technology, communication/network technology, application service and mashup technology, and thus IoT can be considered as a new result of ICT development.

Apart from a complexity of divergent technologies, IoT carries some distinguishable features. Unlike conventional application services, which operated in a vertical application market by a single business, IoT service lies in a horizontal market where various subjects co-exist together. Because of its diversity, each technology has to be checked for vulnerabilities. In order to eliminate such discomfort, a service that can enable effective search of security vulnerability is utterly needed.

2.2 IoT Vulnerability Information Sharing System Diagram



Search History Analysis Display: IoT Vulnerability Information Sharing system analyses search keywords user entered to GUI and Display the result as current vulnerability trend.

Data Transformation: The device scan result file from collection System and vulnerability information from Analysis System to transform into STIX Data Set for Information Sharing purpose and Sharing DB Schema type for management purpose.

IoT Device – Vulnerability Mapping: Mapping each IoT device's vulnerabilities (CVE, CVSS, CWE, Microsoft Windows Security Updates, etc.) form Sharing DB.

Search History Analysis: Take Search keyword change trend in GUI by user and store it to Sharing DB.

STIX Transformation: Transform IoT device analysis result and vulnerability mapping information into STIX format (Observable, Exploit Target, COA, etc.) in order to share them between users.

TAXII Transformation: Transform STIX format (Observable, Exploit Target, COA, etc.) information into TAXII protocol in order to share them between users.

3 IoT Security Issues

According to a recent survey on internet issue experts by Korean Internet Professionals Association (KIPFA), the foremost condition of IoT is a complete security, whereas the most likely risk of it is a hacking. So far the security threats such as hackings only caused personal information leakage or monetary damage at most. However, any security breach of IoT may lead to a social disaster or physical damages. Plausible IoT security issues can be categorized as follow.

First of all, since IoT utilizes conventional internet environment for already-existing things and devices, every risk and vulnerability of internet can also happen in IoT service. This means that things with limited resources and networks with power-efficient technology are exposed to all of the threats from the internet. Thus, for those things with limited computing or storage resources, it is necessary to have an imbedded security technology such as safe S/W coding, power efficient Crypto S/W, and etc.

Secondly, open platform source is easier for attackers to find vulnerabilities which can then lead to the exploits of those security vulnerabilities. As IoT environment consists of various devices, it is predicted to have much higher attack opportunity such as malicious S/W or virus infection.

Thirdly, Symantec's analysis in 2015 showed smart home devices lacked protection of accounts in terms of password complexity, mutual accreditation and so on. Intel McAfee research lab's prediction report in 2015, which involved security tests on the 10 most used devices, also provided some statistical results that emphasized the current security situation. In short, smart home devices and smart office products have already exposed security vulnerabilities and do not satisfy password complexities.

Last but not the least, IoT security issues can develop into a life threatening matter. Various medical devices including mobile healthcare, and any internet connected vehicles such as automatically driving cars can cause serious damage, if not death, to human beings when faultily used or maliciously attacked. Therefore, as long as IoT possesses abusive potential, preventive security procedure must be established.

4 Conclusion

This article discussed about how to construct an IoT Vulnerability Information Sharing system that enables referring and searching of vulnerabilities of various technology in one service, mainly due to the complexity of the compartments of IoT technology. In the future, such information sharing system much be considered for other IoT service areas such as smart factory, home IoT, smart car, etc.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No.B0717-16-0099, IoT Security Vulnerabilities Search, Sharing and Testing Technology Development).

A Study on the Service Identification of Internet-Connected Devices Using Common Platform Enumeration

Sarang Na, Taeeun Kim, and Hwankuk Kim(✉)

Security Technology R&D Team 2, Korea Internet and Security Agency Seoul, Seoul, Republic of Korea
{no.1.nasa, tekim31, rinyfeel}@kisa.or.kr

Abstract. Internet-connected device information can be acquired through the open ports of a network host. It is also possible to determine whether a particular host is vulnerable by associating publicly known vulnerabilities with this information. Currently, the analysis of the device information to identify the security vulnerability is carried out manually; therefore, automatic analysis technology is necessary in order to deal with a huge number of devices. In this paper, we propose a method that automatically generates the Common Platform Enumeration (CPE) of Internet-facing devices based on banner information to discover security vulnerability information such as Common Vulnerabilities Exposures (CVE).

Keywords: Common platform enumeration (CPE) · Common vulnerabilities and exposures (CVE) · Service identification · OS fingerprinting · Security vulnerability analysis

1 Introduction

National Vulnerability Database (NVD) provides approximately 80,000 Common Vulnerabilities and Exposures (CVE) entries [1]. The CVE system provides a reference-method for the publicly known security vulnerabilities. A CVE entry is composed of vulnerability overview, Common Vulnerability Scoring System (CVSS), Common Platform Enumeration (CPE), and Common Weakness Enumeration (CWE), etc. CVE entries express the information of vulnerable products in the CPE format which presents the vendor, product name, version, etc., of released software packages. If the CPE name for a software product is discovered, CVE entries can be easily identified by associating the relevant CPE.

A computer system on a network is searchable by using an Internet-wide scanning tool, such as Shodan [2]. The device OS/App information can be found through the banner texts at open ports such as HTTP, SMTP, and TELNET. Using this information, the known security vulnerabilities of the device can be identified. Currently, such a process is manually carried out and it is very inconvenient to match it with vulnerability information after searching items of device information one by one. Therefore, an automatic analysis technology for converting collected device information into the CPE

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_40

format is needed to enable security experts to easily find out CVE entries and quickly cope with threats.

It is difficult to obtain the entire OS/App information of a network host through an Internet scanning. Therefore, it is necessary to acquire additional software information from already-identified OS or App information. In the case where just a device's OS information is known (assuming, for instance, that it is Windows 10), the basically-installed application, e.g., Internet Explorer 11 can be additionally analyzed in order to identify the application vulnerability.

In this paper, we propose a method of automatic analysis for device software information based on the CPE dictionary for linkage with CVE vulnerability information. This method works as following: device OS/App information is identified through banner and the basically-installed OS/App information of an already-identified software is added to identify more security vulnerabilities.

2 Related Work

John Matherly developed the Shodan search engine to acquire device information of Internet-connected hosts [2]. Shodan provides responsive scan information (e.g., handshake) and banner information for open ports such as HTTP. Durumeric et al. developed Censys, a search engine, which can identify potentially vulnerable devices [3]. Censys provides responsive scan information including banner for fourteen protocols such as HTTP and POP3, and partial vulnerability information of Heartbleed, Poodle, etc. Genge and Enăchescu developed ShoVAT, which is a vulnerability assessment tool based on Shodan [4]. Using banners acquired through Shodan API, it analyzes the CPE and identifies CVE vulnerability information.

Through banners provided by Shodan and Censys, we can discover device OS/App information. However, the device information has to be manually converted into the CPE format, because it exists in a format that simply delivers banner texts without processing. In the case of ShoVAT, banner information is processed to express the CPE format, but since CPE information is created by searching product information, based on the product's version pattern ("Integer.Integer.Integer"), there is a limitation in analyzing diverse CPE candidates. Our proposed method can create CPE information automatically by conducting an entire keyword analysis on banner texts based on the CPE dictionary, and also identify the basically-installed software information of the relevant CPE information.

3 Methodology

The CPE is composed of vendor name, product name and product version as follows. [part, vendor, product, version, update, ~edition~sw_edition~target_sw~target_hw~other, language] [5]. By using CPE dictionary, we can generate device software information in the CPE format from the banner texts.

3.1 Device Information Analysis Based on the CPE Dictionary

We can search the OS/App information of a particular device through open port banner information, as shown in Fig. 1. In order to run an automatic analysis of device information, we generate a CPE tree based on the CPE dictionary, conduct a keyword analysis of the banner, and convert discovered keywords into the closest CPE.

```
[root@prober] nc www.targethost.com 80
HEAD / HTTP/1.1

HTTP/1.1 200 OK
Date: Mon, 11 May 2009 22:10:40 EST
Server: Apache/2.0.46 (Unix)  (Red Hat/Linux)
Last-Modified: Thu, 16 Apr 2009 11:20:14 PST
ETag: "1986-69b-123a4bc6"
Accept-Ranges: bytes
Content-Length: 1110
Connection: close
Content-Type: text/html
```

Fig. 1. HTTP response containing service information of the host [6].

We create the keyword list used in the banner text analysis by generating a CPE tree with a total of 6 levels according to the CPE format, as shown in Fig. 2. Level 1 is comprise of vendor information, and the vendor's node creates product information as a child node (level 2). Likewise, levels 3, 4, 5, and 6 are built in the same manner. For level 5, as one node contains various items, the character strings can be separated with the '~' character. In each level, if there is the '_' character, the character strings are separated. For example, in the case of "cpe:/a:microsoft:ftp_service:7.0", level 1 is microsoft; level 2 is ftp, service; level 3 is 7.0; and information on levels 4, 5, 6 is absent.

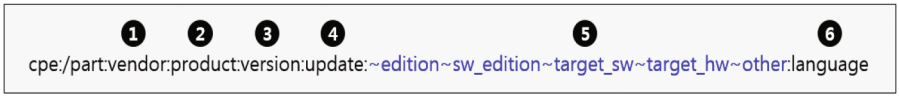


Fig. 2. CPE composition.

Keyword analysis for the banner is run with a character string at each level of the CPE tree. Levels 1 and 2 run the entire keyword search, whereas levels 3, 4, 5 and 6 only search the keyword relevant to levels 1 and 2. For version information, substring combination (e.g., version 2.10.2) which can separate in unit of '.' is taken into account to search the higher version (e.g., version 2 or 2.10).

By combining the discovered keywords, they are converted into the CPE from level 1 in order. The keywords are combined so that the CPE with the longest character string among the CPE candidates is converted. Since two or more CPEs can exist, if

there is another keyword other than the keyword that uses the longest character string, the CPE is created additionally. Through the abovementioned process, product information is identified in the CPE format in order to be easily utilized to analyze CVE vulnerability information.

3.2 Identification of Target OS

In cases where only the use information of a device's application is known, the application's installation OS information can be added at CPE tree level 5, the information target_sw, target_hw. As shown in Fig. 3, target_sw information is acquired, and therefore the OS information can be found through the CPE information without the need for an additional search process.

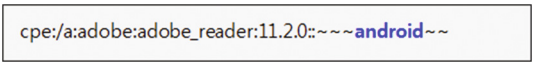


Fig. 3. Example of target OS information.

3.3 Identification of Basically-Installed Application

In cases where only the OS information of a device is known, the information of an application basically-installed in the OS can be added by analyzing the product catalogue provided by the vendor's website. The product information of the OS basically-installed application or the specific application supported by the OS is automatically structured and the application can be identified. As shown in Fig. 4, for instance, Internet Explorer which is one of the basically-installed applications of Microsoft Windows can be found.

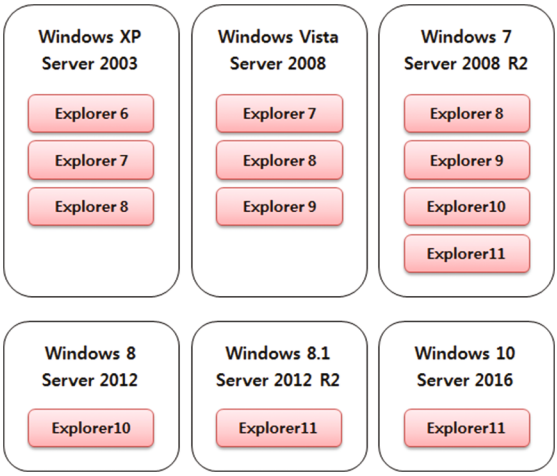


Fig. 4. Example of basically-installed application information.

4 Conclusion

We propose a method of identifying the device OS/App information through an analysis of open port banner information based on the CPE dictionary. Using this method, we can obtain the software information for a network host in the CPE format; and the information of the basically-installed software can be additionally identified so that it can be utilized in a vulnerability information analysis of Internet-connected devices.

Acknowledgments. This work was supported by a grant from the Institute for Information & Communications Technology Promotion (IITP), funded by the Korean government (MSIP). (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing, and Testing Technology Development).

References

1. National Vulnerability Database. https://nvd.nist.gov
2. Shodan. www.shodan.io
3. Durumeric, Z., Adrian, D., Mirian, A., Bailey, M., Halderman, J.A.: A search engine backed by internet-wide scanning. In: SIGSAC Conference on Computer and Communications Security (CCS 2015), pp. 542–553 (2015)
4. Genge, B., Enăchescu, C.: ShoVAT, shodan-based vulnerability assessment tool for internet-facing services. In: Security and Communication Networks, pp. 1–19 (2015)
5. NIST: Common Platform Enumeration: Naming Specification Version 2.3. NIST Interagency Report 7695 (2011)
6. Banner Grabbing. https://en.wikipedia.org/wiki/Banner_grabbing

A Design of IoT Protocol Fuzzer

Dae-il Jang, Taeeun Kim, and HwanKuk Kim(✉)

Security Technology R&D Team 2, Korea Internet and Security Agency,
IT Venture Tower, 135 Jungdae-Ro, Songpa-gu, Seoul 05717, Republic of Korea
{dale, tekim31, rinyfeel}@kisa.or.kr

Abstract. IoT Devices is increased rapidly, but the tool or framework for security check of IoT device is an insufficient. To improve the security of IoT device finds a weak part of source using the Source Code Auditing Tool or check the error or termination part of protocol using Protocol Fuzzer. But normal Protocol Fuzzer does not find an internal problem through fuzzing process. So we design a blackbox fuzzer combining firmware dynamic analysis platform and IoT protocol fuzzer.

Keywords: Fuzzer · Protocol fuzzing · Dynamic analysis

1 Introduction

While we develop software, we face unintentional security vulnerabilities. To solve this problem, there have been many researches on how to verify the developed software. Many researches attempt to solve the problems of software through various methods such as conducting stress tests, auditing the source codes and fuzzing. While stress tests can verify the stability of software [3], this method cannot find security vulnerabilities. Source code audit is only subjected to source codes [4], and thus has difficulties in finding security vulnerabilities of dependent environments such as binaries or protocols. Fuzzing, or fuzz tests, can be applied to various environments, but it is also difficult to find the spot where errors occur [5].

This paper proposes a black-box fuzzer for IoT protocol which uses open-source protocol fuzzer and firmware dynamic analysis tool. To solve the problem of previous protocol fuzzer regarding the difficulty of finding the location of error, crash and termination, the subject software is put on the dynamic analysis system and dynamically analyzed through the fuzzing process to easily find the location where error and termination action occur.

2 Related Works

2.1 Protocol Fuzzing

Many researchers have thrived to detect in advance the vulnerability that might occur in software. One of the methods they've used is the method of finding vulnerabilities through fuzzing. Fuzzing changes input values and specific field values to find the

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_41

possibility of errors and vulnerabilities in software. Fuzzers are often used to verify the input values of file formats or applications, as well as the stability and security of protocols. Especially, many tools exist to verify the stability of IoT protocol. Some of the open source tools are conduct fuzzing for protocols such as Bluetooth, Zigbee, ZWave, or for web applications on some IoT protocols. 6LowFuzzer provides the function of message, mutation, scenario fuzzing for 6Lowpan protocol [6].

2.2 Firmware Dynamic Analysis

Various dynamic analysis tools exist to analyze malicious codes which are expanding not only to Windows environment but also to Linux and IoT environments, and many methods are being proposed to analyze such malicious codes. There are tools for dynamically analyzing firmware such as IoT devices. To analyze the action of firmware, all kinds of information are provided. Based on such information, analysis on the response, action and error regarding the input value is available, and this helps the tools trace the location where errors occur when they perform protocol fuzzing.

3 Design of Blackbox Fuzzer

3.1 Overview

The proposed fuzzing system includes a fuzzing tool based on open-source and a firmware dynamic analysis system based on QEMU. The system conducts fuzzing on the various protocols used in the IoT device and installs a penetration test tool to verify the vulnerabilities of protocols used in the IoT device. When verifying the vulnerabilities, dynamic analysis takes place to find the location of errors (Fig. 1).

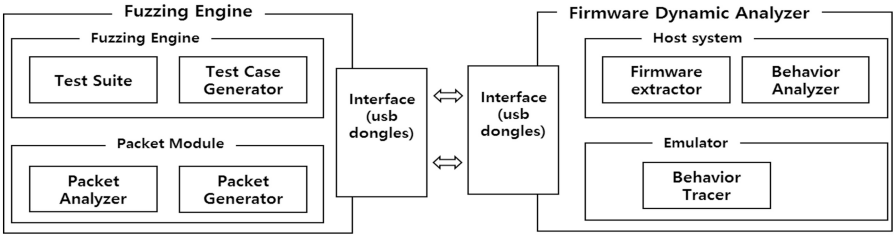


Fig. 1. Design of fuzzer and dynamic analyzer

3.2 Protocol Fuzzing

Structure of Protocol Fuzzing Module in proposed system are shown in the following figure (Fig. 2).

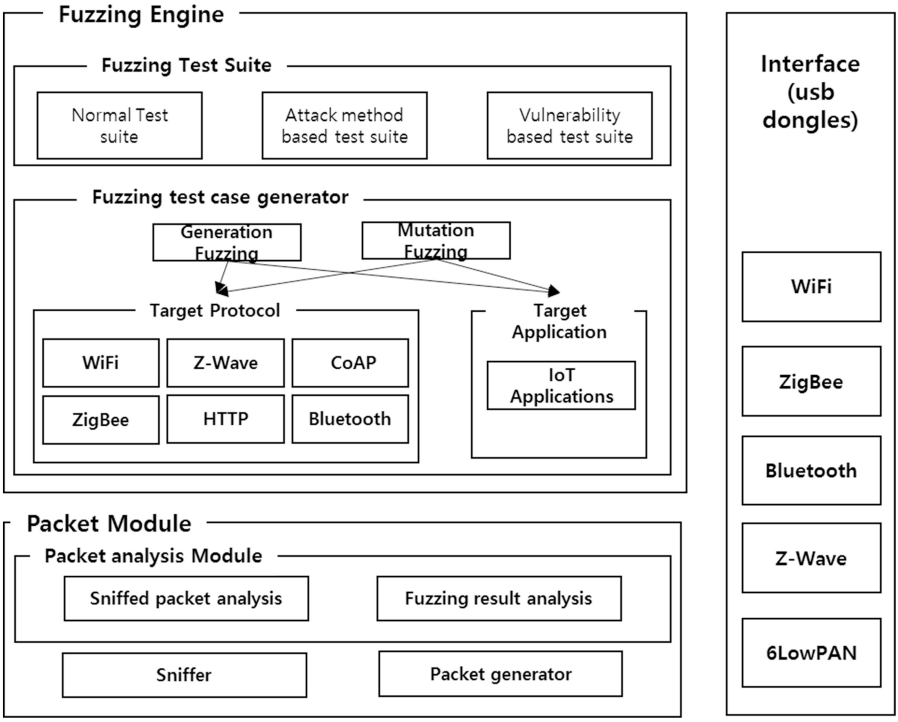


Fig. 2. Structure of protocol fuzzing module

To communicate with the subject, the system has communication interface using usb dongles such as WiFi, Zigbee and Bluetooth. Through the communication interface, packets are created, transmitted and received. Then, the result is analyzed. The protocol fuzzing engine is mainly composed by generation fuzzing and modification fuzzing. Modification fuzzing changes the values of all fields except for protocol fields where modification is unavailable. Then, it conducts security inspection on abnormal input values. Generation fuzzing module uses the pre-defined test cases of attack and weakness to conduct fuzz test. Here, test cases used in previous protocol fuzzing and application fuzzing are classified based on OWASP Top 10 to conduct fuzzing. For an example, the following input values are for web applications [1].

Request-URI
- /[path]/[page].[extension]?[name]=[value]&[name]=[value]…
- Path : Path traversal,
- Page : predictable resource location, Directory indexing, Information leakage
- Extension : filter bypass, DoS
- Value : SQL Injection, XSS, file inclusion, command injection and etc

Protocol
- HTTP/[major].[minor]
- Unsupported protocol version

Headers
- [Header name]:[Header value]
- BoF, DoS, Script injection, Code injection, Sql Injection

PostData
- [Name1]=[Vaule1]& [Name2]=[Vaule2]…
- Name : abuse of functionality
- Value : Sql injection, XSS, File inclusion, Command injection, BoF

Cookies
- Cookie: [Name1]=[Vaule1]; [Name1]=[Vaule1]…
- Value : CSRF, insufficient session expiration, insufficient authentication

Various attack codes such as Directory Traversal, Bypass, DoS, SQL Injection, XSS, File inclusion, Command injection and unsupported value are inputted in these input values to conduct fuzzing.

3.3 QEMU Based Dynamic Analysis Module

Among the proposed system, FIRMADYNE [2] is used as the dynamic analysis system. Studied by CMU, FIRMADYNE is an automated system which conducts dynamic analysis on imbedded firmware based on Linux. From extracting the firmware file system to performing virtual execution and dynamic analysis, FIRMADYNE automates the process and provides the result of analysis.

In this research, it can be used in extracting the execution file from IoT device subject for tests or firmware, and in tracing the location where error, crash and device termination event occur.

4 Conclusion

This paper suggests a fuzzing system which uses open-source. To solve the problem of previous protocol fuzzing system regarding the difficulty of estimating the location of error, the suggested system adds a dynamic analysis module to the open-source protocol fuzzer. When conducting fuzz tests in the subject binary, dynamic analysis is conducted on the action to analyze API where errors occur and the reason for such occurrence. More researches will be made in the future to realize the designed system. Researches will also be made on fault localization based on static and dynamic analyses.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing and Testing Technology Development).

References

1. Sutton, M.: Smashing Web Apps: Applying Fuzzing to Web Applications and Web Services. Black Hat DC (2007)
2. Chen, D.D., Egele, M., Woo, M., Brumley, D.: Towards automated dynamic analysis for Linux-based embedded firmware. In: ISOC NDSS (2016)
3. Musuvathi, M., Qadeer, S.: Chess: systematic stress testing of concurrent software. In: Puebla, G. (ed.) LOPSTR 2006. LNCS, vol. 4407, pp. 15–16. Springer, Heidelberg (2007)
4. Viega, J., Bloch, J., Kohno, T., McGraw, G.: ITS4: a static vulnerability scanner for C and C++ code. In: The 16th Annual Computer Security Applications Conference (ACSAC 2000), New Orleans, Louisiana (2000)
5. Sutton, M., Greene, A., Amini, P.: Fuzzing: Brute Force Vulnerability Discovery. Addison-Wesley, Reading (2007)
6. Security Monitoring/IoT Fuzzer - User and Programmers Guide, https://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/Security_Monitoring_/_IoT_Fuzzer_-_User_and_Programmers_Guide

A Study on the Management Structure of Vulnerability Information in IoT Environment

Eunhye Ko, Taeeun Kim, and Hwankuk Kim(✉)

Security R&D Team 2, Korea Internet and Security Agency, Seoul, Republic of Korea
{keh, tekim31, rinyfeel}@kisa.or.kr

Abstract. Due to the recent development of wireless network technologies, the use of IoT devices is increasing as well as security threats caused by vulnerable security. To prevent security threats and accidents, it is necessary to share information on cyber security threats. This paper proposes a platform of sharing vulnerability information regarding IoT devices, and contents of shared information.

Keywords: Cyber threat intelligence (CTI) · Structured threat information eXpression (STIX) · Trusted automated eXchange of indicator information (TAXII) · Vulnerability information sharing platform

1 Introduction

As we are recently entering the era of Internet of Things (IoT), all kinds of devices such as smart phones, smart home appliances and wearable devices are being introduced. According to recent market survey conducted by the consulting company Gartner, the number of IoT devices is increasing worldwide, and by 2020, the number will reach 20.8 billion.

Although the use of IoT devices is increasing as mentioned above, security level regarding IoT devices is low, and attacks using vulnerabilities of IoT devices are increasing [1] (Fig. 1).

Instead of attacking unknown vulnerabilities, many of the attacks take advantage of vulnerabilities or weak security configurations which are already known. These attacks can be blocked by prompt fixing such vulnerabilities. For this reason, technologies related to Cyber Threat Intelligence (CTI) are actively being developed recently to share information on cyber threats quickly.

This paper takes a look at the standards of sharing information on cyber threats, and proposes a platform of sharing information. The structure of this paper is as follows; in Sect. 2, the paper describes the standards of sharing information on threats and domestic and international status of using such standards; in Sect. 3, the paper proposes a platform of sharing information on device threats in IoT environment, and the information that can be shared using this platform; finally, in Sect. 4, the paper describes future research plans and makes a conclusion.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_42

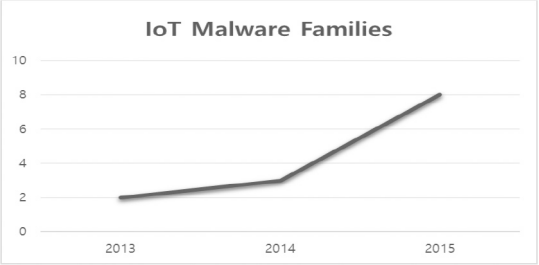


Fig. 1. New IoT malware families by year (Symantec)

2 Domestic and Overseas Status

Recently, to automatically share information on cyber threats, there are lots of standards are developed. Based on the expressed information, those standards can be divided into threat information and intrusion indicator. Threat information is comprehensive information which includes the analysis information related to attack and threat, the strategies and tactics which is used, and the motive of attack. Meanwhile, intrusion indicator is information which includes hash values and registry of files related to the accident.

Structured Threat Information eXpression (STIX) and Open Indicator Of Compromise (OpenIOC) can be examples of standards for sharing information mentioned above. STIX is a standard which expresses threat information. It shows information on vulnerability and some events related accident. On the other hand, OpenIOC is a standard which expresses intrusion indicator. It can express intrusion indicator such as the trace of intrusion accidents like traffic, hash value of file as well as information on the rules for firewall, intrusion detection system, and intrusion prevention system.

However, it is difficult to express and understand the details of threat information because OpenIOC focuses on expressing observation information. Therefore, the platform proposed by this paper will use STIX standard to share information on vulnerability.

2.1 STIX and TAXII Standard Specification

Together with MITRE, STIX is a standard developed by US Department of Homeland Security (DHS) to efficiently and safely share information and transmit to cyber threats. Using STIX, information of accident, vulnerability and observed event can be systemized and expressed. As shown in Table 1, STIX is composed by eight elements to express all kinds of information [3].

TAXII is an automatic transmission standard of sharing information on cyber threats expressed by STIX in real-time. Currently, TAXII supports HTTP and HTTPS protocols. TAXII provides services such as Push, Pull, Discovery and Feed Management. The services can be used to requests and transmit information between the producer and consumer of information.

Table 1. Elements of STIX structure

Element	Description
Observable	What has been or might be seen in cyber
Indicator	Patterns for what might be seen and what they mean if they are
Incident	Instances of specific adversary actions
TTP	Attack patterns, malware, exploits, kill chains, tools, infrastructure, victim targeting, and other methods used by the adversary
Exploit target	Vulnerabilities, weaknesses, or configurations that might be exploited
Course of action	Response actions that may be taken in response to an attack or as a preventative measure
Campaign	Sets of incidents and/or TTPs with a shared intent
Threat actor	Identification and/or characterization of the adversary

3 Platform of Sharing Information on Vulnerability

In this chapter, a platform of sharing information on IoT devices' vulnerability is proposed. This platform attempts to prevent security accidents and threats caused by the security vulnerability of IoT devices.

The proposed platform of sharing information on vulnerability scans IoT devices and collects vulnerability information such as Common Vulnerabilities and Exposures (CVE), Common Platform Enumeration (CPE) and Common Weakness Enumeration (CWE). Then the platform shares the analyzed information with external agencies or users.

The structure of platform sharing vulnerability information is shown in Fig. 2 below.

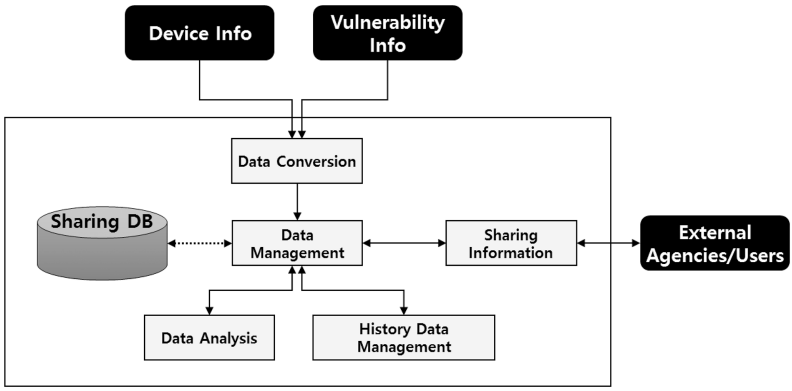


Fig. 2. Concept of vulnerability information sharing system

3.1 Input and Output Data of Platform

The platform of sharing vulnerability information needs the scan information of IoT devices and vulnerability information collected.

The scan information of IoT device is subject to devices with officially approved IP. And scanning of device is conducted using scanning tool such as Zmap. Also for extracting information some technologies are needed, such as Banner Grabbing [2]. Meanwhile, vulnerability information each can be collected using officially approved databases such as CVE, CPE and CWE [4].

3.2 Operating Flow of Platform

The function of vulnerability information sharing platform can be mainly divided into three; managing, analyzing and sharing information. By using the platform of sharing vulnerability information, information on threats and vulnerabilities can be stored and managed. Also information can be produced through association analysis using various information. The information above can also be shared with external agencies or users. The platform of sharing operates as follows.

In the process of managing information, information of devices and vulnerabilities mentioned above are converted and stored in the shared DB, after scanning and collecting information is processed. In addition, certain history data can be stored using cloud. In the process of analyzing information, IoT devices with many vulnerabilities can be pointed out through analysis. Using history data, variety of statistical data can also be created. In the process of sharing information, relative information is converted into an appropriate format for sharing, and it can be shared based on specific policies.

Information converted into STIX format is finally transmitted to external agencies or users. In this stage, the platform uses TAXII transmission mechanism.

3.3 Converting Data to STIX Structure

Information to share is that scan data of IoT device, vulnerability information and additional solution for vulnerability. To convert information, it uses STIX components. They are Observable, Exploit Target, and Course of Action. These components can express some information like information of devices, vulnerability, and possible solution.

Observable may express information such as an event or a property of device. Because of the Object type, it can be expressed a variety of information. There are 88 Observable Objects (based on Cybox 2.1 version). Proposed sharing system only uses 9 Observable Objects such as Address Object, Device Object, and Product Object.

Exploit Target can express detail information about vulnerability. When used with the Course of Action (COA) may represent responses to the vulnerability.

STIX structured information will be sent to external agencies or users. To transfer, by utilizing TAXII transport mechanism used for services such as Push and Pull. With this TAXII mechanism, it is possible to process automatically when transmitting the data.

4 Conclusion

Due to the recent development of wireless network technologies, the use of IoT devices is increasing as well as various threats caused by vulnerable security. Most of these security threats can be solved by fixing some known vulnerabilities. Accordingly, the need of sharing information on cyber security threats is increasing these days.

This paper proposed a platform of sharing vulnerability information in IoT environment, as well as the format of shared data with using standard. In the future, the proposed platform will be implemented and tested.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP). (No. B0717-16-0099, IoT Security Vulnerabilities Search, Sharing and Testing Technology Development).

References

1. Symantec, IoT devices being increasingly used for DDoS attacks. http://www.symantec.com/connect/blogs/iot-devices-being-increasingly-used-ddos-attacks
2. Durumeric, Z., Adrian, D., Mirian, A., Bailey, M., Halderman, J.A.: A search engine backed by internet-wide scanning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 542–553. ACM (2015)
3. Barnum, S.: Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX™). MITRE Corporation 11 (2012)
4. Na, S., Kim, T., Kim, H.: A study on the classification of common vulnerabilities and exposures using naïve bayes. In: Proceedings of the 11th International Conference on Broad-Band Wireless Computing, Communication and Applications. BWCCA (2016)

Assessing the Impact of DoS Attacks on IoT Gateway

Yungee Lee, Wangkwang Lee, Giwon Shin, and Kyungbaek Kim(✉)

Department of Electronics and Computer Engineering,
Chonnam National University, Gwangju, South Korea
negeel564@naver.com, kwang9092@gmail.com,
giwonie@gmail.com, kyungbaekkim@jnu.ac.kr

Abstract. Internet of Things (IoT) becomes more popular, and things are connected to each other through wired or wireless communication methods. Though things are connected with various methods easily, it attracts network attackers who exploit these open and convenient network connections in order to obtain unjustified information and benefits or to subvert various IoT systems. Especially, Denial of Service (DoS) attack becomes a serious problem on IoT system where huge number of devices are connected to. These devices are usually connected to IoT gateways in order to send packets to Internet. However, currently the impact of DoS attack on an IoT gateway, which has various interfaces such as wireless LAN interface and wired LAN interface, is not well examined. In this paper, we assess the impact of DoS attack on an IoT gateway with various scenarios. We implemented a prototype of an IoT gateway which has wired and wireless network interfaces by using Raspberry Pi, OpenWRT, and OVS (Open vSwitch). With this prototype, we evaluated various DoS attack scenarios on this IoT gateway. Through this evaluation, we observed the severity of DoS attack on IoT gateways, especially for wireless connections.

Keywords: Internet of things · Gateway · Denial of service attack

1 Introduction

Everything is connected to Internet in anytime and anywhere. The concept of ubiquitous era is not a story of future life any more, but current technology of Internet of Things (IoT) becomes a popular and powerful tool to realize it.

Internet of Things lets everything connect to Internet, and provides communication methods between humans and devices or between devices and devices. Through this rich communication methods and intelligence services with these open and collaborative connections, IoT is considered as one of the essential technologies such as Artificial Intelligence (AI) which will lead the forth industry innovation. An example of intelligent IoT system is a smart streetlamp which monitors human activities, analyze the activities on the centralized server and turns on the lamp automatically under given conditions. Not only this simple example, various life appliances such as smart phones, refrigerators and televisions can collaborate to each other and more intelligence

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_43

services through network connections. Consequently, IoT technology let people control everything with a small device in their palm.

However, this easy and convenient control of everything may be a serious problem of network security. That is, the IoT intelligence system with basic vulnerability of network security may be exploited by network attackers in order to obtain unjustified information and benefits or to subvert the intelligence system. In practice, a general IoT based smart home service uses a gateway in order to connect the various devices in a house. In this case, if the gateway is manipulated by attackers, every device connected to the gateway becomes target of network attacks or means for various network attacks such as bots for DoS (Denial of Service) attack [1]. Even though the gateway is not manipulated, if some malicious devices are connected to the gateway, they can initiate DoS attack to hamper the communication between the normal devices. Lately, the source code of Mirai, which is a powerful DDoS tool managing over 300 K IoT device bots easily, is released, and the possibility of new types of bot nets for IoT DoS attack increases significantly.

Along with the development of IoT industry, the number of devices connected to IoT system and the volume of data traffic in IoT system increases substantially. This increase of complexity of IoT system lead more vulnerability of DoS attacks. In this paper, we evaluate the impact of DoS attacks under IoT gateway in the aspects of communication ability between IoT devices. Through this evaluation we show the importance of DoS attack detection and prevention on IoT gateway.

2 Backgrounds

2.1 Internet of Things (IoT)

Internet of Things (IoT) lets everything in our life environment connected to each other through wired or wireless communication method, and enables these things to exchange data for collaboration. The concept of IoT has been already used in various area such as wearable health care devices and self-managing refrigerators. Recently Korean government selects IoT as a key technology for leading future industry, and early this year Korean government represents a stimulating strategy supporting tax credits to companies for conducting the research and development of IoT up to 30%. Also, many companies including big-size as well as medium size companies focus on developing IoT technologies to preempt the share of IoT Industry.

2.2 Denial of Service (DoS) Attack

Denial of Service (DoS) attack is an attack to make an online service unavailable by overwhelming it with traffic from tens or hundreds of PC in short time. Attackers makes the tons of traffic which is cannot afforded by a server or a network, so general users cannot use the server or the network normally. Generally, attackers install malicious bot control program to other PCes, and control these manipulated PCes in remote. By using this ability of remote control, attackers initiates DoS attack in remote easily by using an automation program for generating heavy traffic from these zombie

pc at the same time. The severity of DoS attack is also that it is easy to generate meaningless heavy traffic, and in IoT environment without basic security concerns it is easy that attackers get network connections to many other devices [3].

2.3 OpenWRT

OpenWRT is a Non-Real-Time Linux based open source operating system for a wireless LAN router, and it supports most router platforms [2]. OpenWRT provides various Linux features required by wireless LAN router conveniently as Linux packages. Also, OpenWRT supports ipkg (or opkg) as a package management system for users to install various kinds of software on OpenWRT, and it provides very high flexibility in its usage.

3 Implementation of a Prototype of an IoT Gateway

To evaluate the impact of DoS attack on IoT gateway, we implemented a prototype of IoT gateway by using Raspberry Pi 2, OpenWRT and Open vSwitch (OVS). The basic architecture of the prototype implementation of our IoT gateway is shown in Fig. 1. The implemented IoT gateway has three Ethernet interface and one Wifi interface. Among these interfaces, two Ethernet interfaces and one WiFi interface are bound in a bridge of OVS, and one Ethernet interface is used for connecting to Internet. With this IoT gateway, we connect four devices (n1, n2, n3, and n4). Two devices, n1 and n2, are connected with Ethernet (wired) interfaces, and other two devices, n3 and n4 are connected with WiFi (wireless) interface.

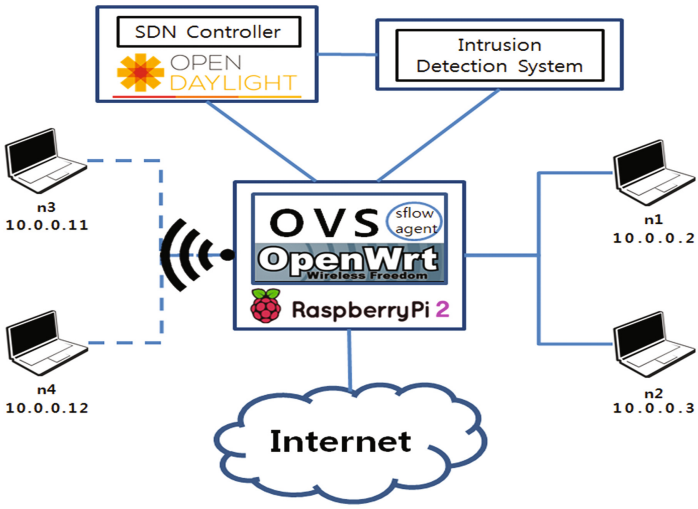


Fig. 1. Architecture of a prototype IoT gateway and connected devices.

In this implementation, we consider the ability of DoS detection and prevention discussed in past efforts [4–6] and attach an intrusion detection system such as SNORT and a SDN controller for controlling flow traffic to our IoT gateway.

For the implementation of IoT gateway and devices, we use the same WiFi interface module and its detail characteristics are described in Table 1. And for the devices connected with wireless interface, n3 and n4, the distance between the device and the IoT gateway set to 5 m.

Table 1. Setup of 802.11 Wireless LAN

Category	Specification
Chipset	Realtek 8188
Bandwidth	20 MHz
Channel	11 (2.462 Ghz), single channel
TX Power	20 MHz

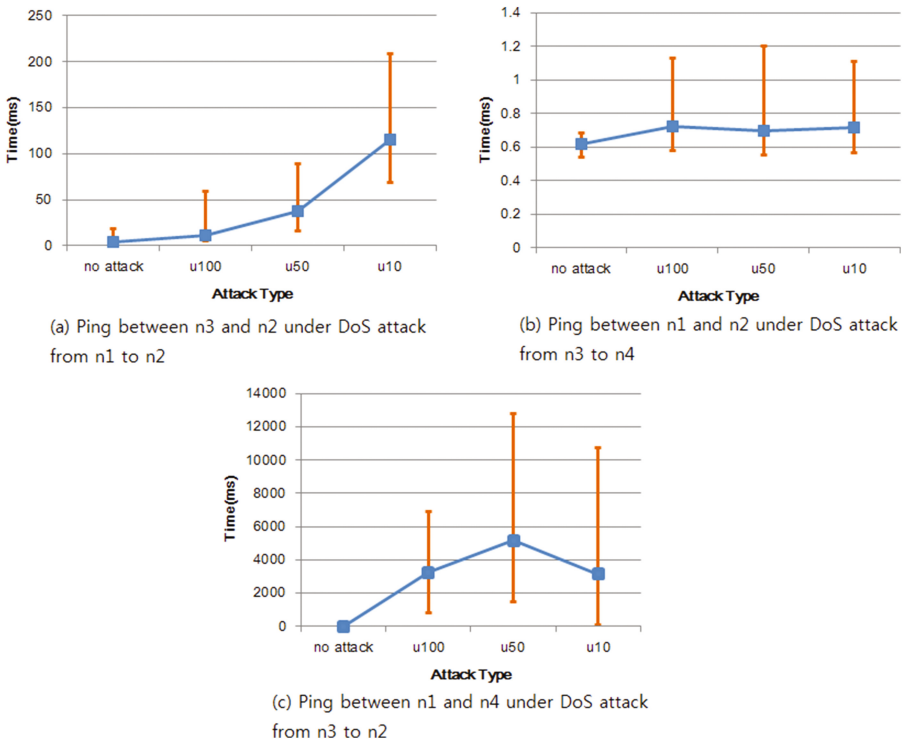
4 Evaluation of the Impact of DoS Attack on an IoT Gateway

By using the implemented IoT gateway, we emulated various scenarios of DoS attack on IoT gateways, and evaluated the impact of different kinds of DoS attack. We consider for difference DoS attack scenarios in the aspect of interfaces of an attacker and a target; wired to wired attack, wireless to wireless attack, wireless to wired attack and wired to wireless attack. We generated the SYN flooding traffic as DoS attack traffic by using hping3 program. Under this DoS attack, we confirm that our intrusion detection system can detect the attack traffic and SDN controller can prevent the flow successively.

In this evaluation, we focus on assessing the impact of DoS attack, and we turn off the DoS detection functionality of our IoT gateway. Then, we measure the average round trip time of ping messages for 1 min between legitimate devices attaches to IoT gateway under various kinds of DoS attack. We tested each DoS attack in three times and obtain the average value of the round trip time.

To evaluate the detail impact of DoS attack, we change the rate of DoS attack by using different rate such as 1000 packets in a sec (u100), 2000 packets in a sec (u50) and 10000 packets in a sec (u10).

Figure 2 illustrate the average round trip time between legitimate devices under various kinds of DoS attacks. In this figure, we use device notations such as n1, n2, n3 and n4 described in Fig. 1.



(a) Ping between n3 and n2 under DoS attack from n1 to n2

(b) Ping between n1 and n2 under DoS attack from n3 to n4

(c) Ping between n1 and n4 under DoS attack from n3 to n2

Fig. 2. Average round trip time between legitimate devices under various DoS attack

4.1 DoS Attack from Wired Interface to Wired Interface

For evaluating wired to wired DoS attack, we initiated a DoS attack from n1 to n2, then measured the round trip time of ping between n3 and n2 like Fig. 2(a). We observed that the ping time increases along with the increase of DoS attack traffic.

4.2 DoS Attack from Wireless Interface to Wireless Interface

For evaluating wireless to wireless DoS attack, we initiated a DoS attack from n3 to n4, then measured the round trip time of ping between n1 and n2 like Fig. 2(b). Interestingly, DoS attack between wireless devices does not affect the performance of ping between wired devices. It is because the attack traffic does not bother the legitimate traffic.

4.3 DoS Attack from Wireless Interface to Wired Interface

For evaluating wireless to wired DoS attack, we initiated a DoS attack from n3 to n2, then measured the round trip time of ping between n1 and n4 like Fig. 2(c). In this case,

even though only u100 case DoS attack affects the round trip time between legitimate devices significantly. It is because the packets from wireless interface bother the packets from wired interfaces. Also, we observed lots of packet loss from DoS attack traffic (more than 70%), it is because of the difference of bandwidth between wireless interface and wired interface.

4.4 DoS Attack from Wired Interface to Wireless Interface

For evaluating wired to wireless DoS attack, we initiated a DoS attack from n1 to n3, then measured the round trip time of ping between n3 and n4. But in this case, we cannot measure the time for ping because n3 and n4 lose their connection to IoT gateway. In this case, the DoS attack traffic which is generated by wired interface overwhelm the queue of wireless interface and drop the heartbeat packets for maintaining wireless connections.

5 Conclusion and Future Works

IoT technology gains huge attentions for future industry, and more and more devices are attached to IoT gateway for collaborating each other with more intelligent services. In this case, DoS attack on IoT gateway may be a critical challenge to maintain intelligent IoT services. Through the implementation based evaluation of DoS attack on IoT gateway, we observed that the wired to wireless type DoS attack is most severe attack on IoT gateway, and realized that it is important to monitor wireless traffic carefully in order to providing effective DoS attack detection and prevention in IoT gateways.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (R0110-16-1001, Secure hardware containers technology to protect IoT devices from Denial of Service attack).

References

1. ETNEWS: IoT devices are used for DDoS attack (2014). http://www.etnews.com/20141029000125
2. OpenWRT: What is OpenWRT? https://openwrt.org
3. Rudman, L., Irwin, B.: Characterization and analysis of NTP amplification based DDoS attacks. In: Information Security for South Africa (ISSA), 12 August 2015, pp. 1–5. IEEE (2015)
4. Raza, S., Wallgren, L., Voigt, T.: SVELTE: real-time intrusion detection in the internet of things. Ad Hoc Netw. **11**(8), 2661–2674 (2013)
5. Lee, Y., Kim, S., Vu, D.T., Kim, K.: Sampling based network flooding attack detection/prevention system for SDN. KISM Smart Media J. pp. 24–32, 31 December 2015
6. Vu, D.T., Kim, K.: Evaluation of network flooding attack detection/prevention system for SDN in KOREN network. In: Proceedings of 2016 KISM, April 29–30, 2016. Silla University, Busan (2016)

Low-Cost Infrared Video-Oculography for Measuring Rapid Eye Movements

Youngsun Kong, Hyeonsoo Lee, Namik Kim, Seungyeon Lee, Jihwan Park, Taeyang Han, and Yunyoung Nam(✉)

Department of Computer Science and Engineering, Soonchunhyang University, Asan 336-745, South Korea
{yskong, ynam}@sch.ac.kr, nohf2002@gmail.com, {skadlr321, park_jh6444}@naver.com, lsy950919@gmail.com, tyh.sunhan@gmail.com,

Abstract. In this paper, we developed a low-cost video-oculography device that diagnose neural diseases such as nystagmus by the vestibular function test using an infrared camera. An infrared camera and LEDs were attached in a pair of developed goggles, which was located in front of an eye for experimenters. Collected videos were converted to gray channel from RGB channel, and each pupil was extracted using morphology operation. Rotatory chair tests were conducted with our device. Gain, asymmetry and phase were calculated from obtained video.

Keywords: Video-oculography · Infrared camera · Videonystagmography

1 Introduction

Eye movement is separated into voluntary movement for tracking an interest target and involuntary movement for relocation of eyes. Involuntary movement of eyes makes it possible to fix on a target in head rotating by moving eyes to opposite direction [1]. This is induced by the semicircular canals in the vestibule by sensing angular acceleration [2]. However, disorder of the semicircular canals can cause involuntary movement of eyes while no rotating head, which is called nystagmus [3].

In order to diagnose nystagmus, rotatory chair test, which was introduced in 1907, is conducted to potential patients [4]. The rotatory char test measures dizziness by rotating chair at 0.01 to 0.64 Hz increased by two times with sinusoidal velocity and by measuring eye movement signal [4]. In the past, electronystagmography (ENG) was frequently implemented to detect eye movement signal [5]. ENG has advantages of being able to measure while closing eyes and of cost, yet has disadvantages of being caused motion artifact by eye blink, perspiration or light, and of difficulty measuring vertical eye movement. With the advance in the computer vision technology, videonystagmography (VNG) is used in general, although ENG is still valuable data [6]. VNG is the technology that can test whether dizziness is caused by inner ear disease, which uses an infrared camera for tracking pupil in dark. Advantages of VNG that are high accuracy rate and non-invasive method make it being used, despite its

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_44

high cost. In order to analyze eye movement signals with rotatory chair test, three factors should be calculated, which are gain, phase and asymmetry.

Various researches that use VNG have been conducted [7–9]. In [7], a new method that resolves the problem of the estimation of eye position in the analysis of VNG was presented. Horizontal and vertical signals were analyzed based on position, amplitude and duration. In [8], the researchers proposed a medical characteristics analysis method for displacement vectors of nystagmus by using Gaussian mixture models. In [9], a vestibular diseases analysis method was proposed from VNG applications by measuring pupil movement. The researchers selected features using a Fisher linear discriminant analysis that can extract the most significant features based on the inter intra variance deviation. Most researches extracted pupil from video data and analyzed the eye movement signals. However, the analyzed data are not enough as clinical diagnosis, and cannot use with rotatory chair test.

In this paper, a pair of head-mounted goggles was built with an infrared camera and used with rotatory chair to obtain eye movements. Then, pupil coordination was extracted from the video. From the pupil data, three factors, which are gain, phase and asymmetry, were calculated for measuring rapid eye movements.

2 Video-Oculography (VOG) System

Figure 1 shows a pair head-mounted goggles that can obtain infrared video. The system is composed of a Raspberry Pi 3 model B with Pi Camera v2 and gyroscope MPU-6050. In addition, three 850 nm infrared LEDs were attached to the goggles.

(a) (b)

Fig. 1. A head-mounted goggle with an infrared camera. (a) goggle inside, (b) measuring eye movement.

2.1 Feature Extraction

Figure 2 shows a flow chart of the pupil extraction method. First, the gray channel is extracted from the infrared images. Then, binary images were generated from the gray channel. After image binarization, the morphology operation was implemented to reduce noise in binary images. Finally, Hough transform algorithm was used for extraction pupil contour. Figure 3 shows examples of the pupil tracking method.

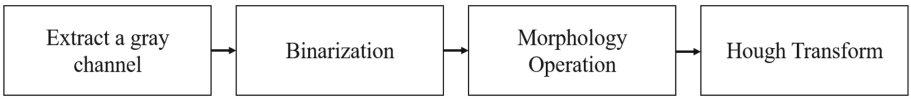


Fig. 2. Flow chart of pupil contour extraction

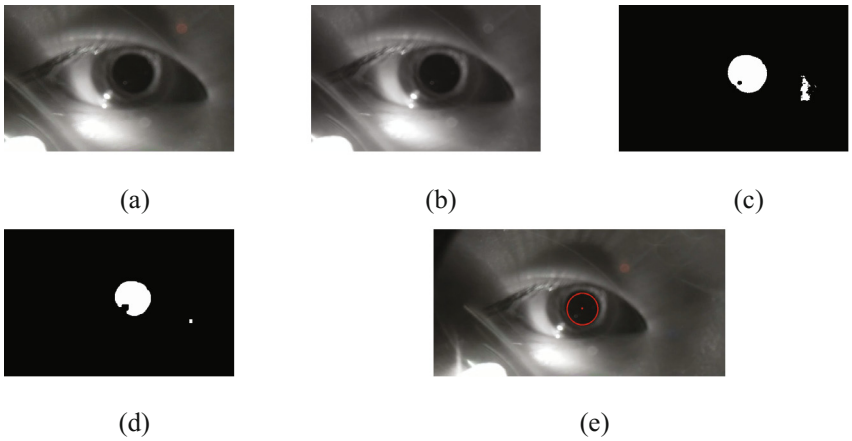


Fig. 3. Examples of pupil extraction (a) raw image, (b) gray channel, (c) binary image, (d) noise reduced image by morphology operation, (e) results

In order to calculate gain, asymmetry and phase, velocity of pupil should be calculated. Figure 4 shows a flow chart of the pupil velocity calculation. First, pupil movement signal is normalized. Then, velocity of the movement signal was calculated by derivative of normalized signal. In addition, fast phase of signal is removed using frequency analysis with interpolation method. Finally, moving average filter is applied.

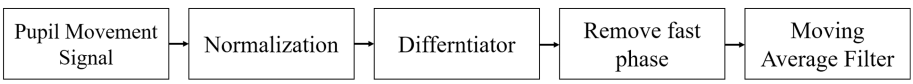


Fig. 4. Flow chart of pupil velocity calculation

Gain is proportion of velocity between of the maximum slow phase eye and the maximum stimulus velocity (rotatory chair), which is estimated as follows:

$$\text{Gain} = \frac{\textit{Amplitude of the maximum slow phase eye velocity}}{\textit{Amplitude of the maximum stimulus velocity}}. \quad (1)$$

Phase indicates time difference between the maximum stimulus velocity and the maximum slow phase eye velocity, which is calculated in degree. Finally, asymmetry compares the maximum slow phase eye velocities between clockwise and counterclockwise.

3 Experiments

Data were collected while a subject seated on a rotatory chair [10] (*System 2000*, *micromedical*) in the upright position with wearing the developed goggles. The subject was instructed to bend his head forward at approximately 15 degrees. Video data were recorded with 600 × 420 resolution at 30 Hz. Moreover, 3-axis gyro data were recorded at 80 Hz. Figure 5 shows calculated velocities of head and slow phase eye at 0.01 Hz and 80°/s. The y-axis represents averaged velocities in degree and rotation direction.

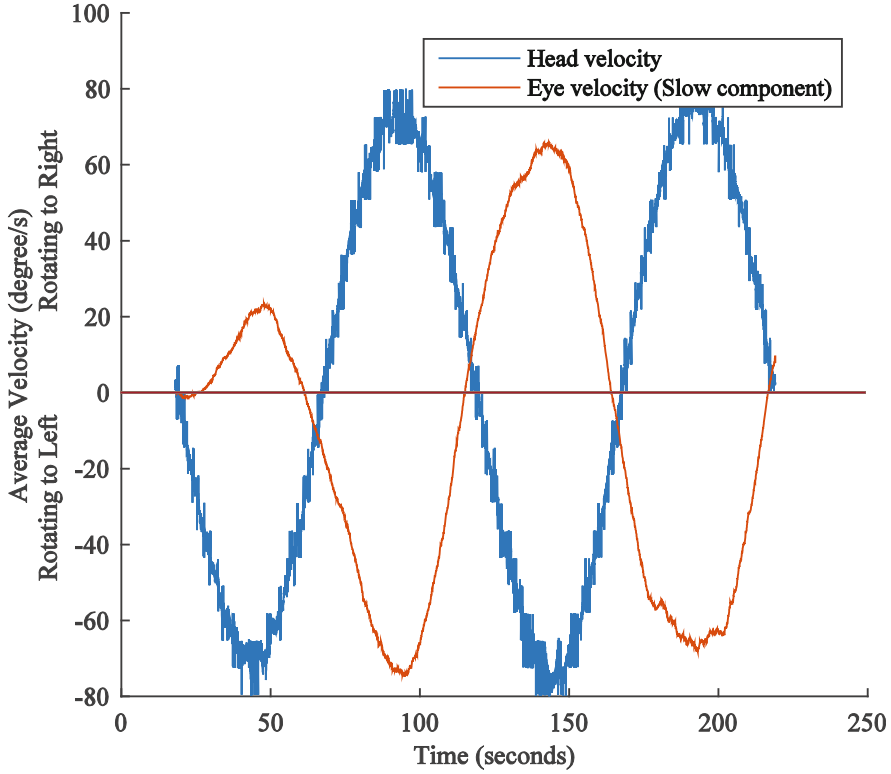


Fig. 5. Averaged velocities (head and eye) at 0.01 Hz and 80°/s

4 Conclusion and Discussion

This paper has presented low-cost VNG using infrared camera. The device was developed using a pair of head-mount goggles with infrared camera based on pupil tracking method. When a subject blinks he/her eyes, we found that pupil tracking signals were missed. Thus, the system should adjusted the missing signals using interpolation algorithms. Moreover, the method of eye velocities should be improved

because the velocities in degree is required, not in pixel. In the future, the more experiments will be conducted on more subjects for clinical test.

Acknowledgments. This study was partially supported by a grant from the Korean Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (HI12C0110), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1004033), and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-H8601-16-1009) supervised by the IITP(Institute for Information & communications Technology Promotion, and was also supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (NRF-2015M3A9D7067219).

References

1. Ditchburn, R.W., Ginsborg, B.L.: Involuntary eye movements during fixation. J. Physiol. **119**(1), 1 (1953)
2. Meiry, Jacob Leon. The vestibular system and human dynamic space orientation. Diss. Massachusetts Institute of Technology (1965)
3. Epley, J.M.: Positional vertigo related to semicircular canalithiasis. Otolaryngol. Head Neck Surg. **112**(1), 154–161 (1995)
4. Hain, T.C.: Rotatory Chair Testing (2016)
5. Jongkees, L.B.: Electronystagmography. Hno **12**, 325–329 (1964)
6. Ganança, M.M., Caovilla, H.H., Ganança, F.F.: Electronystagmography versus videonystagmography. Braz. J. Otorhinolaryngol. **76**(3), 399–403 (2010)
7. Slama, A.B., Machraoui, A.N., Sayadi, M.: Pupil tracking using active contour model for videonystagmography applications. In: 2014 First International Image Processing, Applications and Systems Conference (IPAS). IEEE (2014)
8. Mao, Y., et al.: An application of Gaussian mixture models for medical characteristics analysis of nystagmus signals. In: 2013 6th International Conference on Biomedical Engineering and Informatics. IEEE (2013)
9. Slama, A.B., et al.: Features extraction for medical characterization of nystagmus. In: 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE 2016
10. Micromedical, System (2000). http://www.micromedical.com/Products/Rotational-Chairs/

VM Relocation Method for Increase the Resource Utilization in Cloud Computing

Sangwook Han[1], MinSoo Chae[1], and Hwamin Lee[2(✉)]

[1] Department of Computer Science and Engineering,
SoonChunHyang University,
Soonchungyang-ro 22, Asan, Chungnam 336-745, Korea
sanguk@sch.ac.kr, cmspr90@gmail.com
[2] Department of Computer Software Engineering, SoonChunHyang University,
Soonchungyang-ro 22, Asan, Chungnam 336-745, Korea
leehm@sch.ac.kr

Abstract. Virtual machines are enabled to many of physical server can be integrated into fewer physical server. Integration of the server using virtual server technology induces the efficient use of resources to bring the cost benefits. Power consumption of the data center has been increased by 45% or more every year. More than 60% of the maximum power consumption is wasted on the physical server idle state, one way to reduce energy consumption is to minimize the number of physical servers. In this paper, VM usage time (running time) applied to 0–1 Knapsack algorithm. This method is the VM arrangement technology that can minimize the use of energy.

Keywords: VM relocation · Knapsack algorithm · Server integrated · VM running-time

1 Introduction

Development of virtualization technology is supports a server integration in the cloud data center, which is a physical machine to be able to have a high utilization rate [1]. The entire world, take advantage of smart devices and big data analysis and spread the Internet of things, cloud data center has been fixed in the necessary infrastructure.

As data center utilization is increased, the power consumption to be used by the data center is also increasing. In the case of the South Korea of NAVER data center, 120,000 of the server is running without a break for 24 h. During the past year, NAVER data center power consumption used was 52,710,000 kWh [2]. And the operation of the server discharges about 22,352t CO_2. This is the amount that occurs when a round-trip distance of 400 km 178,000 times.

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support programs (IITP-2016-R0992-16-1006 & IITP-2016-H8601-16-1009) supervised by the IITP (Institute for Information & communications Technology Promotion).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_45

In the case of cloud data center, more than 60% of the maximum power consumption is wasted on the physical server idle state. Therefore, increase the resource utilization in the cloud data center. It is the best way to reduce power consumption.

In this paper, 0–1 Knapsack problem is applied to the virtual machines that are distributed multiple physical server, we propose an algorithm to improve the resource utilization.

2 Related Works

Server integration is to relocate the running virtual machines to several physical machine and to conserve energy using off the power supply of the remaining physical server, and. Therefore energy-saving efficiency of server integration can increase the fewer the number of requests to the physical machine.

First Fit Decreasing (FFD) algorithm shows the high speed and good placement performance. Therefore, a lot of research of the server integration uses FFD [3, 4].

However, FDD algorithm is the larger CPU or memory deviation of the VM, the lower placement performance. Using the 0–1 Knapsack algorithm in order to solve these drawbacks. In addition, take advantage of CloudSim simulation by compare the results, the conclusion of this paper.

3 Optimization Method of Placement of the VM

3.1 VM Placement Method Using the 0–1 Knapsack Algorithm

Knapsack problem is a famous problem of combinatorial optimization. Briefly, the maximum value of the weight that can be saved in one of the bags have been fixed. When you put the thing that there is a certain 'value' and 'weight' to the bag, the sum of value is a method to choose the one that maximize [5].

0–1 Knapsack problem cannot split things, each unique things one by one. Therefore, it is suitable for applying an algorithm to placement the VM on the physical machines.

Table 1 is the source code that implementation the CloudSim simulation. Easy to understand was make the pseudo code.

$$
\begin{gathered}
\mathrm{i} = \{0 \leq \mathrm{i} = \mathrm{CPU}\} \\
\mathrm{j} = \{0 \leq \mathrm{i} = \mathrm{MEM}\} \\
\mathrm{k} = (0 \leq \mathrm{k} \leq \text{the number of VM}) \\
\mathrm{matrix[i][j]} = \max(\mathrm{matrix[i - vCPUk][j]},\ \mathrm{matrix[j - vMEMk]} + \mathrm{VMiv})
\end{gathered}
$$

The results are returned in the form of matrix[CPU][MEM], With the maximum value at matrix[500][1000] when CPU = 500 and MEM = 1000. Therefore, returned to VM allocation list when is matrix[500][1000] (Table 2).

The value of VMiv is determines the priority due to the calculated by vCPUi * vMEMi * runningtime. Here, runningtime is running-days for VM. In other words,

Table 1. Our VM placement algorithm using Knapsack algorithm

Knapsack algorithm

```
Output : matrix[CPU][MEM]

for i = 0 to n {
  for j = CPU to 1 {
    for k = MEM to 1 {
      if(vCPUi > j || vMEMi > k)
        matrix[j][k] = matrix[j-1][k-1]
      else if( matrix[j][k].value <
                    matrix[j-vCPUi][k-vMEMi].value+VMiv)
        matrix[j][k]=matrix[j-vCPUi][k-vMEMi].value+VMiv)
        matrix[j][k].selectVM += VMi
    }
  }
}
return matrix[CPU][MEM]
```

Table 2. Variable notation

Variable	Explan
CPU	CPU performance of the PM (Unit : MIPS)
MEM	Memory capacity of PM (Unit : MB)
VMi	The i-th VM
VMiv	The i-th VM value
vCPUi	The i-th VM cpu requirements
vMEMi	The i-th VM memory requirements
runningtime	running-days for VM
VMiv	The i-th VM value of which determines the priority of the VM allocation
matrix[i][j]	When using a CPU and memory of about i, j, the maximum value

when VM is running for three days, the value of runningtime is 3. Therefore, running-time is longer VM has a high value, it is first assigned.

3.2 VM Placement in Consideration of the Running Time

VM running time is very important in order to assign the VM to a physical machine. The algorithm proposed in this paper, In consideration of the VM driving time, It is possible to increase the resource utilization of the data center at the time of relocation of the VM.

Figure 1 is VM placement that does not take into account the runningtime. When considering only resource utilization, When you assignment a VM, resources are

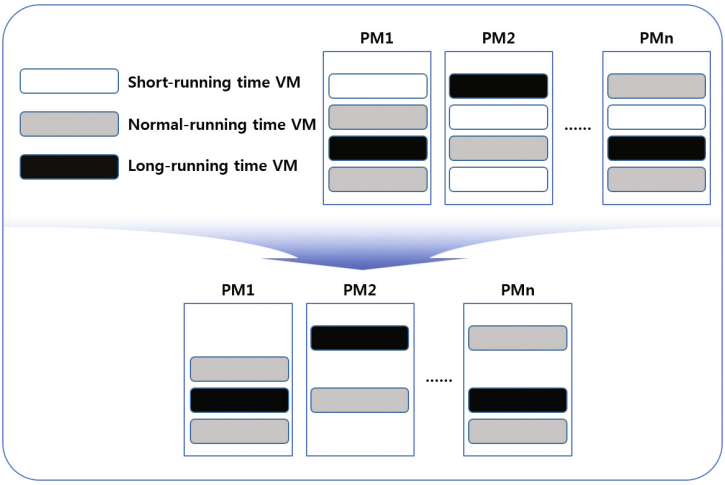


Fig. 1. Doesn't take into account the runningtime

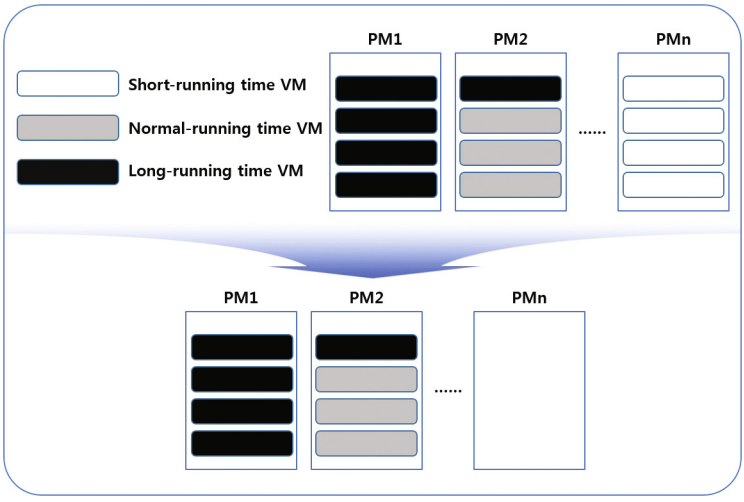


Fig. 2. Considering VM placement running time

released to the running time is short order. as time goes by resource utilization has deteriorated.

As shown in Fig. 2, when the arrangement by increasing the priority to have long running time VM, Even running time is short VM resources are released, it is possible to maintain the resource utilization.

3.3 Simulation

Using the CloudSim simulator, we were compared FFD and with proposed in this paper. Physical server was set equal to the high-spec PC. CPU usage and memory is respectively assumed as 8000Mips and 10000 MB.

Table 3. Range of VM random generation

CPU	Memory	runningtime	The number of VM
100 ∼ 2000 Mips	100 ∼ 2500 MB	1 ∼ 100days	100

Table 3, when creating a VM, a range of set randomly CPU and memory and the drive time. When you create a large number of the VM to the actual data center, We are take into account VM size not constant. To definitely compare FDD algorithm and proposed in this paper, the simulation was five times.

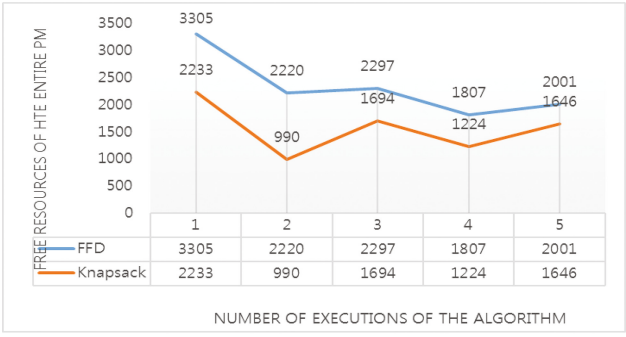


Fig. 3. Free resources of the entire PM

Fig. 4. A number of PM generated

Figure 3 is comparing the free resources of the entire PM, Fig. 4 is the number of PM requires FDD and Knapsack. A algorithm proposed in this paper can know better than FFD. Averagely, 49% of resource utilization has increased.

4 Conclusion

To summarize the results of this paper are as follows. Arranging method of the virtual machine using the existing FFD, when assigned a uniform size of the VM to the physical machine looks relatively good performance, but in other case when we assigned to VM size is not uniform in the physical machine then we can found to be inefficient. And also when we admeasure it we don't consider VM's driving time so FFD algorithm is unsuitable in Optimizing resource utilization.

As shown from the results of the present paper, our suggestion about algorithm can useful in efficiency of energy in large-scale of data center and other IT field that using cloud server.

References

1. Kim, H., Kim, Y., Youn, C.: Cost-efficient VM placement with VM interference control in cloud environment. **20**(02), 0192–0194 (2013). 11
2. Song, G., Shin, T.: A study on the introduction of green IT based on the cases of implementing green internet data center. Inf. Syst. Rev. **11**(02), 147–166 (2009). 8
3. Wang, M., Meng, X., Zhang, L.: Consolidating virtual machines with dynamic bandwidth demand in data centers. In: Proceedings of IEEE INFOCOM, pp. 70–75 (2011)
4. Bobroff, N., Kochut, A., Beaty, K.: Dynamic placement of virtual machines for managing SLA violations. In: Proceedings of IFIP/IEEE International Symposium on Integrated Network Management, pp. 119–128 (2007)
5. Knapsack Problem, Wikipedia. https://en.wikipedia.org/wiki/Knapsack_problem

Challenges and Experiment with LoRaWAN

Seung-Kyu Park[1], Kwang-il Hwang[1(✉)], Hyo-Seong Kim[2], and Byoung-Sup Shim[2]

[1] Department of Embedded Systems Engineering, Incheon National University, Incheon, South Korea
hkwangil@inu.ac.kr
[2] Leotek R&D Center, Incheon, Korea

Abstract. Low-Power Wide-Area Network (LPWAN) or (LPWA) is a type of wireless telecommunication network designed to allow long range communications at a low bit rate among things (connected objects), such as sensors operated on a battery. The characteristics of extremely long communication range and low power provide more feasible, reasonable wireless connectivity with a lot of researchers, field engineers, and application designers, who have suffered from weak wireless connectivity, range limitations, and energy efficiency. Therefore, in this paper, key applications and challenges of LPWA will be handled first, and several existing (emerging) new LPWA standards will be discussed. In particular, LoRaWAN, which is regarded as one of the most effective LPWA solutions, is mainly focused. So, several functionalities and characteristics of LoRaWAN will be presented. In addition, network architecture different from legacy short range wireless communications such as IEEE 802.15.4, IEEE 802.11, etc., will be overviewed and state-of-the-art off-the-shelf LoRa chipsets and modules will be introduced. Then, in order to help understand how to develop LoRaWAN application, our LoRaWAN experiments will be presented. At the end of the keynote, design considerations of LoRaWAN application and network, and several types of design methodologies will be discussed.

Keywords: LPWA · LoRaWAN · Gateway

1 Introduction

Low-Power Wide-Area (LPWA) [1, 2] is a generic term for a group of technologies with the following key characteristics: long battery life (often in excess of 10 years whilst supporting a benchmark smart metering application), low cost chipsets and networks, and limited data communications throughput capacity.

Major advantages in using LPWA technology are as follows. First, One technology cannot serve all of the projected applications and volumes for IoT. Second, WiFi and BLE are widely adopted standards and serve the applications related to communicating

This work was supported by the R&D Program (S2392966) funded by the Small and Medium Business Administration (SMBA, Korea).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_46

Feature	LoRaWAN	Narrow-Band	LTE Cat-1 2016 (Rel12)	LTE Cat-M 2018 (Rel13)	NB-LTE 2019(Rel13+)
Modulation	SS Chirp	UNB / GFSK/BPSK	OFDMA	OFDMA	OFDMA
Rx bandwidth	500 - 125 KHz	100 Hz	20 MHz	20 - 1.4 MHz	200 KHz
Data Rate	290bps - 50Kbps	100 bit/sec 12 / 8 bytes Max	10 Mbit/sec	200kbps – 1Mbps	~20K bit/sec
Max. # Msgs/day	Unlimited	UL: 140 msgs/day	Unlimited	Unlimited	Unlimited
Max Output Power	20 dBm	20 dBm	23 - 46 dBm	23/30 dBm	20 dBm
Link Budget	154 dB	151 dB	130 dB+	146 dB	150 dB
Batery lifetime - 2000mAh	105 months	90 months		18 months	
Power Efficiency	Very High	Very High	Low	Medium	Med high
Interference immunity	Very high	Low	Medium	Medium	Low
Coexistence	Yes	No	Yes	Yes	No
Security	Yes	No	Yes	Yes	Yes
Mobility / localization	Yes	Limited mobility, No loc	Mobility	Mobility	Limited Mobility No Loc

Fig. 1. Comparison of LPWA technologies

personal devices quite well. Lastly, LPWAN is designed for sensors and applications that need to send small amounts of data over long distances a few times per hour from varying environments [3–6].

As shown in Fig. 1, there are several state-of-the-art LPWA standard technologies including LoRa, Sigfox, LTE-m, and NB-IoT. Among them, at this moment, LoRaWAN shows superior characteristics, especially in terms of power efficiency, security and battery life aspects. Therefore, in this paper we focus on LoRaWAN standard.

2 Overview of LoRaWAN

LoRaWAN (Long Range Wide Area Networks) is a PHY and MAC protocol for a high capacity long range and low power star network. The LoRa Alliance is standardizing for Low Power Wide Area Networks (LPWAN). First of all, major advantage of LoRa is long range capability (e.g., a single gateway or base station can cover entire cities or hundreds of square kilometers). Actually, communication range highly depends on the environment or obstructions in a given location, but LoRa and LoRaWAN™ have a link budget greater than any other standardized communication technology [7–10].

In addition LoRa is the physical layer utilized to create the long range communication link. Many legacy wireless systems use frequency shifting keying (FSK) modulation as the physical layer because it is a very efficient modulation for achieving low power. However, LoRa is based on chirp spread spectrum modulation, which maintains the same low power characteristics as FSK modulation but significantly increases the communication range. Chirp spread spectrum has been used in military and space communication for decades due to the long communication distances and robustness to interference, but LoRa is the first low cost implementation for commercial usage.

There are a number of important design factor for LPWA. So, LoRa fulfills several requirements as follows:

- Network architecture
- Communication range
- Battery lifetime or low power
- Robustness to interference
- Network capacity (maximum number of nodes in a network)
- Network security
- One-way vs. two-way communication
- Variety of applications served.

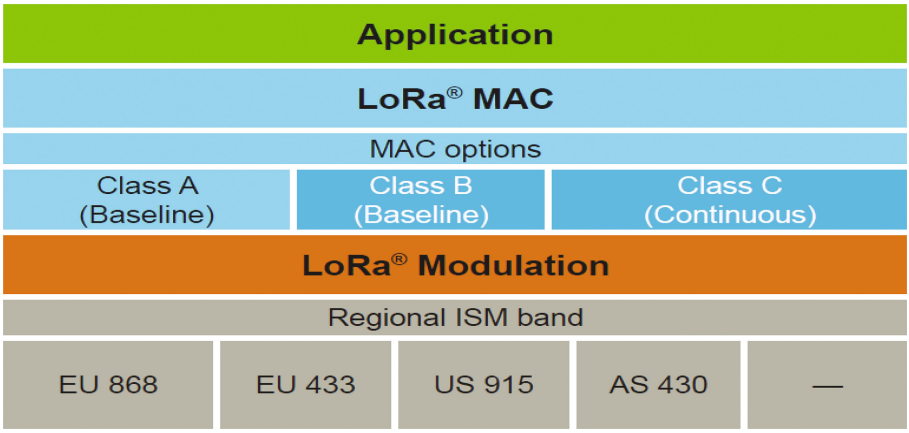


Fig. 2. LoRaWAN protocol stack

Basically, LoRa standard define LoRa physical layer and MAC layer as shown in Fig. 2. The frequency band used is varied according to region and LoRa has three different types of MAC options: class A, B, and C. Figure 3 summarizes the characteristics of each class.

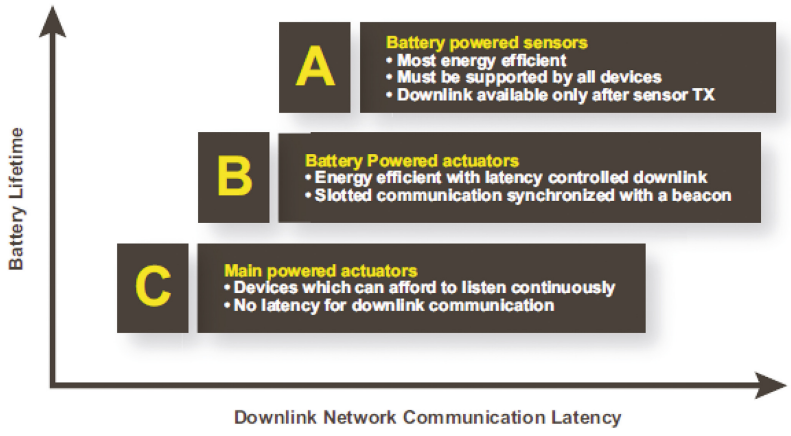


Fig. 3. LoRa class

End Devices serve different applications and have different requirements. In order to optimize a variety of end application profiles, LoRaWAN utilizes different devices classes.

In order to participate in a LoRaWAN network, each end-device has to be personalized and activated. Therefore, for join process, LoRa support two modes as follows:

- Over-The-Air Activation (OTAA) when an end-device is deployed or reset.
- Activation By Personalization (ABP) in which the two steps of end-device personalization and activation are done as one step (Fig. 4).

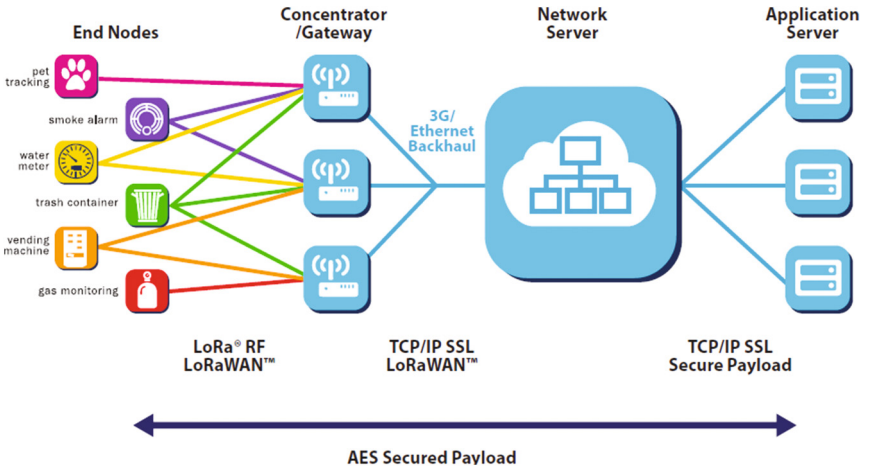


Fig. 4. LoRaWAN network architecture

Unlike legacy short range wireless communication technologies such as IEEE 802.15.4, or IEEE 802.11, LoRaWAN requires Internet-based network server management. Therefore, In a LoRa network, gateway plays a role in packet forwarder between LoRa network and IP network. Instead, Network server should directly manage each node connected to gateways. Finally, user can receive several service from network server.

3 Experiments

In order to test full functionality of LoRaWAN, we need all components for LoRaWAN including gateway, network server and end device. Therefore, in this section we describe our test environments and results using off-the-shelf LoRa products and network server services.

For logical network test, we utilize MultiTech conduit gateway wth 915 MHz LoRa mCard and for end device MultiTech mDot 915 nodes. In addition, to adapt Korean

regulation for LoRa frequency, we also configured mDot and Conduit Gateway's environmental variables including Network ID, FSB, Network Key. After configuration to gateway and end device, respectively, we tested communication range test to verify the long range performance of LoRa.

Figure 5 shows the result of range test. The test was conducted in Songdo, Incheon, Korea, the result proved that maximum communication range of LoRa is up to 5 km. The distance is sufficient to cover entire area of the city with a single gateway.

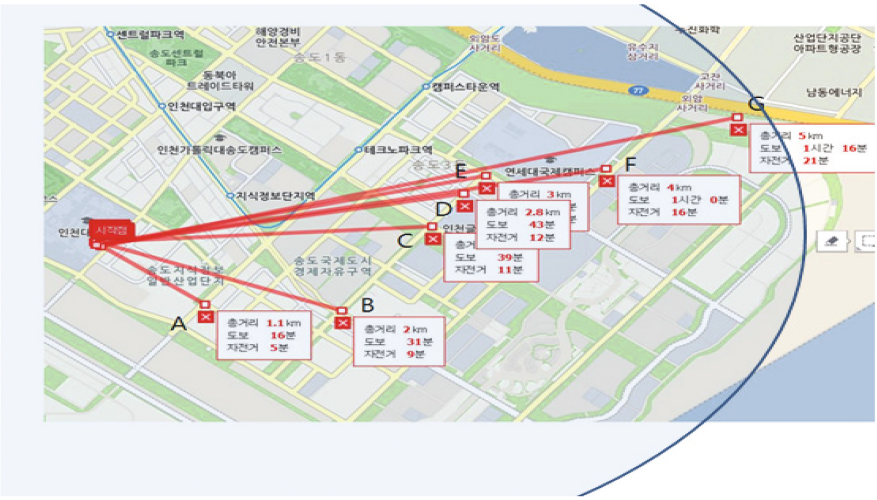


Fig. 5. Communication range test result.

4 LoRaWAN Network Server

LoRa network server is used to maximize the full functionality of LoRaWAN and most LoRa application developers use LoRaWAN cloud service. Therefore, we tested open LoRa network server services, TTN (The Things Network), and Loriot (Limited free account) sites Fig. 6 shows TTN network architecture.

Figure 7 also shows application data access using REST API via web browser.

Our test environment is as follows.

- The number of gateway: 1
- The number of End Devices: 2
- Data: Periodic data @ every 2 s
- Network Server: TTN server
- Success ratio: 100% (obtained from network server).

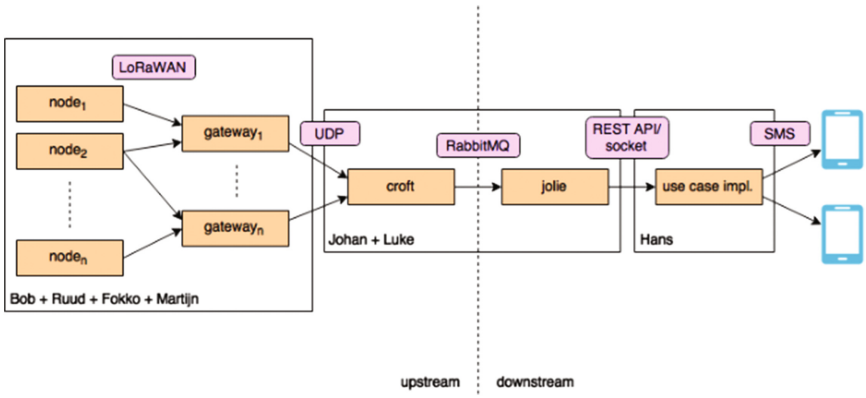


Fig. 6. Architecture of TTN

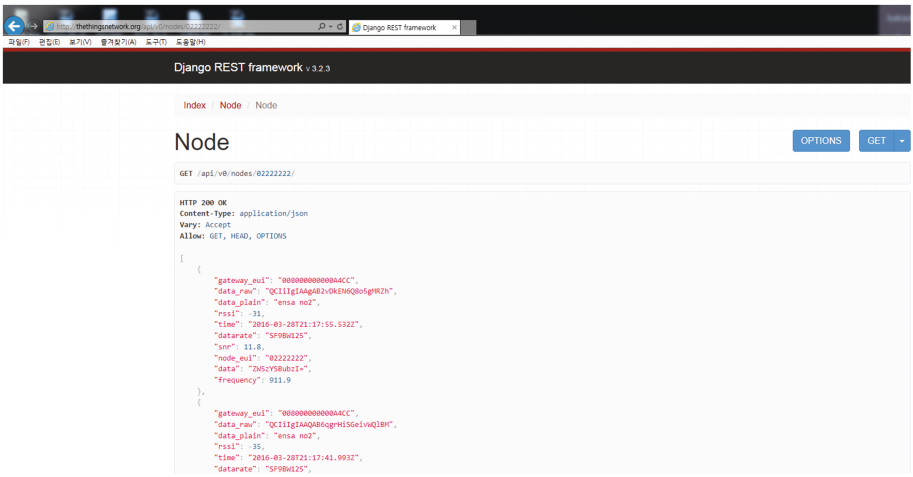


Fig. 7. TTN Web browser

However, The TTN cloud service only supports one-way (Uplink: ED → gateway → network server → user) and Text-based RESP api debugging is so uncomfortable (Fig. 8).

Therefore, we also tested the other network server service, Loriot Cloud Service. The Loriot service provides:

- Good visibility using dash board
- Easy to check several status such as connectivity, data channel, history, etc.
- For free account (1 gateway, 10 devices).

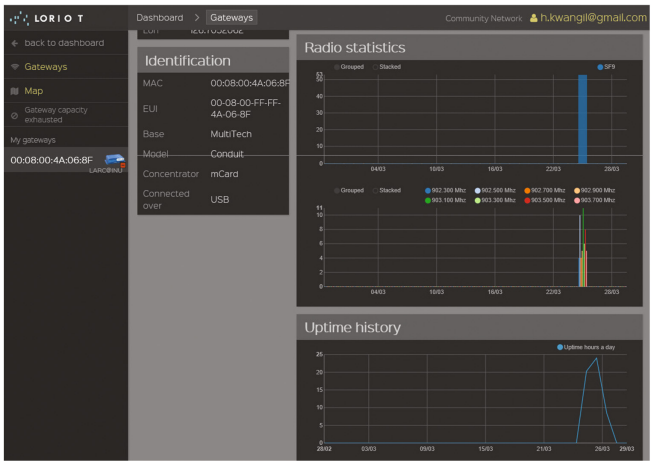


Fig. 8. Loriot cloud service

5 Conclusion

In this paper, we introduced Low-Power, Wide-Area (LPWA) technology which complements existing cellular mobile network and short range technologies, enabling lower costs and better power consumption. In particular, long range, low power characteristics are leading to the extremely rapid growth rate driven by a wide range of applications, each of which is the product of a rapidly evolving technological landscape. In particular, LoRa is an integrated total solution covering all layers, softwares, and hardware platforms. However, users (developers) need to choose a design method appropriate for their LoRa application.

References

1. LPWA Technology for IoT. Nable Communications (2016)
2. LPWA Technologies (Unlock New IoT Market Potentiol). LoRaAlliance (2015)
3. A technical overview of LoRa and LoRaWAN. LoRaAlliance (2015)
4. LoRaWANSpecification, LoRaAlliance. https://www.lora-alliance.org
5. LoRa vs LTE-M vs Sigfox, N. hunn. http://www.nickhunn.com/lora-vs-lte-m-vs-sigfox/
6. Margelis, G.: Low throughput networks for the IoT: lessons learned from industrial implementations. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), UK, 21 January 2016
7. So, J.: LoRaCloud: LoRa platform on OpenStack. In: 2016 IEEE NetSoft Conference and Workshops (NetSoft), South Korea, July 2016
8. Aref, M.: Free space range measurements with Semtech Lora technology. In: Wireless Systems within the Conferences on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS-SWS), Germany, September 2014

9. Petäjäjärvi, J.: On the coverage of LPWANs: range evaluation and channel attenuation model for LoRa technology. In: International Conference of ITS Telecommunications, Finland, January 2016
10. Mikhaylov, K.: Analysis of capacity and scalability of the LoRa low power wide area network technology. In: European Wireless 2016, Finland, June 2016

Consumer's Behavioral System of Approach and Avoidance Investigating Generic Medicine Distribution and Logistics in Japan

Takefumi Hosoda and Hongsik J. Cheon(✉)

Department of IT Logistics, Soongsil University, Seoul 06978, Korea
takefumi.hosoda@takeda.com, hcheon@ssu.ac.kr

Abstract. This study will first select the feature of consumer behavioral system and build up its model then will evaluate the goodness-of-fit to confirm its suitability through the second-order factors analysis. There are six dimensions of risk correlated with consumer BIS – Functional, Financial, Social, Physical, Psychological, Time – and four dimensions of factors correlated with consumer BAS – Quality, Efficacy, Safety, Cost effectiveness.

Keywords: Generic medicine · BIS · BAS · Perceive risk · Consumer behavior

1 Introduction

In order to examine consumer behavior dealing with generic medicine, we selected Quality, Efficacy, Safety and Cost Effectiveness as the factors of BAS, and Functional risk, Financial risk, Social risk, Physical risk, Psychological risk, and Time risk as the factors of BIS. Considering that the medical cost does not go any smaller in Japan, this research thus is to examine and diagnose potential variable controlling a consumer's "acceptance" or "avoidance" in order to improve benefits of welfare for Japanese citizens. To examine and diagnose variables properly, which is also the main goal of this research, we need to evaluate reliability and adequacy of measuring method then it will allow us to extract factors and items composing the method.

2 Literature Review

The Japanese government, recognizing their rapidly aging population and the associated increase in healthcare, began to be concerned about how to contain these costs, specifically pharmaceutical costs, since the mid-1970s (Oliver et al. 1997). As a means to control costs, the Japanese government introduced a number of pharmaceutical price-regulation policy starting in 1981, however, eight years later the per capita drug spending was still significantly higher than the U.S. (Ikegami et al. 1994). However, Japan falls far short of the successes that other countries have had in using generic drugs to lower health care costs (Thomson Reuters 2014). The Japanese pharmaceutical market, the world's 2nd largest, is traditionally renowned for its brands' domination and weakest generics share among major established economies (Jakovljevic et al. 2014).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_47

2.1 Consumer's Decision Making

If BAS and BIS do control human motives, then those system differentials between each human has something important to do with the personal welfare and health. For instance, the BAS sensitivity determines "How would a person react to an anticipated reward?" while the BIS sensitivity measures "How would a person react to an imminent danger or punishment?" Davidson (1992, 1998, 2000) and his team used the data measuring laterality of left and right prefrontal cortexes in resting phase then proved that low BAS sensitivity can cause depression and it can be the person's basic character. Also, he points out the sensitivity differential between BAS and BIS causes differentials between each effective styles, then it directly correlated with inheritance (Kagan et al. 1988), personality (Gross et al. in press), and vulnerability as in psychopathology (Meehl 1975).

2.2 Perceived Risk

Perceived risk is a concept being discussed in economics and Bauer (1960) first introduced it to explain consumer behaviors. Jacoby and Kaplan (1972) and Brooker (1984) classify it into five risks: financial, functional, physical, social and psychological risk; Peter and Ryan (1976) add time/comfort loss; Zikmund and Scott (1977) adds the risk of opportunity loss. Stone and Gronhaug (1993) divides risks into two big categories: financial/functional/physical/time risk and social/psychological risk. Schiffman and Kanuk (1994) types the perceived risk into financial, safety, functional, social, psychological, and time loss risk. Kurtz and Clow (1997) determines the risk in three types: social/psychological, financial, and consequence risk.

The understanding Quality and Efficacy can be also found in cue theory. A consumer or a patient can use two values: expected, assured value to recognize the quality (Olsen 1972). The expected value is to map certain cue on the product's quality, which provides reliability on solving problems. The assured value is a certainty of consumer's ability to use the cue in order to evaluate the product. These judgments pay significant roles on the evaluation process. The worst point of using generic medicine is that consumers generally recognize its quality lower than branded (Richardson et al. 1994).

Therefore, based on the preceding studies, we will examines consumer practical behaviors towards generic drugs with five causing factors: Quality, Efficacy, Safety, Reputation of Manufacturer, Price (Bearden and Mason 1978) in order to confirm their BAS property, and with six risks: Functional, Financial, Social, Physical, Psychological, Time (Mason and Bearden 1980) in order to figure their BIS property. Functional risk is the probable fact that consumers could question about quality of generic drugs; Financial risk is the probable fact that consumers find no value on purchasing generic drugs; Social risk is the probable fact that consumers worry if others would view their purchase of generic drugs in negative ways; Physical risk is the probable fact that consumers worry if buying generic drugs could harm them or their families physically; Psychological risk is the probable fact that consumers get uncomfortable with purchasing generic drugs; Time risk is the probable fact that consumer could regard purchasing generic drugs as wasting of time (Mieres et al. 2006).

3 Method

We grasped the concepts and subordinate factors of BIS and BAS through preceding researches, and based on this, we obtained the specific behaviors and characters of consumers shown towards generic medicine. Variables used in this study are reconstructed through the research of previous literatures and works. The primary dimension, BIS, comprises six risks, the secondary dimensions: Functional Risk, Financial Risk, Social Risk, Physical Risk, Psychological Risk, Time Risk, which are established from the basic theories and previous studies. We should examine whether those six secondary dimensions are converged to Behavioral inhibition system (BIS), the primary dimension. In order to study this matter, we carry out the second-order confirmatory factor analysis. The results are as follows (See Fig. 1). Likewise, we need to examine whether those five dimensions: Quality, Efficacy, Safety, Cost Effectiveness, are converged to Behavioral Activation System (BAS). We also proceed with second-order confirmatory factor analysis (See Fig. 1).

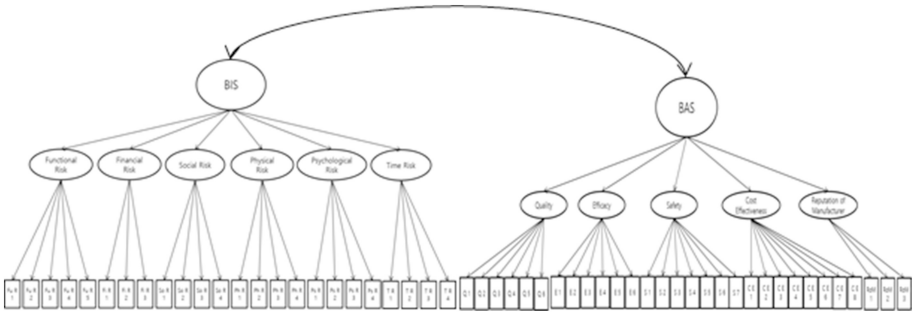


Fig. 1. Second-order confirmatory factor analysis of BIS & BAS

4 Result of Factor Analysis

4.1 Exploratory Factor Analysis

The result of the exploratory factor analysis per each dimension was as follows. KMO and Barlett's Test were carried out to confirm the validity and adequacy of items measuring the secondary factors of BAS and BIS, and its value turned out .72 in this work. On Barlett's test, Chi-square value was 5198.34 (df = 1431), which verified that the selected items for factor analysis were significant.

4.2 Confirmatory Factor Analysis and Model Fit

Each latent variable was observed 23 indicators for six dimensions (Functional Risk (t-value = 21.16–23.69), Financial Risk (19.65–23.32), Social Risk (20.01–25.05),

Physical Risk (24.90–26.07), Psychological Risk (24.43–25.61), Time Risk (23.60–25.70)) interacting with BIS, and 17 indicators for four dimensions (Quality (15.23–18.87), Efficacy (17.62–23.27), Safety (18.24–22.51), Cost Effectiveness (18.44–24.82)) interacting with BAS. The observed variables that represent the relationships between the measuring indicators and conducts turned out statistically significant. The correlations between primary and secondary dimensions in both groups (Functional Risk, Financial Risk, Social Risk, Physical Risk, Psychological Risk & Time Risk with BIS, Quality, Efficacy, Safety & Cost Effectiveness with BAS) were also verified as significant (see Fig. 2).

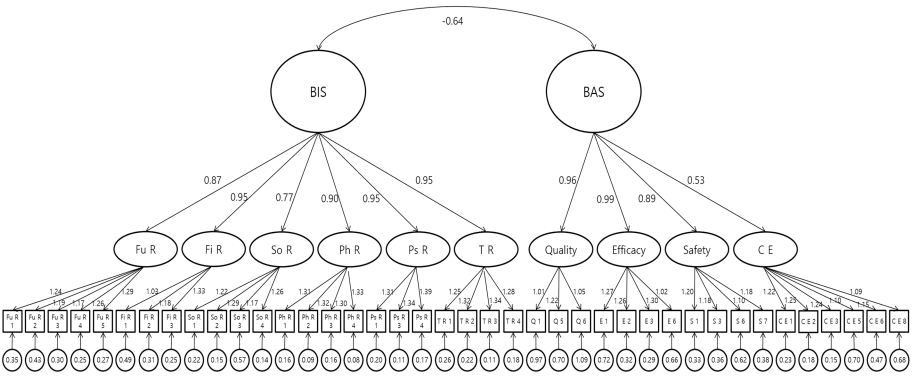


Fig. 2. Second-order confirmatory factor analysis for behavioral system

The overall value from the goodness-of-fit test in the confirmatory factor analysis was 2242.97 Chi-square (df = 729). As shown in results, the model thus is supported well by the data as the p-value < .01 with respect to the Chi-square statistics. Goodness-of-fit index (GFI) was 0.75; Adjusted goodness-of-fit index (AGFI) was 0.72. These two indices were over 0.7, which satisfies the acceptance standard, meaning the model adequacy is satisfying. Normed fit index (NFI) and Non-normed fit index (NNFI) were 0.98 and Comparative fit index showed 0.98. The closer these indices are to 1, the more the results are significant (Kline 1998). Since each of our indexes is 0.98, the model adequacy appears very satisfying. Root mean square error of approximation (RMSEA) shown 0.08 and this means also the model adequacy fits in a favorable range according to Cudeck and Browne (1993). RMSEA < 0.05 means satisfactory, 0.05 $\leq$ RMSEA < 0.08 is an acceptable range, 0.08 $\leq$ RMSEA < 0.10 is an average level, and RMSEA $\geq$ 0.10 would be out of acceptable range. The overall adequacy index from the hierarchical factor structure fits in the favorable range, so that the structure obtains the satisfying model adequacy as six factors in BIS and four factors in BAS.

Description below suggests the structural model of BIS and BAS scales that pertain to Japanese consumers dealing with generic medicine. We found that the correlation coefficient between BIS and BAS (r = −0.64) and the one between two factors from Batra and Ahtola (1990) (r = 0.55) were peculiarly different. The meaning of −.64

between BIS and BAS in (Fig. 2) is that two primary concepts are contrary one another and they hold completely different implications.

5 Conclusions

5.1 Results Summary

When a generic drug comes out in the market, not only it triggers lowering the price of patent-expired-drugs, but also the consumers can save money. Additionally, the research labs in the existing drug companies would more focus on generating new medicine rather than competing with generic medicine companies. Overall, governments in most countries have been encouraging consumers to take generic drugs to manage health security expenditure since they realized that the more use of generic drugs, the lower healthcare payout for drugs (Shin and Choi 2008). However, not all consumers have the same positive opinions about the copied drugs despite that generic drug markets keep expanding. Therefore, the aim of this study is to suggest the alternate plan for consumers to choose more of generic medicine by classifying their behaviors into BIS and BAS.

5.2 Limitation and Future Study

We find several important meanings from this study, which are described as follows. Firstly, BIS and BAS, are correlated with consumer behaviors towards generic medicine by clarifying how each system plays a role on their distinct behaviors towards generic medicine. Also, this work can be the contribution to future researches since it provides the solid theoretical proof supporting the tool developments to measure BIS and BAS. Secondly, this work is beneficial to the marketing divisions of pharmaceutical companies interested in specific and various causes of consumer acceptance and approach towards generic medicine through studying secondary factors of BIS and BAS. Third, this work is also beneficial to the government when it comes to establishing policies to cut pharmaceutical expenditure, as the high rate of elderly population and new pharmaceutical technologies concern them, by learning how to balance the factors of BIS and BAS. Finally, this study has ultimately examined the factors of BIS and BAS that consumers possess towards generic medicine, and brought the suitable and adequate research model. Especially in Japan, they have national and social health insurances which can carry out the survey to find consumer recognition about generic medicine so other countries that have similar healthcare system could use the same methods applying BIS and BAS.

Acknowledgments. This research was supported by the Ministry of Trade, Industry and Energy (MOTIE), KOREA, through the Education Program for Creative and Industrial Convergence (Grant Number N0000717).

References

Bearden, W.O., Barry Mason, J.: Consumer-perceived risk and attitudes toward generically prescribed drugs. J. Appl. Psychol. **63**(6), 741–746 (1978)

Ikegami, N., Yoo, B.K., Hashimoto, H., Matsumoto, M., Ogata, H., Babazono, A., Kobayashi, Y.: Japanese universal health coverage: evolution, achievements, and challenges. Lancet **378**(9796), 1106–1115 (2011)

Jakovljevic, M.B., Nakazono, S., Ogura, S.: Contemporary generic market in Japan–key conditions to successful evolution. Expert Rev. Pharmacoeconomics Outcomes Res. **14**(2), 181–194 (2014)

Kuribayashi, R., Matsuhama, M., Mikami, K.: Regulation of generic drugs in Japan: the current situation and future prospects. AAPS J **17**(5), 1312–1316 (2015)

Oliver, A.J., Ikegami, N., Ikeda, S.: Japan's aging population. Pharmacoeconomics **11**(4), 306–318 (1997)

Olson, J.C.: Cue utilization in the quality perception process: a cognitive model and an empirical test. Doctoral Dissertation, Purdue University (1972)

Richardson, P.S., Dick, A.S., Jain, A.K.: Extrinsic and intrinsic cue effects on perceptions of store brand quality. J. Mark. **58**(4), 28–36 (1994)

Shin, J.Y., Choi, S.E.: The volume and the price of generic drug consumption in Korea prescription drug market. Korean J. Health Econ. Policy **14**(1), 103–125 (2008)

Simoens, S., de Coster, S.: Sustaining Generic Medicines Markets in Europe. Palgrave Macmillan, New York (2006)

Vogler, S., Zimmermann, N.: The potential of generics policies: more room for exploitation–PPRI Conference (2012)

U.S. Food and Drug Administration Center for Drug Evaluation and Research, 19 August 2013. Generic. http//www.medterms.com

Cross-Conforming Approaches of ICT Functionality Design for Smart City

Jae-Young Ahn[1], Eunjun Rhee[1], Hyun-Woo Lee[1], and Dae Joon Hwang[2(✉)]

[1] Smart Platform Research Department, ETRI, Yusung, Deajeon City, Republic of Korea
{ahnjy, ejrhee, hwlee}@etri.re.kr
[2] Department of Information and Communiaction, SungKyunKwan University, Suwon, Republic of Korea
djhwang@skku.edu

Abstract. This paper deals with the design approaches of ICT functionality for smart city in an attempt to identify cross-conforming aspects between extreme approaches (IIFA, CSFA). Based upon a general modelling of city spaces and infrastructures, properties of those two approaches are compared to identify the chance to compensate the weak points and to combine the preferable features of those models each other. The result gives us a direction to combine those approaches for the design of ICT structure for smart city, that is synergistic enough for the planning of a smart city not only at a manageable level of fast installation but also for accommodating heterogeneity in city subsystems.

Keywords: Smart city · IoT · City infrastructure · Collaboration platform · Trust management · Interactive media · Interoperability

1 Introduction

Smart city is basically an effort to resolve many problems on the way of urbanization, by applying ICT as the vehicle of innovation. However, there are no definitive references for the design of ICT for smart city, and rather diverse approaches are suggested in many cases based upon each city's ideals and circumstances. [1–3]. In order to start the design of smart city under a specific ideals and goals, we need to define the city model first, as for the target system to deploy ICT functions on it. Though there are very many criteria to define the property of a city, this paper relies on the general model of a city as with the high level definitions of national land space development schemes in many countries [4]. The model accommodates two different viewpoints; 'citizens centric' and 'city operators centric'. At in the citizen centric viewpoint, a city is a collection of ***eco-systematic space***s that supports social and economic activities of their life. On the other side of it, for the city operators' or governors' viewpoint, a city is a collection of ***city infrastructures*** that can be layered into two; non-physical and physical supply infrastructures. Those spaces and infrastructures are shown as simple boxes at the central part of Fig. 1.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_48

Among many approaches to deploy ICT functions over a city, this paper depicts two extreme cases; one is to install ICT-based infrastructure alongside the traditional utilities and services of a city. It will be a rational solution for the development of smart city especially in viewpoint of city developers'. In that case the interoperability of ICT-based infrastructure with traditional city subsystems is essential to activate its ability of orchestration and integration. This type of installation is often represented as an ***(ICT) infrastructure-focused approach*** [4] in light of smart city-planning to add ICT as an additional infrastructure into traditional city structure.

Another approach that gives priority to the new socio-economic opportunities of convergence services exists; In that case the collaboration of various city subsystems is crucial to give synergistic innovation to the socio-economic spaces of a smart city. This type of installation is represented as ***(convergence) service-focused approach*** [4] or ***a holistic view on smart cities*** [12] since it focuses primarily on the innovative cross-governance among city service domains, from the early stage of designing ICT platforms, to create practical synergies through convergence.

This paper attempts to unveil the compatibility of those models introduced – an abstracted city model, a new ICT infrastructure, and service platforms – despite that they are dealing with city domains with different properties such as space, abstracted city asset flows, hardware based ICT system, and software platforms, etc. Those two extreme approaches were selected, to compare the typical properties and configurations of them at a manageable level of granularity.

2 ICT Infrastructure Focused Approach (IIFA)

The first case of the approach, adding the ICT function domains pertaining to ICT infrastructures (see the left most side domain of Fig. 1) alongside the traditional city structures (see the central part of Fig. 1), serves to form an integrated picture of a smart city at a stroke. A promising feature of an ICT infrastructure–focused approach (IIFA) is that to minimize the work for resolving non-technical issues, such as for the cross governance and regulatory issues, at the time of design and installation of a smart city. The aspect contributes the simplicity of ICT system design so as to help the easier installation of ICT infrastructures. However, it potentially accompanies a drawback that the ICT functions of a smart city are only being realized in isolation from the traditional city subsystems, while the approach advocates the city operators' viewpoint and the installation of new ICT infrastructure for new services. It is highly likely that newly launched smart city services will not focus on the convergence of new ICT functionality and city subsystems but seek to rely only on newly installed ICT infrastructure functions since it is much easier. To counter such an outcome, we can consider two countermeasures: (a) provisioning of a smart city interoperability framework that encourages smooth interworking of ICT infrastructure with traditional city subsystems to develop convergence services, and (b) the fine selection of citizen-preferred services that compensate the effects to the eco-systematic space innovations. Since intervening to the business of city infrastructure operators' is not much practical, rather a process of selecting citizen-preferred service prioritization [4] is usually applied to improve the effect of the ICT infrastructure. It seeks to exploit emerging technical means in ICT

infrastructure, such as IoT, cloud computing, big data, and 5G [5–11] as innovation vehicles. However, this approach still remains a risk of limited effects and weak innovation in traditional city infrastructure functions because of its structural limits. Most of the installation in this approach implements an integrated ***smart city control centre*** and advanced public services on the city management and surveillance, utilizing newly deployed ***smart devices and facilities***.

3 Convergence Service Focused Approach (CSFA)

In contrast, the convergence service–focused approach (CSFA) seeks to support an amalgamative realization of a smart city services, as such an approach would focus on providing new socioeconomic opportunities in a city. Those opportunities that occur as the result of new services based on interactions and coordination, from interactive-media in smart city such as multiscreen digital signage in eco-systematic space [17] to inter-departmental interactions and sharing of resources in city infrastructures, are assumed as to expedite the innovation of a city. This approach is represented, in some cases of city renovation project [12], by the term ***holistic view*** that implies the city-wide redesign challenges to interconnect the city services and infrastructures holistically for synergistic effects. It naturally builds a frame that focuses on the creation of convergence services in the eco-systematic spaces first, since the practical collaboration of hard systems in city infrastructure is generally more difficult than service level works. This aspect drives the innovation in the eco-systematic space of city to build innovative information services first, and it eventually moves to bring the city space and services into digital/cyber city space along with the ***city service platform*** that is created and operated by smart city operators (see the right most side domain of Fig. 1). The platform should be based on the collaborative functionality that realizes the inter-departmental interactions and coordination of resources in a city, In order to provide synergistic innovation to the socio-economic spaces of a smart city, as defined in Sect. 1. However, this ICT functions for the collaboration should resolve the heterogeneity of city subsystems, at the same time, to harmonize the management strategies and business goals of such subsystems. A ***collaboration and governance platform*** installed and operated by a smart city operator is required, which composes the core of cross-governance aspect, as experimented in a number of European smart city projects [12–14]. The right most side of Fig. 1 gives the illustration of smart city collaboration control center that deploys and operates the platform. At in the bottom line of the picture, a ***city data platform*** is drawn to show the capability of sharing data among physical and non-physical city infrastructures. There are many infrastructure operators and their facilities in a city, and those are generating data based on their own data platforms; however, sharing of the data is not generally activated, as those data collected from the physical end of infrastructures are often critical or sensitive, thus being used for their own usages exclusively. A standardized or harmonized way of sharing city infrastructure data is one crucial challenge to achieve innovation of convergence in a smart city [15, 16].

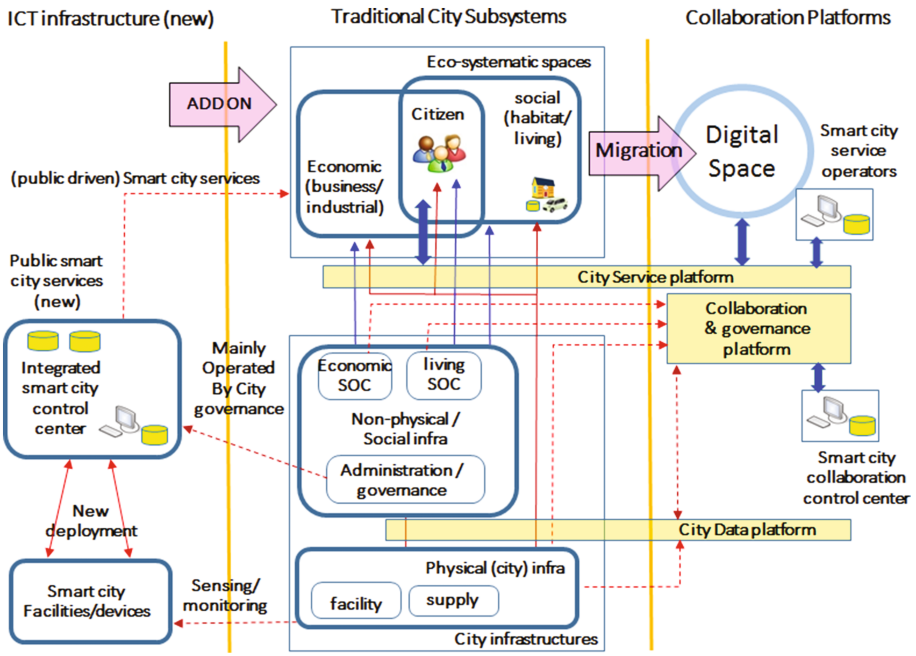


Fig. 1. ICT Infrastructure focused approach (left) and Convergence service focused approaches (right) for a smart city

4 Cross-Conforming Aspects of the Approaches

Those two approaches introduced in previous sections are the extreme cases of smart city planning models, in terms of structural design and technical concepts that they have respectively. From the combined drawing of those approaches in Fig. 1, we can identify the different features of their structures and technology at a glance. IIFA stands for the installation of new ICT functions vertically to provide integrated public services, and CSFA deploys the layers of platforms over a city to integrate city services horizontally. There are more like this, but the evidence is that they are standing in the opposite side of the functional aspect with their own drawbacks, that it gives a chance to compensate each other's weak points at the same time. IIFA deploys new devices network as smart city facilities; as newly deployed devices apply emerging standard technology of IoT which adopts data platform for horizontal integration, it gives a chance to address the challenge of 'sharing city infrastructure data' mentioned in Sect. 3. IIFA allocates the function domains at the same level of city infrastructures, so that it contributes the integration of city data platforms layer that is identified as a long term challenge in CSFA. On the other hand, CSFA preferentially deploys city service platform to provide amalgamative realization of a smart city services, so that it drives the innovation of city eco-systematic spaces where the IIFA wants to deploy citizen-preferred integrated services. In between those two layers cross-conforming, there are 'control centers' in both approaches. Those are basically for 'city control'

regardless of the span of controls. Public smart city service and city information management are specified in IIFA, as well as the collaboration and the cross management of city infrastructures service are designated in CSFA. Those functionality are much based on the specific scheme of a city planning; however, they can be merged or composed to take advantage of both approaches.

5 Conclusion and Further Issues

There are huge number of attempts to design a smart city, based on their own modeling concepts for a city and ICT functionality for it; however, not many study has been done to categories the design approaches to identify the dominant aspect of them and to cross-correlate their advantages and weak points for synergies. This paper depicted two extreme approaches (IIFA and CSFA) of modeling ICT structures to install a smart city, to apply it onto the simplified city model. The outcome was an understandable level of illustration for the cross-conforming aspects of the approaches, though they are in different levels of abstraction and properties, which expedites the potential of those approaches for a coordinated work towards the ultimate goals of smart city.

It has provided a conviction that we can develop a combined structure of ICT for smart city that takes advantages of both extreme approaches and resolves major drawbacks hindering the efficient and effective realization of them in a smart city. Designing the cross-conforming structure and functionality of city control center is another issue as it forms a core part of ICT deployment in smart city. A reasonable approach to identify high level structure and functional configuration of the center can be identified through the validation review process of the design. We believe our work to develop the city model and to identify the cross-conforming approaches of designing ICT for smart city will contribute the development of another noble approach to develop smart city.

Acknowledgments. This work was supported by the ICT R&D program of MSIP/IITP. [B0101-16-0429, Development of Open Screen Service Platform with Cooperative and Distributed Multiple Irregular Screens].

References

1. Mosannenzadeh, F., Vettorato, D.: Defining smart city - a conceptual framework based on keyword analysis. published in TeMA Journal of land use, mobility and environment – special issue on smart city, June 2014. print ISSN 1970-9889/online ISSN 1970-9870
2. Xixi, L., Hua, Q., Hong, Z., Yinghua, H.: The 5I model of smart city: a case of Shanghai, China. In: IEEE First International Conference on Big Data Computing Service and Applications, 978-1-4799-8128-1/15 (2015)
3. Anthopoulos, L., Fitsilis, P.: Exploring architectural and organizational features in smart cities. In: ICACT2014, February 2014. ISBN 978-89-968650-3-2

4. Ahn, J.Y., Lee, J.S., Kim, H.J., Hwang, D.J.: Smart city interoperability framework based on city infrastructure model and service prioritization. In: The 8th International Conference on Ubiquitous and Future Networks (ICUFN2016), July 2016
5. Khan, Z. Anjum, A., Kiani, S.L.: Cloud based big data analytics for smart future. In: 6th International Conference on Utility and Cloud Computing, IEEE/ACM (2013)
6. Bonino, D., et al.: ALMANAC: internet of things for smart cities. In: 3rd International Conference on Future Internet of Things and Cloud. IEEE (2015)
7. Clohessy, T., Acton, T., Morgan, L.: Smart city as a service (SCaaS) - a future roadmap for e-government smart city cloud computing initiatives. In: IEEE ComSoc, 978-1-4799-7881-6/14, IEEE (2014)
8. Lea, R., Blackstock, M.: CityHub: a cloud based IoT platform for smart cities. In: 6th International Conference on Cloud Computing Technology and Science. IEEE (2014)
9. Skouby, K.E., Lynggaard, P.: Smart home and smart city solutions enabled by 5G, IoT, AAI and CoT services. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I), 978-1-4799-6629-5/14. IEEE (2014)
10. Zanella, A., et al.: Internet of things for smart cities. IEEE Internet Things J. **1**(1), 22–32 (2014)
11. Vilajosana, I., et al.: Bootstrapping smart cities through a self-sustainable model based on big data flow. IEEE Commun. Mag., **51**(6), 128–134 (2013)
12. Peelen, J.: A Dutch perspective on smart cities. In: Global City Team Challenge (2016). http://www.slideshare.net/US-Ignite/presentations/2
13. IBM Smarter Cities Challenge. http://smartercitieschallenge.org/city_negeri_sembilan_malaysia.html
14. Ong, E.A.: Montpellier – replicable solution: smart technology for smart water. In: IBM, 7th World Water Forum MF4.4, April 2015
15. Sanchez, L., Elicegui, I., Cuesta, J., Munoz, L., Lanz, J.: Integration of utilities infrastructure in a future internet enabled smart city framework (2013). ISSN 1424-8220, Sensors, http://www.mdpi.com/journal/sensors
16. Monzon, A.: Smart cities concept and challenges bases for the assessment of smart city projects. In: International Conference on Smart Cities and ICT Systems (SMART GREENS), May 2015
17. Moon, S.-W., Lee, J., Lee, J., Yoon, K.: Interactive UHD digital signage system for multi-screen displays. Int. J. Inf. Electron. Eng. **5**(5), 366–369 (2015)

GUI-Based Korean Font Editing System Using Font Parameterization Technique

Minju Son, Gyeongjae Gwon, Geunho Jeong, and Jaeyoung Choi(✉)

School of Computer Science and Engineering,
Soongsil University, Seoul 06978, Korea
{sonmibz, gwon, choi}@ssu.ac.kr, metaopus@gmail.com

Abstract. When designing Korean fonts, about 2,500 widely used characters should be designed among 11,172 characters. When generating fonts, characters are generally described as 'outline'. On average, it takes more than 1 year to design single set of Korean font with an outline font editing systems. Also, it takes almost the same amount of time to change the style of an already generated font using general outline font editing systems. In this paper, we propose a Korean font editing system which uses font parameterization technique based on METAFONT. Korean characters are composed of basic units of strokes and radicals unlike Roman characters. They have combination rules of the basic units. Therefore, we extracted font parameters for changing font styles by considering these characteristic of Korean characters, and applied them to Korean fonts implemented with METAFONT. In addition, we developed GUI-based Korean font editing system for efficient user interaction.

Keywords: METAFONT · Programmable font · Korean · Font editor

1 Introduction

Fonts are used in many digital contents such as E-book, smartphone applications, TV advertisements etc., so users' attention to typography and calligraphy has increased. Several companies try to give unique impressions of their products to the users by designing and distributing their own specific fonts. Therefore, the usage of fonts has increased in the design and media industry beyond simple usage of fonts in printing.

When producing Roman fonts, about 256 characters should be designed. Whereas designing Korean fonts, around 2,500 widely used characters should be designed among the total 11,172 characters. When generating fonts, characters are generally described as 'outline'. It takes averagely more than 1 year to design one set of Korean font with an outline font editing system. Similarly, it takes almost the same amount of time to change the style of an already generated font using general outline editors. In order to solve these problems, many studies on programmable fonts had been performed since 1980s. METAFONT is a font design system for improving the quality of TeX typesetting in TeX document, and it is the representative programmable font [1]. METAFONT can reduce the cost of generating fonts by deriving various fonts by changing sizes or shape of pen or values that are required to draw the characters.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_49

However, as METAFONT is a programming language, it is very difficult for general font designers to design fonts directly by using it.

In order to solve this problem, we propose a GUI-based Korean font editing system by font parameterization technique, 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.

2 Related Works

METAFONT is a programming language to define fonts [1]. It uses 'handwriting' method to draw skeleton of characters, and fills the track of the skeleton with a pen to express fonts. In other words, METAFONT provides all the processes that we use to write characters on a plain paper with our hand, like selecting a pen, grabbing the pen, and drawing characters with desired directions in a human friendly way. Along with all these processes, METAFONT can also define the pen, curve, and etc. It can enhance the productivity of font design by code reusability.

Metaflop is a web editor which can generate Roman fonts based on METAFONT [2]. Metaflop provides 3 kinds of basic fonts (bespoke, adjuster and fetamont) based on 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 Korean font. Currently, there is no proper GUI-based font editor for Korean fonts using the METAFONT [3].

The programmable characteristics of METAFONT become more obvious when dealing with the Korean font in which characters are constituted through the combination of radicals than single characters such as Latin characters. Based on these characteristics of METAFONT, the structural font generating program has been implemented with METAFONT in our preceding study [4]. 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. This program can produce fonts automatically 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 parameter values. 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 the outline method.

However, since structural font generating program is implemented with METAFONT language, therefore, knowledge for METAFONT is required. 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 address this limitation presenting GUI-based Korean font editing system using METAFONT.

3 Implementation of Font Editing System

We extract font parameters from anatomies of Korean that determine font styles, and applied them to a skeleton of Korean fonts in our proposed editing system. Accordingly, users can easily and efficiently generate various Korean fonts by changing parameter values in GUI-based Korean font editing system.

3.1 Anatomy of Characters and Extraction of Parameters

There are many anatomies to determine styles of fonts, such as width and height of character, weight (thickness), serif and slant [5]. These features sometimes affect legibility of characters, and they are required to generate variety of font shape in design aspects [6]. For example, the serif fonts have high legibility by introducing visual continuity; therefore serif font is used in specific print applications such as newspaper. In this way, anatomies that determine styles of fonts are important. Alphabets are constituted with their own anatomies as shown in Fig. 1. However, Korean characters have different concept of anatomies as shown in Fig. 2 because they are composed by combining several radicals. Korean characters have 67 radicals, consisting of 1 to 9 strokes. Also, there are combining rules according to positions of radicals. Thus, the same radical might have different anatomies depending on their positions. Exploiting these key points in Korean characters, we classified parameters into two groups; parameters for defining overall font styles and parameters for affecting radicals only according to their relative positions.

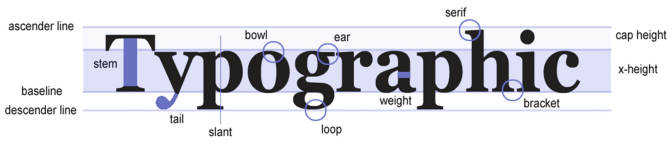


Fig. 1. Anatomies of alphabet

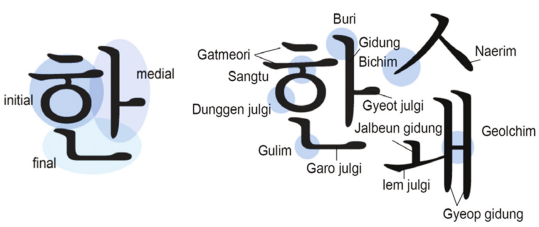


Fig. 2. Anatomies of Korean

We defined parameters that determine font style of Korean as shown in Fig. 3. Figure 3(a) shows some parameters such as pen, serif, slant and distance between radicals that are applied to all characters to determine font styles, whereas Fig. 3(b) show some parameters which are applied to specific radicals to define its width, height, and location depending on their relative positions.

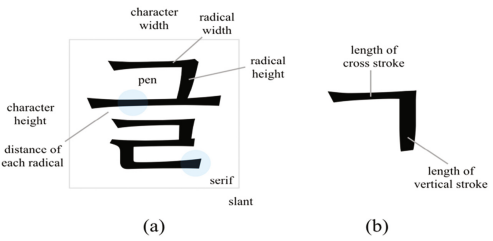


Fig. 3. (a) Parameters applied to all characters (b) Parameters of radical 'ㄱ'

There are four kind of the parameters that can applied to all characters. We need to define 'pen' parameters at first. METAFONT provides various shapes of pen such as circular, tetragonal, and triangular [1]. The skeleton of character is to be filled with a selected pen. Thus, a pen can represent a thickness of characters such as bold, light, etc. In Fig. 4, the style of font has been modified by changing width of a pen. Then, we need parameters to determine overall sizes and styles of fonts, such as 'width' and 'height' of characters. These parameters can change styles of existing fonts to be thick or thin. In addition, we require parameters to determine serif characteristics. Fonts with serifs are quite popular, for example, serif font increases legibility and san-serif font is more sophisticated. Lastly, we need slant parameters that change straight fonts to italic fonts.

We also defined parameters for applying specific radicals as shown in Fig. 3(b), such as width and height of each radical, and location and length of strokes. In Korean, each radical is arranged to its predetermined position according to composing rules of characters. We have extracted parameters which affected specific radicals only. If these parameter values of radicals are changed, the modification affects all characters which have the same radicals.

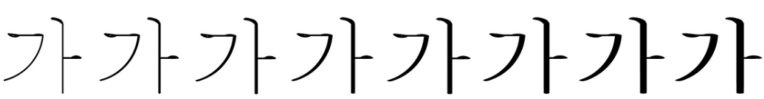


Fig. 4. Shapes of Korean character for which the parameters of pen are applied

3.2 Applying Parameters to METAFONT Program

In order to apply parameters to our font editing system, we used METAFONT as a programming language. We extended our preceding work of 'structural font generating program using METAFONT,' and applied the extracted parameters to METAFONT font for changing all character styles in the font. Most values used in drawing characters can be mapped with these parameters. We provide GUI for easy control of parameters. We modified the structural font generating program by defining two METAFONT files (their format is.mf). A 'global.mf' is for storing input parameter values, which are defined by controlling sliders or buttons, and it's also used for updating in real-time by applying the values to the prototype. This file is frequently referred to draw radicals in the font editing system. As mentioned in Sect. 3, the same radical might have different parameter values according to their locations. For example, '가' and '고' are using same radical of 'ㄱ', but their relative positions are different. A 'distribution.mf' is used for applying different parameter values to each radical according to their relative positions.

3.3 Korean Font Editing System with GUI

Our proposed font editing system is provided with GUI based on HTML5, and it can be used by font designers who are not familiar with the METAFONT programming language. The structure of font editing system is shown in Fig. 5 and font designers can easily modify font styles by changing the extracted parameter values. The 'Ming style' is provided as a prototype Korean font. Users can modify the parameters of a prototype font and can confirm modification of fonts through the web. As mentioned before, the 'global.mf' is updated with new parameter values done by font designers through GUI, and the file is used for the runtime application in METAFONT program.

In order to apply the fonts to the web, we have to convert the METAFONT font to the corresponding outline font such as Type 1, TrueType, etc. For this reason, we use 'mftrace' which converts the bitmap font i.e. the output of METAFONT, into the corresponding outline font [7]. The user interface is developed using HTML5 as shown in Fig. 6; therefore, the result of modified styles can be confirmed directly on the screen. Depending upon the prototype of Ming style' in Fig. 6(a), completely different styled fonts could be generated by changing the thickness of pen as shown in Fig. 6(b), and by changing the slant to the character as shown in Fig. 6(c). Figure 6(d) and (e) show, respectively, changing the size of serif and changing the width and height of character.

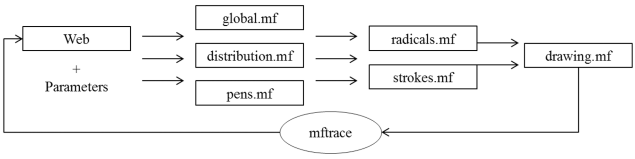
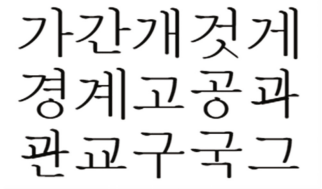
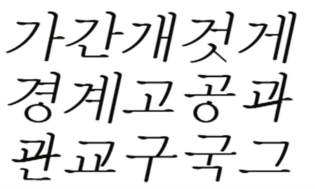


Fig. 5. Structure of GUI-based Korean font editing system

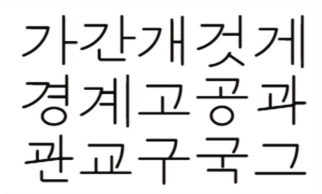
(a) Prototype of 'Ming style'



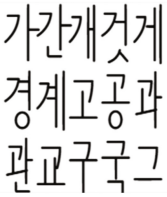
(b) Changing the width of pen



(c) Changing the slant



(d) Changing the serif



(e) Changing the width and height of character

Fig. 6. User Interface of Korean font editing system

4 Conclusions

In this paper, we proposed a GUI-based Korean font editing system, which is very convenient to use for font designers, who don't have any subject knowledge and programming skills. Korean font style can be changed with simple GUI controls, provided by Korean GUI-based font editing system. Lots of repeated efforts were required to change font styles with existing outline font editors. In the GUI-based Korean font editing system; the styles of whole characters can be changed simultaneously by extracting parameters from anatomies of characters, and applying them to whole characters. In this way, our Korean font editing system can reduce the complexity of designing fonts and increase usability of METAFONT.

This GUI-based Korean font editing system enable designers to design fonts using METAFONT, based on the anatomies of characters. As it is implemented with HTML5, it can be used on tablets and smartphones as well as on PCs. Our GUI-based Korean font editing system is currently using parameters to change font style. However, the work for extracting more parameters and applying them effectively is under progress. With this work, it is expected that more convenient and detailed font editing system will be provided to font designers in the near future.

References

1. Knuth, D.E.: METAFONT: The Program. Addison-Wesley Longman Publishing, Boston (1986)
2. Metaflop. http://www.metaflop.com (2012)
3. Crossland, D.: Why didn't mEtaFoNt catch on? In: Proceedings of the 2008 Annual Meeting TUGboat, vol. 29, No. 3 (2008)
4. Gwon, G., Son, M., Choi, J., Jeong, G.: Structured hangul font generator using METAFONT. In: Korea Information Science Society Conference, pp. 1134–1136 (2015)
5. Cao, Y.: Visual translation: a new way to design a chinese typeface based on an existing latin typeface. Faculty of the Louisiana State University and Agricultural and Mechanical College, Hunan Normal University (2012)
6. Boyarski, D., Neuwirth, C., Forlizzi, J., Regli, S.H.: A study of fonts designed for screen display. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 87–94. ACM Press/Addison-Wesley Publishing Co (1998)
7. Mftrace. http://lilypond.org/mftrace (1999)

Spatial Big Data Analysis System for Vehicle-Driving GPS Trajectory

Wonhee Cho[1] and Eunmi Choi[1,2(✉)]

[1] Graduate School of Business IT, Kookmin University, Seoul, Republic of Korea
danylight@gmail.com, emchoi@kookmin.ac.kr
[2] Department of Financial Information Security, Kookmin University, Seoul, Republic of Korea

Abstract. The data collection of vehicle-driving GPS trajectory becomes the basis of big data analysis and prediction for a variety of purposes, such as navigation and movement analysis. In order to properly analyze a large amount of GPS location information, it is necessary to determine the exact road map and location data by matching a digital map and space. We previously discovered the road information of the GPS coordinates using the commonly utilized map-matching technique. However, such a navigation map-matching technique requires a lot of supplementary corrections in order to rapidly and accurately navigate a large amount of data. In this study, we apply geohash indexing and long link vertex dividing preprocessing to spatial data for performance improvement of massive data map matching. Also speed filtering logic is applied together for qualified analysis. We established and implemented a distributed analysis environment for the better big data map-matching with HBase. Altogether we constructed a spatial analysis system using the MapReduce mechanism, which improved its performance. This paper shows that our analysis system provides the 44 times performance achievement compared to traditional mysql DB processing with mesh structure for 5,000,000 cases of GPS trajectory.

1 Introduction

The map-matching technique is essential in navigation. GPS coordinates and various spatial information-gathering and -analysis technologies are essentially used in autonomous intelligent transportation systems [1]. In Korea, a GPS-based built-in digital tachograph (DTG) is mandatorily equipped in every commercial vehicle, and it is used for traffic safety [2]. We can utilize eco-routing, logistics flow analysis, and so on, if we can analyze corrective location information on a large volumes of GPS data generated from DTG. Thus, spatial big data analysis technologies are essential for GPS analysis [10].

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-H0502-13-1071).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_50

On the other hand, the map-matching technique for matching the digital map is important to extract the accurate geographic information for the GPS coordinates. Most conventional map-matching techniques are for real-time navigation on the road, and they are complicated techniques for real-time correction, such as GPS error correction. However, they have issues of slow matching speed and implementing systems.

Thus, in this study, we built a Hadoop ecosystem environment for map-matching of large amounts of GPS coordinates information, and we investigated the spatial big data map-matching technique using Hadoop-based HBase as a representative distributed big data NoSQL DB. We implemented MapReduce spatially distributed big data processing features by using MapReduce for distributed processing in Java. We used a geohash index for each vertex in place of a conventional mesh structure index system, and divided the vertex of the link from a long straight section like a freeway into 90-m units in order to improve searching accuracy and speed. We also added a speed-filtering logic in order to improve matching accuracy.

2 Related Research Work

- **Map-matching Algorithm**

Map-matching is a process of mapping GPS coordinates on a road network in order to find a correct link for coordinates of the digital map. In particular, it is an essential technique in navigation to receive GPS coordinates and guide directions [1]. The main map-matching algorithms are the geometric algorithm and topological algorithm [3]. The geometric algorithm provides matching features using the geometric information. This involves methods of point to point matching, point-to-curve matching, and curve-to-curve matching. The topological algorithm is a matching method regarding topology information on connectivity and accessibility of each link of the road network. This method removes the links with non-connectivity and non-accessibility in the navigation candidates. In addition to the map-matching method, there are the segment-based algorithm [4], probabilistic algorithm [5], fuzzy logic-based algorithm, and genetic algorithm. Recently, many studies have worked on a big data map-matching method using Hadoop MapReduce for matching a large volume of GPS trajectory [6].

- **Spatial Big Data Analysis**

The spatial big data technology has been developing continually for processing large amounts of data in GIS domain. As GPS built-in smart phones have been pervasively used, location and spatial data have been treated as a part of big data. Thus, the importance of spatial big data technology has increased. However, the spatial big data has limitations in processing by a conventional Relational DBMS. Therefore, studies that utilize a distributed NoSQL DB have gained attention [5].

A study of the 'Spatial Hadoop' in University of Minnesota sought to provide a space operator, a spatial index based on Hadoop [7]. SpatialHadoop is a MapReduce framework supporting spatial data in Hadoop. Its features support a spatial index

consisting of Grid File, R-tree, R+-tree. It makes and saves the distributed spatial index. It supports a distributed processing by MapReduce programming. It also provides operators including Range Query, KNN, Spatial Join, and Shortest Path [7]. Emory University and Ohio State University jointly-developed Hadoop-GIS, which is a spatial data analysis system with scalability and performance supporting large-scale spatial query of Hadoop. It is designed to perform fast spatial analysis using MapReduce parallel processing techniques and the R*-tree spatial index. It supports multiple types of spatial query and SQL-like language HiveQL of Apache Hive [8].

- **Geohash Spatial Index**

Geohash was developed by Gustavo Niemeyer. It is a technique for improving spatial information search speed in the spatial big data by converting the coordinates of the two values of longitude and latitude as a single number. GeoHash is represented by a Base32 encoded string. Base32 is represented by 32 characters of B-Z, and 0–9. 6bits are represented by 1 (8bit) byte char. Advantages of GeoHash are very easy calculation and giving a key role to the prefix to find the nearest neighbor. Its disadvantages are problematic in terms of accuracy of the prefix and boundary value [9].

3 Spatial Big Data Analysis System Design

In this study, we used the point-to-curve and topology algorithm, which are most often used for map-matching. For improving big data matching speed and accuracy, we use (1) Geohash as spatial index (2) Vertex pre-processing division for long link, and (3) Additional speed filtering logic.

- **Using Geohash**

In general, Geohash retrieves a large amount of link information by using the index for the vertical and horizontal minimum 1 km or more Mesh unit of map-matching logic, and matching the link information. It retrieves the link information by gathering a large amount of information by matching. In case of navigation route guidance, it isn't matter. But slow speed is estimated for map-matching tens of gigabytes of big data trajectory. Geohash index guides digital map vertex as an index of less than 1 m. In this study, searching speed improvement is acquired by configuring spatial information index in Geohash (Fig. 1).

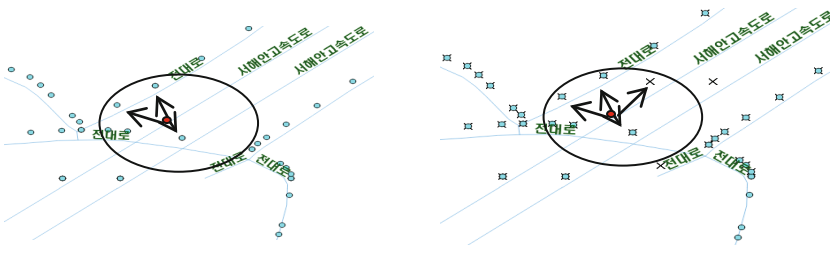


Fig. 1. Vertex division on the point of crossing West Highway and local roads

- **Long link vertex dividing pre-processing**

According to the road links of a conventional digital map, vertex uses metrics from 1 km up to 10 km in straight metrics. Though using Geohash as the map-matching logic matching vertex position is standard, the mesh unit should be widened for searching an any-km range in order to match the long straight metrics of road information, such as in a highway.

In this study, a method is used to divide the vertex of the link from a long straight section into 90-m units for solving the above matter. Thus, it acquires an advantage of improvement in searching speed as the search range decreases. Algorithm 1 shows the pseudo code of dividing pre-process regarding long-link roads.

Algorithm 1: Handling long-link dividing pre-process

```
1:  for each link n
2:    get vertex list by link;
3:    for each vertex v
4:      write vertex[v] // write the starting vertex;
5:      len = distance between vertex[v] and vertex[v+1];
6:      bearing = bearing between vertex[v] and vertex[v+1];
7:      baseX = vertex(x);  baseY = vertex(y);
8:      while (len >90)
                    // Generate 90m-interval vertices
                    // if the interval is larger than 90m.
9:        nvertex = getXYbyDirLen(baseX,baseY,bearing,90);
                   // Calculate 90m-intercal vertices
                    // according to the angle rates.
10:       write nvertex;
11:       len = len – 90;
12:     end while
13:   end for
14: end for
```

- **Speed filtering logic**

Map-matching in the area where several roads are interconnected in a complex pattern generates a lot of matching errors if the links are within the GPS error range. This study corrected the road where the difference is less than can be handled by the running speed of the GPS coordinates and the digital map road attribute information, utilizing road speed limit information of the digital map on these problems.

- **MapReduce design for map matching**

MapReduce programming is required to distributedly perform a large amount of GPS trajectory spatial big data using map-matching. Figure 2 is a MapReduce operation flow for map-matching. When designing MapReduce operation for map matching, the key is a combination of CarNum, Date, Hour, and GPS coordinates.

The Map function is to extract the link edge list in geohash of GPS coordinates, and select the possible three links as the candidates, by finding the lightest weighted values. The weighted values are computed by distance and speed weights. The result of Map operation contains the result values of Time, list of candidate links, the weighted values. The Reduce function explores the incoming Map results again with time series

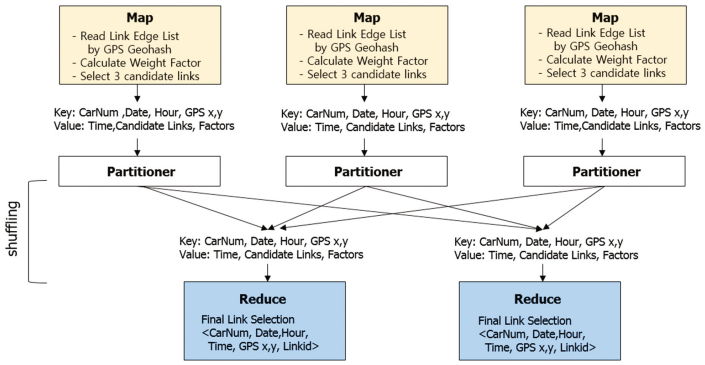


Fig. 2. MapReduce operation flow for map-matching

ordering [11], and gives more values to the links which were formerly matched from historical information. Through the Reduce operation, the link values are revised with the time-series view point. Also, it removes links with fewer opportunities to access by using the topological matching logic. Finally, the results of Reduce operation stores the selected links with link ids, CarNum, Date, Time, GPS x, y coordinates into the HDFS.

In this study, spatial big data analysis processing was implemented using HDFS, a HBase of Hadoop ecosystem for MapReduce functions.

4 Map Matching Result

Experimental environment included five servers as follows: Server specification is ESS-R4200, CPU: Intel Xeon E5-2407(2.2 GHz) 4-core, Memory: 4GBRDIMM*4, HDD 1 TB NLSAS. One Master server remains namenode, and the other four servers are slaves performing datanode functions. Jobtracker is configured at node1, and tasktrackers are configured at node2–node5. HMaster and Zookeeper are configured at node1 for HBase, and Region Server is configured at node2–node5.

Figure 3 is an example of results from map-matching. The map-matching is performed around Gyo wol-ri, Cheong yang-eup, Chung cheong nam-do. Red dots are

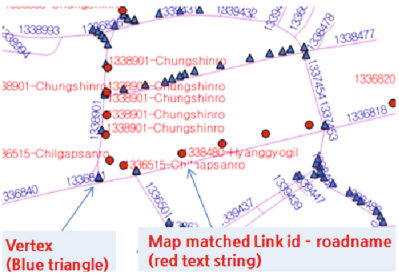


Fig. 3. Map-matching result example

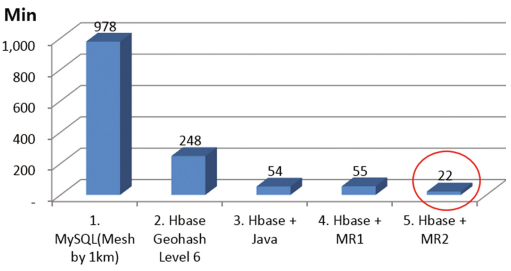


Fig. 4. Map-matching processing time

Table 1. Map-matching processing time

Method applied	Processing time[a]	Implementing and processing case
1. MySQL (Mesh by 1 km)	978	MySQL+One server process
2. HBase geohash level 6	248	HBase+One server process
3. HBase+Java	54	HBase+One server process
4. HBase+MR1	55	HBase+One server process and 4 distributed servers
5. HBase+MR2	22	HBase+One server process and 4 distributed servers (vertex preprocessing+remove duplication)

[a]Elapsed time (in minutes)

GPS coordinates, red numbers are matched link numbers and street names. Light blue numbers are link lengths. In this study, we performed map-matching test with 500,000 GPS coordinates. Table 1 shows the elapse time of map-matching processing with a number of experimental cases (Fig. 4).

5 Conclusion

In this study, we researched the spatial big data system with MapReduce using HBase, a Hadoop-based NoSQL DB, for map-matching of a large collection of GPS trajectory data. We used Geohash for a mesh structure index system in order to improve matching speed, and applied a method to divide the vertex of the link from a long straight section into 90-m units. We also used a speed filtering method for corrected matching accuracy. HBase was equipped with a digital map and the MapReduce method was designed and implemented in Java for map-matching. Applying 500,000 map-matching tests of GPS coordinates in the mesh standard of MySQL 1 km unit-size resulted in a 44 times speed improvement in the map-matching performance. Further research is to improve the analysis performance and map-matching quality.

References

1. http://en.wikipedia.org/wiki/Map_matching
2. Cho, W., Choi, E.: A GPS trajectory map-matching mechanism with DTG big data on the HBase system In: International Conference on Big Data Applications and Services (2015)
3. Velaga, N.R., Quddus, M.A., Bristow, A.L.: Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems. Transp. Res. Part C: Emerg. Technol. **17**, 672–683 (2009)
4. Chawathe, S.S.: Segment-based map matching. In: Intelligent Vehicles Symposium, 2007 IEEE, pp. 1190–1197 (2007)

5. Quddus, M.A., Ochieng, W.Y., Noland, R.B.: Integrity of map-matching algorithms. Transp. Res. Part C: Emerg. Technol. **14**, 283–302 (2006)
6. Tiwari, V.S., Arya A., Chaturvedi, S.: Framework for horizontal scaling of map matching: using map-reduce. In: 2014 International Conference on Information Technology (ICIT), pp. 30–34 (2014)
7. Eldawy, A., Mokbel, M.F.: SpatialHadoop: a MapReduce framework for spatial data. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE 2015). IEEE (2015)
8. Aji, A., Sun, X., Vo, H., Liu, Q., Lee, R., Zhang, X., et al.: Demonstration of hadoop-gis: a spatial data warehousing system over mapreduce. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 528–531 (2013)
9. Geohas Wikipedia, 1 December 2015. http://en.wikipedia.org/wiki/Geohash
10. Cho, W., Choi, E.: Rural traffic map coverage extension using DTG big data processing. J. Inf. Technol. Archit. **12**, 51–57 (2015)
11. Lim, Y., Choi, E.: Time series bigdata processing mechanism of digital tachograph on Hadoop. In: International Conference on Advanced Intelligent Mobile Computing (AIM2015) of World IT Congress 2015, Jeju, Korea, February 2015

i-SHSS: An IoT Based Smart Home Security System

Saurabh Singh, Pradip Kumar Sharma, Seo Yeon Moon, and Jong Hyuk Park(✉)

Department of Computer Science and Engineering,
Seoul National University of Science and Technology (SeoulTech),
Seoul 139-743, Republic of Korea
{singhl989, pradip,
moon.sy062l, jhparkl}@seoultech.ac.kr

Abstract. Smart homes are increasing their popularity as the most promising application of Internet of Things (IoT). Security has becoming an important issue in smart home. There are many security threats and challenges present in smart home. To overcome these security issues, we proposed security system architecture for home automation system. The architecture is divided into three parts: management platform, secure home gateway, and home controller. The proposed system fulfills the security goals such as user and device authentication, protecting communication, and different attacks.

Keywords: IoT · Threats · Gateway · Controller · Authentication · Attacks

1 Introduction

The increasing demand for independent and comfort lifestyle of human beings has motivated the development of Smarthome. It provides smart life atmosphere based on human requirement that enables the convenience for the people. Smart home use and control the home appliances, sensors, actuators, and many more, that helps to assist the unsafe activities and reminding their forgotten their activity. The smart home is a part of IoT service because of rapid development of mobile network, sensing technology, proliferation of smartphone and interest of personal security and safety. However, many incidents happened because they only focus on the services instead of concerning security part too [1].

Security has becoming an important issue in smart home. The works goes on to explain and justify why home automation is more attractive targets to the attacker. Security threats inside the smart home are usually trying to compromise more security goals and make more control on home device [2]. The attackers compromise the security goals such as confidentiality, availability, integrity, authenticity and perform passive and active attacks [3]. To overcome the security issues and challenges in smart home we propose a framework that ensure the security and prevent attacks by attackers. The main aim of our paper is to make secure, reliable and efficient communication between human beings and devices in smart home.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_51

In this chapter, we will discuss the smart home technology, security goals, security threats and challenges. We also discuss some existing research for security development in smart home.

A smart home includes a huge number of wireless connected devices having a place with various application regions. These gadgets are described by heterogeneous equipment and programming assets, and they support various communication protocol technology such as Z-Wave, ZigBee Wi-Fi, Bluetooth etc. By synchronizing, interacting, and cooperating among each other's, these devices form a distributed heterogeneous network and communicate amongst them [4].

To fulfill the aim of this paper, we consider the four commonly adopted goals that described below as the most important for smart home security. These goals are:

Confidentiality: The confirmation that information will be unveiled just to approved people or frameworks.

Integrity: The guarantee that the exactness and consistency of information will be kept up. No unauthorized modifications, obliteration or loss of information will go undetected.

Availability: The assurance that any system resource (information/transfer speed/equipment) will dependably be accessible for any approved element. These resources are protected against any incident that debilitates their accessibility.

Authenticity: The validation that conveying parties are who they guarantee they are, and that messages evidently sent by them are for sure sent by them. The authenticity may be user authentication or device authentication in smart home.

Deploying the smart home it faces many security challenges. Currently in smart home the possible threads such as message modification, DoS attack, eavesdropping, replay attack, repudiation, physical attack are compromised security goal. Additionally, attacks on smart home might target the weakest and vulnerable element to capture identifications of the HAN and elaborate more powerful attacks. For example, scientists have recuperated the Wi-Fi private key from an unsecured home device and could communicate with the system to take control of the smart home [5].

As associated home device depend increasingly on remote get to and cloud innovations, ensuring client information in the cloud and on the device turns out to be progressively essential.

While associated home items guarantee customers comfort, these items may have potential security conditions that welcome programmers to assault. Obviously, most associated home gadget organizations remember this. The genuine test is the way to offer security in a way that does not influence the general client experience furthermore in a cost-effective manner [6].

2 i-SHSS

2.1 Architecture

In this chapter, we discusses the propose system architecture for securing the smart home. The proposed system architecture is divided into three parts: management platform, secure home gateway, and home controller.

Management platform: It manages the home based applications that are trying to connect Home Area Network (HAN).

Secure home gateway: It provides the secure protection to the managed home network. It secures the two way communication between application and HAN. The gateway also enables smart home for secure monitoring and to report alarm.

Home controller: The smart home managed and control device. All managed devices are connected to the controller through secure and less energy power consumption communications (Fig. 1).

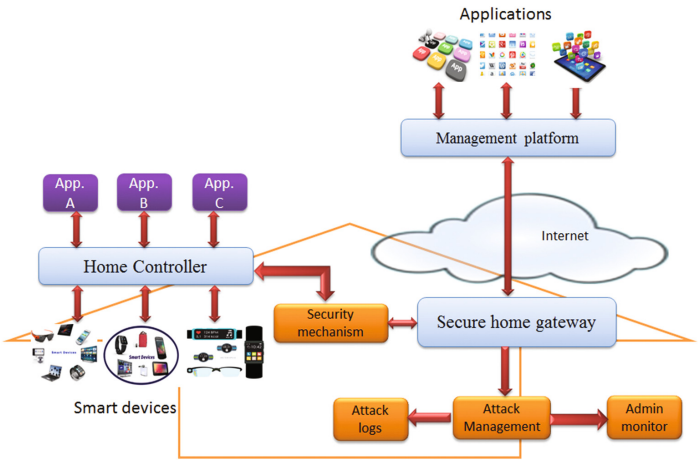


Fig. 1. Proposed secure system architecture in smart home

2.2 Discussion and Analysis

It is very challenging to provide appropriate access permission for authorized users in the open communication environment. The requests from management platform have been identified by the secure home gateway. The gateway checked the request whether it is authorized or non-authorized. After authentication process the two way communication has been established. Many attacks can be admin protected by home gateway security system. Using attack management, it creates logs of attack and report to the admin.

The security mechanism resides the several techniques such as cryptographic encryption, IDS, AMI, hash chain, digital watermarking. These techniques ensure confidentiality, privacy, authenticity, integrity, and availability in the smart home communication system.

3 Conclusion

Smart home as one of the IoT service is growing more interested. The paper analyzed the technology used in smart home and security goals. The paper also discussed the possible threats and challenges in smart home. The proposed security architecture has various security functions that can protect the HAN from attacks. This architecture fulfill the security goals through various security techniques such as encryption, IDS, AMI, hash chain, digital watermarking. With this study, it could provide various Smart home applications safely, securely, and efficiently.

Acknowledge. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B4011069).

References

1. Yang, C., Yuan, B., Tian, Y., Feng, Z., Mao, W.: A smart home architecture based on resource name service. In: IEEE 17th International Conference on Computational Science and Engineering (CSE), pp. 1915–1920 (2014)
2. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. Fut. Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
3. Sang-Hyun, L., Lee, J.G., Kyung-Il, M.: Smart home security system using multiple ANFIS. Int. J. Smart Home **7**(3), 121–132 (2013)
4. Sripan, M., Lin, X., Petchlorlean, P., Ketcham, M.: Research and thinking of smart home technology. In: International Conference on Systems and Electronic Engineering-(ICSEE) (2012)
5. Komninos, N., Philippou, E., Pitsillides, A.: Survey in smart grid and smart home security: issues, challenges and countermeasures. IEEE Commun. Surv. Tutorials **16**(4), 1933–1954 (2013)
6. Jacobsson, A., Boldt, M., Carlsson, B.: A risk analysis of a smart home automation system. Fut. Gener. Comput. Syst. **56**, 719–733 (2016)

The VM Weighted Filter Scheduling Algorithm for OpenStack Cloud

Mohan Krishna Varma Nandimandalam[1] and Eunmi Choi[1,2(✉)]

[1] Graduate School of Business IT, Kookmin University, Seoul, Republic of Korea
{nmohankv, emchoi}@kookmin.ac.kr
[2] Department of Financial Information Security, Kookmin University, Seoul, Republic of Korea

Abstract. OpenStack cloud is an open source private cloud environment and uses the filter scheduling algorithm for scheduling of virtual machines to hosts. Filter scheduler selects the host with highest weight and assigns the virtual machine. In the process of weighing filter scheduler does not consider virtual machine weight. We propose a new scheduling algorithm for OpenStack private cloud environment by considering the virtual machine weight in the weighing process of hosts. In this paper, Round Robin, Greedy, and Filter scheduling algorithms are compared with the VM weighted filter scheduling algorithm.

Keywords: Cloud computing · Virtualization · Scheduling · OpenStack

1 Introduction

Cloud computing [1] is a most extensively used system, where computers are networked to deliver computing, storage, and application services using virtualization technology. Cloud computing satisfies five necessary features, such as on demand service, access network, resource pooling, elasticity and measured services. To achieve these, cloud computing provide three kinds of basic service models, such as Software as a Service (SaaS), Platform as a Service (PaaS) [2] and Infrastructure as a Service (IaaS) [3]. While providing many SaaS applications via webs and apps such as customer relationship management applications, the PaaS delivers application platform for developers by providing development and deployments of libraries. Cloud Foundry, google app engine, amazon elastic beanstalk and Microsoft azure environments can be used for PaaS. IaaS can be used to build private infrastructure or deliver infrastructure for public. Usage of IaaS can reduce the installation budget. IaaS can provide resources such as computation, storage and communication with the help of virtualization. Eucalyptus [4], and OpenStack [5] environments can be used to provide IaaS. In this paper, we study about Virtual Machine (VM) scheduling process of OpenStack cloud. We propose and compare the VM weighted filter scheduling algorithm which applied the VM weights to the VM scheduling algorithms [6] in terms of response time and elapsed time.

In this paper, Sect. 2 explains virtual machine scheduling algorithm of OpenStack cloud, the VM weighted filter scheduling algorithm is given in Sect. 3, Sect. 4 shows

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_52

the performance evaluation of the VM weighted filter scheduling algorithm, and finally Sect. 5 concludes the paper.

2 Virtual Machine Scheduling in OpenStack Cloud

The OpenStack Cloud consists of different components, such as Dashboard, Compute, Networking, Object storage, Block storage, Identity service, Telemetry, Orchestration, Database service and Image Services. Dashboard service is named as Horizon in OpenStack. Horizon provides a web based portal to interact with other OpenStack services, such as starting instances, assigning IP addresses and stopping instances. Compute service is named as Nova and it is the core part of the OpenStack cloud to manage the instances of virtual machines and networking. Networking service is named as Neutron and it enables network connectivity for other OpenStack services such as Nova. Object Storage service is named as Swift and it stores and retrieves unstructured data objects from scalable system. Block storage service is named as Cinder and it provides persistent block storage to running instances. Identity service is name as Keystone, it provides authentication and authorization service for other OpenStack services. Telemetry service is name as Ceilometer, it monitors and meters the OpenStack cloud for billing, scalability, and statistical purposed. Orchestration service is named as Heat and it orchestrates multiple composite cloud applications. Database service is name as Trove and it provides scalable and reliable cloud database as service functionality for bother relational and non-relational database engines. Image Service is named Glance, it stores and retrieves virtual machine disk images. Nova makes use of Glance during instance provisioning. OpenStack can be deployed and runs on different Linux environments. It supports KVM, XEN, and Hyper-V hypervisors. Nova services can be deployed on same host with other OpenStack components or can be installed on different compute node hosts. OpenStack network consists of public and private networks. IP addresses from the public network are associated with instances of virtual machines to be accessed from the Internet, whereas private network is used for internal web service communication. OpenStack components uses Rabbit Message Queue Protocol for internal communication. Nova API processes virtual machine requests with the help of Queue. Virtual machine scheduling done via nova-scheduler. Nova-scheduler maps nova-API calls to the suitable OpenStack components. Nova scheduler is responsible for scheduling of virtual machines in OpenStack. In this section virtual machine scheduling techniques in OpenStack are explained in detail.

Nova scheduler uses filter scheduling algorithm by default. Filter scheduling algorithm uses filtering and weighing process to schedule the virtual machines in nova compute node hosts. OpenStack contains all hosts filter which passes all the available hosts, image properties filter which filters the hosts based on properties defined in the image of the instance, availability zone filter which filters the hosts based on availability zone, compute filter which passes all hosts that are enabled and operational, core filter which filters based on CPU core utilization, ram filter which filters hosts by their RAM, disk filter which filters host by their disk space, and so on. Among the filters compute filter, ram filter, and availability zone filters are default filters. Other filters can

be used along with or instead of default filters. Filtering of compute node hosts will be done based on the filter parameters. Based on the filtered hosts weighing process will be started. Weighing process can compute the weight based on the compute node host metrics. To compute the weights normalization process is applied to each host's metric. Hosts are weighted and sorted with the least weighing compute node host lost and high weighing compute node host first. Highest weighing compute node host will be select to schedule the virtual machine. Filter scheduling algorithm is written in python. Table 1 shows the each stages in filter scheduling, corresponding classes, and functionality of each class.

Table 1. Filter scheduling algorithm classes

Stages	Class	Functionality
Filtering	DiskFilter	Compute node hosts with sufficient disk space are filtered
	RamFilter	Filters the compute node hosts based on the available RAM
Weighing	DiskWeigher	Compute node hosts are weighted and sorted based on the free disk space
	RamWeigher	Computes the weight based on the available RAM on the compute node host
Filter scheduling	HostManager	Passes the compute node hosts which are filtered and weighed
	FilterScheduler	Assigns the virtual machines to appropriate compute node hosts

Filter scheduler takes the compute node hosts that are remained after the filters been applied. Filter Scheduler class used methods of Host Manager Class for filtering and weighing. Host Manager Class uses other required methods in different classes to fulfill the requirement of Filter Scheduler class.

3 The VM Weighted Filter Scheduling Algorithm

We modified the filter scheduling algorithm by considering the virtual machine requirements in the normalization of weighing phase. The VM weighted filter scheduling algorithm's normalization process is given below.

$$N = Hm{-}min\,(Hm\,list) \tag{1}$$

In the above equation N is the temporary variable and Hm is the host metric. In the below equation Fm is the free metric like available ram or disk in each host and Rv is the required virtual machine metric.

$$Hm = Fm - Rv \tag{2}$$

$$Nw = N/Rh \tag{3}$$

In the above equation Nw is the normalized weight and Rh is the range of hosts.

$$Rh = max\,(Hm\,list) - min\,(Hm\,list) \tag{4}$$

After changing the normalization, the VM weighted filter scheduling algorithm is applied to the filter scheduler. The algorithm for virtual machine scheduling in OpenStack cloud is as follows.

Algorithm 1: The VM weighted filter scheduling algorithm

```
1:  Request hosts for VMs
2:  For each VM from the list
3:      Get all host states
4:      Choose the filters to be applied
5:      Filter the hosts based on chosen filters
6:      Select the weighing metrics
7:      Find the weights of the filtered hosts based on metrics
8:      Normalize the weights by considering VM weights
9:      Sort the weighted hosts in descending order
10:     Select the first host
11: end for
```

In VM weighted filter scheduling algorithm at step 8 we consider the virtual machine weight for normalization.

4 Performance Evaluation of the VM Weighted Filter Scheduler

We used four hosts to evaluate the performance of the VM weighted filter scheduler. In the weighing process we consider the Ram and disk parameters. Table 2 shows the host's Ram and disk capacities.

Table 2. Configuration of hosts

Host	Component	Capacity
Host 1	Disk	424 GB
	Ram	3387 MB
Host 2	Disk	424 GB
	Ram	3394 MB
Host 3	Disk	424 GB
	Ram	3387 MB
Host 4	Disk	267 GB
	Ram	3312 MB

We compared the response time and elapsed time for greedy, round robin, filter scheduler and the VM weighted filter scheduling algorithm using OpenStack cloud. Figure 1 can give a clear idea of comparing response times of VM scheduling.

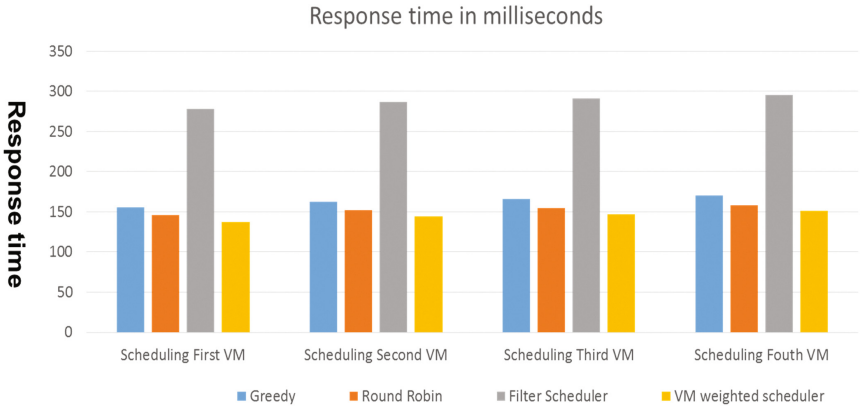


Fig. 1. Response time comparison of VM scheduling

Figure 1 shows the response time while scheduling each virtual machine. In this figure a sample scheduling of four virtual machines is given. Out of all the four scheduler algorithms namely Greedy, Round Robin, Filter scheduler and VM weighted Filter scheduler, the best response time is given while scheduling a virtual machine using VM weighted Filter scheduler. Filter scheduler takes more time to schedule a virtual machine instance. Greedy and round robin scheduler algorithms take almost same time to schedule virtual machines even the response time is more than the time taken by VM weighted Filter scheduler.

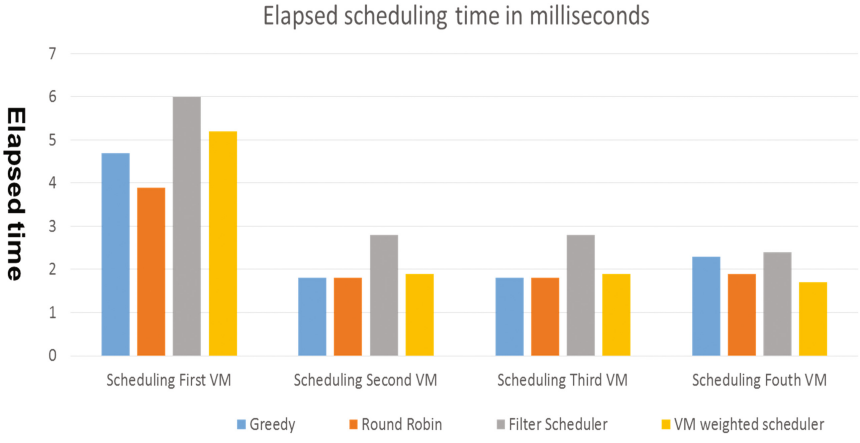


Fig. 2. Elapsed time comparison of VM scheduling

An implementation to calculate the time elapsed during the scheduling of virtual machine instances is done in OpenStack cloud environment. Figure 2 gives a clear idea of which algorithm takes less time to schedule virtual machine instance, so that the performance of an algorithm could be efficient enough to perform tasks.

Through the Fig. 2 we can infer that for scheduling the first virtual machine almost all the algorithms took same time except filter scheduler taking more time than all. The time elapsed to schedule the virtual machines is gradually decreasing for the scheduling of second third and fourth virtual machine using VM weighted Filter scheduler. The elapsed time taken by Greedy First Fit, Round Robin and Filter scheduler vary accordingly. From the above figure we can conclude that the VM weighted Filter scheduler takes less time to schedule the VMs.

5 Conclusions

This paper study about the virtual machine scheduling algorithm of OpenStack private cloud environment. Brief introduction is given about the OpenStack cloud components. Detailed explanation of OpenStack Nova's filter scheduling algorithm is give in this paper. Filter scheduling algorithm is memory aware, because it can filter the host based on the ram filter and weights can be calculated based on the memory. The VM weighted filter scheduling is proposed and compared with existing algorithms like greedy, round robin and OpenStack's filter scheduling algorithm. Based on the results we concluded that the VM weighted filter scheduling algorithm gives better response time and has less elapsed time.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant Number: 2011-0011507).

References

1. Rimal, B.P., Choi, E., Lumb, I.: A taxonomy and survey of cloud computing systems. In: International Joint Conference on INC, IMS and IDC, vol. 5, pp. 44–51 (2009)
2. Hossny, E., Khattab, S., Omara, F., Hassan, H.: A case study for deploying applications on heterogeneous PaaS platforms. In: International Conference on Cloud Computing and Big Data, CloudCom-Asia, pp. 246–253 (2013)
3. Varma N.M.K., Choi, E.: A scalable grid infrastructure in the Eucalyptus cloud environment. In: The Second International Conference on Computers, Networks, Systems, and Industrial Applications, pp. 286–289 (2012)
4. Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., Zagorodnov, D.: The Eucalyptus open-source cloud-computing system. In: IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid), vol. 9, pp. 124–131 (2009)

5. Teixeira, J.: Developing a cloud computing platform for big data: the openstack nova case. In: International Conference on Big Data, pp. 67–69 (2014)
6. Varma, M.K., Choi, E.: Study and comparison of virtual machine scheduling algorithms in open source clouds. In: The 11th International Conference on Future Information Technology (FutureTech2016), 20–22 April 2016, Beijing, China (2016)

A Hierarchical Motion Estimation Based on the Properties of Motion Vectors for Low Complexity in Video Coding

Hyo-Sun Yoon[1] and Mi-Young Kim[2(✉)]

[1] Department of Computer Science, Chonnam National University, Gwangju, South Korea
estheryoon@hotmail.com
[2] Department of Health and Medical, Jeonnam Provincial College, Damyang, South Korea
kimmee@dorip.ac.kr

Abstract. To transmit and to store digital video sequences, the compression is vital. Motion Estimation (ME) is generally used to reduce redundant data in video sequences. ME which limits the performance of image quality, bitrates and encoding time require much complexity. To reduce the huge computational complexity, a hierarchical motion estimation method for multi-view video coding is proposed. The proposed method exploits the properties of motion vectors. The characteristic of the distribution of motion vectors is used to place the search points in the search area and to choose a search pattern for the current block. Experiment results show that the complexity reduction of the proposed method over PBS and TZ can be up to 98% and about 45–76% respectively while maintaining image quality and bitrates.

Keywords: Motion estimation · Motion vector · Multi-view video coding

1 Introduction

Recently, there has been an exponential growth in video services and application and demands in multimedia services and systems grow quickly. Among multimedia data, digital video sequences are transmitted over various networks and are stored in many storage devices. To transmit and store video sequence, these video sequences must be compressed. The international video compression standards exploit motion estimation (ME) technique to remove redundant data in video sequences. ME technique which limits the performance of image quality, generated bitrates and the coding speed of the encoder plays an important role in digital video compression.

Recently, there has been a growing interest in 3D TV and free viewpoint video systems. These systems use multi-view video sequences which are obtained by capturing one three-dimensional scene with many cameras at different positions. For multi-view video sequences, the amount of initial data information is very huge, but a large amount of the data is redundant. To store or transmit these data, efficient compression technique for MVV (Multi-View Video) is needed. Multi-view video coding

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_53

(MVC) exploits motion estimation technique to remove temporal redundancy and inter-view redundancy in MVV [1, 2]. ME techniques have become one of the most important issues and have attracted much attention. Many motion estimation methods have been proposed. FS (Full Search) and PBS (Pel Block Search) which check every point in search area of the block in reference frame to find the best matched point require huge complexity. Therefore, many fast motion estimation methods have been proposed to reduce the computational complexity while maintaining the image quality at the same time. Diamond Search (DS) [3, 4], Hexagon Search (HS) [5], Four Step Search (FSS) [6], Three step search (TSS) [7], New Three Step Search (NTSS) [8] and Predictive Motion Vector Field Adaptive Search Technique (PMVFAST) [9] are well known fast motion estimation methods. These methods are generally used in small size sequences. And also these are effective in the small motion sequences. UMHexagonS [10] in H.264 which practices the early termination concept is a kind of hierarchical motion search strategy. This method achieves high compression rate at the expense of the increases in computational complexity. The main reason of the increase of complexity comes out of complex motion estimation method which uses 7 modes for the estimation block and multiple reference frames. Experiment results show that the ME covers 60% (for 1 reference frame)–80% (for 5 reference frames) of the total coding complexity [10]. This method shows good performance in large and complex motion sequences. However, these search methods [3–10] are generally used for single view video sequences.

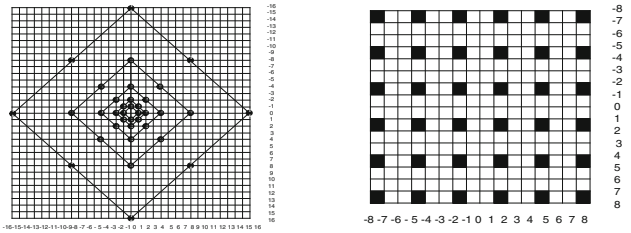
For multi view video sequence, PBS and TZ search algorithm [11] are used. Like FS, PBS requires much complexity. To this complexity, TZ search method performs raster search after initial search. Both raster search and initial search place search points that can cover the overall search area. It implies no need to perform initial search when uiBestDistance is larger than iRaster. It becomes the cause of using the unnecessary computational complexity. To reduce this computational complexity and to maintain the image quality, we proposed a hierarchical motion estimation scheme for multi-view video coding. The proposed search method exploits the properties of motion vectors to decide a search pattern and to place the search points in each search pattern.

This paper is organized as follows. Section 2 PBS and TZ search method in JMVC. The proposed motion estimation scheme is described in Sect. 3. Section 4, reports simulation results and conclusions are given in Sect. 5.

2 PBS and TZ

PBC and TZ search are the motion estimation algorithms for multi view video in JMVC. Pel Block Search checks all points in search area of reference frames to find the optimal motion vector. However, it requires huge computational complexity. To reduce the complexity of PBS, TZ search is used as a fast motion estimation algorithm. TZ search method uses the search patterns in Fig. 1 and is summarized as follows [12].

Step 1: Motion Vector Prediction – motion vectors from median predictor, left predictor, up predictor and upper right predictor and (0, 0) are used to decide a starting point for the current block.



(a) Initial search pattern (b) Raster search pattern

Fig. 1. TZ search patterns

Step 2: Initial Grid Search - run the diamond search with different stride length X in search area [–96, 96] shown in Fig. 1(a) the obtained search center in the previous step. X is 1, 2, 4, 8, 16, 32 or 64. The points shown in Fig. 1(a) are tested to decide the point with the minimum SAD. The point with the minimum SAD is taken as a search center for next search. Calculate the distance between the search center and the point with the minimum SAD. The stride length for this minimum SAD point is stored in variable uiBestDistance (BD). If the BD is equal to 0, terminate the motion estimation search. If BD is larger than iRaster, go to Step 3. Otherwise, go to 4.

Step 3: Raster Search is performed. Raster Search is a simple full search on a down-sampled version of the search window [–96, 96]. Run the raster search with raster length 3(iRaster = 3) shown in Fig. 1(b). The points on raster search pattern are tested. The point with the minimum SAD is taken as a search center for next step. And then go to Step 4.

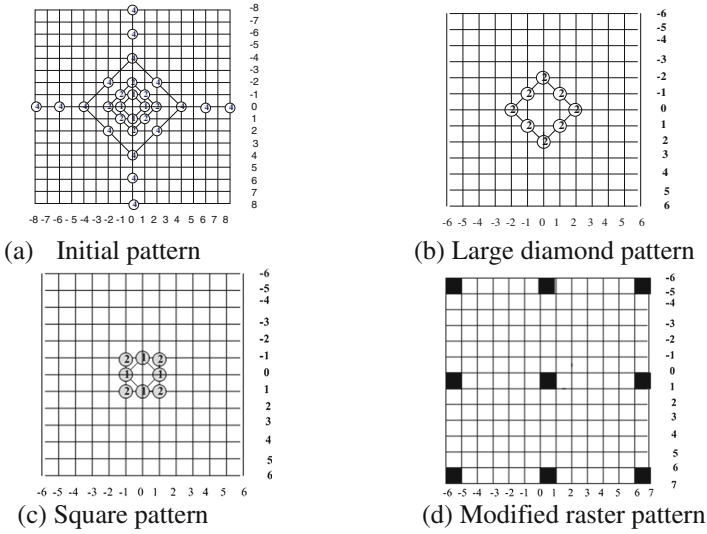
Step 4: Star refinement search is carried out. The search center is moved to the point with the minimum SAD obtained from the previous Step. Repeat the Step 2. The points in this search pattern are tested and decide the point with the minimum SAD. If the BD is equal to 0, decide the point with the minimum SAD as the motion vector and terminate the motion estimation search. Otherwise, repeat this step until BD is 0.

3 Proposed Search Method

There are some problem in TZ. At first, if the motion of the block is large, raster search is performed after initial gird search. Both raster search and initial grid search place search points that can cover the overall search area. It becomes the cause of the unnecessary computational complexity. Secondly, TZ search method used star refinement search which is similar to initial grid search. In star refinement search, there is no need to check the points with stride length 4, 8, 16, 32 or 64. Because there is a strong likelihood that the optimal motion vector exists around the point with the minimum SAD obtained previous step. This refinement search is performed repeatedly until BD is zero. This process needs unnecessary complexity. Lastly, raster search is

performed when the motion of the block is large. Raster search of TZ search configures iRaster as 3 in search area [–96, 96] shown in Fig. 1(b). When the motion of the block is large, iRaster has to be larger than 3. This is because refinement search is performed after raster search.

To reduce the computational complexity of TZ, a new search scheme for motion estimation is proposed. The proposed hierarchical motion estimation search scheme exploits the properties of the distribution of the motion vectors. That is, about 50%–98% of the motion vector are within a radius 2 pixels around the search origin (0, 0). Proposed method also uses the fact that there are more motion vectors in vertical and horizontal direction than in diagonal directions.



(a) Initial pattern (b) Large diamond pattern

(c) Square pattern (d) Modified raster pattern

Fig. 2. Proposed search patterns

These characteristics of the motion vectors are used to decide search pattern and to terminate the search early and to place the search points in the search area. The proposed motion estimation scheme is a hierarchical search strategy. It consists of search patterns in Fig. 2. Initial pattern in Fig. 2(a) is used as a first search pattern to find the motion size of the current block. Large diamond pattern in Fig. 2(b) and square pattern in Fig. 2(c) are used. If the motion of the current block is large, modified raster search pattern in Fig. 2(d) is carried out. Large diamond pattern and square pattern are used as a refinement search pattern to find the best motion vector. The proposed search method is summarized as follows.

Step 1: Motion vector prediction – (0, 0) and motion vectors which come from median predictor, left predictor, up predictor, upper right predictor are used to decide a starting point for the current block.

Step 2: Initial search - Initial search shown in Fig. 2(a) at the obtained search center in the previous step is carried out. The points (①②③ in Fig. 2(a) in initial

pattern are tested and the point with the minimum SAD is decided. Calculate the distance between the search center and the point with the minimum SAD. The distance is stored in BD. If the BD is equal to 0, decide the search center as the motion vector of the block and terminate the motion estimation search. If the BD is equal to 1, go to Step 3. If the BD is equal to 2, go to Step 5, otherwise go to Step 4.

Step 3: Square search - Two neighboring points (② points) of the point with the minimum SAD obtained from the previous Step is carried out. The point with the minimum SAD is decided as the motion vector of the block. And then terminate the search.

Step 4: Modified raster search - Run modified raster search with raster length 5 (iRaster = 6) shown in Fig. 2(d). The points on modified raster search pattern are tested. The point with the minimum SAD is taken as a search center for next search center.

Step 5: Large diamond search - large diamond Search at the point with the minimum SAD obtained from the previous Step is carried out. The point with the minimum SAD is decided. And move from the search center to the point with the minimum SAD. Then, square search is performed. Test all points (①② points) of square pattern and decide the point with the minimum SAD. The point with the minimum SAD is considered as the motion vector of the block.

4 Simulation Results

In this section, we show the experiment results for the proposed motion estimation method. The simulation is carried out on JMVC reference software version 6.0. Multi-view test sequences are used for the experiment; Exit, Uli and Ballroom. The frame size of Exit and Ballroom is 640 * 480 and the frame size of Uli is 1024 * 768. The total number of frames is 100 in each view. QP is varied from 22 through 37 (22, 27, 32 and 37). The search range is 96. We compared TZ to the proposed method in image quality, bit rates and the total encoding time. Table 1 shows the Bjontegarrd Delta (BD) bit rate and BD-PSNR values taken over the four QPs using the Bjntegaard Delta metric. The results shows that there is almost 42 (the motion of block is small)–78 (the motion of block is large)% reduction in the total encoding time shown in Table 2, while

Table 1. BDPSNR and BDBR

Sequence	BD-PSNR (dB)	BDBR (Kbps)
Exit	0.2	13.8
Uli	0.05	1.3
Ballroom	0.4	12.0
Race1	2.0	62.3
Flamenco	0.4	9.2

Table 2. Encoding time

Sequence	TZ (s)	Proposed (s)	ΔT(%)
Exit	77122	41718	46
Uli	386955	113865	71
Ballroom	83309	45344	46
Race1	215806	46724	78
Flamenco	69597	27254	61

obtaining negligible change in the PSNR and bitrates shown in Table 1 using BD-PSNR and BD-bitrates values. In other words, image quality degradation of the proposed method over TZ is about 0.05 (the motion of block is small)–2.0 (the motion of block is large) (dB) and bitrates increment is about 1.3 (the motion of block is small)–62.35 (the motion of block is large) (kbps).

5 Conclusions

Pel Block Search (PBC) and TZ search are the motion estimation algorithms To reduce computational complexity and maintain the image quality, a new motion estimation search method for multi-view video coding is proposed in this paper. The proposed search method which is a hierarchical search method exploits the characteristics of the distribution of motion vectors to place the search points in the search area and terminate motion estimation search early. Simulation results show that the encoding time for motion estimation is reduced.

Acknowledgments. 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 (No. 2010-0024120).

References

1. ISO//IEC JTC1/SC29/WG11 N10357: Vision on 3D Video (2009)
2. Draft ITU-T recommendation and final draft international standard of joint video specification, ITU-T Rec. H.264/ISO/IEC 14496-10 AVC, in Joint Video Team (JVT) of ISO/IEC/ MPEG and ITU-T VCEG, JVT-G050 (2010)
3. Tham, J.Y., Ranganath, S., Kassim, A.: A novel unrestricted center-biased diamond search algorithm for block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **8**, 369–377 (1988)
4. Zhu, S., Ma, K.-K.: A new diamond search algorithm for fast block matching motion. IEEE Trans. Image Process. **9**(2), 287–290 (2000)
5. Zhu, C., Lin, X., Chau, L.P.: Hexagon based search pattern for fast block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **12**, 349–355 (2002)
6. Po, L.M., Ma, W.C.: A novel four-step search algorithm for fast block motion estimation. IEEE Trans. Circuit and Systems for Video Technology **6**, 313–317 (1996)

7. Koga, T., Iinuma, K., Hirano, A., Ishiguro, Y.: Motion compensated interframe coding for video conference. In: Proceedings of the NTC81, pp. G5.3.1–G5.3.5 (1981)
8. Li, R., Zeng, B., Liou, M.L.: A new three step search algorithm for block motion estimation. IEEE Trans. Circ. Syst. Video Technol. **4**(4), 438–441 (1994)
9. Tourapis, A.M., Au, O.C., Liou, M.L.: Fast Block Matching Motion Estimation using Predictive Motion Vector Field Adaptive Search Technique, ISO/IEC/JCTI/SC29/WG11 MPEG2000/M5866 (2000)
10. Chen, Z., Zhou, P., He, Y.: Fast Integer Pel and Fractional Pel motion estimation in for JVT in JVT of ISO/ICE/ MPEG and ITU-T VCEG, 6th meeting (2002)
11. JVT of ISO/IEC MPEG, ITU-T VCEG, MVC software Reference Manual-JMVC8.2 (2010)
12. Tang, X., Dai, S., Cat, C.: An analysis of TZSearch algorithm in JMVC. In: International Conference Green Circuits and Systems (ICGCS), pp. 516–520 (2010)

Research on Method of Technological Evolution Analysis Based on HLDA

Liang Chen, Xiaoping Lei, Guancan Yang, and Jing Zhang(✉)

Institute of Science and Technology Information of China,
National Engineering Research Center of Science and Technology Information,
15 Fuxing Road, Beijing, People's Republic of China
{chenl, leixp, yanggc, janezhang}@istic.ac.cn

Abstract. This paper analyzes technological evolution from viewpoint of change in technology system. As knowledge base, which used to describe technology system conventionally, suffers from heavy dependency on domain experts, this paper replaces knowledge base with hierarchical topic model to analyze the evolution process of technology system. Specifically, we find frequent closed itemsets from terminologies in patent documents at first, then discover association rules and use them to measure the importance of terminologies and semantic relationship between terminologies, afterwards we clean terminologies in corpus and run HLDA model to describe technology system, finally, we analyze technological evolution via changes of technology system. An empirical research on Hard disk drive demonstrates the feasibility of this method.

Keywords: HLDA · Technological structure · Technological evolution · Patent analysis

1 Introduction

In industrial field, it has always been observed that the components contained in technology system keep change as technology develops, incremental innovations lead to minor improvement of components, while disruptive innovations cause change of the whole system [1]. By analyzing technological evolution from viewpoint of change in technology system, we can achieve deeper understanding of how technology develops. Currently, researchers mainly describe technology system via ontology [2–10], which suffers two drawbacks: first, ontology is unsuitable for technology in rapid changing situation, which is quite common during technology life-circle, second, it's unable to quantitatively measure technology development trend, third, manual method is high-cost and unefficient, so limit the usage of domain ontology to technological evolution analysis.

Although researchers proposed many automatic method to generate ontology, due to the low precision problem computer-aided methods of ontology creation are still far from practical application. In this paper we represent technology system with hierarchical topic model instead of domain ontology, thus analyze technological evolution through change of hierarchical topic structure over time.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_54

The rest of this paper is organized as follow: In Sect. 2 we provide background knowledge related to proposed scheme. Sect. 3 elaborates the Proposed framework. In Sect. 4 an empirical research is conducted on hard disk drive to demonstrates the feasibility of the proposed approach. Finally we conclude the paper in Sect. 5 with a brief overview of future efforts that are planned for this project.

2 Background Knowledge

As an unsupervised learning algorithm for text mining, hLDA [11] (Hierarchical Latent Dirichlet Allocation) is a generalization of LDA [12] and it can adapt itself to the growing data set automatically. hLDA can mine latent topics from a large amount of documents and organize these topics into a hierarchy, in which the topics of higher level are more abstractive while the topics of lower level are more specific As we know, technology system evolves over time, if we use topic hierarchy to represent technology system, then time series analysis of the topic hierarchy enable us to get better understanding of technological evolution. Compared to conventional approaches based on ontology or knowledge base, technological evolution analysis based on hLDA relieves researchers of manually intensive activities. But there is a problem when use topic to represent component of technology system, hLDA model takes document as bag-of-word, which means it disregards grammer and word order, as a result each topic is represented by a list of words and the probabilities corresponding to the words. Given the fact that patent documents contain terminologies including single words and compound noun phrases, it's always hard even impossible to interpret technology system by topics consisting of single words.

In order to solve this problem, we use frequent closed itemsets instead of single words to constitute topics. Specifically, we find frequent closed itemsets from terminologies in patent documents at first, then discover association rules and use them to measure the importance of terminologies and semantic relationship between terminologies, afterwards we clean terminologies in corpus and run HLDA model to describe technology system, finally, we analyze technological evolution via changes of technology system.

3 Methodology

The procedure for technological evolution analysis based on hLDA is composed of five steps (Fig. 1). (1) Selecting patent documents, (2) Extracting terminologies, (3) Finding frequent closed itemsets and discovering association rules, (4) recognizing significant terminologies, merging synonyms and reconstructing patent documents, (5) Generating topic hierarchy by hLDA and analyzing technological evolution.

3.1 Selecting Patent Documents

The first step is selection of patent documents. The main objective of this step is to select patent documents from a patent database and to eliminate irreverent patents. To collect patents, a patent retrieval query is used. The query consists of keywords related to the target technology, and bibliographic information such as patent citation, applicants and application data. After collecting patents, we eliminate irrelevant patents by applying a filtering rule.

3.2 Extracting Terminologies

The second step is extraction of terminologies from the selected documents. Terminologies can be extracted from any description in textual format including the abstract, claims and description sections of a patent document. Of these sections, the abstract is precise and has been regarded as the most meaningful part [13]. Thus this paper uses the abstract of patents. Currently there are many automatic methods to extract terms, such as C-value, NC-value [14], tf-idf and chi-square-based method [15], in this paper we use C-value algorithm which is implemented in NLP tools jatetoolkit (https://code.google.com/archive/p/jatetoolkit/).

3.3 Finding Frequent Closed Itemsets and Discovering Association Rules

A problem occurring in automatic term extraction is low precision and high recall. Take C-value algorithm for example, the highest precision on computer science corpus is 35.73%, while the highest recall on it is 97.57% [16]. Thus there are plenty of meaningless multicharacter strings extracted from patents, moreover, on purpose of protecting innovation, patent applicants try to avoid using industry-standard language while writing patent document. As a result, we need remove meaningless strings as well as recognizing synonyms before patent analysis.

In contrast to conventional manual way to clean extracted terminologies, here we use frequent closed itemsets and association rules instead [17]. Specifically, we take terminology as transaction and each word in it as item to find frequent closed itemsets, then discover association rules based on the frequent closed itemsets. The basic idea is, terminologies belonging to a synset are always similar in word-composition, e.g. we have four terminologies to express the same meaning of magnetic head in hard disk drive as shown in Table 1, Since words “thin”, “film”, “head” occur in all 4 terminologies, the frequency of these 3 words is 4, if we set support threshold to be 2, then {thin, film, head} turns out to be frequent itemset, as this frequent itemset has no superset with the same frequency, it is a closed frequent itemset. On the one hand, closed frequent itemset found from terminologies contains significant words occurring in a synset, and these words always form a terminology [18, 19], on the other hand, the association rule discovered from closed frequent itemsets can be used to measure semantic relationship between closed frequent itemsets.

Table 1. Illustration of finding closed frequent itemset

ID	Terminology			
1	Thin	Film	Head	
2	Thin	Film	Magnetic	Head
3	Thin	Film	Inductive	Head
4	Thin	Film	Read/write	Head

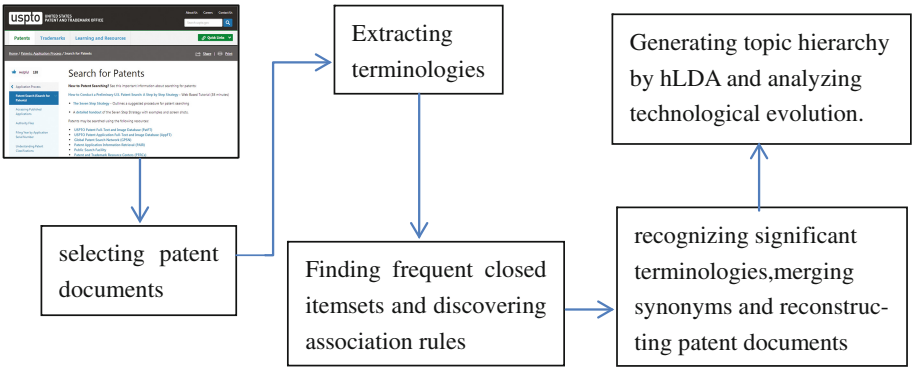


Fig. 1. Process of technological evolution analysis based on HLDA

3.4 Cleaning Terminologies

Since closed frequent itemsets mentioned in Sect. 3.3 always form terminologies, here we take closed frequent itemset as terminology and use its support to indicate the terminology's importance. For association rules between closed frequent itemsets, as its confidence is defined as:

$$C(X \rightarrow Y) \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{1}$$

In which $\sigma(X)$ indicates the frequency of item X we can use the confidence to represent semantic relationship between terminologies. Based on description above, we recognize significant terminologies and combine different terminologies with almost similar or identical meanings into representative terminologies. Then we use these significant terminologies and representative terminologies to represent patent abstracts.

3.5 Generating Topic Hierarchy and Analyzing Technological Evolution

As previous description, hLDA is a hierarchical topic model which can not only detect topics and topic hierarchies from documents collection, but also enable documents collection itself to determine the number of topics. The graphical model representation of hLDA with a nested CRP prior is shown in Fig. 2. As a complex Bayesian networks,

it's not featurable to derive the likelihood function for hLDA, Blei et al. proposed a Gibbs sampling procedure for this model which provides a simple method for simultaneously exploring the spaces of trees and topics, and the result is shown in Fig. 3, where each node represents a topic and each edge points from parent node to child topic, all topics from one document are allocated on one path from root node to leaf (In the rest of this paper we refer path to the path from root node to leaf).

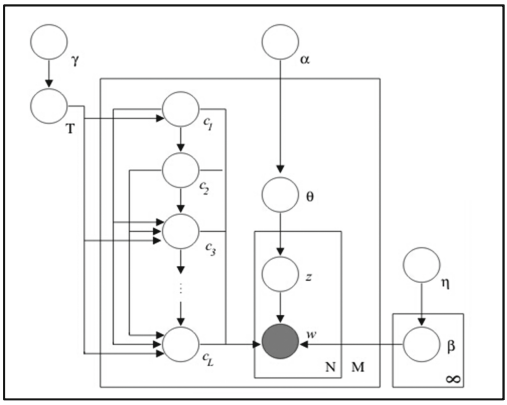


Fig. 2. Plate notation of HLDA model

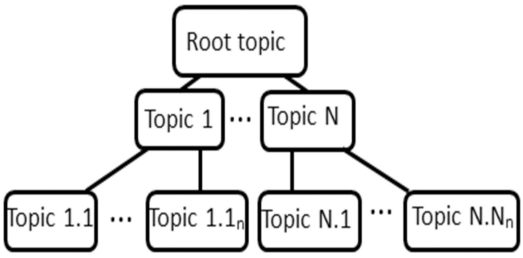


Fig. 3. Hierarchical topic tree generated by HLDA

The process of running hLDA model includes 3 steps:

(1) Initializing topic hierarchy via Nested Chinese restaurant process, during this step each document is randomly assigned to a path in topic hierarchy, and each word from a document is randomly assigned to a node on this document's path;
(2) Keeping topic hierarchy fixed, sample topics for every document using Gibbs sampling, which means for every word in a document, reallocate it to a more possible topic on this document's path;
(3) Keeping words allocation in every topic fixed, sample paths using Gibbs Sampling, which means for every document in dataset, reassign a more possible path to it;

The last 2 steps are repeatedly performed until topic hierarchy and words allocated in each topic become stable.

From every topic in topic hierarchy, we can get two aspects of information: the probability of generating a term given current topic and the number of patents allocated to current topic. Based on the information, we can generate a series of topic hierarchies in different time slices and apply for technological evolution analysis. A common way to do this is, separating patent documents into several time slices according to patent application date, for every time slice generating a topic hierarchy based on patents in this time slice, but then there is a problem that topics from different topic hierarchies will be inconsistent even they are semantically identical, which is a hurdle for technological evolution analysis based on topic; an alternative way is, using patent in the first time slice to generate a topic hierarchy, then inferencing topic allocation for words of patents in other time slices, but in this way topic hierarchy will be fixed once it is generated, thus it is unable to detect new topics. To avoid drawbacks of previous methods, we generate topic hierarchy based on the whole patent collection, then by removing words of patents in the topic hierarchy from the last time slice to the first one, we obtains several topic hierarchies. For example, let's assume we divide the whole period into 4 time slices, if we want to obtain topic hierarchy in 2nd time slice, we need to generate topic hierarchy based on all patent documents first, then remove words of patents in the third and fourth time slice from the topic hierarchy.

After obtaining a series of topic hierarchies, we achieve a variety of information to reflect technology system, which includes not only topic hierarchies in different time slices, but also detailed information about each topic, such as terminologies distribution in a topic and number of patents related to a topic. With the information, we can achieve a deeper understand about technological evolution.

4 Case Study: Hard Disk Drive Head

The hard disk drive (HDD) technology field is selected as an exemplary study, because changes in technology of HDD have been much rapider than that in other fields, so it provides an opportunity of technological evolution analysis in a shorter time (Christensen 2011). America is the birthplace of HDD and it keeps playing an essential role in propelling HDD technology move forward, major HDD companies have paid much attention to intelligence protection and filed amount of patent applications to protect their intelligence right since the beginning. Therefore relevant patents can be treated as a more representative and trustworthy data source for technological evolution analysis. Since the first appearance of HDD in 1956, HDD has undergone 3 generations which includes ferrite head disk drive, thin film head disk drive and magnetoresistant head disk drive (Christensen 2011). As thin film head disk drive has endured the longest period of development, here we focus on thin film head which is a core component of thin film head disk, and collect bibliographic and textual information of patents on HDD from USPTO database.

4.1 Data Collection

(1) Took (TTL/"thin film head" OR ABST/"thin film head") AND APD/1/1/1976->31/12/2013 as search query and ran it on USPTO official website, we achieved 190 patents.
(2) With the assistance of ISTIC Patent Database supported by Institute of Science and Technology Information of China (data updated til 38th week in 2014), we retrieved patents citing or cited by 190 patents. According to these patents' USPC labels, we removed irrelevant ones and got a dataset of 3503 patents.
(3) With citation information of 3503 patents, we generated a citation network consisting of 15 weakly connected components, among which the giant connected component contains 3193 patents. After removing invalidated patents, there are 2876 patents left. We used the bibliographic and textual information of these 2876 patents for our exemplary study. As for patents in other weakly connected components, ignorance of them induces little negative effect on analysis result because of their isolation.

4.2 Measuring Terminologies' Relationship

We used C-value algorithm to extract terminologies from patent abstracts, then applied A-close on these terminologies to find frequent closed itemsets and discover association rules. After trial-and-error learning we set threshold of support to 20, and found 134 frequent closed itemsets, the association rules between closed frequent itemsets is shown in Fig. 4, where node represents closed frequent itemset and node size represents support of the closed frequent itemset, the directed edge between nodes represents association rule between closed frequent itemsets and its width indicates confidence of corresponding association rule. On purpose of highlighting hierarchical structure of the association rules networks, we distinguished nodes on different levels by their color.

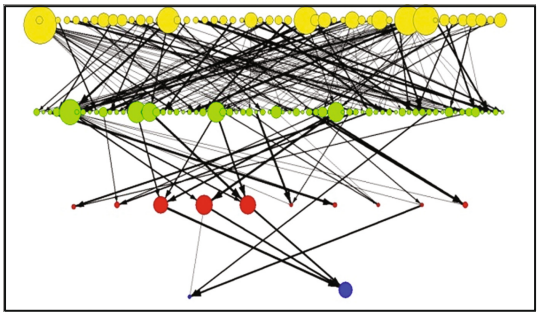


Fig. 4. Networks of closed frequent itemset

Based on significance of terminologies and semantic relationship between terminologies reflected by Fig. 4, we filtered off nonsignificant terms and combined synonyms. Notably, as for synonyms that have no common words, such as giant megnetoresistent head and spin valve, association rules can't recognize them, here we combined them manually. Eventually we got terminology vocabulary consisting of 53 closed frequent itemsets. After replacing terminologies in patent abstract by terminologies in vocabulary, we generated new patent documents which took terminology as the basic unit.

4.3 Building Topic Hierarchy

We used hLDA model implemented in Mallet [20] to generate topic hierarchy. Before running the algorithm, we need set 4 parameters, which include parameter γ of nested Chinese restaurant process, hyperparameter α of document-topic distribution, hyperparameter η of topic-term distribution and k which is levels count of hierarchical topic tree. Through trial-and-error learning we set $\gamma = 10$, $\alpha = 40$, $\eta = 1$, $k = 3$, eventually obtained topic hierarchy with 276 topics in it.

4.4 Technological Evolution Analysis

The patent collection spans 36 years (from 1976 to 2012), here we divided this period into four time slices: 1976–2003, 2004–2006, 2007–2009 and 2010–2012. The reason for longer period of the first time slice is that there are fewer patents in the early days. These four time slices can ensure that number of patents in each time slice is roughly even.

For better visual effect, we removed the nodes with patents fewer than 20, and finally obtained a topic hierachy shown in Fig. 5. Each node contains 3 kinds of information: (1) number of patents, (2) all IDs of time slice the node occurs, (3) top five significant terms in the node.

We used nodes of different color to label topics occurred in different time slices, red node indicates topic not appearing after 2nd time slice, yellow node indicates topic not appearing after 3rd time slice, green node indicates topic appearing all the time. Based on this information, we can observe changes of topics.

As can be seen in Fig. 5, the trend of development for thin film head topic keeps stable, the same situation happens to its three sub-topic which include topic about "ferromagnetic material", "shield layer", "flux guide", topic about "antiferromagnetic layer", "insulation layer" and topic about "head slider". Taking patent number of these topics into consideration, we found that the innovative activity of thin film head during 2003–2012 focused on improvement of coating layer and head slider. As for improving coating layer technology, inventors mainly did research on flux guild, ferromagnetic material and antiferromagnetic material on purpose of bringing better performance of magnetic insulation, as for other sub-topics of thin film head, like spacer layer, free layer, they failed to develop after 2007, however, this doesn't mean these technologies vanished for all in HDD field, on the contrary, these technologies became sub-topics of

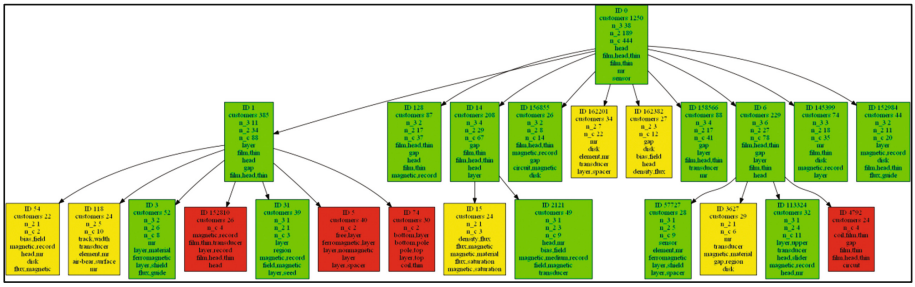


Fig. 5. Hierarchical Topic tree generated by sample data

magnetoresistant head topic, and in this position they developed rapidly. This information indicates that after brief attempt in thin film head, magnetoresistant technologies and their core components, such as spacer layer, free layer, achieved rapid development in magnetoresistant head technology. For topic of storage density, technologies relevant to air-bearing surface and thin film coil developed considerably.

5 Conclusions and Future Research

In spite of significance of analyzing technological evolution from viewpoint of technology system, due to the difficulty of representing technology system automatically, little research has been conducted on their development. Currently researchers use ontology or knowledge base to describe technology system, but these methods suffer labor intensity. This paper provides an approach to represent technology system by topic hierarchy, and analyzes technological evolution via changes of topic hierarchy in different time slice. To demonstrate the feasibility of the proposed approach, we developed a case study for hard disk drive head technology.

The main contributions of this research involve four aspects. First, take advantage of closed frequent itemset and association rule to find significant terminologies and detect synonyms; Second, instead of single word, we take cleaned terminologies as the basic unit to topic hierarchy, thus improve topic's interpretability; Third, we use topic hierarchy to represent technology system; fourth, we proposed a robust method to create topic hierarchy snapshot for each time slice, and based on these topic hierarchy snapshots to analyze technological evolution.

Despite these contributions, some aspect need to be further addressed, such as patents are sparsely scattered through the topic hierarchy, some leaves only contain one patent in it, and hLDA model may not tackle short text so well as it was claimed in [11]. Thus there are a lot of improvement can be made for our method, such as integrating supervised information into hLDA to enhance the interpretability of topics, generating topic hierarchy based on patent claims and descriptions instead of patent abstract, these tasks are listed in our group's further plan.

References

1. Benson, C., Magee, C.: Quantitative determination of technological improvement from patent data. PLoS ONE **10**(4), E0121635 (2013)
2. Dewulf, S.: Directed variation of properties for new or improved function product DNA, a base for connect and develop. Procedia Eng. **9**, 646–652 (2011)
3. Yoon, J., Kim, K.: Trendperceptor: a property-function based technology intelligence system for identifying technology trends from patents. Expert Syst. Appl. **39**(3), 2927–2938 (2012)
4. Moehrle, M., Walter, L., Geritz, A., et al.: Patent-based inventor profiles as a basis for human resource decisions in research and development. R&D Manage. **35**(5), 513–524 (2005)
5. Choi, S., Kang, D., Lim, J., et al.: A fact-oriented ontological approach to SAO-based function modeling of patents for implementing function-based technology database. Expert Syst. Appl. **39**(10), 9129–9140 (2012)
6. Yoon, J., Ko, N., Kim, J.: A function-based knowledge base for technology intelligence. Ind. Eng. Manage. Syst. **14**(1), 73–87 (2015)
7. Wang, X., Qiu, P., Zhu, D., et al.: Indentification of technology development trends based on subject-action-object analysis the case of dye-sensitized solar cells. Technol. Forecast. Soc. Chang. **98**, 24–46 (2015)
8. Choi, S., Kim, H., Yoon, J., et al.: An SAO-based text-mining approach for technology roadmapping using patent information. R&D Manage. **43**(1), 52–73 (2013)
9. Choi, S., Park, H., Kang, D., et al.: An SAO based text mining approach to building a technology tree for technology planning. Exp. Syst. Appl. **39**(13), 11443–11455 (2012)
10. Park, H., Yoon, J., Kim, K.: Identifying patent infringement using SAO based semantic technological similarities. Scientometrics **90**(2), 515–529 (2012)
11. Blei, D.M., Griffiths, T., Jordan, M.I.: Hierarchical topic models and the nested Chinese Restaurant process. In: Sebastian, T., Lawrence, K., Bernhard, S. (eds.) Advances in Neural Information Processing Systems, Cambridge, MA. MIT press (2004)
12. Blei, D.M., Andrew, N., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
13. Chen, L., Tokuda, N., Adachi, H.: A patent documents retrieval system addressing both semantic and syntactic properties [EB/OL], 15 Februrary 2016. http://www.aclweb.org/anthology/W03-2001
14. Frantzi, K., Sophia, A., Hideki, M.: Automatic recognition of multi-word terms: the C-value/NC-value method. Nat. Lang. Process. Digit. Libr. **3**(2), 115–130 (2000)
15. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools **13**(1), 157–169 (2004)
16. Milios, E., Zhang, Y., He, B., et al.: Automatic term extraction and document similarity in special text corpora. In: Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics, Halifax, Scotia N, Canada, pp. 275–284 (2003)
17. Pasquier, N., Bastide, Y., Taouil, R., et al.: Discovering frequent closed Itemsets for association rules [EB/OL], 14 Feburary 2016. http://hal.archives-ouvertes.fr/docs/00/46/77/47/PDF/Discovering_frequent_closed_itemsets_for_association_rules_Pasquier_et_al._ICDT_1999.pdf
18. Chen, L., Zhang, Z.Q., Shang, W.J.: Research method of technological evolution based on frequent closed itemset mining. Lib. Inf. Serv. **57**(9), 107–111 (2013)
19. Chen, L., Zhang, Z.Q.: A method of recognizing technological architecture component based on patent documents. Lib. Inf. Serv. **58**(10), 134–144 (2014)
20. McCallum, K.: MALLET: a machine learning for language toolkit [EB/OL]. http://people.cs.umass.edu/∼mccallum/mallet/

A Mixture Record Linkage Approach for US Patent Inventor Disambiguation

Guan-Can Yang, Cheng Liang, Zhang Jing, Dao-Ren Wang, and Hai-Chao Zhang(✉)

Institute of Science and Technical Information of China,
15 Fuxing Road, Haidian District, Beijing 100038, China
{Yanggc, Chenl, Janezhang, Wangdr, Zhanghc}@istic.ac.cn

Abstract. Inventor name disambiguation is a task that distinguishes each unique inventor from all other inventor records in patent database. This task is essential for processing person name queries in order to get information related to certain inventor. We proposed a mixture approach that applies to the combination of supervised learning, stochastic record linkage and ruled-based method to determine whether each pair of inventor records are from same inventor or not. Our algorithm tested on the USPTO patent database disambiguated 12 million inventor records in 7 h. Evaluation is on labeled dataset from USPTO PatentsView inventor name disambiguation competition and showed our approach have an excellent output.

Keywords: Inventor disambiguation · Machine learning · Patent family · Record linkage

1 Introduction

Interest in innovation micro-data has risen along with the availability of detailed patent databases. However, data quality poses significant barriers to their use. Patent databases do not reliably identify all instances of the same inventor as a unique entity [1]. Inventor disambiguation is the problem of identifying whether two given inventor-patent instances belong to a same inventor or not. The need for disambiguation is accentuated by the fact that USPTO does not have any unique or consistent identifiers for inventors [2]. A disambiguated USPTO database would be helpful to identify how researchers collaborate and what are the possible upcoming areas for new invention.

Name disambiguation algorithms are often used to solve these problems [3–6]. The primary goal of any inventor disambiguation algorithm is to identify all records of all unique inventors in the database (and assign corresponding inventor IDs) with a minimal number of non-systematically occurring errors. Because of the large number of records, manual disambiguation of all records is not feasible and automatic methods are used. An automatic name disambiguation algorithm is typically composed of three parts. The first is features selection, which is perhaps the most important aspect of designing a model for disambiguation because it determines the upper limit of accuracy. The second is a pairwise linkage classifier that determines whether each pair of

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_55

records are from same person or not. The third is a clustering algorithm, grouping records by each unique person using the classifier.

We proposed an inventor name disambiguation algorithm for the USPTO patent database. Our algorithm follows the typical steps of author name disambiguation. We proposed a mixture classifier, which is based on machining learning accompanied by a filtering mechanism to train a pairwise linkage classification. At the same time of optimizing algorithm, we added Inpadoc patent family data and NBER category as external features and simplified the blocking process. Our algorithm tested on the USPTO patent database disambiguated 12 million inventor records in 7 h. Evaluation is on labeled datasets from USPTO Patents View inventor name disambiguation competition and showed our algorithm achieve a good result (http://dev.patentsview.org/workshop) (Fig. 1).

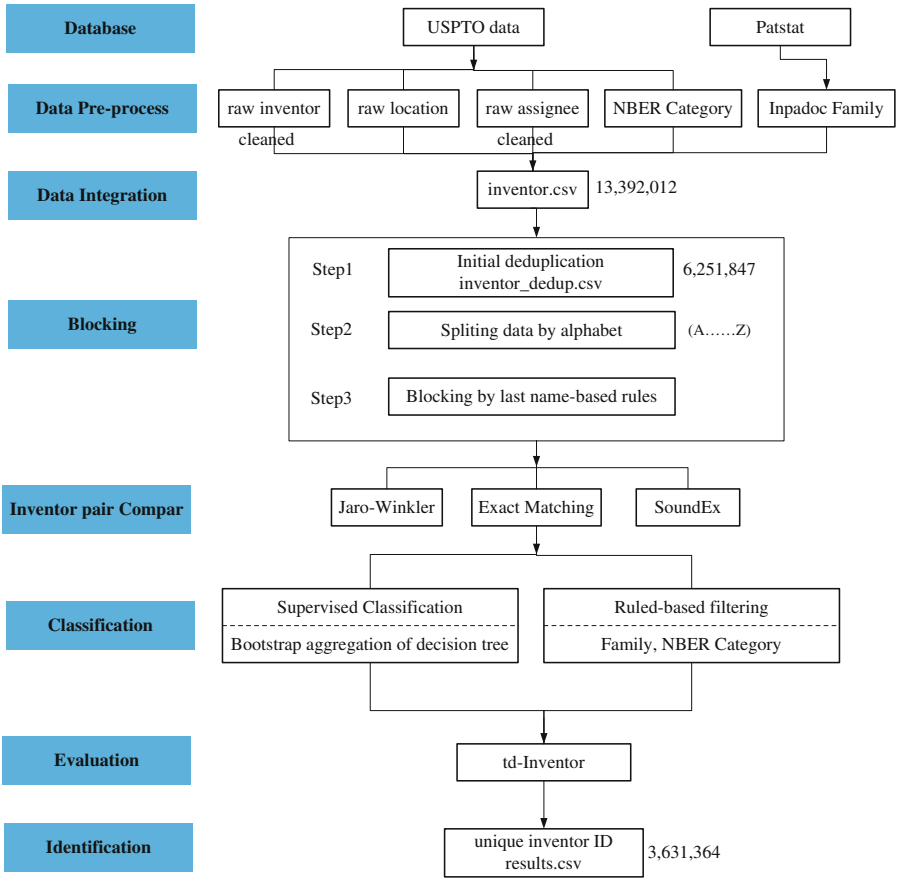


Fig. 1. The inventor disambiguation schema.

2 Disambiguation Process

Our inventor disambiguation process mainly follows the general record linkage schema. Our schema have 3 features different from the general schema: (1) Blocking simplified strategies; (2) Mixture linkage classifier; (3) External information (NBER category and patent family).

2.1 Data Pre-process

Recently there was an inventor name disambiguation competition providing the whole USPTO patent raw data database. This raw data contains all published US patent grants from 1976 to 2014. Raw data is publicly available via the competition's web page (http://dev.patentsview.org/workshop). Although the organizer of this competition made great efforts, there are some missing and errors in some important data fields.

Inventors' names cleansing. Hence a round of pre-processing data is required for late inventor disambiguation process. The original patent inventor dataset does not contain the middle name and full name field, so it is to generate a field name as middle name for each record, which are important fields required by similarity matching. There are some errors in the raw inventor dataset, such as explanatory text, special characters, unnecessary spaces, and so forth. So it is necessary to convert these strings into a standard strings.

Assignee' names cleansing. Assignee is also a very important feature for identifying inventors. However, the raw assignee dataset has a poor quality. Our pre-process of assignee data mainly referenced the 'Patent processor' script, which scripts were generously provided by university of California. We only use the function of 'get_-cleanid' and related modifications are made. The data source is 'rawassignee.csv' and imported local database. Our raw assignee cleansing mainly implement the following 3 steps: (1) Convert to uppercase characters; (2) Condense assignee name by removing spaces and non-alphanumeric characters; (3) Match condensed name with existing list of condensed and harmonized name (using the nber_substitutions library from the Patentprocessor).

2.2 Features Selection

In our approach, the base unit of analysis in inventor disambiguation is an inventor-patent instance. Each record contains attributes used for disambiguation. Such as the inventor's first and last names, the patent assignee, and others as explained below. Each inventor-patent instance occurs only once. In contrast, a patent may appear multiple times, once for each inventor listed on the patent. In our inventor disambiguation, an inventor-patent instance consists of 13 fields, divided into two sets of independent features sets namely inventor name features and patent features.

Besides the above mentioned fields, we consider to add some useful external information to help inventor identification.

(1) Inpadoc patent family. As far as we known, in current, patent family information is rarely used in inventor/assignee disambiguation. In fact, patent family information should be an important indicator for identifying inventor. A patent family is a group of patent publications on a single invention, filed by the same applicant or joint applicants in one or more countries. According the above definition, patent family actually set a boundary of an invention, so the name or address (organization) variations of the same inventor public at the different patent literatures could be traced by the patent family information.
(2) NBER Subcategory. Technical dimension is also important external information. This field provide us a more broadly perspective to discover the name variants of inventors during the technological evolution process.

In sum, patent family-based information added some new dimensions from the technological evolution perspective. The following show some basic descriptive statistics.

2.3 Blocking

Patent database has huge number of records; e.g. the USPTO database contains more than 12 million records of inventors. Clustering everything together is quite difficult, due to the limitation of both memory size and computational power. To solve this problem, blocking is often used. Blocking splits whole records into several blocks, and the clustering is done within each block separately, assuming records from same persons rarely split into different blocks. In this work, we adopted two methods of decreasing amount of computation.

Remove duplications directly according to exact matching rule. We adopt a rule of thumb; with a strict matching rule to initial identify some duplicated inventor's id. We adding a temp field, which is contains five important strings:

"name_first+name_middle+name_last+assignee+city"

We assumed if two inventor-patent instances have an exact matching on these fields, then we have faith in they are the same inventor. This straightforward method have an important role on the whole inventor disambiguation. This step greatly reduced our amount of computation, after this preliminary deduplication, the whole amount of inventor-patent instances reduced from 12,392,012 to 6,251,305, at the same time, this simple way is accompanied by the low error rate.

Blocking simplification. Blocking is a technique used in the field of record linkage to decrease the number of possible comparisons to be done between records. It helps to decrease the computational processing power required. We perform blocking to our dataset, this makes sure that whenever we do a comparison in our algorithm it is done among only potential matches. There are two kind of blocking strategies: (1) the core of blocking is based on name_last_f, which is consist of two part:

"name_last + the first initial of name_first"

As the common cases, the blocking strategy requires that two records must agree in first name and last name, or name_last_f and city, or name_last_f and assignee name to appear in the resulting set of comparison patterns. (2) In terms of some special parts, such as "S", "L", we take a more strict blocking strategy. We consider the two factors, the last name amount of some special letter parts is very large, and many of these last name is consist of fewer characters, which easily lead to duplication of last name and greatly increasing the calculation amount of similarity comparison. At the same time, to further control the memory consumption, we split all the inventor data into 26 parts according the alphabet.

2.4 Mixture Linkage Classifier

Inventor disambiguation is very complex problem, sometimes, a single supervised machining learning approach does not achieve the good performance under special cases: (1) Training data is limited and cannot cover the main variation patterns, or some special name distribution, such as in East Asian names collection, there are an amount of the first character of last name is "L", "W", and "K", according to this names, a more stricter limit should be considered. (2) Sometimes, a classification error may cause a very expensive mistake. Such as there are two inventors group existed, at this time, one wrong inventor pair would cause the two subgroup be connected. Under the precautionary principle, so we proposed a mixed method that applies to the combination of supervised learning, stochastic record linkage and ruled-based method. Finally, we selected to consolidate all these candidate approaches into a new mixture approach.

Bootstrap based supervised classification. Using training data as a validation, we find the bootstrap method embedded in Recordlinkage R's package have good performance at supervised classifications. So we adopt this classification method and predict the whole 26 parts similarity of inventor-patent instances based on the training data. The bootstrap algorithm is based on the idea of generating weak learners that iteratively learn a larger portion of the difficult-to-classify examples in the training data by paying more attention (that is, giving more weight) to often misclassified examples. The aim we choose a bootstrap algorithm as the method of pair similarity classification is to find a boundary contour of the whole inventor network.

Stochastic record linkage. Stochastic record linkage relies on the assumption of conditional probabilities concerning comparison patterns. The probabilities of the random vector $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ having value $\tilde{\gamma} = (\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_n)$ conditional on the match status Z are defined by

$$u_{\tilde{\gamma}} = P(\gamma = \tilde{\gamma}|Z = 0), m_{\tilde{\gamma}} = P(\gamma = \tilde{\gamma}|Z = 1), \tag{1}$$

Where Z = 0 stands for a non-match and Z = 1 for a match. In the Fellegi-Sunter model these probabilities are used to compute weights of the form:

$$w_{\tilde{\gamma}} = \log\left(\frac{P(\gamma = \tilde{\gamma}|Z = 1)}{P(\gamma = \tilde{\gamma}|Z = 0)}\right). \quad (2)$$

The aim we choose a stochastic record linkage as the method of pair similarity scores is to calculate the weight of pair similarity. In our inventor disambiguation, we choose epi weight methods [8].

Rule-based filtering. Some rules are applied in order to decide which matches have to be retained (that is, the two matched inventors are considered to be the same person) and which discarded (the two matched inventors are simply considered homonyms or quasi-homonyms). These rules are often based on "similarity scores". Patent database have some differences from scientific literature database are that patent data including patent family data, which is a native manual interpretation data with great efforts of examiners'. So the similarity based on patent family should be viewed as a strong rule. The remains of candidate patent pairs need to coupling with the weight based on stochastic record linkage to judge (Table 1).

Table 1. Judgments of ruled-based filtering.

Rules	Descriptions	Weight
Rule 1	The highly similarity weight	>= 0.78
Rule 2	The highly similarity weight + the exact matching first character of the middle name	>= 0.70
Rule 3	The highly similarity weight + the exact matching subcategory_id	>= 0.70
Rule 4	The exact matching inpadoc_family_id	Not required

3 Results

The host of competition tested our algorithm. A total 2 training datasets and 5 test datasets were given. The test dataset includes ALS, ALS common, IS, E&S, Phase2. The ALS and ALS common datasets are from Azoulay et al. [9], which consist of inventors from the Association of Medical Colleges (AAMC) Faculty Roster. ALS common is a subset of the ALS dataset with common popular names. The IS dataset is from Trajtenberg and Shiff (2008) [10], containing Israeli inventors in USPTO database. E&S dataset is from Chunmian et al. [11] And consists with patents from engineers and scientists. Phase2 is a random mixture of previous datasets. The training dataset includes Mixture and Common characteristics dataset. Mixture dataset is random mixture of IS and E&S dataset, and Common characteristics dataset is a subsample of E&S dataset which was subsampled according to the match characteristics of the USPTO database, in terms of the mean number of inventors per patent and percentage of missing assignees [12]. Evaluation is done with pairwise precision, recall, and F1 scores calculated by comparing every all calculated possible pairs in the test set of with the given manually labeled dataset.

Table 2 shows the results for each training and test dataset. Results were slightly better with the Common characteristics dataset. Result shows our algorithm performs good results at all different datasets. We can see from the results that the mixture approach, which is effective for inventor name disambiguation for patent database.

Table 2. Disambiguation evaluation

Test Data	Training set	Precision	Recall	F-score
ALS	Mixture	0.998171	0.949989	0.973484
	Common	0.996064	0.961459	0.978456
ALS common	Mixture	0.993201	0.920831	0.955648
	Common	0.982322	0.963280	0.972708
IS	Mixture	0.995252	0.763279	0.863966
	Common	0.995673	0.837911	0.910005
E&S	Mixture	0.999463	0.894823	0.944253
	Common	0.998730	0.903074	0.948496
Phase2	Mixture	0.998488	0.883155	0.937287
	Common	0.991932	0.896134	0.941603

4 Conclusions

This study aims to build an automatic and effective approach for US patent inventor disambiguation. We proposed a mixture approach that applies to the combination of supervised learning, stochastic record linkage and ruled-based method to determine whether each pair of inventor records are from same inventor or not, in addition, we try to integrate some special features such as patent family and NBER category in order to improve the accuracy of approach. Our algorithm tested on the USPTO patent database disambiguated 12 million inventor records in 7 h. Evaluation is on labeled dataset from USPTO PatentsView inventor name disambiguation competition and showed our approach have an excellent output.

Acknowledgments. This research is supported by National Natural Science Foundations of China (Nos. 71403256 and 71303023), and also supported by National Key Technology R&D Program of China (No. 2013BAH21B00).

References

1. Huberty, M., Serwaah, A., Zachmann, G.: A flexible, scaleable approach to the international patent 'name game', Bruegel Working Paper (2014/10i), (2014)
2. Li, G., Lai, R., Amour, D., Doolin, D.M., Sun, Y., Torvik, V.I., Yu, A.Z., Fleming, L.: Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). Res. Policy **43**(6), 941–955 (2014). doi:10.1016/j.respol.2014.01.012
3. Smalheiser, N.R., Torvik, V.I.: Author name disambiguation. Annu Rev Inform Sci **43**(1), 1–43 (2009)

4. Ferreira, A.A., Gonçalves, M.A., Laender, A.H.F.: A brief survey of automatic methods for author name disambiguation. SIGMOD Rec. **41**(2), 15–26 (2012)
5. Fleming, L., King III, C., Juda, A.I.: Small worlds and regional innovation. Org. Sci. **18**(6), 938–954 (2007)
6. Ventura, S.L., Nugent, R., Fuchs, E.R.: Methods Matter: Rethinking Inventor Disambiguation with Classification & Labeled Inventor RecordsAcademy of Management Proceedings, 2013. Academy of Management, p 14537. (2013)
7. Ventura, S.L., Nugent, R., Fuchs, E.R.H.: Seeing the non-stars: (some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. Res. Policy **44**(9), 1672–1701 (2015). doi:10.1016/j.respol.2014.12.010
8. Sariyar, M., Borg, A.: The RecordLinkage package: detecting errors in data. The R Journal **2**(2), 61–67 (2010)
9. Azoulay, P., Michigan, R., Sampat, B.N.: The anatomy of medical school patenting. New Engl J Med **357**(20), 2049–2056 (2007). doi:10.1056/NEJMsa067417
10. Trajtenberg, M., Shiff, G.: Identification and mobility of Israeli patenting inventors Pinhas Sapir Center for Development, Tel Aviv University (2008)
11. Ge, C., Huang, K., Png, I.P.L.: Engineer/scientist careers patents, online profiles, and misclassification bias. Strateg. Manag. J. **37**(1), 232–253 (2016). doi:10.1002/smj.2460
12. Bailey, J.: Evaluation approach and outcomes of the workshop, http://patentsview.org/data/presentations/Bailey_PV_Workshop.pptx

The LTC Framework for Competitive Analysis on Industrial Technology

Hongqi Han[1], Changqing Yao[1], Maoxiang Peng[2], and Yongsheng Yu[1(✉)]

[1] Institue of Scientific and Technical Information of China (ISTIC), No. 15 Fuxing Road, Beijing 100038, China
bithhq@163.com, {yaocq,yuys2015}@istic.ac.cn
[2] Research Center of Science, Technology and Society, Tsinghua University, Beijing 100084, China
pengmaoxiang@tsinghua.edu.cn

Abstract. A framework comprised of three types of analysis is presented for the competitive analysis on industrial technology. The three analysis are technology life cycle analysis, competitive technology analysis and competitor analysis. The first is used to comprehend current technical development stages and developing trend, and the second is used to discover developing technical topics, the development situation and relationship of sub-technologies, while the third is used to explore technical competitors, their relationship and technical strength. A case for analysis the industrial competition of fuel cell technology is introduced. The case shows the three analysis can provide valuable competitive information of industrial technology, and it convince the effectiveness of the framework.

Keywords: Competition analysis · Industrial technology · Technology life cycle

1 Introduction

Technology is considered as one of the important factors in competition analysis, however its value is usually underestimated. When the world enters 21th century, the role of technology innovation in economic and social development becomes more significantly than ever before, especially for the high and emerging technology industry. Therefore, technical competition analysis has been used more frequently and has become more important method in competition analysis. Although traditional competition analysis models and methods can be used in technical competition analysis, such as the five forces model [1] and the diamond model [2], their functions are restricted and the analysis effectiveness is discounted because they are proposed for market competition analysis.

Qualitative analysis and quantitative analysis are two basic methods for technical competition analysis. Along with the study development, researchers have come to accept that the quantitative analysis gives more support for competitive decisions. More and more researches mine competitive intelligence from scientific textual data using

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_56

statistical theory for qualitative analys is [3–8]. Patent data are widely used in competitive analysis owing to its good characteristics. Currently, researches on the application of patent in technical competition analysis are becoming much more, such as core technologies identification [9], technology life cycle analysis [5], technology opportunity analysis [10], technology strength analysis [11, 12]. However, most of these studies are only simple usages of patent analysis [13]. Some studies give the competitive analysis process, while others give research frameworks for special applications. To the best of our knowledge, the systematic analysis framework on technical competition is few.

Wang (2009) [13] presented a three-dimensional model from the aspect of patent analysis (Fig. 1). In the model, technical competition is analyzed from three dimensions: competitive environment analysis, competitive technology analysis and competitor analysis. In first dimension, the analysis includes current technology development trend, technology life cycle, technology situation of different countries or regions, competitor identification. In the second dimension, the analysis includes key technologies and their development trend, core technologies and their development trend, potential or vacant technologies. In the third dimension, the analysis includes R&D trend of competitors, R&D emphasis of competitors, potential alliance of competitors, position of competitors and technical strength of competitors.

Although Wang's model made interesting exploration on the analysis framework of technical competition, the model is just designed for company competition rather than industrial competition. On the one hand, the industrial competition analysis is more complex and broad because it doesn't focus specific company. On the other hand, Wang didn't give further researches, and didn't give case studies or empirical analysis, so the effectiveness of the model can't be testified and whether the intended purpose can be achieved is still in doubt by competitive analysis using patent data.

2 The LTC Analysis Framework

The five forces model and diamond model are two classical models for competitive analysis, which are based on the idea of factor decomposition. Following the idea of factor decomposition, the analysis framework on industrial technology is built based on three basic competitive factor: competitor, competitive environment and competitive purpose [14]. First, competitors of industrial technology refer to institutes involved in research and development. From the viewpoint of companies, competitive analysis on industrial technology is the external environment analysis, and competitors are those companies which develop similar technology. From the viewpoint of government or regions, competitors are those nations or regions which invest into same technology. Second, the competitive environment of industrial technology is different from market environment. The important thing is analyzing the technology readiness level. If the level of industrial technology is in developing stage, competitors will be attracted and join the development; while if the level is in decline stage, competitors will reduce investment, even quit development competition [15, 16]. Finally, let's see what is the main problem in the competitive analysis on industrial technology from the competitive purpose. Standing at the position of companies, the analysis purpose is to find research

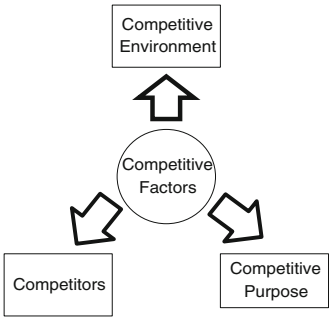


Fig. 1. Three factors of competition

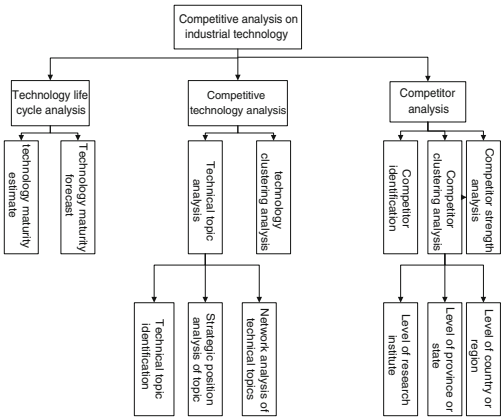


Fig. 2. Framework of competitive analysis on industrial technology.

fields and directions, to understand self position on industrial technology, and thus select potential technologies as future development fields; standing at the position of government or regions, the analysis purpose is to understand the development situation of sub-technologies of industrial technology, and then assess the importance of different sub-technologies, formulate appropriate technology roadmap, make policy to guide investment and technology selection for companies, and thus achieve goals of acquiring industrial technology advantages.

By investigating the requirements of the competitive analysis on industrial technology in three competitive factors, we found that the technical analysis is based on current technical situation and is (1) to comprehend current technical development stages and developing trend, (2) to explore technical competitors, their relationship and technical strength, (3) to discover developing technical topics, the development situation and relationship of sub-technologies. Based on such analysis, the LTC framework is presented to analyze the competitive situation on industrial technology, where L refers to life cycle analysis, T refers to competitive technology analysis, C refers to competitor analysis. The life cycle analysis corresponds to environment factor, while the competitive technology analysis corresponds to purpose factor, and the competitor analysis corresponds to competitor factor. The framework is shown in Fig. 2.

3 Case Study

3.1 Data

In order to test the effectiveness of proposed analysis framework, the patent acquisition system was run to download patent data in the field of fuel cell technology from SIPO (http://www.sipo.gov.cn/). 6346 patents are collected totally. The collected data was preprocessed before the formal analysis.

3.2 Technology Life Cycle Analysis

In the analysis of technology maturity estimate, we count the number of granted and valid patents by year, and make a line chart (Fig. 3). It can be seen the maturity of technology of fuel cell has entered fast developing stage after about 15 years starting stage.

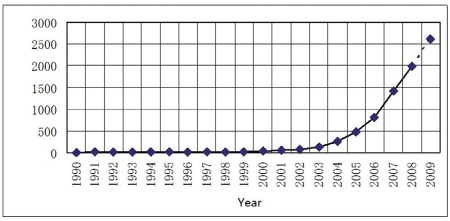


Fig. 3. Technology life cycle estimate.

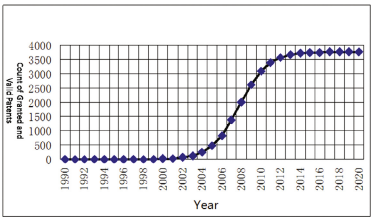


Fig. 4. Technology life cycle forecast.

To forecast future development of fuel cell, we use curve fit method on logistic growth model [17] and get fit coefficients. Then we use logistic growth model to forecast future patent numbers yearly. The results are shown is Fig. 4. The results show the fuel cell technology will go to maturation phase about after 2018. What is necessary to be pondered is the forecast is based on collected data and logistic growth model. It should be scrutinized when using the forecast results because of the problems of S-curve fit method [18]. Nevertheless, it is easy to conclude that the fuel cell technology will go to industrial application phase quickly. The farsighted companies are suggested to increase investment on research, market and product to achieve future competitive advantages.

3.3 Competitive Technology Analysis

In order to identify technical topics and their relationship, we first extract technical terms in patents, and then cluster these terms into topics using two passes algorithm proposed by Coulter [19]. The discovered topics are constituted by densely connected terms internally and connected to each other by co-occurred terms between them [20]. Using discovered topics as nodes, and the connections between them as edges, we can construct a technology network (shown in Fig. 5). In the network, the node size shows occurred patent number for a topic, and the weight of the edge between two nodes is the link numbers, where link represents two co-occurred terms between connected topics. In Fig. 5, the weight threshold is set 4, thus leaving some isolated nodes. Because the threshold guarantees the nodes in a components of the network is strongly connected, we can define each connected sub-network as a sub-technology.

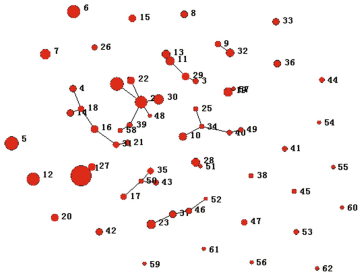


Fig. 5. Discovered technical topics and their relations.

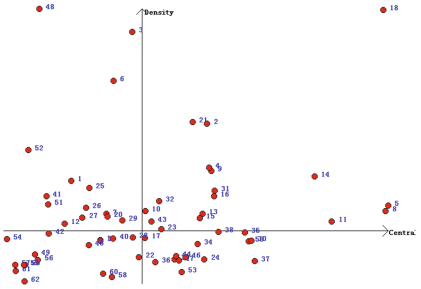


Fig. 6. The strategic diagram of fuel cell technology.

When technical topics are discovered, we can evaluate their developing situation of topics using the strategic diagram method [21]. The strategic diagram use density and centrality index to identify the position of topics [22]. Density is used to measure the strength of the links that tie the keywords of the cluster, while Centrality is used to measure the degree of interactions of a network with other networks. The diagram is drawn in the Cartesian coordinate system, using x-axis for representing centrality, and y-axis density, and the medium of centrality and density as the origin. The topics in quadrant 1 represents central and developed technologies, and the topics in quadrant 2 represents peripheral and developed technologies, and the topics in quadrant 3 represents peripheral and undeveloped technologies, and the topics in quadrant 4 represents central and undeveloped technologies. Figure 6 shows the strategic position analysis of topics in fuel cell technology.

Moreover, we can speculate the developing situation of each sub-technologies using clustering analysis. Figure 7 shows the analysis results. Here, we use k-means algorithm to cluster all collected patents into 6 groups, and each group represents a type of sub-technologies. The ratio of each sub-technology represents the investment quota.

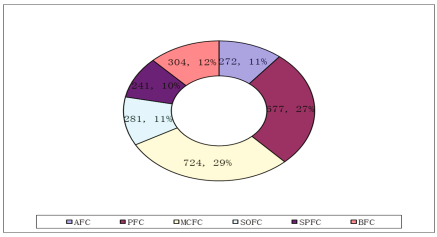


Fig. 7. Classification result of fuel cell technology by electrolyte.

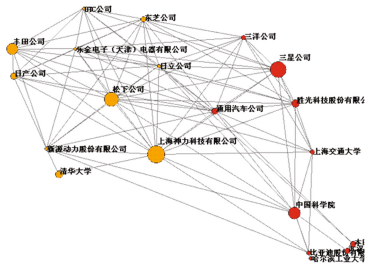


Fig. 8. Clustering result of research institutions. Table 1 shows the Chinese name of institutes in the graph, and their English names and group numbers.

3.4 Competitor Analysis

The competitor identification is the basic of competitor analysis. It is easy to identify technical competitors with patent data, because the bibliographic data contain the name, address and other information of research organizations. It's also easy to identify the organization type by their names, because the institute names usually contain university, company and other specific words. Organizations can be further divided into three levels in the region: research institute, province or state, country or region. After obtaining research organizations, their types and other information, we can cluster them into some groups. Figure 8 shows the results in research institute level. In the case, the research institutes are clustered into 2 groups. The institutes in each group have developed similar technologies.

Technology strength evaluation is another competitor analysis. The analysis will rank research organizations by their strength. Many evaluation indexes of technology strength have been proposed, such as current impact index [23], score [23], technology centrality index [24], etc. Here, we evaluate the strength by using comprehensive indexes combining patenting activity index, patent growth ratio index, technology importance index, and technology value index [25]. Figure 9 shows the results in technology strength. In the case, the Samsung is the most powerful institute, the strongest competitors are the Panasonic from Japan and Shanghai Shen-Li High Tech from China. And the Chinese Academy of Sciences is the potential technology competitor.

Table 1. The Chinese name, English name of institutes and their group numbers.

Chinese name	English name	Group No
清华大学	Tsinghua University	1
新源动力股份有限公司	Sunrise Power	1
上海神力科技有限公司	Shanghai Shen-Li High Tech	1
松下公司	Panasonic	1
日产公司	Nissan	1
丰田公司	Toyota	1
日立公司	Hitachi	1
东芝公司	Toshiba	1
BTC 公司	BTC	1
乐金电子电器有限公司	LG	1
上海交通大学	Shanghai Jiaotong University	2
中国科学院	Chinese Academy of Sciences	2
三星公司	Samsung	2
三洋公司	Sanyo	2
通用汽车公司	General Motors	2
胜光科技股份有限公司	Antiq	2
哈尔滨工业大学	Harbin Institute of Technology	2
比亚迪股份有限公司	BYD	2
本田株式会社	Honda	2
武汉大学	Wuhan University of Technology	2

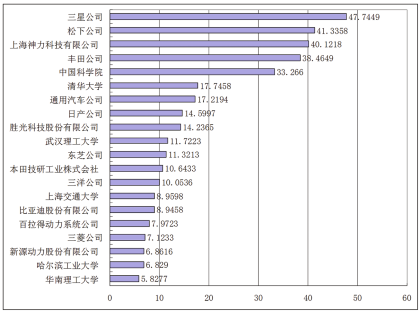


Fig. 9. The rank of research institutions on technical strength. The Chinese name of institutes in the graph, and their English names and group numbers are shown in Table 1.

4 Conclusion

The LTC model is presented for the competitive analysis on industrial technology. There are three types of analysis in the framework of the model. The first is technology life cycle analysis, which is used to comprehend current technical development stages and developing trend. The second is competitive technology analysis, which is used to discover developing technical topics, the development situation and relationship of sub-technologies. The third is competitor analysis, which is used to explore technical competitors, their relationship and technical strength. A case for the competitive analysis on fuel cell technology is introduced using the framework, and it shows the framework can provide valuable competitive information of industrial technology.

Acknowledgments. This work is mainly supported by the National Natural Science Foundation of China (No. 71473237), and partially supported by the Key Technologies Research and Information Service Resource Development Towards to Scientific and Technical Information Analysis (No. 2015BAH25F01).

References

1. Porter, M.E.: Competitive advantage: creating and sustaining superior performance. Competitive Adv. Creating Sustaining Super. Perform. **1**, 94 (2004)
2. Clark, T., Porter, M.W.: The competitive advantage of nations. J. Mark. **312**(1), 108–111 (1991)
3. Teo, T.S.H., WY, Choo: Assessing the impact of using the Internet for competitive intelligence. Inf. Manage. **39**, 67–83 (2001)
4. Alan, P., Scott, C.: Tech Mining: Exploiting New Technologies for Competitive Advantage. Wiley, Hoboken (2005)
5. Reinhard, H., Martin, K., Marcus, L.: Patent indicators for the technology life cycle development. Res. Policy **36**(3), 387–398 (2007)
6. Fai, F., Von Tunzelmann, N.: Industry-specific competencies and converging technological systems: evidence from patents. Struct. Change Econ. Dyn. **12**(2), 141–170 (2010)
7. Gerdsri, N., Daim, T.U.: Generating intelligence on the research and development progress of emerging technologies using patent and publication information. In: 2008 IEEE International Conference on Management of Innovation and Technology, pp. 1–6. IEEE Press, New York (2008)
8. Alessandro, Z.: Competitive intelligence through data mining public sources. Competitive Intell. Rev. **9**(1), 44–54 (2001)
9. Lee, H., Kim, C., Cho, H.: An ANP-based technology network for identification of core technologies: a case of telecommunication technologies. Expert Syst. Appl. **36**(1), 894–908 (2009)
10. Porter, A.L., Roessner, J.D., Jin, X.Y.: Measuring national 'emerging technology' capabilities. Sci. Public Policy **29**(3), 189–200 (2002)
11. Han, H., Gui, J., Xu, S.: Technical strength evaluation method based on patent text data. New Techn. Libr. Inf. Serv. **30**(1), 66–71 (2014)
12. Haupt, R.: Patent Analysis of a Companys Technology Strength. Friedrich-Schiller-Universität Jena, Wirtschaftswissenschaftliche Fakultät, Jena (2005)

13. Wang, X., Sun, J.: Research on application of patent map to technical competition analysis. Libr. Inf. Serv. **53**(53), 79–82 (2009)
14. Wang, Z.: Competition Intelligence. Scientific and Technical Literature Publishing House, Moscow (2005)
15. Liang, G.: Research on technology life cycle based on patent information analysis–taking method for manufacturing graphene of China as an example. Technology Intelligence Engineering (4), (2015)
16. Kim, J., Hwang, M., Jeong, D.H., Jung, H.: Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. Expert Syst. Appl. **39**(16), 12618–12625 (2012)
17. Daim, T.U., Rueda, G., Martin, H., Gerdsri, P.: Forecasting emerging technologies: use of bibliometrics and patent analysis. Technol. Forecast. Soc. Chang. **73**(8), 981–1012 (2006)
18. Christensen Clayton, M.: Exploring the limits of the technology s-curve. Part I - component technologies. Prod. Oper. Manage. **1**(4), 334–357 (1992)
19. Coulter, N., Ira, M., Suresh, K.: Software engineering as seen through its research literature: a study in co-word analysis. J. Am. Soc. Inf. Sci. **49**(13), 1206–1223 (1998)
20. Han H., Xu S., Qiao X.: Mining technical topic network from Chinese patents. In: Proceedings of the First International Workshop on Patent Mining and Its Applications. University of Hildesheim, Germany (2014)
21. Callon, M., Courtial, J.P., Laville, F.: Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. Scientmetrics **22**, 155–205 (1991)
22. Han, H., Gui, J., Xu, S.: Revealing research themes and their evolutionary trends using bibliometric data based on strategic diagrams. In: 2013 International Conference on Information Science and Cloud Computing Companion, pp. 653–659 (2013)
23. Chen, Y., Huang, Y., Fang, J.: Collection and Analysis of Patent Information. Tsinghua University Press, Beijing (2006)
24. Yoon, B., Park, Y.: A text-mining-based patent network: analytic tool for high-technology trend. J. High Technol. Manage. Res. **15**(1), 37–50 (2004)
25. Han, H., Gui, J., Xu, S.: Technical Strength evaluation method based on patent text data. New Technol. Libr. Inf. Serv. **30**(1), 66–71 (2014)

Bayesian Multinomial Naïve Bayes Classifier to Text Classification

Shuo Xu, Yan Li, and Zheng Wang(✉)

Institute of Scientific and Technical Information of China, No. 15, Fuxing Road, Haidian District, Beijing 100038, People's Republic of China
{xush, liyan2015, wangzheng2015}@istic.ac.cn

Abstract. Text classification is the task of assigning predefined classes to free-text documents, and it can provide conceptual views of document collections. The multinomial naïve Bayes (NB) classifier is one NB classifier variant, and it is often used as a baseline in text classification. However, multinomial NB classifier is not fully Bayesian. This study proposes a Bayesian version NB classifier. Finally, experimental results on *20 newsgroup* show that Bayesian multinomial NB classifier with suitable Dirichlet hyper-parameters has similar performance with multinomial NB classifier.

Keywords: Text classification · Multinomial naïve bayes classifier · Fully bayesian

1 Introduction

Text classification [1] is the task of assigning predefined classes to free-text documents, and it can provide conceptual views of document collections. Given the growing volume of online text available through World Wide Web, news feeds, electronic mail, and digital libraries, this task is of great practical significance. Instead of manually classifying documents or hand-crafting automatic classification rules, many machine learning algorithms are trained to automatically classify documents based on human-labeled training documents.

The multinomial naïve Bayes (NB) classifier [2, 3, 4] is one NB classifier variant used for multinomially distributed data like the ones encountered in text classification. It is often used as a baseline in text classification because it is fast and easy to implement. Moreover, with appropriate preprocessing, it is competitive with more advanced methods including support vector machines (SVMs) [5].

However, multinomial NB classifier, as standardly presented, is not fully Bayesian. At least not in the sense that a posterior distribution over parameters is estimated from training documents and then used for predictive inference for new document. The organization of the rest of this paper is as follows. After multinomial NB classifier is briefly described in Sect. 2, Bayesian version is proposed in Sect. 3. In Sect. 4, experimental results on *20 newsgroup* data show that Bayesian multinomial NB classifier has similar performance with multinomial NB classifier, and Sect. 5 concludes this work.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_57

2 Multinomial NB Classifier

As a matter of fact, multinomial NB classifier is a generative model. It assumes that a corpus of documents is generated by selecting a class $c_m \sim \vec{\vartheta}$ for a document \$m\$ then generating each word of that document independently according to a class-specific distribution $w_{m,n} \sim \vec{\varphi}_{c_m}$. The bag-of-words assumption is clearly violated by natural language texts.

In multinomial NB classifier, every word w_i gets a say in determining which label $c \in \{1, \ldots, C\}$ should be assigned to a unseen document $\vec{\tilde{w}} = (\tilde{w}_1, \ldots, \tilde{w}_{\tilde{N}})$. In order to choose a label c for $\vec{\tilde{w}}$, multinomial NB classifier begins by calculating the prior probability $\Pr(c)$ of each label c, which is determined by assuming equiprobable classes, or checking the frequency of each label in the training set. The contribution from each word is then combined with this prior probability, to arrive at a likelihood estimate for each label. This is known as the maximum a posteriori (MAP) decision rule. It can be defined formally as follows.

$$c = \arg\max_c \Pr(c) \prod_{\tilde{n}=1}^{\tilde{N}} \Pr(\tilde{w}_{\tilde{n}}|c) = \arg\max_c \vartheta_c \prod_{\tilde{n}=1}^{\tilde{N}} \varphi_{c,\tilde{w}_{\tilde{n}}} \tag{1}$$

Given a training document set $\mathcal{D} = \{(\vec{w}_m, c_m)\}_{m=1}^{M}$, ϑ_c and $\varphi_{c,v}$ are usually estimated by a smoothed version of maximum likelihood (ML) as follows. In fact, these are both MAP estimates given uniform Dirichlet priors (see further).

$$\hat{\vartheta}_c = \frac{n^{(c)} + \alpha}{M + C\alpha} \tag{2}$$

$$\hat{\varphi}_{c,v} = \frac{n_c^{(v)} + \beta}{n_c + V\beta} \tag{3}$$

where V is the number of unique words, $n^{(c)}$ is the number of documents with class c in $\mathcal{D}$, $n_c^{(v)}$ is the number of times word v appears in a document of class c in $\mathcal{D}$, and $n_c = \sum_{v=1}^{V} n_c^{(v)}$. The smoothing priors $\alpha \geq 0$, $\beta \geq 0$ prevent zero probabilities in further computations. Setting $\alpha = 1$, $\beta = 1$ are called Laplace smoothing, while $\alpha < 1$, $\beta < 1$ is called Lidstone smoothing [6].

3 Bayesian Multinomial NB Classifier

The graphical model representation for Bayesian multinomial NB classifier is shown in Fig. 1. Bayesian multinomial NB classifier can also be viewed as a generative process, which can be described as follows.

1. Draw a multinomial $\vec{\vartheta}$ from Dirichlet ($\vec{\alpha}$);
2. For each class $c \in \{1, \ldots, C\}$:
   (a) Draw a multinomial $\vec{\varphi}_c$ from Dirichlet ($\vec{\beta}$);

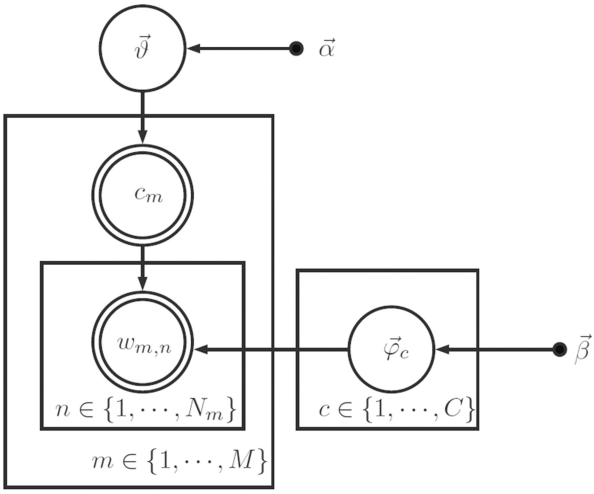


Fig. 1. The graphical model representation of Bayesian multinomial NB classifier.

3. For each document $m \in \{1, \ldots, M\}$:
   (a) Draw a class c_m from Multinomial($\vec{\vartheta}$);
   (b) For each word $n \in \{1, \ldots, N_m\}$ in document m:
      i. Draw a word $w_{m,n}$ from Multinomial ($\vec{\varphi}_{c_m}$).

Given a training document set $\mathcal{D}$ and priors $\vec{\alpha}$, $\vec{\beta}$, MAP estimates of $\vec{\vartheta}$ and $\{\vec{\varphi}_c\}$ can be calculated formally:

$$\begin{aligned}\overrightarrow{\hat{\vartheta}} &= \arg\max p(\vec{\vartheta}, \{\vec{\varphi}_c\}|\mathcal{D},\ \vec{\alpha}, \vec{\beta}) \\ &= \arg\max \Pr(\mathcal{D}|\vec{\vartheta}, \{\vec{\varphi}_c\})\, p(\vec{\vartheta}|\vec{\alpha})\, p(\{\vec{\varphi}_c\}|\vec{\beta}) \\ &= \arg\max \text{Dirichlet}(\vec{\vartheta}|\vec{n} + \vec{\alpha}) \prod_{c=1}^{C} \text{Dirichlet}(\vec{\varphi}_c|\vec{n}_c + \vec{\beta})\end{aligned} \tag{4}$$

Following the mode of Dirichlet distribution, MAP parameter estimates can be expressed as follows.

$$\hat{\vartheta}_c = \frac{n^{(c)} + \alpha_c - 1}{M + \sum_{c'=1}^{C}(\alpha_{c'} - 1)} \tag{5}$$

$$\hat{\varphi}_{c,v} = \frac{n_c^{(v)} + \beta_v - 1}{n_c + \sum_{v'=1}^{V}(\beta_{v'} - 1)} \tag{6}$$

It is not difficult to see that Eqs. (5 and 6) are equivalent to Eqs. (2 and 3) when $\alpha_c = 2$ ($c \in \{1, \ldots, C\}$) and $\beta_v = 2$ ($v \in \{1, \ldots, V\}$). In order to assign a class c to a given document $\overrightarrow{w}$, fully Baysian inference should be used as follows.

$$
\begin{aligned}
c &= \arg\max \Pr(c|\overrightarrow{\tilde{w}}, \mathcal{D}, \vec{\alpha}, \vec{\beta}) \\
&= \arg\max \iint p(c, \overrightarrow{\tilde{w}}, \vec{\vartheta}, \vec{\varphi}_c|\mathcal{D}, \vec{\alpha}, \vec{\beta}) d\vec{\vartheta} d\vec{\varphi}_c \\
&= \arg\max \iint \Pr(c|\vec{\vartheta}) p(\vec{\vartheta}|\mathcal{D}, \vec{\alpha}) \Pr(\overrightarrow{\tilde{w}}|\vec{\varphi}_c) p(\vec{\varphi}_c|\mathcal{D}, \vec{\beta}) d\vec{\vartheta} d\vec{\varphi}_c \\
&= \arg\max \int \vartheta_c \mathrm{p}\left(\vec{\vartheta}|\mathcal{D}, \vec{\alpha}\right) d\vec{\vartheta} \int \prod_{i=1}^{n} \varphi_{c,\tilde{w}_i} p\left(\vec{\varphi}_c|\mathcal{D}, \vec{\beta}\right) d\vec{\varphi}_c \\
&= \arg\max \frac{\Gamma\left(\sum_{c'=1}^{C} \alpha_{c'} + M\right)}{\prod_{c'=1}^{C} \Gamma\left(\alpha_{c'} + n^{(c')}\right)} \frac{\prod_{c'=1}^{C} \Gamma\left(\alpha_{c'} + n^{(c')} + I(c = c')\right)}{\Gamma\left(\sum_{c'=1}^{C} \alpha_{c'} + M + 1\right)} \\
&\quad \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v + n_c\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v + n_c^{(v)}\right)} \frac{\prod_{v=1}^{V} \Gamma\left(\beta_v + n_c^{(v)} + \tilde{n}^{(v)}\right)}{\Gamma\left(\sum_{v=1}^{V} \beta_v + n_c + \tilde{n}\right)} \\
&= \arg\max \frac{\alpha_c + n^{(c)}}{\sum_{c'=1}^{C} \alpha_{c'} + M} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v + n_c\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v + n_c^{(v)}\right)} \frac{\prod_{v=1}^{V} \Gamma\left(\beta_v + n_c^{(v)} + \tilde{n}^{(v)}\right)}{\Gamma\left(\sum_{v=1}^{V} \beta_v + n_c + \tilde{n}\right)}
\end{aligned} \tag{7}
$$

where $\Gamma(\cdot)$ and $I(\cdot)$ is the Gamma and indicator function respectively, and $\tilde{n}^{(v)}$ is the number of times the word v appears in the document $\overrightarrow{\tilde{w}}$.

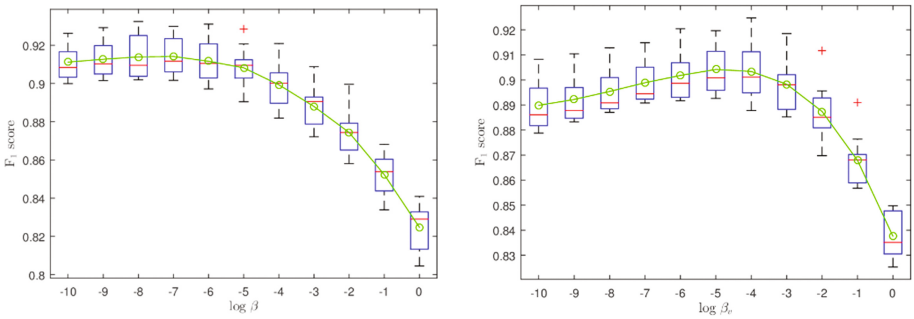
4 Experiments and Discussions

In this study, 20 newsgroup data[1], which was collected and originally used for text classification by Lang [7], is utilized to evaluate the performance. It contains 18,828 non-empty documents evenly distributed across 20 categories, each representing a newsgroup, where the number of documents in training data is 13,180. The *TfidfVectorizer* in scikit-learn [8] is used to convert 20 newsgroup data to a matrix of TF-IDF features. The number of features is 130,274.

In multinomial NB classifier, α is fixed to 1, and β is tuned. For simplicity, the symmetric Dirichlet priors are utilized in Bayesian multinomial NB classifier, where $\alpha_c = 2$ ($c \in \{1, \ldots, C\}$) and β_v ($v \in \{1, \ldots, V\}$) is tuned. In order to identify proper parameters, the grid searcher [9] with 10-fold cross validation is used. Let $\log\beta$, $\log\beta_v \in \{-10, -9, \ldots, 0\}$. The performance in terms of F_1 score with β or β_v is reported in Fig. 2. From Fig. 2, one can see that the optimal values for β and β_v are 2^{-7}, 2^{-5}, respectively.

According to tuned parameter above, the experimental results on test data are reported in Table 1, which indicates that Bayesian multinomial NB classifier has

[1] 20 newsgroup data can be available online from http://mlcomp.org/datasets/379.



(a) Multinomial NB classifier (b) Bayesian multinomial NB classifier

Fig. 2. The performance of 10-fold cross validation on *20 newsgroup* training data with $\log\beta$ or $\log\beta_v$ in terms of F_1 score. (a) Multinomial NB classifier (b) Bayesian multinomial NB classifier

similar performance with multinomial NB classifier. This observation is consistent with Rennie [10].

Table 1. Experimental results on *20 newsgroup* data.

Category	Precision (%)		Recall (%)		F_1 score (%)	
	Standard	Bayesian	Standard	Bayesian	Standard	Bayesian
alt.atheism	92.68	91.13	93.83	92.24	93.25	92.68
comp.graphics	80.06	83.28	88.93	86.91	84.26	80.06
comp.os.ms-windows.misc	97.44	84.29	96.51	80.82	96.97	97.44
comp.sys.ibm.pc.hardware	90.78	80.27	74.85	79.73	82.05	90.78
comp.sys.mac.hardware	89.59	88.48	83.1	92.97	86.23	89.59
comp.windows.x	97.08	91.26	96.03	87.88	96.55	97.08
misc.forsale	89.7	91.13	86.26	77.93	87.95	89.7
rec.autos	95.48	92.07	97.05	93.21	96.26	95.48
rec.motorcycles	92.04	95.97	88.14	97.28	90.04	92.04
rec.sport.baseball	91.09	98.38	91.8	96.19	91.44	91.09
rec.sport.hockey	97.94	95.83	96.94	99.01	97.44	97.94
sci.crypt	95.41	92.33	98.31	97.31	96.84	95.41
sci.electronics	97.09	87.99	99.34	86.58	98.2	97.09
sci.med	91.59	97.09	96.59	96.39	94.02	91.59
sci.space	90.03	93.4	88.22	97.38	89.12	90.03
soc.religion.christian	90.04	88.85	92.24	97.95	91.13	90.04
talk.politics.guns	95.29	87.5	95.29	96.75	95.29	95.29
talk.politics.mideast	81.17	95.41	83.06	98.31	82.1	81.17
talk.politics.misc	90.42	94.14	95.93	88.56	93.1	90.42
talk.religion.misc	87.32	95.2	82.53	69.59	84.86	87.32
average	91.61	91.2	91.25	90.65	91.35	91.61

5 Conclusions

Text classification is the task of assigning predefined classes to free-text documents, and it can provide conceptual views of document collections. Instead of manually classifying documents, many machine learning algorithms are trained to automatically classify documents based on annotated training documents.

The multinomial naïve Bayes (NB) classifier is one NB classifier NB variant, and it is often used as a baseline in text classification. However, multinomial NB classifier is not fully Bayesian, so a Bayesian version is proposed in this study. Finally, experimental results on *20 newsgroup* show that Bayesian multinomial NB classifier with suitable Dirichlet hyper-parameters has similar performance with multinomial NB classifier.

Acknowledgments. We thank the financial support from National Science Foundation of China (ID: 71403255), and Key Technologies R&D Program of Chinese 12th Five-Year Plan (2011–2015) (ID: 2015BAH25F01). Our gratitude also goes to the anonymous reviewers for their valuable comments.

References

1. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining Text Data, pp. 163–222. Springer (2012)
2. McCallum, A., Nigam, K.: A comparison of event models for naive Bayes text classification. In: ICML/AAAI-98 Workshop on Learning for Text categorization, AAAI, pp. 41–48 (1998)
3. Rish, I.: An empirical study of the naïve Bayes classifier. In: IJCAI Workshop on Emprical Methods in AI (2001)
4. Bird, S., Klein, E., Loper, E. (eds.): Natural Language Processing with Python. O'Reilly, Springfield (2009)
5. Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th International Conference on Machine Learning (2003)
6. Manning, C.D., Raghavan, P., Schütze, H. (eds.): Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
7. Lang, K.: Newswenews: learning to filter netnews. In: Proceedings of the 12th International Conference on Machine Learning, pp. 331–339 (1995)
8. Pedregosa, F., Varoquaus, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(2), 2825–2830 (2011)
9. Xu, S., Ma, F., Tao, L.: Learn from the information contained in the false splice sites as well as in the true splice sites using SVM. In: Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering, Amsterdam, pp. 1360–1366. Atlantis Press, Netherlands (2007)
10. Rennie, J.D.M.: Improving multi-class text classification with naive Bayes. Master's thesis, Massachusetts Institute of Technology (2001)

CERIF: A Research Information Model for Electric Vehicles Decision Support System

YingJie Zhang[1] and Na Qi[2(✉)]

[1] National Engineering Research Center of Science and Technology Inofrmation, ISTIC, Fuxing Road No. 15, 100038 Beijing, China
edwardzhangjie@163.com
[2] Research Department, ISTIC, Uxing Road No. 15, 100038 Beijing, China
qinall20@163.com

Abstract. This paper presents the application of CERIF on the electric vehicles decision support system. Firstly, it discusses why we choose the electric vehicle industry and what's its special requirements; how do we use CERIF to respond the challenge for the data-driven decision support and how do we carry out our data management. At last, we propose a list of summary for the application of CERIF in the electric vehicle decision support activities.

Keywords: CERIF · Electric vehicles · Decision support system · Multi-source data management

1 Introduction

Nowadays, the research is becoming more data-intensive. Common European Research Project Information Format (CERIF) is considered a standard recommended by the European Union to its Member States as reference model for the development of research information system [1]. It has a broad coverage including all aspects of research infrastructure, such as projects, persons, organizations, funding, publications, datasets, patents, products, bibliometrics, impact indicators, equipment, etc. As we know, Current Research Information System (CRIS) activities and developments in Europe are tightly interrelated with CERIF, which can go back to early 1984 starting from the IEDAS project [2]. More practices later has improved the CERIF. In 1997 a working group of the EC was set up to produce CERIF2000. Now the latest version is CERIF 1.6, which accommodates metadata of research datasets and is based on the input from the CERIF for Datasets (C4D) [3]. Also, the development of CRIS systems based on the CERIF model is rapidly increasing in recent years.

As for the electric vehicles decision support system, it was initiated as a pilot project to build a data-expert-tool decision model by the Ministry of Science and Technology of the People's Republic of China from the national science and technology supporting plan of the twelfth five-year. In China, the new energy vehicles is a key breakthrough for the vehicles industry to become stronger. What's more, it is becoming urgent when emission reduction has been commonly recognized by nations in the world. All countries have taken positive measures to strengthen the regulations

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_58

on the vehicle's fuel consumption. The target has been set by the EU, America and Japan. So the new energy vehicles is the focus of global vehicle industry transformation and update.

From the industry level, the new energy vehicle has been formulated as strategic emerging industry in China. According to statistics from EVI (Electric Vehicles International), until the end of 2015, the global sale of electric vehicle has reached 950 thousand, among which the battery electric vehicle has surpassed 350 thousand with the total fund rebate of 16 billion [4]. As for China, the total sale is coming to 490 thousand, which occupies almost 30% of the global [5]. The electric vehicle is playing a magnificent role in the future.

So in the following section, we will discuss why we choose the electric vehicle industry and what's its special requirements; how do we use CERIF to respond the challenge for the data-driven decision support and how do we carry out our data management. At last, we propose a list of summary for the application of CERIF in the electric vehicle decision support activities. We would like to draw your attention to the fact that it is not possible to modify a paper in any way, once it has been published. This applies to both the printed book and the online version of the publication. Every detail, including the order of the names of the authors, should be checked before the paper is sent to the Volume Editors.

2 The Data Management Requirement of EV Decision Supporting

Electric vehicles has an important position in China's high-tech industry development. It has taken for almost ten years to develop the electric vehicles technology in China. During the twelfth five-year-plan, it encounters the important strategic opportunity period for the EV R&D and industrialization. In order to fulfill the series of technology innovation indicators, cultivate the EV market, the decision management of EV project needs to be more precise, more comprehensive and more predictive.

2.1 The Strategic Decision Requires to Integrate Multi-sourced Information

In the 12th five-year-plan of electric vehicles, it adheres to the "three vertical and three horizontal" layout research structure [6]. The strategic decision has becoming more complex involving with technology roadmap, market development, policy and regulations, funding scale, etc. From the experience of Japan NDEO's best practice and US evidence-based decision [7], we need to establish the subject database, which shall collect the data produced in the full life cycle and data from the stakeholders in the downstream industries.

2.2 The Process Management Requires More Technology Frontier Detection and More Middle Output Monitoring

In the procedure management of nowadays EV project management, node control is the normal management requirement. There are apparent drawbacks to slowly respond the lots of project in-time output. Meanwhile, the technology frontier is becoming faster and difficult to detect timely. So the funding organizations and the administrative decision-makers found it was necessary to find a decision data mall so as to support the in-time decision-making and collect the feedbacks after the projects are released.

2.3 The EV Experts Requires More Kinds of Data to Sustain the Decision Support Activates

The expert advisory board for the MOST electric vehicles are from universities, research institutions and enterprises in different branch department. Those experts main aim is to provide high-level advices for the EV strategic direction. Some of their advices are based on the survey, some are academic research results, some are products characteristics, some are city promotion statistics and some are patents mining. So it's found that experts cannot be familiar with every technology fields. They require more evidence to offer decision-making efficiently and comprehensively.

In this section, we can find the requirements from the EV decision-making are involving more data from the start point of project funding, to organizations and researchers carrying out and to all kinds of outcomes including patents, journal papers, products, etc. The demands of decision-makers in the EV fields needs to be fulfilled in a comprehensive way. After a thorough survey, the CERIF model comes to meet the challenge and localize according to Chinese EV data management.

3 The Application on EV Data Management and EV Decision Supporting

CERIF was under draft of version 1.5 while the EV pilot decision supporting was initiated three years ago. Now the CERIF 1.6 has been released [8]. A formal data model is described by CERIF standard which enables interoperability between research management systems and contains information about people, projects, organisations, publications, patents, events, prizes, equipments.

The CERIF entities are divided into the groups. The core group contains Project, Person and Organization entities. Entities ResultPublication, ResultaPatent and ResultProduct are the result entities group and these entities contain metadata about scientific research results. Also, there are CERIF 2nd entities which can be linked with CERIF core and result entities. The semantic layer enables classification of entities and relations between entities according to some classification scheme. Other entities of the CERIF data model are linked with semantic layer through the cfClass entity [9].

In the project of electric vehicles decision support system, all CERIF data are stored in the Microsoft SQL database management system. The first application version was

finished by October, 2013. Until now, about ten versions have been improved in order to facilitate the EV data management including EV journal, conference paper, dissertations, patents, products, organisations, persons and sale datasets from 8 different sources with different formats (Fig. 1).

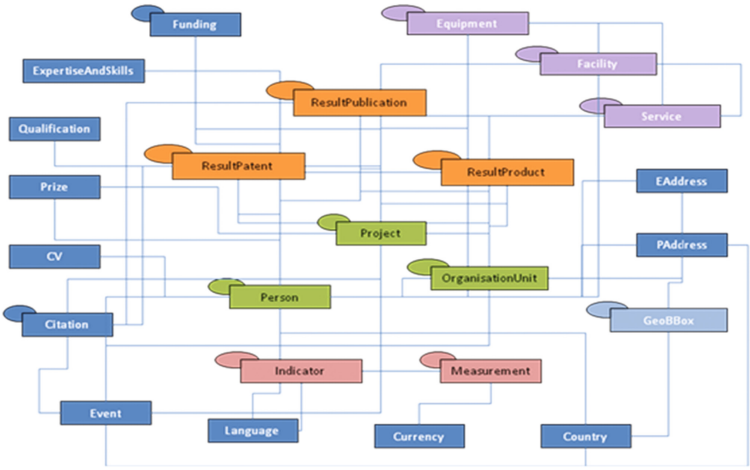


Fig. 1. CERIF 1.5 entities and their relationships

3.1 The Principles of CERIF Application

(1) To be consistent

All data elements are introduced from CERIF data model. Additional are descried by the international or national standards accordingly. Also in the revised versions, the data elements shall be consistent.

(2) To be extensible

Under the framework of CERIF, it shall be extended as the EV data reality. For example. The Result product entities has been extended into two sub-entities, ResultEVModels, ResultEVkeparts. In the projects, lots of extensions have been made to fulfil the expectation.

(3) To be operable

Since the CERIF has been improved and updated by experts in a wide variety of EU, it is quite different from the China's practice. The CERIF model needs to be operational under local circumstances. Take the example of data element granularity, the local version has taken into account to fit Chinese current practice.

3.2 The Generic Mapping of CERIF to Electric Vehicle Data

The electric vehicles data in the pilot system consists of 19 different data object. Some of them, such as journal paper, conference, has the mature structure. Others, such as EV keyparts, EV sale datasets and demo city datasets, are building their own data elements and shall be "sliced and diced" frequently according to aspects of requirements. So the CERIF can be generically translated and matched into the EV conceptual logic shown as Table 1.

Table 1. The CERIF data model matching to electric vehicle data table.

Groups	Entities	Remarks
Base entities	Person	
	Project	
	OrganisationUnit	Including research institutions and enterprises
Result entities	ResultPatent	
	ResultProduct	Including EV models; keyparts;
	ResultPublication	Including EV policies; journals; conferences; dissertations; monograph; standards; web news; S&T reports
2nd entities	Funding	
	Qualification	
	ExpertiseAndSkills	
	Prize	
	CV	
	Event	
	Language	
	Currency	
	Country	
	EAddress	
	PAddress	
	Citation	
Indicators and measurements	Indicator	
	Measurement	
Infrastructure entities	Equipment	
	Facility	
	Service	
Geographic entities	Geobbox	
Semantics entities	Term and vocabulary	

As for the semantic layer of CERIF, 1264 EV terms have been formed to describe the EV subject in the project. In addition, for the needs of mappings EV data model entities the cfClass entity is used for classification of scientific domains and categories of data sets. Finally, it sums up 9 kinds of 65 relationships to describe the linkage of entities.

Furthermore, the classification schema covers some Chinese classification, such as criteria for classification of national economy trade (GB/T 4754-2011), Chinese library classification, International Classification for Standards, International Patent Classification, etc. The above classification do help the CERIF data model to localize and fit to the EV decision support system.

3.3 The CERIF Data Model Application—Taking the Example of Project Entities

The project entity is one of the base entities. Usually, the Research managers in funding organisations need know the detailed description, such as ProjectID, ProjTitle, ProjSubTitle, ProjAbstr, ProjKeyWord, ProjectAcro, ProjectURI, Classification, Funding and DCStatus, which offer basic description for the project.

In addition, the decision-makers need to be clear about the complex linkage among projects. So the link entities of project cover all possible aspects including project-project relationship, project-organization relationship, project-person relationship in the base entity level shown in the Fig. 2; project-Resultpublicaiton relationship, project-Resultpatent relationship, project-Resultproduct relationship, project-equipment relationship, etc. shown in the Fig. 3.

Furthermore, the logical model is designed for further description. In the process of application, several extensions has been made. (1) cfClass entity is used for Chinese S&T project classification and the semantic term has been revised accordingly. 134 semantic terms has been updated into the term catalogue. Such as the project organizing unit. (2) the subject indicators and measurements has been updated, especially some indicators related to outputs of the research or the impact of the research, such as the outputs transformation indicators. (3) several data elements have been

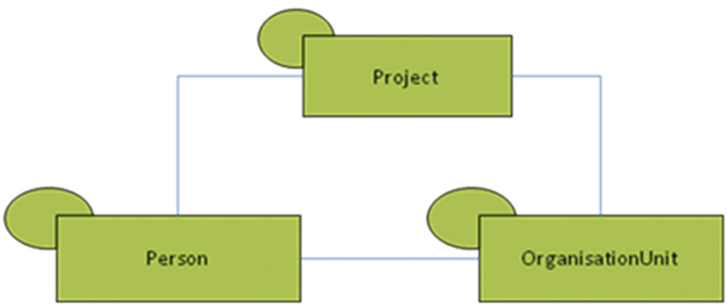


Fig. 2. The base entities

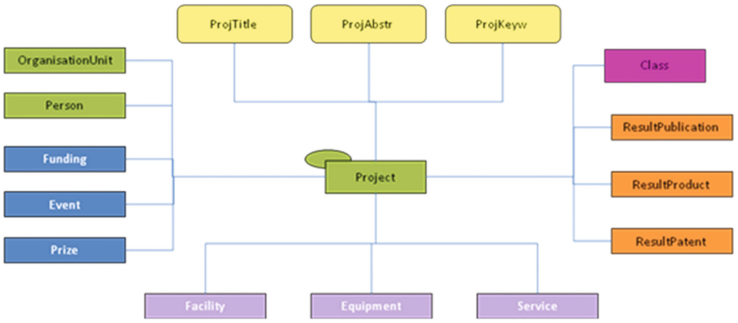


Fig. 3. Project link entities

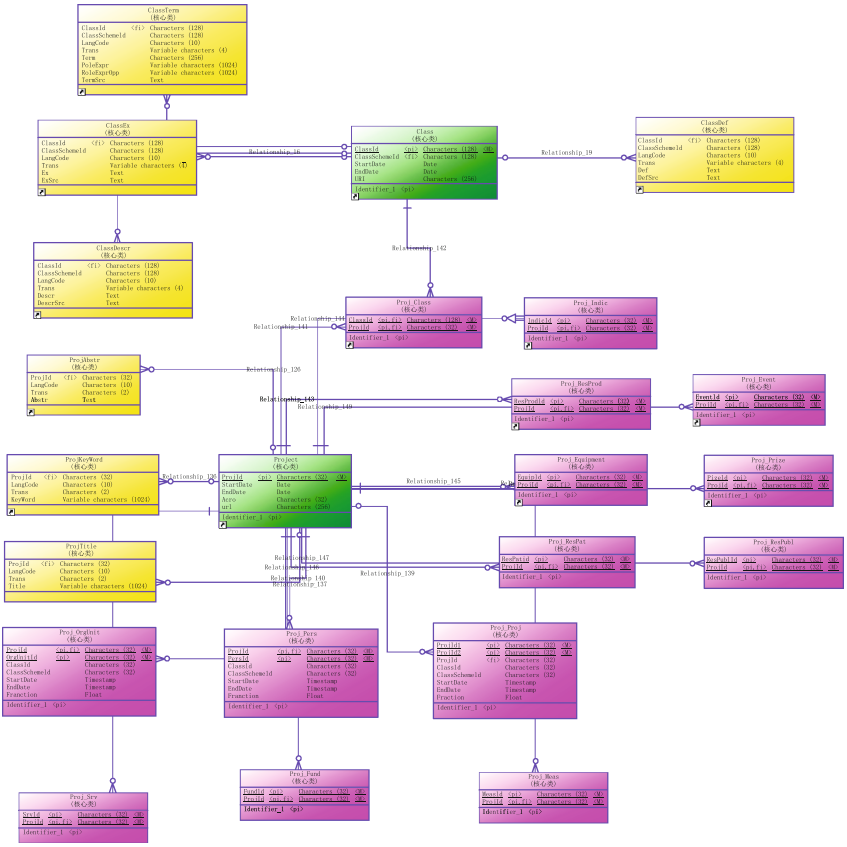


Fig. 4. CERIF example taken as project in the logical level

modified to meet the requirement from the EV project administrative branch, such as Fundingsource has been added to illustrate the matching fund from the bank or the local governments (Fig. 4).

4 Conclusion

It is feasible that CERIF data model can be introduced into the electric vehicle decision support system with its extensions to link different entities and dataset. Also, it's obvious that CERIF data model is semantically rich and can be used to describe the detailed description of electric vehicle R&D activities.

Through the CERIF introduction, we found that Policy and decision makers use the data to monitor research activities, identify strong and weak points, define strategic priorities and make informed decisions on funding allocation at the macro level. The researchers can easily get access to the electric vehicle subject database without searching different portals and the advisory experts use it in a similar way to find the 360° scenario of the interested fields or targeted experts or organizations due to the strong link between different entities from CERIF.

Acknowledgments. The authors would like to acknowledge the support from The National Key Technology R&D Program of the Ministry of Science and Technology of China (Grant No. 2015BAH25F02).

References

1. Pinto, C.S., Simões, C., Amaral, L.: CERIF – Is the Standard Helping to Improve CRIS? Procedia Comput. Sci. **33**(01), 80–85 (2014)
2. An introduction to CERIF. http://www.ukoln.ac.uk
3. CERIF for Datasets. https://cerif4datasets.wordpress.com
4. Global EV Outlook 2015. http://www.iea.org
5. Speech by wan gang in the China EV100 2016. http://auto.sina.com.cn
6. Energy saving and new energy automobile industry development planning (2012–2020). http://www.gov.cn
7. The science of science policy roadmap - The White House. http://www.whitehouse.gov
8. CERIF-1.6. http://eurocris.org/cerif/feature-tour/cerif-16
9. Bailo, D., Jeffery, K.G.: EPOS: a novel use of CERIF for data-intensive science. Procedia Comput. Sci. **33**(33), 3–10 (2014)

On Bypassing Page Cache for Block Devices on Storage Class Memory

Jin Baek Kwon(✉)

Department of Computer Science,
Sun Moon University, Asan, Republic of Korea
jbkwon@sunmoon.ac.kr

Abstract. The class of memory technologies with best features from memory and storage are called storage-class memory or SCM. In the hybrid usage model of SCM, it is physically attached to the memory bus just like DRAM but it is logically shown as a block device just like a storage. In this paper, we question the effectiveness of the page cache in the hybrid model, because SCM has the read and write performance comparable to DRAM. Therefore, I implemented a way to bypass the page cache in the Linux kernel, and show the effectiveness the page cache by thorough experiments with various file I/O benchmarks.

Keywords: Disk cache · Nonvolatile memory · Operating systems · Page cache · Storage class memory

1 Introduction

Resistive memory technologies such as phase change memory (PCM) [1], and spin-transfer torque RAM (STT-RAM) [2] has received attention for recent years due to their persistency, byte-addressability, low latency close to DRAM, and high density [3]. The class of memory technologies with best features from memory and storage are called storage-class memory or SCM. This class of techniques have become one of the major issues in the memory systems area. Since the SCM is placed on the memory bus just like a DRAM, CPU can directly access via the load and store instructions. Studies on SCM are mainly on enabling SCM in the computer systems from architecture perspective [4–7] and exploiting SCM from system or application perspective [8–11].

However, those work is not trivial to deploy in practice. The current operating systems have been evolved on the base of two-level memory hierarchy of a fast, volatile, small, and byte-addressable main memory, and a slow, persistent, large, and block-addressable storage. However, the premise does not fit well to SCM, which is a fast, large, persistent, and byte-addressable memory. Thus, directly adopting SCM into the computer systems requires that operating systems should be almost redesigned. And, any sudden changes will meet strong resistance in industry due to the large amount of intellectual property associated with the traditional systems. In addition, it is

This work was supported by the Sun Moon University Research Grant of 2014.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_59

very difficult to make users change their heavily tuned systems or rewrite existing applications. Therefore, Chen et al. took an evolutionary approach to adopt the SCM technology into the computer systems by proposing a hybrid model, where the logical architecture is separated from the physical one [12]. They devised a block device driver, called *pmbd* (persistent memory block device), which directly accesses SCM attached to the memory bus and provides a logical block I/O interface to users. However, since the I/O mechanism in current operating systems still assume much slower block devices than main memory, there is a limitation to exploit the benefit of SCM.

The page cache or disk cache has been used to mitigate the big difference between the two memories and hide the slow access to the secondary storages, such as hard disks or flash memory. However, when it comes to the secondary storage would SCM, it's a different story. In this paper, I study the effect of the page cache in the hybrid model, where users access SCM as a block device and the SCM is on the memory bus. I modified the kernel to enable to bypass the page cache for the I/O requests from or to SCM. I conducted the experiments with *FileBench* [13] to show the effect.

2 Bypassing Page Cache for Data on SCM

2.1 Overview of Page Cache

Figure 1 shows the typical I/O stack of operating systems. As shown in the figure, the application requests a disk I/O to the kernel through a systems call such as read or write provided through VFS. As a general buffered I/O, it first tries to access data in the page cache. If it does not exist in the page cache, it is processed through the block I/O layer. After the I/O is completed, the data is stored in the page cache and it is copied to user mode buffers of the application. For example, suppose that application *A* reads a file *F* 1,024 bytes at a time and stores it in a buffer allocated to the heap. First, *A* requests the kernel to read 1,024 bytes from the beginning of F by read() system call. The kernel looks into a 4 KB chunk containing the requested data from the page cache. If it is not

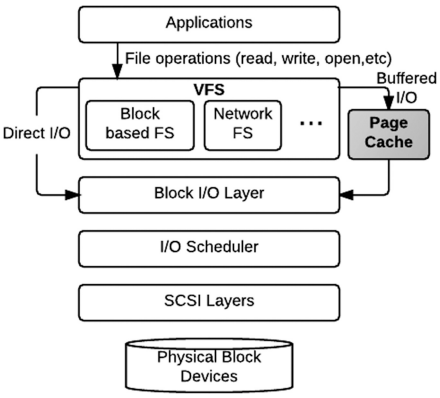


Fig. 1. Block I/O Stack

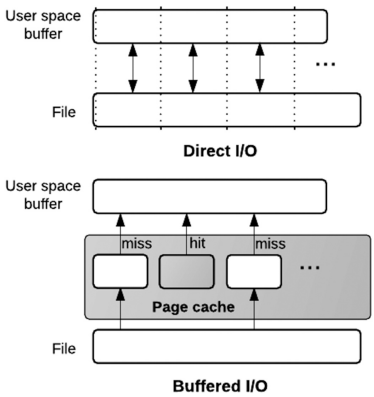


Fig. 2. Direct I/O vs. Buffered I/O

present in the cache, the kernel allocates a page frame to the page cache and requests disk I/O for the 4 KB chunk, which corresponds to offset 0 of file *F*. When the I/O for the page is finished, the kernel copies 1,024 bytes requested by *A* from the page cache to the user buffer and ends the system call. Of course, the block I/O is skipped if the chunks containing data requested was present in the page cache. In the case of Direct I/O, block I/O occurs directly in the buffer of the application without going through the page cache (Fig. 2).

In buffered I/O, if the requested data does not exist in the page cache, the data is copied twice, from the disk to the page cache and from the page cache to the user space buffer, and additionally the page cache management cost is added, On the other hand, since Direct I/O only takes one data copy from the disk to the user buffer, Direct I/O is advantageous compared to the buffered I/O on the cache miss, but the hit time gain is substantial and the hit rate for most of applications is generally quite high. Buffered I/O contributes significantly to system performance, except for special applications, e.g., DBMS.

2.2 Page Cache Bypass: Design and Implementation

Compared with DRAM, the access speed of HDD is about 10^6 times, and the efficiency of page cache is enough to offset the cost of page cache management. That is, the gain of I/O omission according to the cache hit is very high. Now, can we recognize the effectiveness of page caching for SCM in our hypothesized hybrid model? Since the access speed of SCM is close to that of the DRAM, it is clear that the efficiency of the page cache would be lowered, and under the special conditions, the overhead of the page cache would exceed the profit. That is, it can be better to bypass the page cache in the I/O path for SCM. In this section, I present and implement a way to bypass the page cache in the Linux kernel.

As mentioned in Sect. 2.1, the kernel provides a way to bypass the page cache via Direct I/O. Since Direct I/O only transfers data directly between the application buffer and the block device, it can eliminate the page cache overhead as well as the memory copy from the page cache to the user buffer. However, there are limitations to using Direct I/O. First, the application's data access must be aligned to the sector boundaries of the block device. Second, the start address of the user buffer storing the data to be read from or written to the block device must also be aligned at the sector boundaries. That is, the applications should be implemented such as to use Direct I/O if the file is on SCM. Since it is not transparent to the applications and then it requires a careful consideration from the developers. Thus, I suggest a way to bypass the page cache at kernel level to be transparent to the application.

If the above conditions of Direct I/O are satisfied and the request file is on the SCM, the page cache can be bypassed just by performing Direct I/O regardless of whether O_DIRECT is specified or not. If the condition for executing Direct I/O is not satisfied, two data copies through the kernel buffer are inevitable. Nevertheless, the advantage of bypassing the page cache is obvious. The page cache keeps all the pages read to and written from the block device on a page-by-page basis. Suppose that the page cache is full and the entire file is sequentially read and then closes it. These read operations

cause a large page swapping. If the pages being replaced are the *dirty* pages, large number of write operations to the backing store occur, which would not have occurred if we bypassed the page cache. Moreover, if the replaced pages belong to a file on the hard disk, the write cost would be very high, and since the new cached pages are not read again, there is no benefit from the page cache. Since the future hit ratio of the SCM data is uncertain and the hit gain is not large unlike HDD, the benefit of page cache bypass is clearly present. In addition, through bypassing the page cache for files on SCM, the more data chunks from files on HDD can occupy more space in the cache, and in turn we can expect the higher efficiency of page cache.

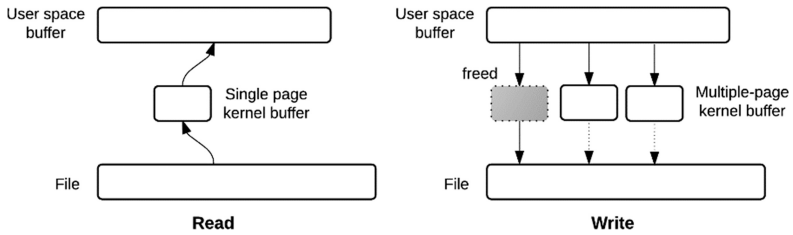


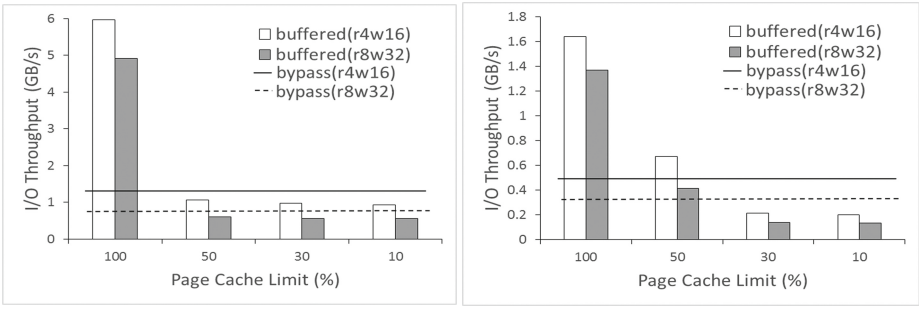
Fig. 3. Kernel Buffers for Page Cache Bypass

Figure 3 shows the basic concept of our page cache bypass with kernel buffers. A read operation uses a single page frame as a buffer to read page by page from the block device, and copies each page to the user buffer. If the application has requested an *n*-page read, the kernel reads one page at a time from the SCM into the kernel buffer and then to the user buffer. It repeats it *n* times. A single page buffer is sufficient because it does not keep pages to be read as in the page cache. To implement a write operation without the page cache, a write-through policy should be used. Theoretically, it can be implemented as a fixed single buffer as the page cache bypass for reading. However, because of subtle synchronization problems in practice. I implemented it as a dynamic single page buffer, where a new page frame is allocated as a buffer for each page write, and immediately returned after writing to the backing store.

3 Experimental Results

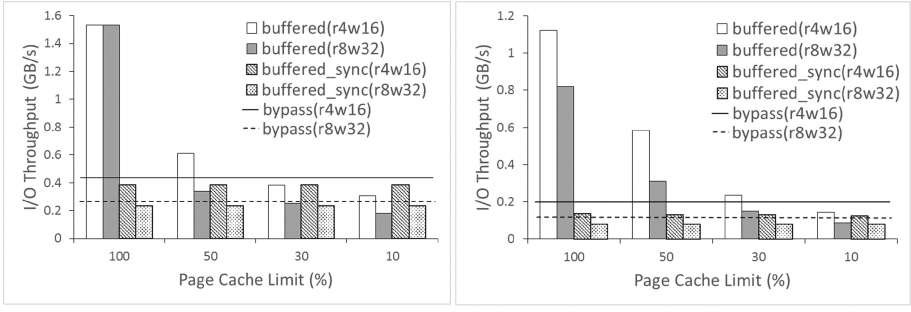
For the experiments, I used a well-designed DRAM based SCM simulator, *pmbd* [12], which is a hybrid model simulator. A portion of DRAM is reserved as an SCM, and read and write speed can be configured in proportion to those of DRAM. If the memory space is ample, it is almost always advantageous to utilize the page cache. For a fair evaluation of the page cache, we should consider a situation where the system load is large and therefore applications compete for the page cache. To simulate such a situation, I applied Lameter's patch to our kernel [14] so as to limit the ratio of page cache. I conducted the experiments on HP Compaq 8200 Elite with Intel *i*5 processor and Linux 3.2.1 with 12 GB of DRAM. 4 GB of DRAM is set as main memory and 8 GB is set as an SCM area. I/O performance test was performed using *Filebench* [13].

In order to show the basic I/O performance of our page cache bypass, depicted as "bypass" in figures, I measured the read and write throughput of sequential and random accesses to a file of 2 GB for ten seconds, while reducing the maximum of page cache size. And I used two configurations about the read and write latencies of SCM: the configuration four times slower than DRAM for reading and 16 times slower for writing, depicted "r4w16" in figures, and the configuration eight times slower for reading and 32 times for writing, depicted "r8w32". Figure 4 shows the read throughput for the buffered I/O and the page cache bypass. Under the condition competing to the page cache space, the page cache bypass performs better for both configurations. Since DRAM is 4 GB and the file is 2 GB, the contention occurs when the page cache ratio is 50%. Without a size limit for the page cache, the entire file is cached by the first read, and then all read occurs only in the cache, so the read performance is very high. Particularly, in the sequential reading, when the competition for the cache starts, the hit ratio becomes zero, so that the bypass without the cache overhead is more efficient. Figure 5 shows the results on the write operation with the additional experiments on synchronous writes ("buffered_sync"). As expected, synchronous writes do not take advantage of the cache, but rather perform worse than the PCB due to the overhead. The other of the results for the writing are similar to the reading.



(a) Sequential Read (b) Random Read

Fig. 4. Read Throughput (a) Sequential Read (b) Random Read



(a) Sequential Write (b) Random Write

Fig. 5. Write Throughput (a) Sequential Write (b) Random Write

4 Conclusion

I studied the effect of the page cache in the hybrid model of SCM by implementing the page cache bypass with small in Linux kernel. Through the experiments, it is demonstrated that bypassing the page cache performs better under a page cache contention. As a future work, I am planning additional experiments on the macro benchmarks, through which I will suggest the favorable conditions of the page cache bypass and present an adaptive scheme for it.

References

1. Simpson, R.E., Fons, P., Kolobov, A.V., Fukaya, T., Krbal, M., Tominaga, J.: Interfacial phase change memory. Nat. Nanotechnol. **6**, 501–505 (2011)
2. Chen, E., Apalkov, D., Diao, Z., et al.: Advances and future prospects of spin-transfer torque random access memory. IEEE Trans. Magn. **46**(6), 1873–1878 (2010)
3. Xue, C.J, Zhang, Y., Chen, Y., Sun, G., Yang, J.J., Li, H.: Emerging non-volatile memories: opportunities and challenges. In: Proceedings of the International Conference on Hardware/Software Codesign and System Synthesis (CODES + ISSS) (2011)
4. Dhiman, G., Ayoub, R., Rosing, T.: PDRAM: a hybrid PRAM and DRAM main memory system. In: Proceedings of the Design Automation Conference (DAC) (2009)
5. Qureshi, M., Srinivasan, V., Rivers, J.A.: Scalable high performance main memory system using phase-change memory technology. In: Proceedings of the International Symposium on Computer Architecture (ISCA), Austin, TX (2009)
6. Meza, J., Chang, J., Yoon, H., Mutlu, O., Ranganathan, P.: Enabling efficient and scalable hybrid memories using fine-granularity DRAM cache management. IEEE Comput. Archit. Lett. **11**(2), 61–64 (2012)
7. Lee, B.C., Ipek, E., Mutlu, O., Burger, D.: Architecting phase change memory as a scalable DRAM alternative. In: Proceedings of the International Symposium on Computer Architecture (ISCA), Austin, TX (2009)
8. Bailey, K., Ceze, L., Gribble, S.D., Levy, H.M.: Operating system implications of fast, cheap, non-volatile memory. In: Proceedings of the Workshops on Hot Topics in Operating Systems (HotOS) (2011)
9. Mogul, J.C., Argollo, E., Shah, M., Faraboschi, P.: Operating system support for NVM + DRAM hybrid main memory. In: Proceedings of the Workshops on Hot Topics in Operating Systems (HotOS) (2009)
10. Meza, J., Luo, Y., Khan, S., Zhao, J., Xie, Y., Mutlu, O.; A case for efficient hardware/software cooperative management of storage and memory. In: Proceedings of the Workshop on Energy-Efficient Design (WEED) (2013)
11. Moraru, I., Andersen, D.G., Kaminsky, M., Tolia, N., Ranganathan, P., Binkert, N.: Consistent, durable, and safe memory management for byte-addressable non-volatile main memory. In: Proceedings of the Conference on Timely Results on Operating Systems (TRIOS) (2013)
12. Chen, F., Mesnier, M.P., Hahn, S.: A protected block device for persistent memory. In: Proceedings of the Symposium on Mass Storage Systems and Technologies (2014)
13. Tarasov, V., Zadok, E., Shepler, S.: Filebench: a flexible framework for file system benchmarking. login: USENIX Mag. **41**(1), Spring 2016
14. Lameter, C.: Limit the size of the page cache. LWN.net, January 2007. https://lwn.net/Articles/218890

Applying Tensorflow with Convolutional Neural Networks to Train Data and Recognize National Flags

Hoang Huu Duc and Keechul Jung(✉)

Media Department, Soongsil University, Seoul, Korea
duc93t@gmail.com, jungkeechul@gmail.com

Abstract. In the recent years, machine learning and deep learning has been becoming hot research titles. In many human life's fields, AI takes big roles like auto driving a car, automatically working robots or vacuum cleaner using image recognition techniques. Tensorflow is a machine learning system with open source code was introduced and provided by Google on November 9, 2015. It has been being famously used in images recognition field. In our work, we recognize an image and classify it using tensorflow based on Convolutional Neural Networks (CNNs) and determine what it is. We train 5-layers CNNs by supervised learning from a database. After training process, trained data files are generated. In the next steps, we use this data to recognize input image and classify it. Finally, we test the results by a testing program.

Keywords: Tensorflow · CNN · Deep learning

1 Introduction

Images recognition using deep learning is now very popular and widely applied into AI system. To work like a person, a machine needs to recognize object via it's shape by catching its images in the around environment [1–4].

Although Google co-operation has distributed open source of tensorflow, but how to apply this potential resource and how to use tensorflow with high effect is not ease.

Firstly, we introduce tensorflow and how to apply it to develop your own software. In the URL: "https://www.tensorflow.org", there are open source samples introducing how to recognize an indicated image with the particular ratio. If you haven't yet download their trained data file, the program will download necessary trained data and show out the name of given image with ratios of five highest score things- default with the "cropped_panda.jpg" to recognize [5–8].

Just change this indicated file to classify your own images one by one. The better way is that re-programming to read arguments and run with your own options via arguments (Figs. 1 and 2).

In this case, not only mortarboard, but also the infant is the main object of the image, or the background with flower is important, too. That mean, this program trained data have some problems, maybe input data must be things only. You must work wrong with their

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_60

```
IPython console
Console 1/A
Python 3.5.2 |Anaconda 4.2.0 (64-bit)| (default, Jul  5 2016, 11:41:13) [MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 5.1.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.

In [1]: runfile('D:/1ImageRecog/ImageRecognition_original.py', wdir='D:/1ImageRecog')
giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca (score = 0.89233)
indri, indris, Indri indri, Indri brevicaudatus (score = 0.00859)
lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens (score = 0.00264)
custard apple (score = 0.00141)
earthstar (score = 0.00107)
C:\Program Files\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:2889: UserWarning: To exit: use
'exit', 'quit', or Ctrl-D.
  warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)
An exception has occurred, use %tb to see the full traceback.

SystemExit


In [2]:
```

Fig. 1. "cropped_panda.jpg" and the recognition result.

```
IPython console
Console 1/A
  warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)
An exception has occurred, use %tb to see the full traceback.

SystemExit


In [2]: runfile('D:/1ImageRecog/ImageRecognition_graduated-infant.py', wdir='D:/1ImageRecog')
mortarboard (score = 0.95902)
academic gown, academic robe, judge's robe (score = 0.03893)
envelope (score = 0.00002)
bonnet, poke bonnet (score = 0.00002)
velvet (score = 0.00001)
C:\Program Files\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:2889: UserWarning: To exit: use
'exit', 'quit', or Ctrl-D.
  warn("To exit: use 'exit', 'quit', or Ctrl-D.", stacklevel=1)
An exception has occurred, use %tb to see the full traceback.

SystemExit


In [3]:
```

Fig. 2. "graduated-infant.jpg" and the recognition result.

fix trained data and trained definitions. Sometime, it works wrongly with another topic of the images.

Secondly, to address above problem, in the URL:

"https://www.tensorflow.org/how_tos/image_retraining", they perform the re-training progress with five classes of flowers (daisy, dandelion, roses, sunflowers, tulips) and then recognize the type of an indicated flower. The default image is "tulip.jpg". You can change you data in the "flower_photos" folder. Add or remove a class of flowers by adding/removing a subfolder contains flowers (named by this subfolder) in the main folder "flowers". Each subfolder in "flowers" is the label for a class of flowers and flower's file name has no meaning. In this program, although you can change your training data, but you still could not change the structure of the trained data's structure. This program concentrates to flowers only due to get better result although it limits the input topic (flowers only). That mean, the recognition's result gives out the unique one.

Actually, if you download the source files from that link, you couldn't run it. By the default, you should download data to "/tmp" folder from both URL:

'http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz'

'http://download.tensorflow.org/example_images/flower_photos.tgz'. Because that, in training process, this example uses data from those two source (Fig. 3).

```
In [1]: runfile('D:/1ImageRecog/ReTrainFlowers1.py', wdir='D:/1ImageRecog')
Looking for images in 'daisy'
Looking for images in 'dandelion'
Looking for images in 'roses'
Looking for images in 'sunflowers'
Looking for images in 'tulips'
Creating bottleneck at d:/huuduc/bottleneck\roses\10090824183_d02c613f10_m.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\102501987_3cdb8e5394_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\10503217854_e66a804309.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\10894627425_ec76bbc757_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\110472418_87b6a3aa98_m.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\11102341464_508d558dfc_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\118974357_0faa23cce9_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\11944957684_2cc806276e.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12165480946_c4a3fe182d_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12238827553_cf427bfd51_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12240165555_98625b1e88_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12240303_80d87f77a3_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12240577184_b0de0e53ea_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12243068283_ee4c2683e2_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\123128873_546b8b7355_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12323085443_8ac0cdb713_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12338444334_72fcc2fc58_m.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12395698413_c0388278f7.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12406229175_82e2ac649c_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12434194695_a7c4e73c6b_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12562723334_a2e0a9e3c8_n.jpg.txt
Creating bottleneck at d:/huuduc/bottleneck\roses\12572786553_634868f7f2_n.jpg.txt
```

Fig. 3. Re-train the input data with 5 classes of flowers.

Actually, if you download the source files from that link, you couldn't run it. By the default, you should download data to "/tmp" folder from both URL:

'http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz'

'http://download.tensorflow.org/example_images/flower_photos.tgz'.

Because that, in training process, this example uses data from those two source.

In our work, we collect sets of flags separated by nation's feature to prepare for training data step. We store whole of them in "flags" folder with sub folders are named by nation's name. Then we run the training program to create trained data (Fig. 4).

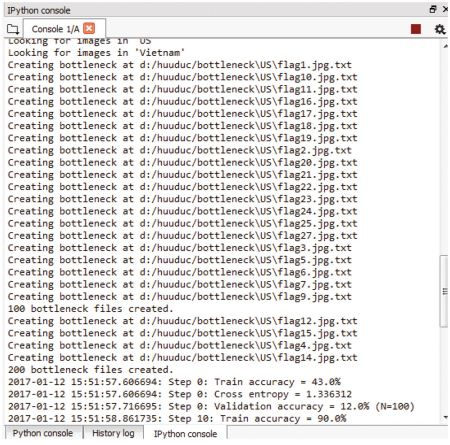


Fig. 4. Re-train the input data with classes of nation flags.

2 Training Procedure and Architecture of CNN

First of all, we resize the mage to 2048-dimensional vector for each image. We describe the details of the training procedure and the precise network architecture. The input to a convolutional layer is an 2048 × 2048 × r image where 2048 is the height and width of the image and r is the number of channels (e.g. an RGB image has r = 3). The convolutional layer have k filters (kernels) of size n × n × q where n is smaller than 2048 and q can either be the same as the number of channels r or smaller (Fig. 5).

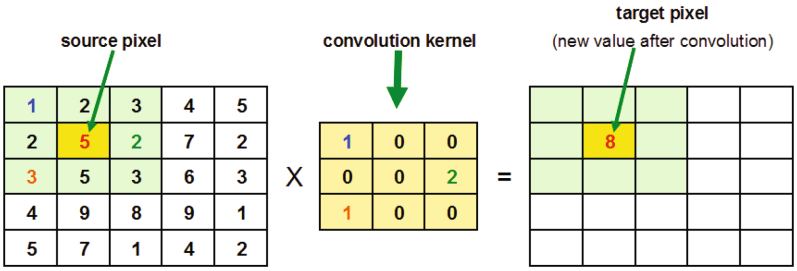


Fig. 5. A kernel operates on one pixel.

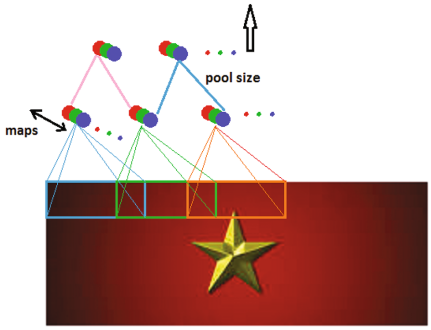


Fig. 6. One layer of the CNN in the pooling stage. Units of the same color have tied weights and units of different color represent different filter maps.

A CNN contains a number of convolutional and subsampling layers followed by fully connected layers shown in Fig. 6. The numbers in the kernel of CNN represent the amount by which to multiply the number underneath it. The number underneath represents the intensity of the pixel over which the kernel element is hovering.

3 Illustrations

We temporary use sample images downloaded from Internet and separate into nation's subfolders in the "flags" folder. Each nation's folder has at least a minimum number of flags to confirm the basic training input (we set minimum number = 20) files to make sure that the calculating accuracy is not too low. If there is a subfolder inside "flags" but it contains less than 20 image files, this folder will be ignored. We use the nation's name to label for each set of flags relating to that nation. The name of subfolders in "flags" folder are very important, they represent for the nation's name as label roles.

During convolution, the center of the kernel passes over each pixel in the image. The process multiplies each number in the kernel by the pixel intensity value directly underneath it. The final step of the process sums all of the products together, divides them by the amount of numbers in the kernel, and this value becomes the new intensity of the pixel that was directly under the center of the kernel.

While training images, the resources spent for processing is limited but the necessary GPU and CPU is as much as the number of images, training steps, convolutional layers. So, avoiding bottleneck constrains is very important, if we do not resolve this before training, the resources will quickly exhausted and system will be halted. In our program, we create bottleneck file for each image first and store in the given folder called "bottleneck". When reading an image, if the related bottleneck file is not existed, it will be created. On the country, skip this step and process the image with exist bottleneck file. Of course, at the first time, all images need their new bottleneck files; when you add some new images, only those new related bottleneck files are necessary to be created. In case you want to re-create all bottleneck files, just remove/delete the "bottleneck" folder- as shown in Fig. 7.

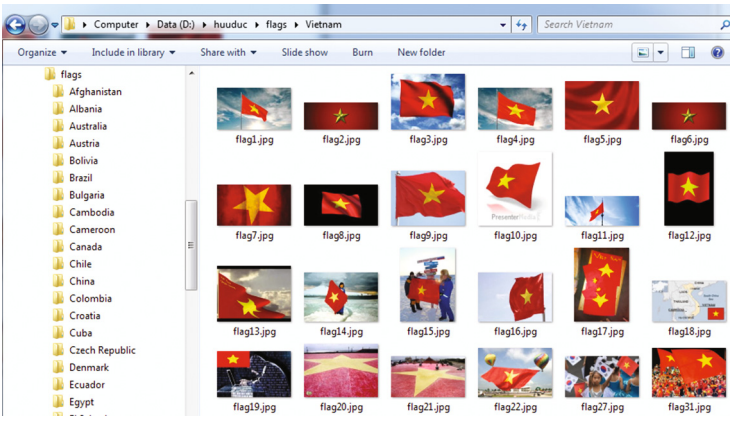


Fig. 7. Structure of folders and files for training data.

4 Results

In the training process, the program always displays the training accuracy and the validation accuracy of each step. We train many times with different numbers of training step. The more training step, the more time it takes. After required steps, the final test accuracy show the training accuracy (Figs. 8 and 9).

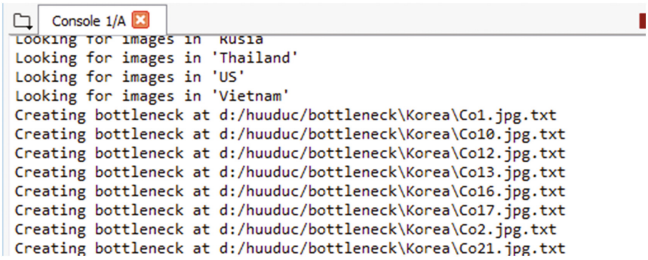


Fig. 8. Creating bottleneck files process.

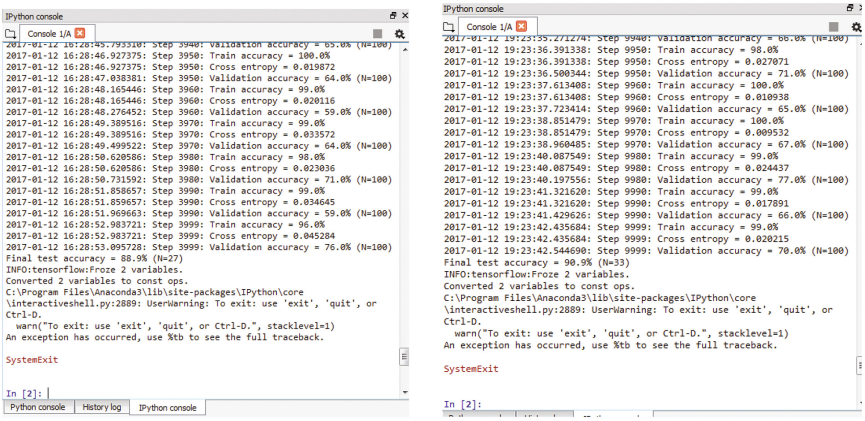


Fig. 9. The different results of training using 4000 steps and 10.000 steps

There are two main factor influent to the training accuracy:

(1) Quantity and quanlity of input images;
(2) Number of training steps.

We have run many times with different number of training steps, with less than 4000 steps, the results have large distance with different paces. From about 3000 steps and over, there are little bit better results when we increase number of steps.

5 Discussion

First of all, we suggest a method to take advantages from applying tensorflow into deep learning and machine learning field. This system is provided by Google co-operation who is very famous in images processing and recognizing with a largest world-wide network with free-email system and social network system. In our industrial century, image recognition has a very big role in automatic system and robotic machine. Although many methods and researches had been provided about images recognition but this trend is still attractive; time by time, many newer and newer intelligent accessories and devices being created. So, how to apply AI- specially images recognition technique into a new system or device more effect is the massage we want to contribute.

Secondly, with our work, it is very useful to specify the relation of any picture or image to a nation. In advance, we need a better tool to resolve the problem that: "is there a flag in a picture?" first. The picture can be from any source: video clips, social network, Internet, paper news… After that, we can specify what nation flag it is. With this result, combine with social network like facebook, twitter, LinkedIn… we can connect any events of social network user via their picture with some main information: What nation they are, the seen or event relate to what country.

Finally, using deep learning in image recognition is very useful with high effect. Combine with tensorflow to build new of your own application or integrate into your devices/machine can have a big advantage from open source. This can be more easy and cheap method not only in research but also in producing intelligent devices.

We hope our work can be useful apply widely into many intelligent devices. By this way, many people can develop their own software and integrate into many AI devices in the future.

References

1. Abelson, H., Sussman, G.J., Sussman, J.: Structure and Interpretation of Computer Programs. MIT Press, Cambridge (1985)
2. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M.A., Senior, A., Tucker, P., Yang, K., Le, Q.V., Ng, A.Y.: Large scale distributed deep networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1223–1231. Curran Associates, Inc., Red Hook (2012)
3. Gelly, S., Silver, D.: Combining online and offline learning in UCT. In: 17th International Conference on Machine Learning, pp. 273–280 (2007)
4. Kocsis, L., Szepesvári, C.: Bandit based Monte-Carlo planning. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) Machine Learning: ECML 2006, pp. 282–293. Springer, Heidelberg (2006)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
6. Sutskever, I., Nair, V.: Mimicking go experts with convolutional neural networks. In: Kůrková, V., Neruda, R., Koutník, J. (eds.) Artificial Neural Networks-ICANN 2008, pp. 101–110. Springer, Heidelberg (2008)
7. Website: Computer Science Department of Stanford University (2016). http://ufldl.stanford.edu
8. Website (2016). https://www.tensorflow.org

An Empirical Study on the Effect of the Interpretability of Metaphors in UI on the Learnability of Mobile Apps

Wonjin Jung[1,2] and HyungRok Yim[1,2(✉)]

[1] The School of Business and Economics, Dankook University, 152, Jook-Jun-Ro, Soo-Ji-Goo, Young-In, Kyung-Ki-Do 16890, Korea
jintari@hanmail.net, hryim@hanyanag.ac.kr
[2] The School of Business, Hanyang University, 222, Wangsimli-Ro, Sung-Dong-Goo, Seoul 04763, Korea

Abstract. Mobile devices, such as smartphones and tablet PCs, have evolved continuously from the time when they debuted in the late 1990s. At the same time, the structure and usage of mobile applications have also become increasingly complex. As a result, it is often found to be difficult to understand the user interface (UI) of applications. In addition, the low interpretability of metaphors in UIs makes the problem worse. These conditions and user environments inhibit smooth learning of applications. Accordingly, it can be inferred that the low interpretability of metaphors is expected to eventually negatively affect the learnability of applications. However, prior studies in the information systems (IS) field have not shown much interest in the effect of the interpretability of metaphors in UIs of mobile applications on the learnability of the applications. The main research goals of this study are as follows: (1) to examine the effects of the interpretability of metaphors in UIs of mobile applications on the mental model of users of the applications and on the learnability of the applications, and (2) to find the effect of the mental model of users on the learnability of the applications. The data was collected through a survey and structural equation modeling (SEM) was used for the analysis. The results showed that the interpretability of metaphors has significant effects on the mental model of users as well as on the learnability of applications.

Keywords: Mobile · Application · Interpretability · Mental · Model

1 Introduction

A User Interface (UI), the part of a system through which users interact with the system, consists of diverse design components to support the seamless interaction of the user with the system [1, 3, 6]. Any study about UI designs of applications for mobile devices such as smartphones and tablet PCs are bound to address the simplicity of the design [7]. The UI of mobile applications are significantly affected by the environmental and physical conditions of the mobile devices because the applications run on the mobile devices. Since the small displays of mobile devices cannot show much information, the UI of applications for mobile devices has to limit the use of

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_61

texts, words, content, and other various design components as much as possible. By doing so, UIs can realize minimalism to maximize user convenience. When too many texts, words, graphics, icons, and unnecessary design components occupy the small display space of mobile devices, the readability of the UI will decrease. Consequently, it would also be difficult to precisely discern and understand the meanings of the design components. Therefore, UI designs of applications for mobile devices require simplicity. Some researchers have stated that simplicity is one of a UI's necessary design attributes for a system. According to the associated studies, simplicity has a significant impact on the usability of a system, including interactions with the system [2, 7].

Meanwhile, metaphors are frequently used in the UI of applications for mobile devices to realize UI design simplicity. Metaphors refer to literary schemes or figurative speech, which can be represented by tools, used to express concepts or ideas. They are easily found throughout the UIs of computers, information systems, mobile applications, or even home electronic appliances. Wastebaskets, floppy diskettes, binoculars, scissors, and carts are good examples [5, 11]. Using such metaphors frees a UI from hundreds of thousands of texts and words, which in turn, results in the utilization of the small display space [2]. In addition, users can intuitively understand the implied information or functions represented by the metaphors. Accordingly, users can learn, interact, and use the devices and applications more easily, quickly, and precisely; thereby, users would perceive higher usability of the devices and applications.

Many empirical studies in e-learning fields have addressed that the ambiguity of metaphors in the UI of applications affects learning. Many researchers have clarified that high interpretability of metaphors in the UI of applications enables users to learn the use of the applications easily and, furthermore, to improve academic performance [11]. Similarly, other researchers have also claimed that educational applications need to be equipped with highly interpretable metaphors to ensure the interactivity and success of the system. To support the argument, they explained that the interpretability of metaphors in a UI enables users to quickly and accurately formulate their mental model of the applications, which is the knowledge of and experience in the use of the applications [6].

Rouse and Morris [8] mentioned that a mental model can be viewed as a cognitive structure comprised of specific knowledge and experience. Similarly, Johnson-Laird [4] defined it as a kind of internal representation of complicated external reality. Rouse and Morris [8] also noted that individuals understand their surrounding world with their mental models. These views imply that mobile application users are able to not only understand, but also learn and use applications based upon their mental models of the applications.

Furthermore, these views also imply that if the metaphors in the UI of an application are well designed, then the users of the application can easily formulate high quality mental models of the application. If their mental model of the application is formed accurately, then it can make the user better able to learn the usage of the application, consequently leading to a positive attitude toward using the application and intention to use the application. On the other hand, a poorly developed mental model of an application makes users confused, which leads to the reluctance of learning the application. If users do not want to learn the application, then they will not use it.

Based upon the prior research, it may be possible to predict that ensuring the interpretability of metaphors in the UI of mobile applications will also significantly affect not only the mental model of users of applications, but also the learnability of the applications. In addition, it is also possible to predict that the mental model of users of applications has an effect on the learnability of the applications. However, it is difficult to guarantee these effects and relationships. This is because the types and usage of metaphors in UIs are quite diverse. Various metaphors are used and utilized in the UIs of mobile applications.

A comprehensive IS literature review found that there is little empirical research on the effects of the interpretability of metaphors in the UI of applications on the mental model of users and the learnability of the applications in the context of mobile applications. Therefore, it is necessary to examine these effects and relationships. The main research goals of this study are as follows: (1) to examine the effects of the interpretability of metaphors in the UI of mobile applications on the mental model of users of applications and on the learnability of the applications, and (2) to find the effect of the mental model of users on the learnability of the applications. This study will attempt to verify these effects and relationships through the following hypotheses.

Hypothesis 1: The interpretability of metaphors in UI of mobile applications positively affects the learnability of the applications.
Hypothesis 2: The interpretability of metaphors in UI of mobile applications positively affects the mental model of users in the use of the applications.
Hypothesis 3: The mental model of users in the use of mobile applications positively affects the learnability of the applications.

2 Research Methodology, Data Analysis, and Results

This study empirically explored the effects that the interpretability of metaphors in the UI of mobile applications has on the mental model of users and the learnability of the applications. This study also examined the effect of the mental model of users on the learnability of mobile applications. Thereby, the mental model of users was considered a mediating variable in this research model. A survey was conducted to collect data. A total of 258 university students and practitioners participated in the survey. One hundred and seventy two students (66.7% of the participants) answered all of the questions. The participants majored in economics, business administration, computer science, and graphic design at three universities in Korea. The gender ratio of the participants was 53.9% male to 46.1% female. The majority of the participants (78.7%) were in their twenties. The application types that the participants used most frequently, just before the survey, were social networking and communications-related applications (61.7%).

This study employed Structural Equation Modeling (SEM) to analyze the proposed hypotheses and used SPSS statistics and AMOS ver. 18 as the statistical software. Path coefficients were examined to check the causal effects between the variables. To do so, the significance of the relationships between the variables in the proposed model was analyzed. As predicted in the hypotheses, interpretability had a significant influence on

Table 1. Hypothesis test

	Paths	Coeff.	Stand. Coeff.	P value	Results
H1	Interpretability → Learnability	.270	.295	.0001	Accept
H2	Interpretability → Mental Model	.283	.307	.0001	Accept
H3	Mental Model → Learnability	.319	.357	.0001	Accept

the mental model of users ($\beta = .307$, $p < .0001$). The mental model of users also had a positive impact on learnability ($\beta = 357$, $p < .0001$). Finally, interpretability also had a direct significant effect on learnability ($\beta = 295$, $p < .0001$). Therefore, the hypotheses H1, H2 and H3 were all supported. Table 1 shows the results of the test of the research model.

3 Discussion and Conclusion

In short, this study empirically explored the relationship between the interpretability of metaphors in the UI of mobile applications and the learnability of the applications. Not only did the study explore this relationship but it also examined the effects that the interpretability of metaphors in UIs had on the mental model of users and the effects the mental model of users had on the learnability of mobile applications. The results showed a significant direct effect of the interpretability of metaphors on the learnability of applications. The results also showed a significant effect of the interpretability of metaphors on the mental model of users of the applications. In addition, the mental model of users had a positive impact on the learnability of applications. These results addressed the importance of the interpretability of metaphors in the UI of mobile applications and in understanding and explaining the factors that influence intention to use mobile applications. Especially, this study addressed how important it is for mobile applications to use interpretable, simple metaphors in UIs.

The Technology Acceptance Model (TAM) proposed in the IS field has been widely referred to in understanding the user's attitudes associated with the acceptance of information technologies [9, 10]. Specifically speaking, the ease of use, one of the variables in the TAM formed on the basis of the learnability of an information technology, determines user's attitudes towards the information technology. The user's attitudes affect the user's behavioral intentions, that is, the acceptance of the information technology. Looking at studies on the behavioral intentions of users in mobile application fields, the impact of the perceived ease of use on the behavioral intentions to use applications was identified. Therefore, it can be inferred that mobile application users form their attitudes and intention to use applications on the basis of experiences with the applications. When developing mobile applications, practitioners may use the findings of this study to improve and enhance users' experiences with mobile applications, not only by improving the interpretability of metaphors in the UI, but also, the learnability of applications.

References

1. Bostrom, R.P., Olfman, L., Sein, M.K.: The importance of learning style in end-user training. MIS Q. **3**(1), 101–119 (1990)
2. Black, M.: More about metaphor. In: Ortony, A. (ed.) Metaphor and Thought. Cambridge University Press, Cambridge (1988)
3. Chu, C., Chan, B.K.: Evolution of Web Site Design: Implications for Medical Education on the Internet. Comput. Biol. Med. **28**, 470–472 (1998)
4. Johnson-Laird, P.N.: Mental models. In: Posner M.I. (ed.) Foundations of Cognitive Science. MIT Press, Cambridge (1989)
5. McWilliam, G., Dumas, A.: Using metaphor in new brand design. J. Mark. Manage. **13**, 265–284 (1997)
6. Moran, T.: An applied psychology of the use. ACM Comput. Surv. **13**, 1–12 (1981)
7. Moshagen, M., Thielsch, M.T.: Facets of visual aesthetics. Int. J. Hum.-Comput. Interact. **68**, 689–709 (2010)
8. Rouse, W.B., Morris, N.M.: On looking into the black box: prospects and limits in the search for mental models. Psychol. Bull. **100**, 349–363 (1986)
9. Schmidt, K.E., Liu, Y.L., Sridharan, S.: Webpage aesthetics, performance, and usability: design variables and their effects. Ergonomics **52**, 641–643 (2009)
10. Szajna, B.: Empirical evaluation of the revised technology acceptance model. Manage. Sci. **42**(1), 85–92 (1996)
11. Wolfe, C.R.: Plant a tree in cyberspace: metaphor and analogy as design elements in web-based learning environments. Cyberpsychol. Behav. **4**, 67–76 (2001)

A Density-Aware Data Encryption Scheme with Query Auditing Index for Secure Mobile Services

Moon-Hwan Kang, Min Yoon, and Jae-Woo Chang(✉)

Dept. of Computer Engineering, Chonbuk National University,
Jeonju, Republic of Korea
{calvin, myoon, jwchang}@jbnu.ac.kr

Abstract. In database outsourcing, because a service provider might be untrusted or compromised, two issues of data security emerge: data confidentiality and data integrity. Motivated by these issues, we propose a density-aware data encryption scheme and a query processing algorithm with query result auditing for secure mobile services. To guarantee the data confidentiality, our density-aware data encryption scheme utilizes a grid index to transform the original data space into a bitmap signature. In addition, to reduce the transmission overhead of verification data, we propose a query result authentication index that stores an encrypted signature for each anchor, which is the concatenated hash digest of clustered data. Through performance evaluation, we show that the proposed scheme guarantees the high level of privacy preservation to users while providing better query processing performance, compared with the state-of-the-art schemes.

Keywords: Data encryption · Query authentication · Database outsourcing

1 Introduction

In database outsourcing, because a service provider might be untrusted or compromised, two issues of data security emerge: data confidentiality and data integrity [1]. First, for the sake of data protection in outsourced databases, there have been few encryption techniques studied. Among them, spatial technique is widely used. In the spatial transformation techniques, M.L. Yiu et al. [2] proposed a Flexible Distance-based Hashing (FDH) that re-distributes the location space with randomly selected anchors and encrypts each groups by using bitmap representation. This technique boots up the computation cost of data encryption by encrypting each divided area through bitmap. However, this method has a critical drawback that FDH technique divides the areas with anchors that are randomly selected, so that data can be skewed towards partial areas. So, if an attacker is able to find out the skewed distribution of partial area, then he/she can easily reveal real points as well.

On the other hand, in order to verify both data correctness and completeness, query authentication should be provided in database outsourcing. For this, the verification information is sent to users with query result so that the result can be verified by using

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_62

data owner's signature. Most popular data authentication research is bucket-based authentication approach. The bucket-based authentication approach was proposed J. Wang et al. [3] where a bucket contains a bucket id, a data range (upper-lower bound), a checksum, and the number of tuples in a bucket. In bucket-based index, security lies in the anonymity of tuples within the same bucket. The larger the bucket size, the less the information disclosure and more secure the bucket index is. In bucket-based authentication, the authentication index contains lower and upper range of each bucket. Thus, applying existing methods to the database is valuable to the data distribution disclosure attack.

Motivated by the problems, in this paper, we propose a new density-aware data encryption scheme and a query processing algorithm with query auditing. Our density-aware data encryption scheme can guarantee both the data confidentiality and the query result integrity of mobile data. First, to solve the original data leakage problem, we devise a bitmap-based encryption scheme that selects anchors by using a data histogram. It means that many anchors can be chosen in dense area, while relatively less number of anchors can be chosen in sparse area. Based on a grid index, our scheme can construct clusters with relatively the same number of POIs, by splitting or merging a cluster area with a given threshold. Secondly, to reduce the transmission overhead of verification data, we devise a query result authentication index that stores an encrypted signature for each anchor and compares the anchor signature with the verification data from the data owner. Hence, we can reduce data transfer overhead for checking the query integrity while enhancing the data privacy.

The rest of this paper is organized as follows. In this section, we briefly review the related work. In Sect. 2, we propose a density-aware data encryption scheme and our query result auditing scheme. An empirical evaluation is presented in Sect. 3. Finally, Sect. 4 concludes our work with future research direction.

2 Density-Aware Data Encryption Scheme with Query Authentication

Figure 1 illustrates the architecture of our spatial database outsourcing system with three main components: a mobile user (i.e., authorized user), a location-based service provider and a database owner. The overall data processing is performed as follows. In pre-processing step, the DO creates a secret encryption key, and then applies the encryption method to converts the original point set P into a transformed points set P`. Then, the transformed dataset is sent to the SP. At this time, the DO sends the encryption key to the trusted users over a secure communication channel, e.g., SSL. In the query processing step, a user (U) issues a query q to SP. For this, U encodes q to q` by using the secret key from DO and sends q` to the SP. The SP evaluates q` over P` and returns the candidate results to U. Finally, U decrypts the encrypted query results and selects the genuine query results.

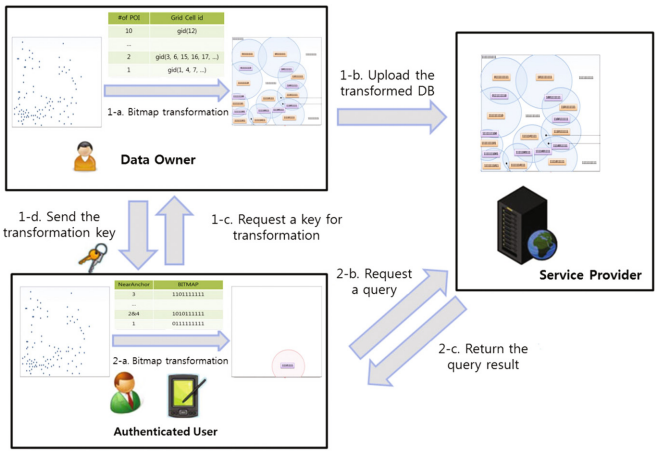


Fig. 1. System architecture

2.1 A Density-Based Data Encryption

Our algorithm is processed in four phases; anchor selection, cluster optimization, bitmap encryption, and index construction. In the anchor selection phase, we use a grid-based histogram to select anchors according to the data distribution. In the cluster optimization phase, all clusters can be adjusted for storing data into its clusters uniformly. For protecting the original data from the attacker, our algorithm transforms the clusters into a bitmap in the bitmap encryption phase. In the index construction phase, a hash-based index is constructed to support efficient kNN query processing.

2.1.1 Phase 1: Anchor Selection

The main problem of the existing method [2] is that the anchors can be chosen in a specific area because the existing method selects anchors randomly. If an attacker is able to find out the skewed distribution of the specific area, he/she can easily infer the real points. To solve this problem, we propose an anchor selection algorithm using a grid-based histogram to consider the data distribution. Our anchor selection algorithm is performed as follows. First, the original dataset P is inserted in the $n \times n$ grid index. P is represented as $<id, x, y>$, where id means an identifier of the original data, x and y are the coordinates of the original data. Secondly, the algorithm counts the number of data within each grid cell for making a histogram, and sorts the cells in the descending order of data counts. Thirdly, by using the histogram, the data owner computes the number of anchors based on Eq. (1), and then randomly selects anchors from the cell. If a cell does not satisfy the threshold, the minimum number of data for a cluster, the algorithm expands the cell to its nearest cells to satisfy the threshold. The threshold can be given by a user or a system. The latter one is calculated by using Eq. (2).

$$AnchorRatio = \frac{(\#\ of\ Anchor) \times (\#\ of\ Data\ in\ Anchor\ Range)}{\#\ of\ data\ in\ P} \tag{1}$$

$$Threshold = \frac{\#\ of\ Anchor}{\#\ of\ data\ in\ P} \tag{2}$$

2.1.2 Phase 2: Cluster Optimization

Even if we perform the density-aware anchor selection phase, the data can be skewed in some clusters. To solve this problem, we propose a merge and split algorithm to store the data into clusters uniformly. Our algorithm is processed as follows. First, our algorithm calculates the merge and split thresholds of all clusters which are constructed from phase 1. Second, if the number of data in a cluster is greater than the split threshold, the algorithm divides the cluster into two groups. In order to split the cluster, the farthest point from the cluster anchor is chosen as an anchor for the new cluster. Third, if the number of the data in a cluster is lower than the merge threshold, the cluster should be merged with the nearest cluster. To merge two clusters, the midpoint between two anchors is calculated to select a new anchor. The new anchor becomes a point which is the nearest to the midpoint. The algorithm repeats until all clusters satisfy the merge and split thresholds.

2.1.3 Phase 3: Bitmap Encryption

Let the selected anchor objects be a_1, a_2,…, a_i, a_n. For each anchor object a_i, we need to calculate a distance value r_i, which represents the cover range of the anchor. Given an object $p \in$ P, we convert the coordinates of the data point into an A-length bitmap where the ith bit of the bitmap is defined in [2]:

$$Bitmap(p)[i] = \begin{cases} i = 0,\ if\ dist(a_i, p) \le r \\ i = 1,\ otherwise \end{cases} \tag{3}$$

Once the bitmaps for all anchors are generated, we build a bitmap-based index for query processing. In order to enhance the performance of query processing, we devise a hash index based on algebraic coding for retrieving the nearest anchor bitmap rapidly. The proposed hash index uses a bit array as the coefficient of a polynomial. The address of a hash table is calculated based on the prime number that is close to the hash table size (Eq. 4). We divide the bit array value by the prime number and assign its remainder as a hash table address. In Eq. 4, bit means an anchor bitmap and *Hsize* indicates a hash table size. Hence, we insert the bitmap of anchors into the hash table by applying their bitmap to the proposed hash function. The query processing performance can be improved because we can directly access the hash table at the query time.

$$hashAddress = bit\% \left(1 + \sum_{m-1}^{2^{Hsize}} \left[\sqrt[Hsize]{Hsize} \left(\sum_{x=1}^{n} \left[\cos^2 \pi \frac{(x-1)! + 1}{x} \right] \right)^{\frac{-1}{n}} \right] \right) \tag{4}$$

The algorithm for the bitmap-based hash index construction is as follows. First, we calculate the hash addresses of the generated anchor bitmaps. Secondly, the anchor bitmap is inserted into the hash table.

2.2 Range Query Processing Algorithm with Result Authentication Index

Our k-NN query processing algorithm efficiently retrieves a query result on the encrypted data. However, in database outsourcing, it is also import to ensure the originality of result data that is genuine and complete. In order to provide query result authentication, we devise a result auditing index in this section. Based on the cluster information generated during data encryption phase, we generate a private data authentication index in order to provide privacy-preserving range query processing. Each data group is signed by the data owner using Condensed-RSA [4] with data ids within the group. By this means, the private authentication index is generated without revealing the partitioning information to unauthorized accesses. At query processing time, our method employs a novel technique for searching nearest data groups from a query point, in order to maximize the utility of the transformed data. In the literature, the Hamming distance measure has been employed for approximate NN search. Once the nearest anchor group is retrieved, the service provider sends the encrypted data within the group and its signature. Since the query user was given the transformation key and signature function from the data owner, he/she can decrypt the result data and generates signatures of them. If the generated signature is identical to the signature from the service provider, the client confirms that the query result is correct and genuine.

After constructing a data partition, a signature for each small-group is generated by using Condensed-RSA. An RSA signature is computed on the hash of the input message. Let *h()* denote a cryptographically strong hash function (such as SHA-1) which takes a variable-length input and produces a fixed-length output. For an input message *m*, *h(m)* denotes the hash of m and a standard RSA signature on message *m*, and computed by using a given formulation from E. Mykletun et al.'s work [5]. In this scheme, each tuple signature is generated by using RSA with its id (e.g., POI id). Condensed-RSA signature is computed as a product of individual signatures.

3 Performance Analysis

In this section, we present the extensive experimental results of our encryption scheme. For this, we compare our encryption scheme with the existing FDH in terms of data encryption time, degree of data distribution, and query processing time. For the evaluation of the query result authentication scheme, we compare the proposed algorithm with S. Balpande et al.'s algorithm [5]. In our experiment, we use four spatial datasets: Uniform (100,000 points), Gaussian (100,000 points), Skewed (100,000 points) and the real dataset of Northern East America (NE) containing 119,898 point of interests (POIs). We did our performance analysis on the Window 7 Enterprise K system with Intel Core2 Quad 2.4 GHz and 4 GB memory.

Figure 2 shows data encryption time with varying the number of anchors. The data encryption time can be increased as the number of anchors increases. This is because our scheme generates clusters by using a grid index while the existing scheme uses a tree structure which has the expensive cost to build in large data. In case of Gaussian and Skewed data, the encryption time of our scheme is higher than Uniform and Real data as shown in Fig. 2(b). This is because for densely populated data, merge and split operations are performed many times to store the data in a uniform way.

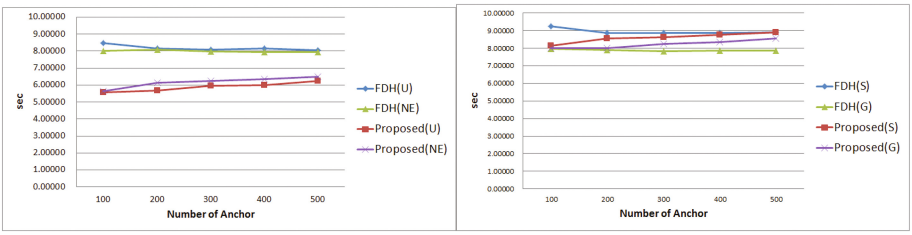
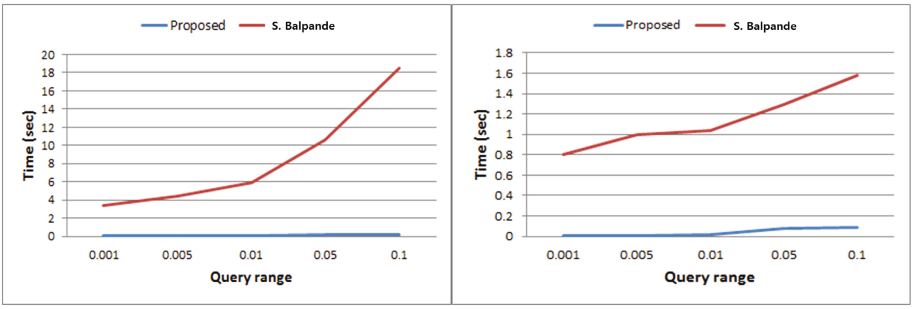


Fig. 2. Data encryption time with data distribution

Figure 3(a) describes the query processing times with varying query range. From the result, it is proven that our algorithm outperforms the existing work up to 15 times, in terms of query processing times. This is mainly because our algorithm reduces the data transmission and verification overheads by using bit operation which easily calculates candidate anchors within query range. Figure 3(b) shows the query result auditing time with varying query range. For auditing the query result, the client decrypts the result data and generates a signature for comparison. Therefore, the performance of this term is closely related to the number of result data. As shown in the figure, the auditing time is increased as the query range increases.



(a) Query processing time (b) Query result auditing time

Fig. 3. Query processing and auditing time

4 Conclusion and Future Work

In this paper, we propose a data encryption scheme that groups data with anchors and transforms them into bitmap information. Through performance evaluation, it is shown that proposed method outperforms the existing method up to 15 times while providing similar performance in returning number of false positive data. And, to reduce the transmission overhead of verification data, we devise a query result authentication index that stores an encrypted signature for each anchor and compares the anchor signature with the verification data from the data owner.

As a future work, we will extend the proposed algorithm to support variety of query types (e.g., range skyline queries).

Acknowledgement. This work was partly supported by the Human Resource Training Program for Regional Innovation and Creativity through the Ministry of Education and National Research Foundation of Korea (NRF-2016H1C1A1065816). Also, 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 number 2016R1D1A3B03935298).

References

1. Pathak, A.R., Padmavathi, B.: Analysis of security techniques applied in database outsourcing. Int. J. Comput. Sci. Inf. Technol. **5**(1), 665–670 (2014)
2. Yiu, M.L., Assent, I., Jensen, C.S.: Outsourced similarity search on metric data assets. IEEE Trans. Knowl. Data Eng. **24**, 338–352 (2010)
3. Wang, J., et al.: Bucket-based authentication for outsourced databases. Concurrency Comput. Pract. Experience **22**(9), 1160–1180 (2010)
4. Mykletun, E., Narasimh, M., Tsudik, G.: Authentication and integrity in outsourced databases. J. ACM Trans. Storage **2**(2), 107–138 (2006)
5. Balpande, S., et al.: Data integrity and confidentiality in outsourced database. In: International Conference and Workshop on Recent Trends in Technology (TCET) (2012)

A Kernel Density Estimation Model for Moving Object Detection

Yulong Qiao(✉) and Wei Xi(✉)

College of Information and Communication, Harbin Engineering University, 145 Nantong Street, Harbin 150001, China
{qiaoyulong, savexi}@hrbeu.edu.com

Abstract. Moving object segmentation is an important component of many vision systems, especially in the non-static background. This paper proposes an approach based on Kernel Density Estimation that can handle situations where the background of the scene is not completely static but contains significant stochastic motion (e.g. water). To get the initial results, a higher dimensional KDE model using the observing pixel intensity values and the information of optical flow is built. Then a KDE observing model based on the Hidden Markov Random Field Model and the Expectation Maximization frame work, is used for segmented the moving object. Experimental results show that the proposed approach can accurately detect moving objects and use less video frames.

Keywords: Moving object detection · KDE model · Optical flow · HMRF-EM frame work

1 Introduction

Moving object detection is the task of identifying the physical movement of an object in a given region or area [1]. It is very important in application areas such as visual surveillance, content-based video coding, and human computer interaction. Traditional Approaches for moving object detection can be broadly categorized into four forms as Background Subtraction, Frame Differencing, Temporal Differencing and Optical Flow [2]. Background Subtraction is considered to be one of the most reliable method.

There are two non-parametric estimator approaches, the Gaussian mixture model [3] and the kernel density estimator [4], are used widely. And the improvements based on them are still updating, like [5–8], etc. Even some parametric estimator approaches like [9] are very effective, they need a lot of calculation.

The non-static background subtraction often need a large data to compute. So we proposed an approach base on KDE model in this paper, using less frames, to balance the computational complexity and the accuracy of detection.

1.1 Kernel Density Estimator

As a non-parametric estimator, the kernel density estimator is under appropriate conditions the estimate it produces is a valid probability itself. Let $o_1, o_2, \ldots, o_N$ be a recent

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_63

sample of intensity values for a pixel. If the kernel estimator function is a Normal function $N(0, \Sigma)$, we can estimate the probability density function that this pixel will have intensity value o_t at time t

$$P_r(o_t) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(o_t - o_i)^T \Sigma^{-1}(o_t - o_i)} \tag{1}$$

Where Σ represents the kernel function bandwidth. Let $l_1, l_2, \ldots, l_N$ be the labels of corresponding pixels. We can get them through the threshold $th1$. $l_t = 1$ means that the pixel is object. And $l_t = 0$ means background.

1.2 Bandwidth Estimation

If we use the gray image sequence, the bandwidth will be $\Sigma = \{\sigma^2\}$. we assume that this local-in-time distribution is Normal $N(\mu, \sigma^2)$, so the deviation $|o_{i+1} - o_i|$ is Normal $N(0, 2\sigma^2)$. Then the standard deviation can be estimated as

$$\sigma' = \frac{m}{0.68\sqrt{2}} \tag{2}$$

Where $m = median(|o_{i+1} - o_i|)$, in case that $|o_{i+1} - o_i| = 0$ leads to $\sigma = 0$, we estimate $\sigma = \max(1, \sigma')$.

2 Optical Flow-KDE Model

Although the KDE model is simple and fast, it needs more video frames and the detection is interfered easily by background. So we consider the information of optical flow to improve the model. It is the distribution of apparent velocities of movement of brightness patterns in an image [10].

Let the image brightness be denoted by $E(x, y, t)$ at the pixel (x, y) in the image plane at time t. Considering when the pattern moves, the brightness of a particular pixel is constant. Respectively, we let E_x,E_y and E_t equal the partial derivatives of image brightness with respect to x,y and t. We get the two velocities of optical flow through the equation

$$(E_x, E_y) \bullet (u, v) = -E_t \tag{3}$$

Combining the intensity values $o_1, o_2, \ldots, o_N$, $u_1, u_2, \ldots, u_N$ and $v_1, v_2, \ldots, v_N$, we build a new KDE model in this paper. And the kernel bandwidth equals

$$\Sigma_n = \mathrm{diag}(\sigma_o^2, \sigma_u^2, \sigma_v^2) \tag{4}$$

Let $I_i = [o_i, u_i, v_i]^T$, the density can be written as

$$P_r(I_t) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_n|^{\frac{1}{2}}} e^{-\frac{1}{2}(I_t - I_i)^T \Sigma_n^{-1}(I_t - I_i)} \tag{5}$$

So, a detected pixel I will be considered to be a part of the moving object only if $P_r(y_t) < th2$.

3 Hidden Markov Random Field-KDE Model

Through the optical flow-KDE model we get an initial detection. But it is unsatisfactory. Because the HMRF model is useful in image segmentation [11], we improve it for reducing the interferences from the dynamic background.

In the video sequence, if one of the frames' size is $M \times N$, it has $S = M \times N$ pixels. Let $y_i (i \in S)$ be the intensity value for a pixel [12]. Every pixel has its label $x_i = l$, $l \in L$. In this paper, $L = \{0, 1\}$ means that the pixel is background or object respectively. The label x_i can't be observed directly, it is the hidden state.

Dependent on the parameter set θ and the neighborhood configuration x_{N_i} of x_i, we computer the marginal probability distribution of y_i

$$P(y_i|x_{N_i}, \theta) = \sum_{l \in L} f(y_i; \theta_l) P(l|x_{N_i}) \tag{6}$$

Given $x_i = l$, y_i follows a conditional probability distribution $P(y_i|l) = f(y_i; \theta_l)$. According to the MAP criterion, we seek the estimated label $\hat{x}_i$ of the true label x_i

$$\hat{x}_i = \arg \max_{i \in S} \{P(y_i|x_i)P(x_i)\} \tag{7}$$

3.1 Model Building

Usually the observation field is modeled through Gaussian distribution in the image segmentation. In fact, both the distributions of object and background in the video frame are not the standard Gaussian distribution. But Gaussian mixture model needs much parameters to be estimated. We propose a KDE model to build the conditional probability distribution of the observable random variables

$$f(y_i; \theta_l) = \frac{1}{M_l}\sum_{j=1}^{M_l} \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{-\frac{(y_i - z_j^l)^2}{2\sigma_l^2}} \tag{8}$$

Where M_l is the amount of the label l in one frame, let $z_j^l (j \in M_l)$ be the intensity value for a pixel of label l. Only one parameter $\theta_l = \sigma_l$ is simpler than the Gaussian model and Gaussian mixture model.

The label filed model can be build according to the Hammersley-Clifford theorem [13], a MRF is characterized by a Gibbs distribution equivalently. So

$$P(X = x) = \frac{1}{Z}\exp\left(-\sum_{c\in C} V_c(x)\right) \tag{9}$$

Where Z is the partition function, $V_c(x)$ is the clique potential which C is all possible cliques. To define the cliques, we use the Multi-Level Logistic Model. Considering the spatial and temporal connections, the x_{N_i} shows in Fig. 1.

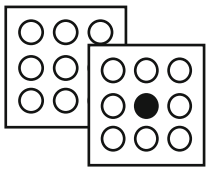


Fig. 1. Let the neighborhood configuration N_i of x_i be modeled using the current frame and previous frame which including 17 points.

So the energy function of a point i can be compute from the posterior probability $p\left(x_j^l|y,\lambda'\right)$ of point $x_j(j \in N_i)$.

$$\begin{aligned} V\left(x_i^l\right) &= -\sum_{j\in N_i} \delta\left(x_i^l - x_j\right) p\left(x_j^l|y,\lambda'\right) \\ p\left(x_j^l|y,\lambda'\right) &= p\left(y_i|x_i^l,\lambda'\right)\exp\left(-\beta_i V\left(x_i^l\right)\right) \Big/ \sum_{l=1}^{L} p\left(y_i|x_i^l,\lambda'\right)\exp\left(-\beta_i V\left(x_i^l\right)\right) \end{aligned} \tag{10}$$

The posterior probability is computed by iterative operation easily. And the parameter set λ includes σ_l and β_i. According the Bayes' theorem $P(y_i|x_i)P(x_i) = P(x_i|y_i)P(y_i)$, where $P(y_i)$ is constant, we estimate the label

$$\hat{x}_i = \arg\max_{i\in S}\{P(x_i|y_i)\} \tag{11}$$

3.2 Parameter Estimation

Expectation-Maximization Algorithm is a parameter estimation method. The E-step calculate the conditional expectation, and the M-step maximize it [14]. We can get

$$\begin{aligned} \sigma_l &= \sum_{i=1}^{S}\left(\sum_{j=1}^{M_l}(y_i - z_j)^2 P(x_i|y_i)\right) \Big/ \sum_{i=1}^{S}\left(\sum_{j=1}^{M_l} P(x_i|y_i)\right) \\ \beta_i &= \operatorname{argmax}\sum_{i=1}^{s}\sum_{l=1}^{L}\left[p(x_i^l/y,\lambda')\log\left(\exp(-\beta_i v(x_i^l)) \Big/ \sum_{l=1}^{L}\exp(-\beta_i v(x_i^l))\right)\right] \end{aligned} \tag{12}$$

4 Results and Conclusion

In order to test the proposed approach, we have used two grayscale video sequences with non-static background. Both the videos have 30 frames. And our approach is compared to the Gaussian mixture model. The threshold in the improved KDE model is the empirical value usually. Let the bandwidth of HMRF model be $\sigma_l = 1$.

Figures 2 and 3 show that the initial results are better than the GMM, but there are a lot of points of background detected falsely. The next step of HMRF model, we get the satisfactory detections showed in Tables 1 and 2. And it proves that we reduce the calculation only using 30 frames. By iterative operation, the interferences of background are decreased step by step. But the points of objects, which were not detected, interfered the error ratios.

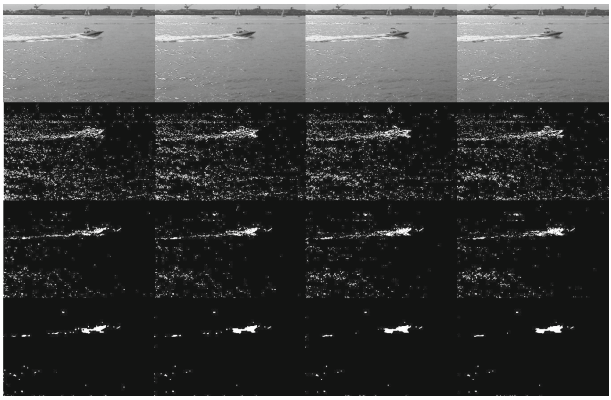


Fig. 2. The four continuous frames' detection of video 'boats'. The first line shows the original images. The second line shows the GMM. The third line shows the initial result. The fourth line shows the final detection.

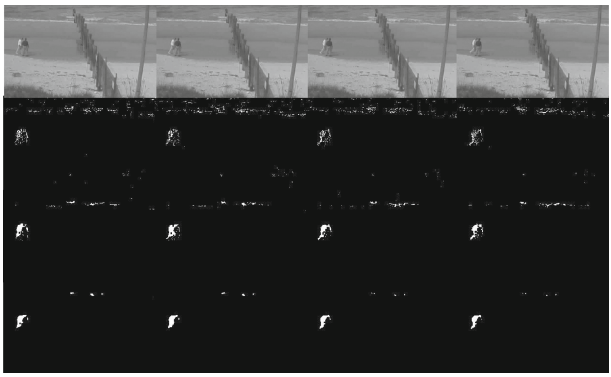


Fig. 3. The four continuous frames' detection of video 'ocean'. The first line shows the original images. The second line shows the GMM. The third line shows the initial result. The fourth line shows the final detection.

Table 1. The detection error ratios of the video 'boats'.

	Frame 1	Frame 2	Frame 3	Frame 4
GMM	2.70%	2.64%	2.44%	2.35%
Initial result	1.35%	1.26%	1.17%	1.22%
Proposed approach	0.70%	0.59%	0.52%	0.49%

Table 2. The detection error ratios of the video 'ocean'.

	Frame 1	Frame 2	Frame 3	Frame 4
GMM	0.43%	0.63%	0.49%	0.39%
Initial result	0.13%	0.13%	0.14%	0.12%
Proposed approach	0.049%	0.054%	0.041%	0.056%

The main shortcoming of the proposed approach is that the object which has a similar value with background be detected hardly. Especially in the Fig. 3, the right human's detection is unsuccessful.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China 61371175 and Fundamental Research Funds for the Central Universities HEUCFQ20150812.

References

1. Kulchandani, J.S., Dangarwala, K.J.: Moving object detection: review of recent research trends. In: International Conference on Pervasive Computing (2015)
2. Shaikh, S.H., Saeed, K., Chaki, N.: Moving Object Detection Using Background Subtraction. Springer, Cham (2014)
3. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 747–757 (2000)
4. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000). doi:10.1007/3-540-45053-X_48
5. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction, vol. 2, pp. 28–31 (2004)
6. Sheikh, Y., Shah, M.: Bayesian modeling of dynamic scenes for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1778–1792 (2005)
7. Bhaskar, P.K., Yong, S.P.: Image processing based vehicle detection and tracking method. In: International Conference on Computer and Information Sciences, pp. 1–5 (2014)
8. Xiang, J., Fan, H., Liao, H., Xu, J., Sun, W., Yu, S.: Moving object detection and shadow removing under changing illumination condition. Math. Prob. Eng. **2014**, 10 (2014)
9. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 171–177 (2009)
10. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
11. Wang, Q.: HMRF-EM-image: implementation of the hidden markov random field model and its expectation-maximization algorithm. Comput. Sci. **94-b**, 222–233 (2012)

12. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging **20**, 45–57 (2001)
13. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. **36**, 192–236 (1974)
14. Piccardi, M., Perez, O.: Hidden markov models with kernel density estimation of emission probabilities and their use in activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (2007)

Detecting Bases of Maximal Cliques in a Graph

Fei Hao[1], Doo-Soon Park[2(✉)], and Zheng Pei[3]

[1] School of Computer Science, Shaanxi Normal University, Xi'an, China
feehao@gmail.com
[2] Department of Computer Software Engineering,
Soonchunhyang University, Asan, Korea
parkds@sch.ac.kr
[3] Center for Radio Administration and Technology Development, Xihua University, Chengdu, China

Abstract. Maximal Cliques Enumeration (MCE), as a fundamental problem, has been extensively investigated in many fields, such as social networks, and biological science and so forth. However, the existing research works usually ignore the formation principle of maximal cliques which can help us to speed up the detection of maximal cliques in a graph. This paper pioneers a novel problem on detection of bases of maximal cliques in a graph. We propose a formal concept analysis based approach for detecting the bases of maximal cliques and detection theorem. It is believed that our work can provide a new research solution and direction for future topological structure analysis in various complex networking systems.

Keywords: Maximal cliques enumeration · Bases · Formal concept analysis · Graphs

1 Introduction

Recent years witnessed the booming development of graph data model and its widely used applications. In practice, many applications can be represented with graph data model, such as social networks, web networks, and protein interactive networks. Therefore, analyzing and mining the useful knowledge from graphs is significant meaningful. In particular, Maximal Cliques Enumeration (MCE) is an important research issue in graphs. In graphs, a clique refers to a complete sub-graph where any two vertices are connected each other. Meanwhile, a maximal clique is a clique, such that there is no clique with more vertices. At present, detection of maximal cliques or

This research was supported by Basic Science Research program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2014R1A1A4A01 007190) and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-H8601-16-1009) supervised by the IITP (Institute for Information & communications Technology Promotion and was also supported by the Fundamental Research Funds for the Central Universities (GK2017 03059). Z. Pei's work was partially supported by National Nature Science Foundation of China (Grant No. 61372187).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_64

MCE is mainly to identify all maximal cliques because these cliques or maximal cliques containing more valued knowledge and information. Thus, MCE is widely used in community detection, topological analysis of web network and so forth.

MCE, as a fundamental problem in graph theory, has been extensively investigated by many researchers [1, 2]. They mainly focus on approximate algorithm devising (since MCE is an NP-hard problem) for extracting all maximal cliques. The existing of algorithm for addressing MCE is categorized as: (1) sequential in-memory algorithms [3, 4]; (2) sequential I/O efficient algorithms [5, 6] which concentrate on reducing the high cost of random disk I/Os for processing graphs that cannot fit in main memory [2]; (3) aiming to reduce the running time, the third type approach is parallel and distributed algorithms [7, 8] are proposed.

However, these existing studies usually ignore the formation principle of maximal cliques. As a matter of fact, the formation principle of maximal cliques is a kind of useful information for better MCE process. That is to say, if we can obtain the partial topological structures which can further form the maximal clique, then this issue can be addressed efficiently. To this end, this paper pioneers to study the bases of maximal cliques in graphs. We firstly present the concept of base of maximal clique, then we point out that the maximal clique can be formed based the detected bases. Therefore, the main contribution of this work are twofold: (1) formalize an interesting problem about detection on bases of maximal cliques in graphs; (2) exploiting the formation procedure of maximal cliques and then present an efficient approach for obtaining the bases of maximal cliques.

The remainder of this paper is structured as follows. Section 2 provides the problem statement. Then, a formal concept analysis based detection approach for bases of maximal clique in a graph is presented in Sect. 3. Section 4 concludes this paper.

2 Problem Statement

To study the problem about detection on bases of maximal cliques in a graph, the graph model and basic concept of maximal cliques are firstly presented and the problem statement is then formalized.

2.1 Graph Model and Maximal Clique

This work focuses on undirected graph and managing the maximal cliques on it; hence graph model here is defined with an undirected graph model G = (V, E), where V is the set of vertices and E is the set of edges of G. Particularly, for as a set of vertices $C \subseteq V$, is a clique if each vertex in C are connected each other. If there is no any other set of vertices $C' \supset C$ such that C' is a clique in G, then C is a maximal clique.

2.2 Problem Descriptions

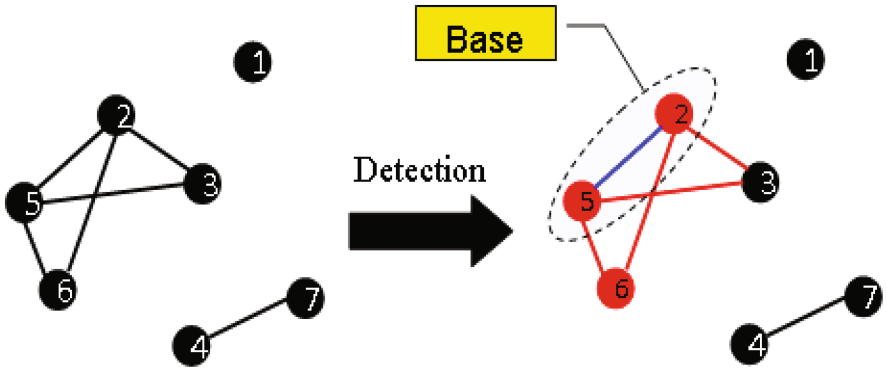
According to our previous work [9], we know that the k-clique community can be formed with skeleton sub-graphs. Similarly, finding the bases of maximal cliques could

help us to detect maximal cliques quickly. Note that, the base of maximal clique refers to the common sub-graph (can be line, can be other sub-graph) among maximal cliques.

Problem Definition*:** Given a graph *G = (V, E)*, this paper proposes a novel topic and the corresponding approach for finding the bases of maximal cliques from *G*, denoted as ***B(maximal_clique(G)).

To better understand the above problem statement, an illustrative example is shown as follows.

Clearly, Fig. 1(a) shows an input of the problem, i.e., graph G, it is composed of 7 vertices. We can easily get the maximal cliques {2, 3, 5}, {2, 5, 6}, {4, 7}, {1} from G using the existing algorithm. However, we found that the common edges {2, 5} between maximal cliques {2, 3, 5} and {2, 5, 6} as shown in Fig. 1(b). Actually, the edge {2, 5} is a base of maximal cliques {2, 3, 5} and {2, 5, 6} since they can be formed by simply adding vertex 3 or 6.



(a) Graph G (b) Base of maximal cliques

Fig. 1. An illustrative example on bases of maximal cliques

3 Detecting Bases of Maximal Cliques Based on Formal Concept Analysis

To address the above problem, this section is devoted to presenting our proposed approach for detecting bases of maximal cliques by using formal concept analysis. First, we analyze the reason for bases of maximal clique for interpreting how maximal cliques can be formed via their bases. Then, a new formal context and its concept lattice are generated by aggregating the attributes which have the common objects. Finally, we extract the extents from the maximal cliques associated concepts and from the new formal concept lattice and then make the intersection.

Detection Approach: Suppose G is a graph, denoted as $G = (V, E)$. We firstly construct the formal context $K = (V, V, I)$ using the approach presented in [9]. Then, a new formal context is induced by granulating some vertices (granulating in terms of equivalence relation R on original attribute) as an attribute, i.e., $K' = (V, VR, IR)$ The K' has the same objects with but the different attributes from the original formal context.

Here,

$$V_R = \{[v]_R | v \in V\},\ I_R = \{(v, [v]_R) \in V \times V_R; \exists n \in [v]_R((v, n) \in I)\}.$$

Let us take Fig. 1(a) as our the input graph G, then, the converted formal context and its induced formal context are constructed as follows (Tables 1 and 2).

Table 1. The formal context of G, K

	1	2	3	4	5	6	7
1	X						
2		X	X		X	X	
3		X	X		X		
4				X			X
5		X	X		X	X	
6		X			X	X	
7				X			X

Table 2. The induced formal context of G, K'

	1	[2,5]	3	[4,7]	6
1	X				
2		X	X		X
3		X	X		
4				X	
5		X	X		X
6		X			X
7				X	

The corresponding concept lattices are generated by existing concept lattice generation algorithm. The relationship between concept lattices of original formal context K, i.e., $C(K)$ and induced formal context K', *i.e.*, $C(K')$ is shown in Fig. 2. According to the findings about the equivalence between equiconcept and clique [9, 10], it is clearly to observe that the maximal cliques in G including {2, 3, 5}, {2, 5, 6}, {4, 7}, {1}.

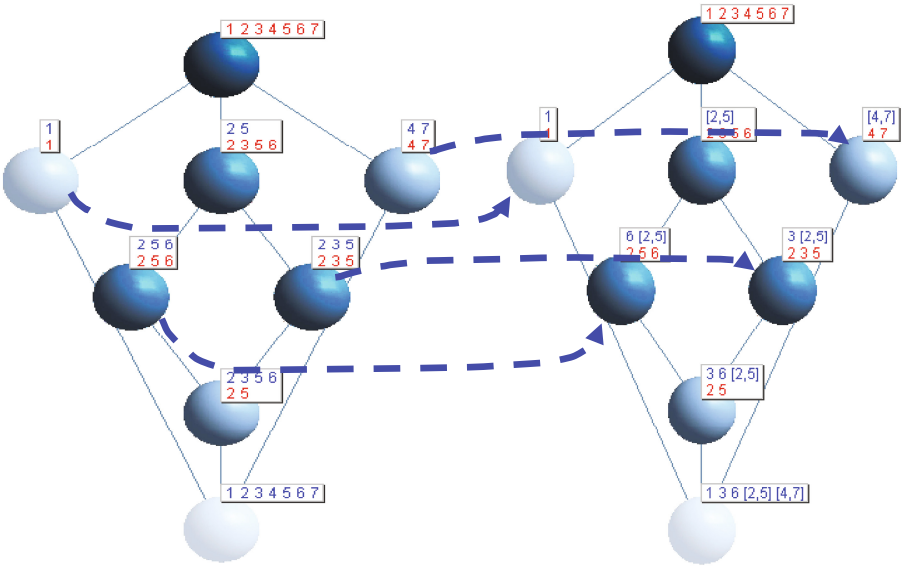


Fig. 2. The relationship between *C(K)* and *C(K')*

Interestingly, these maximal cliques are aggregated then represented as the relevant concepts in Fig. 2. For example, the common attribute of maximal cliques {2, 3, 5} and {2, 5, 6} is {2, 5}, that is to say, {2, 5} is the formation base for maximal cliques {2, 3, 5} and {2, 5, 6}.

Therefore, the following detection theorem is concluded with the above detection approach.

Detection Theorem: Given a graph $G = (V, E)$, the formal context of G is K, the concept lattice of K is denoted as $C(K)$, the bases of maximal cliques ***B(maximal_-clique(G))*** can be obtained from maximal cliques associated formal concepts in concept lattice $C(K')$, where K' is an induced formal context from K based on equivalence relation R over attributes.

4 Conclusions

This paper formalizes a novel problem on detection of bases of maximal cliques from a graph. Aiming to address this problem, this paper proposed a formal concept analysis based approach for detecting the bases of maximal cliques in a graph. We mathematically present a detection theorem according to our proposed approach. Hence, this detection theorem is ubiquitous and can be applied into various complex networks. The proposed detection approach reveals the formation principle of maximal cliques by investigating the relationship between original concept lattice and aggregated concept lattice. We believe that this work can pave the way for future topological structure analysis of social networks and other complex networking systems.

References

1. Conte, A., De Virgilio, R., Maccioni, A., et al.: Finding all maximal cliques in very large social networks. In: Proceedings of EDBT, pp. 173–184 (2016)
2. Xu, Y., Cheng, J., Fu, A.W.C.: Distributed maximal clique computation and management. IEEE Trans. Serv. Comput. **9**(1), 110–122 (2016)
3. Modani, N., Dey, K.: Large maximal cliques enumeration in large sparse graphs. In: Proceedings of 15th International Conference on Management of Data, pp. 1377–1378 (2009)
4. Eppstein, D., Loffler, M., Strash, D.: Listing all maximal cliques in sparse graphs in near-optimal time. In: Proceedings of 21st International Symposium on Algorithms and Computation, pp. 403–414 (2010)
5. Cheng, J., Zhu, L., Ke, Y., Chu, S.: Fast algorithms for maximal clique enumeration with limited memory. In: Proceedings of 18th ACM SIGKDD, pp. 1240–1248 (2012)
6. Goodrich, M.T., Pszona, P.: External-memory network analysis algorithms for naturally sparse graphs. In: Proceedings of 19th Europe Conference on Algorithms, pp. 664–676 (2011)
7. Du, N., Wu, B., Xu, L., Wang, B., Xin, P.: Parallel algorithm for enumerating maximal cliques in complex network. In: Proceedings of Mining Complex Data, pp. 207–221 (2009)
8. Schmidt, M.C., Samatova, N.F., Thomas, K., Park, B.H.: A scalable, parallel algorithm for maximal clique enumeration. J. Parallel Distrib. Comput. **69**(4), 417–428 (2009)
9. Hao, F., Min, G., Pei, Z., et al.: K-clique communities detection in social networks based on formal concept analysis. IEEE Syst. J. **11**(1), 250–259 (2015). doi:10.1109/JSYST.2015.2433294
10. Hao, F., Park, D.S., Min, G., Jeong, Y.S., Park, J.H.: K-clique mining in dynamic social networks based on triadic formal concept analysis. Neurocomputing **209**(C), 57–66 (2016)

Forecasting Cultivable Region-Specific Crops Based on Future Climate Change Utilizing Public Big Data

Yoo-Jin Moon[1], Won Whee Cho[1(✉)], Jieun Oh[1], Jeong Mok Kim[1], Sang Yub Han[1], Kee Hwan Kim[1], and Sungkap Cho[2]

[1] Hankuk University of Foreign Studies, Seoul 02450, Korea
yjmoon@hufs.ac.kr, {woni8879, masulilll, rlghks3232, sangyub2001, lynn0723}@naver.com
[2] Korea University, Seoul 02841, Korea
skcl777@naver.com

Abstract. The study designs and implements a database system for predicting small region-specific cultivable crops based on future climate change utilizing integrated public Big Data. For this study, regional temperature factors, regional precipitation factors, land acidity, solar radiation, cloud amount, and appropriate climatic factors for each crop were utilized. The database system could extract the information of each small region such as kinds of currently cultivating crop, kinds of regional cultivable food crop, kinds of regional cultivable fruit, kinds of regional cultivable medicinal crop, kinds of regional cultivable vegetable, and changing trends of each crop production quantity. Based on these small region-specific crop information, it is possible for the farmers to increase future profits of farm households by providing information of medicinal crops, food crops, vegetables, and fruits that can be produced in each regional farmhouse. It is also possible to present future recommended crops to individual business operators by utilizing these public big data, to suggest the need for development and research on crops that can be cultivated in each region, and to suggest marketing plans for present and future crops.

Keywords: Big data · Climate change · Database system · Cultivable region-specific crops

1 Introduction

According to the climate change scenario, it is expected that the climate of the Korean Peninsula will change drastically in the future, and so that the cultivation area, the production quantity and the quality of the crops will be greatly changed. From 1910 to 2000, changes were observed in average temperature and average precipitation every 10 years at six observation points—in Seoul, Incheon, Gangneung, Daegu, Mokpo and Busan. The result showed that the temperature was increased by 1.7° and the average precipitation increased by 19% in the six cities for last about 100 years.

At the end of the 21st century compared to the end of the 20th century, the average annual temperature of the Korean Peninsula is expected to increase by 4°, and the

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_65

annual range of the temperatures is expected to decrease due to the rise of the daily lowest temperature more than the daily highest temperature. And the annual average precipitation of the Korean Peninsula is expected to increase by 17%, and the increase of precipitation in August and September is expected to be large.

Through the implemented database system, changes in the annual mean precipitation at the end of the 21st century versus at the end of the 20th century were analyzed through analysis of climate-related big data [1, 2]. The main purpose of this research is to propose small region-specific crops in response to climate change, to secure income sources of farm households and to contribute to the revitalization of rural economy.

2 Preliminaries

To get the existing public big data for the region-specific crops database system, the study refers to NCDSS (National Climate Data Service System [3]), Jeollanamdo Agricultural Research & Extension Services [4], Jeju Agricultural Research & Extension Services [5], Gyengsangnamdo Agricultural Research & Extension Services [6], Rural Development Administration [7], and so on. The study also utilizes the MySQL program to build the database system, and the R package for performing graphically plotting of each information with making intuitive grasp possible [1, 2].

The previous studies performed the forecast of cultivable province-specific crops utilizing each big data [8–10]. This study extends the forecast of cultivable province-specific crops to that of small region oriented crops utilizing integration of each big data.

3 The Proposed Scheme

3.1 E-R Diagram and Relation

Public big data of NCDSS [3] provides regional highest and lowest temperatures, regional precipitation factors (highest precipitation, lowest precipitation), land acidity, solar radiation, cloud amount, etc. for all regions in Korea. Public data of Jeollanamdo Agricultural Research & Extension Services [4] provides information of Jeollanamdo Province for regional highest and lowest temperatures, land acidity, solar radiation, cloud amount etc. for each crop. Public data of Jeju Agricultural Research & Extension Services [5] provides information of Jeju Province for regional highest and lowest temperatures, land acidity, solar radiation, cloud amount for each crop. Public data of Gyengsangnamdo Agricultural Research & Extension Services [6] provides information of Gyengsangnamdo Province for regional highest and lowest temperatures, land acidity, solar radiation, cloud amount for each crop. Public data of Rural Development Administration [7] provides information of appropriate temperatures, land acidity, solar radiation, precipitation for each crop.

Figure 1 illustrates the Entity-Relationship Diagram of the region-specific crops database system and tables, respectively. The <Region> table has attributes for highest and lowest temperatures, regional highest precipitation, regional lowest precipitation,

land acidity, solar radiation and cloud amount, which are real big data existing in the Agricultural Development Administration, the Agricultural Technology Center, the National Weather Service of NCDSS, the Jeollanamdo Agricultural Technology Institute, the Jeju Island Agricultural Technology Institute, the Jeollabukdo Agricultural Technology Institute. The <Crop> table has attributes for appropriate climatic factors—temperature, land acidity, precipitation, etc. for each crop. The Relation <RegionalCrop> is created from the tables <Region> and <Crop>.

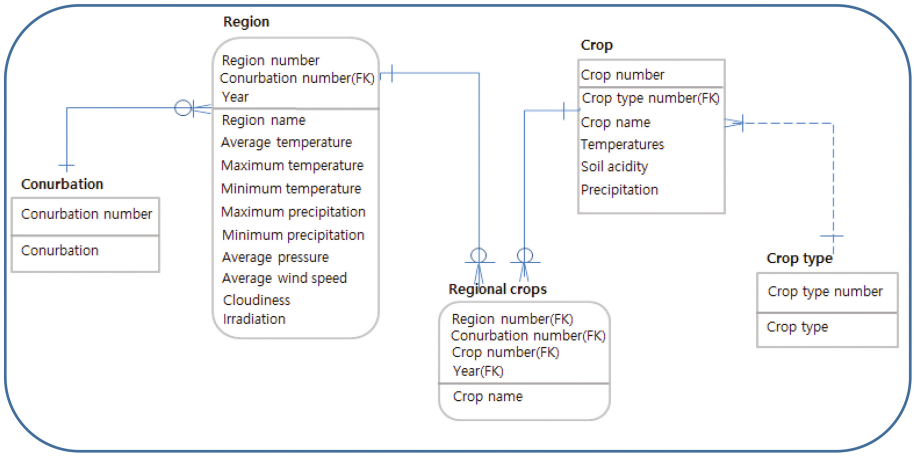


Fig. 1. System architecture

3.2 Representative Information Available

In the region-specific crops database system, crop information were extracted such as crops currently being cultivated, cultivable food crops by the small region, cultivable medicinal crops by the small region, cultivable vegetables by the small region, cultivable fruit trees by the small region, and quantity changes of crop yields.

(1) Cultivable Food Crops Based on Climate Change by Region
The result of prediction provided information on present and future food crops that can be produced locally. Table 1 illustrates that in Gochang area the food crop "barley for brewing" was expected to be produced well in the future, and that in Boseong area the food crop "wheat" to be produced well in the future. The result will provide information on food crops that can be cultivated in each small region in the future and allow planning of mass production for the food crops.

(2) Cultivable Medicinal Crops Based on Climate Change by Region
The result of prediction provided present and future medicinal crops that can be produced locally. Table 2 shows that in Gochang region, the medicinal crops produced from 2016 to 2018 are not suitable for their production in 2020, and that the Gochang region has a good environment to produce many other medicinal crops in the future.

Table 1. Cultivable food crops based on climate change in Gochang and Boseong

Region name	Year	Crop type	Crop name
Gochang	2016	Food crop	kidney bean
Gochang	2016	Food crop	barley for brewing
Gochang	2017	Food crop	kidney bean
Gochang	2017	Food crop	barley for brewing
Gochang	2018	Food crop	kidney bean
Gochang	2018	Food crop	barley for brewing
Gochang	2019	Food crop	barley for brewing
Gochang	2020	Food crop	wheat
Gochang	2020	Food crop	barley for brewing
Boseong	2016	Food crop	wheat
Boseong	2016	Food crop	barley for brewing
Boseong	2017	Food crop	wheat
Boseong	2018	Food crop	wheat

Table 2. Cultivable medicinal crops based on climate change in Gochang

Region name	Year	Crop type	Crop name
Gochang	2016	Medicinal crop	Lycium
Gochang	2016	Medicinal crop	Senna tora
Gochang	2016	Medicinal crop	rapeseed
Gochang	2016	Medicinal crop	Adlay
Gochang	2017	Medicinal crop	Lycium
Gochang	2017	Medicinal crop	rapeseed
Gochang	2017	Medicinal crop	Adlay
Gochang	2018	Medicinal crop	Lycium
Gochang	2018	Medicinal crop	rapeseed
Gochang	2018	Medicinal crop	Adlay
Gochang	2019	Medicinal crop	rapeseed
Gochang	2020	Medicinal crop	Bupleurum falcatum

(3) Cultivable Vegetable Crops Based on Climate Change by Region
The result of prediction provided information on present and future vegetable crops that can be produced locally. For example, the result showed that the cultivable vegetable crops varied over time due to climate change in Jeongseon region, where in 2017 vegetables such as strawberry, garlic, lettuce, cabbage and bellflower can be grown, and in 2018 and 2019 broccoli can be grown. Based on these results, it is possible to increase profitability of future farms by providing vegetable information that can be produced in Jeongseon region.

(4) Cultivable Fruit Crops Based on Climate Change by Region
The result of prediction provided information on present and future fruit crops that can be produced locally. For example, the result showed that as time passed, the fruit crops that can be produced varied due to climate change in Icheon region.

In Icheon region fruit trees can be cultivated such as plum, apple, plum and apricot in 2016, plum, apple, and apricot in 2017, apricot in 2018, and berry in 2020. Based on these results, it is possible to generate profits of farmers by providing fruit information that can be produced in the future in Icheon area.

(5) Trend Analysis of Apple Production
The result shows that for the apple production it is predicted that there will be a decrease of future cultivable areas. The climate change due to global warming will reduce the apple cultivation area in Korea in the future. In Jeonbuk province, it was possible to grow in four places of Sunchang, Imsil, Jeongeup, Gochang in 2016, but in 2020 the number of cultivable areas will be decreased to two places of IImsil and Jeongeup.

(6) Trend Analysis of Pepper Production
The result shows that for the pepper production it is predicted that there will be an increase of future cultivable areas. The climate change due to global warming will lead to increase of growing the pepper in Korea in the future. In Gyeongsang-namdo province, the cultivable area of pepper will increase four times in 2020 compared to 2016.

(7) Trend Analysis of Pear Production
The result shows that for the pear production it is predicted that there will be an increase of future cultivable areas. The climate change due to global warming will lead to increase of growing the pear in Korea in the future. In Gyeonggido province, the cultivable area of pear will increase four times in 2020 compared to 2016.

4 Conclusions

The purpose of this study is to design and implement a database system for predicting small region-specific cultivable crops based on future climate change utilizing integrated public Big Data. The actual big data utilized by the system includes regional temperature factors of maximum temperature and minimum temperature, regional precipitation factors, land acidity, solar radiation, cloud volume, and appropriate climatic factors for each crop, provided by the Agricultural Development Administration, the Agricultural Technology Center, the National Weather Service Korea National Climate Data Center, Jeollanamdo Agricultural Technology Institute, Jeju Island Agricultural Technology Institute.

The database system could extract the information of each small region such as kinds of currently cultivating crop, kinds of regional cultivable food crop, kinds of regional cultivable fruit, kinds of regional cultivable medicinal crop, kinds of regional cultivable vegetable, and changing trends of each crop production quantity. Based on these small region-specific crop information, it is possible for the farmers to increase future profits of farm households by providing information of medicinal crops, food crops, vegetables, and fruits that can be produced in each regional farmhouse. It is also possible to present future recommended crops to individual business operators by utilizing these public big data, to suggest the need for development and research on

crops that can be cultivated in each region, and to suggest marketing plans for present and future crops.

Acknowledgements. This work was supported by Hankuk University of Foreign Studies Research Fund of 2016.

References

1. Silberschatz, A., Korth, H.F., Sudarshan, S.: Database System Concepts. McGraw-Hill, New York (2010)
2. Kroenke, D., Auer, D.: Database Concepts. Pearson, Boston (2015)
3. NCDSS (National Climate Data Service System). http://sts.kma.go.kr/jsp/home/contents/statistics/newStatisticsSearch.do?menu=SFC&MNU=MNU
4. Jeollanamdo Agricultural Research & Extension Services. http://www.jares.go.kr/home/sub.do?m=20
5. Jeju Agricultural Research & Extension Services. http://jeju.agri.jeju.kr/region/
6. Gyengsangnamdo Agricultural Research & Extension Services. http://db.gba.go.kr/sub01/sub04_01.php
7. Rural Development Administration. http://www.nongsaro.go.kr/portal/ps/psb/psbk/kidofcomdtyPrdlstCode.ps?menuId=PS00067&sStdPrdlstCode=FT010601&sStdTchnlgyCode=GP01&totalSearchYn=Y
8. Kim, M., Woo, S., Ha, Y.: Application of sloped upland field machinery in the Ulleung areas. In: Proceedings of Korea Agricultural Machinary Conference (2016)
9. Kim, C., Lee, H.: Analysis of agricultural specialization in Daegu & Gyungbook Province. J. Rural Econ. **33**(1), 99–114 (2010). Korea Rural Economic Institute
10. Sung, K.C.: The study on regional adaptation and demand forecast of tropical and subtropical vegetables. Report of National Institute of Horticultural and Herbal Science (2014)

Statistical Analysis of Determinants of Intention to Use Virtual Reality Services and Moderating Effects

Young-Ho Hwang[1] and Yoo-Jin Moon[2(✉)]

[1] Kunsan National University, Gunsan 54150, Republic of Korea
yhwang@kunsan.ac.kr
[2] Hankuk University of Foreign Studies, Seoul 02450, Republic of Korea
yjmoon@hufs.ac.kr

Abstract. This research statistically analyzes factors that influence consumer intention to use virtual reality services. The result shows that the main predictors of intention to use virtual reality services, in the order of importance, are hedonic motivation, personal innovativeness, effort expectancy and performance expectancy. And it shows that the higher were the impacts of effort expectancy, social influence, performance expectancy, and hedonic motivation on intention to use the services the higher was a customer's personal innovativeness. According to the results, marketing strategies for virtual reality services should appeal to consumers by positioning the using experience as an adventure or a way to reduce their stress and change a negative mood. Also they should be reputation-building and target early adopters.

Keywords: Virtual reality services · Effort expectancy · Hedonic motivation · Social influence · Performance expectancy · Moderating effect

1 Introduction

In the intellectual information society areas of virtual reality (VR) and artificial intelligence are being magnified toward development of the 4th industry innovation. In the past VR had been imagined and described in science fictions or films, but now is becoming a reality. Originally the term "VR" was devised by Jaron Lanier in 1989 and then in 1992 Steuer defined VR as the realistic and simulated environment for the perceptor to experience the telepresence. Until now VR has been utilized for the various real life areas such as the military, entertainment, medical, learning, movie, shopping, architectural design, and tourism etc. [1].

Generally, VR focuses on the head-mounted goggle, which enables for the user to interact with the perception in the tridimensional situation as in the reality, and electronic technology which makes user experience the electronic environment simulation through accessories and clothes connected with networks [2]. Differentiated from other

This work was supported by Hankuk University of Foreign Studies Research Fund of 2016.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_66

media, VR with immersion, the experimental attribute, brings the changeover of the paradigm from pictures to places, from observation to experience, from use to participation, and from interface to inhabitation. And augmented reality (AR) focuses on the functions for complementing the reality people see. For example, AR can set the reality added with the graphic design screen. Gradually VR and AR combined with software technology are being in the limelight and will demand differentiated marketing strategies for the business. The research is going to call VR as representation of both the virtual reality and the augmented reality.

This research aims to analyze factors that influence consumers' intention to use virtual reality services, for which we utilize a new adapted and extended version of the UTAUT. Through this research we can find how consumers behave regarding use of VR services.

2 Research Model and Methods

2.1 Research Model

The research model for this study was designed to investigate main factors that affect intention to use (I-tU) virtual reality services. Our research model expands to adapt on Venkatesh et al.'s research [3] of user acceptance focusing on UTAUT variables and adds key dimensions: personal innovativeness (P-I), effort expectancy (E-E), social influence (S-I), performance expectancy (P-E), hedonic motivation (H-M), and facilitating conditions (F-C). This research model is illustrated in Fig. 1.

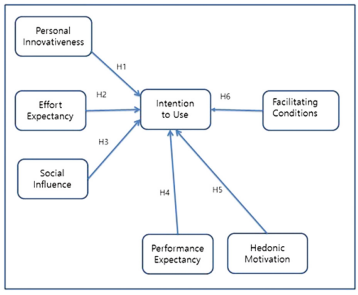


Fig. 1. Research model

2.2 Hypothesis Setting

Hypothesis 1: A customer's personal innovativeness (P-I) would have a positive impact on intention to use (I-tU) VR services.
Hypothesis 2: A customer's effort expectancy (E-E) would have a positive impact on intention to use (I-tU) VR services.
Hypothesis 3: A customer's social influence (S-I) would have a positive impact on intention to use (I-tU) VR services.

Hypothesis 4: A customer's performance expectancy (P-E) would have a positive impact on intention to use (I-tU) VR services.
Hypothesis 5: A customer's hedonic motivation (H-M) would have a positive impact on intention to use (I-tU) VR services.
Hypothesis 6: A customer's facilitating conditions (F-C) would have a positive impact on intention to use (I-tU) VR services.
Hypothesis 7: A customer's effort expectancy (E-E) moderates the relation between his/her personal innovativeness (P-I) and intention to use (I-tU) VR services.
Hypothesis 8: A customer's social influence (S-I) moderates the relation between his/her personal innovativeness (P-I) and intention to use (I-tU) VR services.
Hypothesis 9: A customer's performance expectancy (P-E) moderates the relation between his/her personal innovativeness (P-I) and intention to use (I-tU) VR services.
Hypothesis 10: A customer's hedonic motivation (H-M) moderates the relation between his/her personal innovativeness (P-I) and intention to use (I-tU) VR services.

2.3 Methods

2.3.1 Measurement

The study performed the exploratory factor analysis to ensure the content validity of the scales. Items selected of the questionnaire for measuring the constructs in our research model were adapted from prior studies. Items were measured by seven-point Likert scales.

2.3.2 Data Collection and Analytical Methods

To define the demographics of respondents, a frequency analysis was performed based on a total of 217 samples. The samples chosen for this study are undergraduate students in H University in Seoul and K University in Jeonbuk. Drennan et al. argued that university students are "representative of a dominant cohort of online users" including Virtual Reality users [4]. This study verified reliability and validity of the model. And frequency analysis, T-test and multiple regression analysis was conducted.

3 Results

3.1 Frequency Analysis

We used a non-probability sampling. Those questioned completed self-reported questionnaires and voluntarily participated in responding the questionnaires. A total of 240 questionnaires were distributed, of which 217 questionnaires were collected and used in the analysis with the response rate of 90%. In the demographic distribution of the sample, 55.8% of the respondents were male, and 44.2% were female. Almost of the respondents were between 20 and 30 years old and undergraduate students. In terms of experiences with virtual reality services, 24% had experience. And 58.1% of respondents lived in Seoul and 41.9% lived in the local area.

3.2 Verification of the Research Model

This research verified validity and reliability of the model (n = 217). We conducted the exploratory factor analysis. A factor extraction method was based on principal components analysis and Varimax rotation [5]. Table 1 showed that all seven factors were extracted. Each factor showed that an Eigen value was above 1 and the rate of cumulative variance showed 79.68% of total variance. There was no single factor that accounted for the majority of the covariance. And we found that multi-collinearity did not exist. This study also tested reliability between multi-item scales on 25 measurement variables. Table 1 showed the standardized Cronbach's α ranged from .928 to .932, and all values were greater than the recommended value of 0.7, suggesting adequate measurement reliability.

Table 1. Results of exploratory factor analysis and reliability analysis

	Item	Factor loading	Cronbach α	Standardized Cronbach α	Eigen value	Explained variance (%)
P-I	V1	.833	.931	.931	9.785	39.141
	V2	.830				
	V3	.856				
	V4	.778				
E-E	V5	.764	.929	.930	2.894	11.577
	V6	.759				
	V7	.845				
	V8	.636				
S-I	V9	.837	.929	.930	2.361	9.443
	V10	.793				
	V11	.760				
	V12	.774				
P-E	V13	.859	.931	.932	1.506	6.024
	V14	.873				
	V15	.882				
H-M	V16	.885	.928	.929	1.447	5.789
	V17	.814				
	V18	.897				
	V19	.876				
F-C	V20	.752	930	.931	1.249	4.996
	V21	.815				
	V22	.829				
I-tU	V23	.620	.927	.928	1.109	4.454
	V24	.613				
	V25	.755				
Cumulative variance (%)						81.42

3.3 T-test and Regression Analysis

The result of T-test showed that significant differences between non experiencers and experiencers at $\alpha = .05$ did not exist, males were more likely to be personally innovative than females, and those in the metro area were more likely to use virtual reality services.

This study used multiple regression analysis by setting Intention to Use (I-tU) as a dependent variable and other six variables (P-I, E-E, S-I, P-E, H-M and F-C) as independent variables. The results of multiple regression analysis showed that three of the six suggested hypotheses turned out to be significant in Table 2.

Table 2. Results of multiple regression analysis

Dependent variable	Independent variables	B	Standard error	ß	t	α	Accept/Reject
I-tU	Constant	−.387	.427		−.905	.367	
	P-I	**.046**	**.046**	**.122**	**2.483**	**.014**	**Accept**
	E-E	.083	.058	.077	1.427	.155	Reject
	S-I	**.144**	**.062**	**.128**	**2.317**	**.021**	**Accept**
	P-E	**.088**	**.051**	**.088**	**1.743**	**.083**	**Accept**
	H-M	**.568**	**.052**	**.562**	**11.027**	**.000**	**Accept**
	F-C	.076	.049	.075	1.536	.126	Reject
R^2	.644						
F-value	41.580						

H-1 was accepted because I-tU was significantly determined by tendency of P-I (ß = .122, α = .014). This result did support the previous studies [5, 7].
H-2 was rejected because E-E did not influence on I-tU significantly (ß = .077, α = .155). This result did not support the prior studies [5–7].
H-3 was accepted because S-I did have a significance on I-tU (ß = .128, α = .021). This result supported the prior studies [2, 5, 6].
H-4 was accepted because P-E had a significance on I-tU at the level of α = .10 (ß = .088, α = .083). This result supported the prior studies [5–7].
H-5 was accepted because H-M had a significance on I-tU (ß = .562, α = .000). This result supported the previous studies [2, 5, 6, 8].
H-6 was rejected because F-C did not influence on I-tU significantly (ß = .075, α = .126). This result did not support the prior studies [5, 8].

3.4 Moderating Effects of Personal Innovativeness

To test the moderating effects of personal innovativeness proposed by H-7, H-8, H-9, and H-10, we followed Chin et al.'s Partial Least Squares Product-Indicator approach [9]. We created the moderating variables by cross multiplying the items of E-E and P-I, SI and P-I, P-E and P-I, H-M and P-I. When the predictor has four measures and the

moderator has one indicator we have 4 measures for representing the construct of moderating effects.

As shown in Table 3, all moderating effects were significant: interaction between E-E and P-I (ß = 1.001, p < .000), interaction between S-I and P-I (ß = 1.400, p < .000), interaction between P-E and P-I (ß = 3.191, p < .05), and interaction between H-M and P-I (ß = 1.245, p < .000). Based on these results, hypothesis H-7, H-8, H-9, and H-10 were supported. We could conclude that a consumer's personal innovativeness moderated the impacts of E-E, S-I, P-E, and H-M on intention to use VR services, which implies that the higher were the impacts of E-E, S-I, P-E, and H-M on intention to use VR services the higher was a customer's personal innovativeness.

Table 3. Results of moderating effects

Variables	B	Standard error	ß	t	α	Accept/Reject
E-E→P-I→I-tU	1.078	.164	1.001	6.575	.000***	Accept
S-I→P-I→I-tU	1.581	.312	1.400	5.070	.000***	Accept
P-E→P-I→I-tU	3.200	1.453	3.191	2.203	.029*	Accept
H-M→P-I→I-tU	1.259	.162	1.245	7.748	.000***	Accept

*p < .05, **p < .01, ***p < .001

4 Discussion and Conclusions

H-1 was accepted. This means that the higher the personal innovativeness of new technology is, the higher the intention to use VR services is. H-2 was rejected. This means that ease of use of VR services and ease of learning how to use them have nothing to do with intention to use them. H-3 was accepted. This means that intention to use VR services is positively affected by recognition of family, influential people, and important people who believed that I should use VR services. H-4 was accepted. This means that intention to use VR services was positively affected by helping you improve your finance management and save your time through the VR services. H-5 was accepted. This means that the higher is the hedonic motivation that the consumers enjoy when using the VR services, the higher is the intention to use VR services. H-6 was rejected. This means that consumer perceptions of the support and the resources for using VR services do not influence intention to use VR services. H-7, H-8, H-9, and H-10 were accepted. This means that the effects of E-E, S-I, P-E, and H-M on intention to use VR services were higher at higher levels of a customer's personal innovativeness respectively.

Implications of the two factors of hedonic motivation and performance expectancy can be headed for the management and marketing strategies of VR services providers, which is that consumers should experience the services with enjoyment and get benefits by utilizing them. Thus, the VR services have to be elaborated in the pleasant and beneficial way. Marketing strategies should appeal to consumers by positioning the using experience as an adventure or a way to reduce their stress and change a negative mood. The positive effects of social influences on VR services imply that one of the marketing strategies for VR services should be reputation-building, in order to gain a

favorable opinion from referents, whether they are existing users or not. Thus these persons can actively recommend others to use the services. Another marketing strategy should target the early adopters. Early adopters may not be your first choice, but when targeted correctly, they can build buzz within exclusive inner circles and eventually entice mainstream customers to give your VR product a try.

References

1. Nguyen, M.-T., Nguyen, H.-K., Vo-Lam, K.-D., Nguyen, X.-G., Tran, M.-T.: Applying virtual reality in city planning. LNCS, vol. 9740, pp. 724–735 (2016)
2. Almusawi, A., Dülger, L.C., Kapucu, S.: Robotic arm dynamic and simulation with virtual reality model. In: International Conference on CoDIT, pp. 335–340 (2016)
3. Venkatesh, V., Morris, M., Davis, G., Davis, F.: User acceptance of information technology: toward a unified view. MIS Q. **27**(3), 425–478 (2003)
4. Drennan, J., Mort, G.S., Previte, J.: Privacy, risk perception, and expert online behavior: an exploratory study of household end users. J. Organ. End User Comput. **18**, 1–22 (2006)
5. Hwang, Y.-H., Moon, Y.-J.: Analysis of factors influencing intention to use the online-only bank and interaction effects among the factors. Adv. Sci. Lett. **22**(9), 2588–2591 (2016)
6. Yoon, H.S., Steege, L.M.B.: Development of a quantitative model of the impact of customers' personality and perceptions on internet banking use. Comput. Hum. Behav. **29**, 1133–1141 (2013)
7. Parameswaran, S., Kishore, R., Li, P.: Within-study measurement invariance of the UTAUT instrument: an assessment with user technology engagement variables. Inf. Manag. **52**(3), 317–336 (2015)
8. Escobar-Rodriguez, T., Carvajal-Trujillo, E.: Online purchasing tickets for low cost carriers: an application of the unified theory of acceptance and use of technology (UTAUT) model. Tour. Manag. **43**, 70–88 (2014)
9. Chin, W.W., Peterson, R.A., Brown, S.P.: Structural equation modeling in marketing: some practical reminders. J. Mark. Theory Pract. **16**(4), 287–298 (2008)

Real-Time Human Depression Diagnosis System Using Brain Wave Analysis

Dongmin Shin, Yunjin Nam, Dongil Shin, and Dongkyoo Shin(✉)

Department of Computer Engineering, Sejong University, 98 Gunja-Dong, Gwangjin-Gu, Seoul 143-747, Korea
{gentletiger, namyunjin0912}@gce.sejong.ac.kr,
{dshin, shindk}@sejong.ac.kr

Abstract. This study has the goal of developing a diagnosis system to detect human depression in real time to assist in the diagnosis of a doctor. The developed system may grasp the concentration and depression level of a patient using brainwave data acquired in real time. The depression detection index used in the system is the frontal brain asymmetry (FBA), which is based on the asymmetric phenomenon of depressed patients. In this study, an experiment was conducted with 40 depressed/normal subjects in order to verify the reliability of the developed system. The results proved that the system diagnosed the depression level in real time. It can be used to develop therapy programs for various nervous and mental disorders.

Keywords: Brainwave · EEG · Depression · Real time processing · Detection

1 Introduction

The most thoroughly studied brain computer interface (BCI) is interface technology using electroencephalography (EEG). Wolpaw developed an interface for controlling a cursor using an EEG [1–4] and Zhao proposed a BCI model for controlling driving in a 3D virtual environment [5, 8]. In the case of an EEG, since it is very sensitive to noise such as eye blinking and movement, preprocessing is important. Woestenburg proposed a method for removing the eye movement noise via a regression analysis [6].

Recent studies using brainwaves did not simply aim to understand the cerebral electrophysiological mechanism by analyzing the measured brainwaves, but also to indirectly measure human cognition, emotions, or psychology and diagnose diverse neuropsychiatric disorders such as anxiety disorder, schizophrenia, and depression [7].

This study aimed to develop a real time brainwave measurement system that can detect human depression by analyzing brainwave data acquired in real time.

2 Affective Computing Including Depression Diagnosis

Affective computing is an integrated field that combines human psychological analysis, computer science, and cognitive science to understand the information processing in the human intellectual system [9]. The classification of emotions was offered through

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_67

the emotion tree structure of Parrott and the six basic emotions defined by Ekman [10, 11]. Plutchik and Russell converted these basic emotions into a multidimensional scale utilizing an emotion wheel and valence-arousal, and much background research can still be performed [12, 13].

In order to recognize and interpret human emotion, the autonomic nervous system activity due to emotional changes can be identified using numerous monitoring systems. A monitoring system observes the human internal and external autonomic nervous system activity using environmental, video, audio, and biological signals. It then recognizes the emotion corresponding to the result.

An emotion recognition method using a biological signal can observe the actual human autonomic nervous system activity through the internal observation of the user, which is typically accomplished using an EEG and electrocardiography (ECG). Koelstra conducted an experiment to place user emotion on a valance-arousal plane using a DEAP dataset. A DEAP dataset contains emotion data obtained by showing 120 music videos to 16 males and 16 females, for a total of 32 subjects [14].

Frantzidis also conducted an experiment to match the emotions caused by visual stimulation on the valance-arousal plane, and classified the emotion that was mapped with an accuracy of 81% using the support vector machine (SVM) [15]. Lee Dong-hoon used LAPS consisting of 480 transvestites and showed a 73% accuracy for the eight classified emotions [16]. This study used an EEG for the two-dimensional interpretation of user emotion mapping on the valance-arousal plane.

On the other hand, research using an ECG has been actively conducted to identify the user emotional state using ECG feature points corresponding to the P-QRS-T and heart rate variability. Jing conducted 391 tests on the emotions caused by image data corresponding to joy and sorrow. In addition, he acquired user ECG data in this experiment. Jing detected the feature point corresponding to the P-QRS-T of the ECG using a wavelet transform, and compared the accuracy of the k-Nearest Neighbor (KNN) and Fisher-KNN. As a result, he found an accuracy of 85% using the Fisher-KNN [17]. Christie's studied the autonomic nervous system by mapping the ratio of the negative-positive plane to classify seven emotions. While the existing ratio of the autonomic nervous system showed an average accuracy of 40%, the classification accuracy of the negative and positive plane proposed by Christie showed an average value of 80% [18].

2.1 Depression Diagnosis Using Brainwave

Brainwaves are electrical flows that are created when signals are delivered from the nervous system to the cerebral nerve. They vary according to the person's psychological state, and are the most important indices of brain activity [19–21]. Table 1 shows the characteristics and classification of brain waves.

Neurofeedback is a kind of brain workout during which the users can monitor his or her brainwaves and feedback the information to improve their homeostatic self-regulation skills [22]. The brain function significantly improves. Neurofeedback has been available since Hans Berger developed EEG equipment which made it possible for the users to measure their own brainwaves [23].

Table 1. Classification and characteristics of brainwaves

Type	Frequency band	State of mind
Delta	0–4 Hz	Generating deep sleep
Theta	4–8 Hz	Meditation, being sleepy
Alpha	8–12 Hz	Relaxation, calm state
Low Beta (SMR)	12–15 Hz	Attention, concentration
Mid Beta	15–20 Hz	The active awareness
High Beta	20–30 Hz	Italic stress, tension, mental strain

Depression is diagnosed through surveys and consultation with a doctor. In 1983, Zigmond and Snaith developed the Hospital Anxiety and Depression Scale (HADS) survey for anxiety and depression [24]. In 1988, Beck developed the 21-question Beck depression inventory (BDI) to diagnose depression [25]. The results of surveys have been used as important indicators of the level of depression [26–28]. Asymmetry in brain activity has been hypothesized to be a potential marker for Vladas Valiulis 1, 2 depression [29]. In particular, an asymmetric frontal alpha band is a typical symptom of depression [30]. Therefore, the diagnosis and treatment of depression involves the calculation and settling of the ratio of frontal brain asymmetry (FBA) [31].

3 System Design and Implementation

The real time depression diagnosis system is composed of the brainwave measurement device and diagnosis software, which analyzes the brain waves transmitted from the device (Fig. 1).

We developed a portable brainwave measurement device that includes 12 channels, as shown in Fig. 2. There are 12 electrodes in the back of the device (FP1, FP2, F7, F3, Fz, F4, F8, FC3, FCz, FC4, Cz, and Pz in Fig. 2), which acquire the raw data. Then, each channel is transmitted by WiFi. The acquired raw data pass through an FFT and a power spectrum and each obtained waveform can be seen in the platform.

Fig. 1. Brainwave measurement device

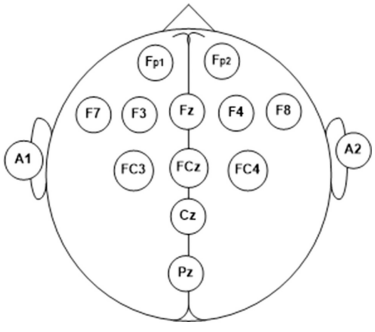


Fig. 2. Channels included in device

The EEG measurement device has twelve electrodes, one ground, and one reference to extract raw brainwaves, and has the shape of a headset, which allows it to be worn around the head. In addition, the device uses a wireless data transmission method, which helps prevent noise from entering the EEG raw data. Power is supplied using an on/off switch, and a rechargeable battery allows semi-permanent usage.

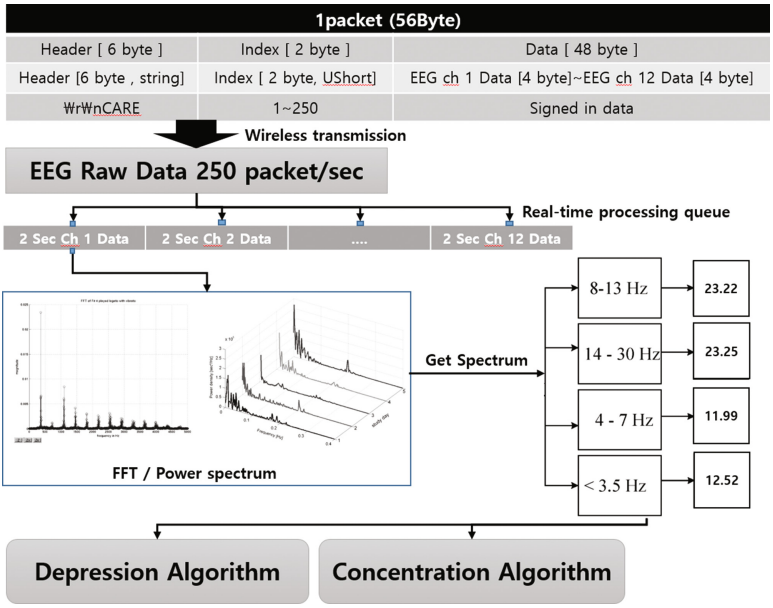


Fig. 3. Data analysis process

The device transmits raw EEG data from the human brain to the connected analysis software system in real time. It uses WiFi communication and can receive raw EEG data via the internal hyper terminal, based on the agreed-upon protocol. The data analysis process in the diagnosis software is shown in Fig. 3.

EEG data consist of 56-byte packets received 250 times per second, and the software stores data in the queue corresponding to 2 s in order to perform real time processing. Two seconds is the minimum window size to find meaningful data from the EEG. The brainwave analysis uses the Fast Fourier Transform (FFT) analysis method, which is designed to divide the brain-wave into several forms.

Depression detection Index used in the data analysis process is described as follows. Brainwave signals generated by frontal brain asymmetry (FBA) have something to do with alpha waves and cerebral activities. The characteristic of the alpha waves is the inactive or awake state, and when the brain activities increase, specific area and alpha waves will be reduced. Activities created by the brain are generally inversely proportional to alpha waves, and this also applies to alpha inactivity mechanism (AIM). The FBA of EEG is measured and digitized by the difference between the right and left brain activities, and is as follows.

$$\text{Depression index} = \frac{1}{p_{aL}} - \frac{1}{p_{aR}} \Big/ \frac{1}{p_{aL}} + \frac{1}{p_{aR}} = \frac{p_{aR} - p_{aL}}{p_{aR} + p_{aL}} \quad (1)$$

The diagnosis software analyzes the raw data for each channel obtained in real time. It calculates the results of a BDI or an HADS survey and the depression level of the patient. The results of the analysis show the frontal depression level for FP1 and FP2 during the measurement, and the average result is displayed.

4 Experiments and Results

We conducted the experiments with normal and depressed subjects (20 of each sex) in order to verify the reliability of the diagnosis software. The experiment methods acquired the EEG of the subjects for 5 min using the developed software. We then compared the survey results and the FBA, which was calculated in the results view. The HADS survey used in experiments was the parameter for analyzing the depression level.

Table 2 lists the results of the experiments. The two experimental groups were clearly divided based on the HADS scores for normal and depressed subjects. We can see that the depression indicators FBA used in the developed software show the difference between the normal and depressed subjects. We defined a depression threshold to allow the diagnosis software to analyze the depression level of the patients in real time. The threshold may be set to ±0.1, and a value between ±0.1 and ±0.2 can be defined as mild depression.

Table 2. Experiment result

Type	HADS	Difference	Proportion	FP1	FP2	FBA	Avg
Depression	33	15.439	2.087	29.643	14.203	-0.352	0.153
	28	13.622	0.512	14.295	27.917	0.323	
	22	6.224	2.327	10.913	4.689	-0.399	
	21	10.705	0.617	17.257	27.963	0.237	
	17	12.319	3.552	17.145	4.826	-0.561	
	15	4.473	1.390	15.942	11.469	-0.163	
Normal	8	0.391	0.948	7.276	7.667	0.026	0.022
	8	1.092	1.139	8.930	7.838	-0.065	
	6	0.211	0.963	5.580	5.792	0.019	
	7	0.842	1.064	13.942	13.100	-0.031	
	9	7.960	1.185	50.880	42.920	-0.085	
	10	0.130	0.991	15.890	16.020	0.004	

5 Experiments and Results

This paper describes the design and implementation of a real-time brain-wave measurement system for measuring EEG to confirm a degree of depression and degree of concentration. The system was designed to ensure diagnosis in real time through numeric values and graphs. Also, we prove that the system diagnoses depressed patients in real time by experiments.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2015R1D1A1A01059253).

References

1. Sung, Y.S., Cho, K.G., Um, K.H.: A framework for processing brain waves used in a brain-computer interface. J. Inf. Process. Syst. **82**, 315–330 (2012)
2. Sulaiman, N., Taib, M.N., Aris, S.A.M., Hamid, N.H.A., Lias, S., Murat, Z.H.: Stress features identification from EEG signals using EEG asymmetry and spectral centroids techniques. In: 2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 417–421 (2010)
3. Wolpaw, J.R., McFarland, D.J., Neat, G.W., Forneris, C.A.: An EEG-based brain-computer interface for cursor control. Electroencephalogr. Clin. Neurophysiol. **78**, 252–259 (1991)
4. Wolpaw, J.R., McFarland, D.J.: Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. Proc. Natl. Acad. Sci. U.S.A. **101**, 17849–17854 (2004)
5. Zhao, Q., Zhang, L., Cichocki, A.: EEG-based asynchronous BCI control of a car in 3D virtual reality environments. Chin. Sci. Bull. **54**, 78–87 (2009)
6. Woestenburg, J., Verbaten, M., Slangen, J.: The removal of the eye-movement artifact from the EEG by regression analysis in the frequency domain. Biol. Psychol. **16**, 127–147 (1983)
7. Knott, V., Mahoney, C., Kennedy, S., Evans, K.: EEG power, frequency, asymmetry and coherence in male depression. Psychiatry Res. Neuroimaging **106**, 123–140 (2001)
8. Picard, R.W.: Affective computing. Tech. rep. no. 321, MIT Media Lab., Massachusetts, USA (1995)
9. Tao, J., Tan, T.: Affective computing: a review. In: International Conference on Affective Computing and Intelligent Interaction, pp. 981–995. Springer, Heidelberg (2005)
10. Ekman, P.: Facial expressions. In: Handbook of Cognition and Emotion, pp. 226–232. Wiley, Chichester (2005)
11. Parrott, W.G.: Emotions in Social Psychology: Essential Readings. Psychology Press, Philadelphia (2001)
12. Plutchik, R.: The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**, 344–350 (2001)
13. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
14. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**, 18–31 (2012)

15. Frantzidis, C.A., Bratsas, C., Papadlis, C.L., Konstantinidis, E., Pappas, C., Bamidis, P.D.: Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli. IEEE Trans. Inf Technol. Biomed. **14**, 589–597 (2010)
16. Lee, D.H., Sim, K.B.: Development of emotion recognition model based on multi layer perceptron. J. Korean Inst. Intell. Syst. **16**, 372–377 (2006)
17. Cai, J., Liu, G., Hao, M.: The research on emotion recognition from ECG signal. In: International Conference on Information Technology and Computer Science (ITCS 2009), Kiev, Ukraine, pp. 497–500 (2009)
18. Christie, I.C., Friedman, B.H.: Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. Int. J. Psychophysiol. **51**, 143–153 (2004)
19. Lee, J.H., Kim, K.H.: A study of biosignal analysis system for sensibility evaluation. J. Korea Soc. Comput. Inf. **15**, 19–26 (2010)
20. Dobashi, N., Magatani, K.: Development of the EEG measurement method under exercising. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minnesota, USA, pp. 380–383 (2009)
21. Park, J., Woo, I., Park, S.: Application of EEG for multimodal human-machine interface. In: 12th International Conference on Control, Automation and Systems (ICCAS), pp. 1869–1873 (2012)
22. Tanaka, J., Kimura, M., Hosaka, N., Sawaji, H. Sakakura, K., Magatani, K.: Development of the EEG measurement technique under exercising. In: 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Shanghai, China, pp. 5971–5974 (2005)
23. Lorig, T.S., Herman, K.B.: EEG activity during administration of low-concentration odors. Bull. Psychon. Soc. **28**, 405–408 (1990)
24. Zigmond, A.S., Snaith, R.P.: The hospital anxiety and depression scale. Acta Psychiatr. Scand. **67**, 361–370 (1983)
25. Beck, A.T., Steer, R.A., Carbin, M.G.: Psychometric properties of the beck depression inventory: twenty-five years of evaluation. Clin. Psychol. Rev. **8**, 77–100 (1988)
26. Choi, K.M., Jang, K.M., Jang, K.I., Um, U.H., Kim, M.S., Kim, D.W., Shin, D.K., Chae, J. H.: The effects of 3 weeks of rTMS treatment on P200 amplitude in patients with depression. Neurosci. Lett. **577**, 22–27 (2014)
27. Segrave, R.A., Cooper, N.R., Thomson, R.H., Croft, R.J., Sheppard, D.M., Fitzgerald, P.B.: Individualized alpha activity and frontal asymmetry in major depression. Clin. EEG Neurosci. **42**, 45–52 (2011)
28. Kline, J.P., Allen, S.: The failed repressor: EEG asymmetry as a moderator of the relation between defensiveness and depressive symptoms. Int. J. Psychophysiol. **68**, 228–234 (2008)
29. Valiulis, V., Gerulskis, G., Dapsys, K., Vistartailte, G., Siurkute, A., Maciulis, V.: Electrophysiological differences between high and low frequency rTMS protocols in depression treatment. Acta Neurobiol. Exp. **72**, 283–295 (2012)
30. Micoulaud-Franchi, J.-A., Richien, R., Cemolacce, M., Loundou, A., Lancon, C., Vion-Dury, J.: Parieto-temporal alpha EEG band power at baseline as a predictor of antidepressant treatment response with repetitive transcranial magnetic stimulation: a preliminary study. J. Affect. Disord. **137**, 156–160 (2012)
31. Niemiec, A.J., Lithgow, B.J.: Alpha-band characteristics in EEG spectrum indicate reliability of frontal brain asymmetry measures in diagnosis of depression. In: 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Shanghai, China, pp. 7517–7520 (2006)

Deep Belief Network Based on Double Weber Local Descriptor in Micro-expression Recognition

Xiao-li Hao and Miao Tian(✉)

College of Computer Science and Technology,
Taiyuan University of Technology, Taiyuan 030024, China
4l49l966l@qq.com

Abstract. Face micro-expression is crucial for feeling perception and yet demanding due to the high dimension nature and the increasingly request for the recognition accuracy. The tradeoff between accuracy and efficiency by Deep Belief Network is a challenging. This paper shows that a two-stage strategy can achieve both speedup and high accuracy. With it, an efficient facial micro-expression algorithm is proposed that consists of Double Weber Local Descriptor devised in this paper firstly for extracting initial texture local features, and Deep Belief Net for more global feature and less computation dimension. The experiments with JAFFE database show that the average recognition rate by the new algorithm is up to 92.66%, and the rate of neutral facial expression is nearly 100%. Compared with LBP, LDP, PCA, Gabor wavelet and Weber local descriptor combined with DBN, the new algorithm of the introduction of Double Weber Local Descriptor into DBN has higher recognition rate.

Keywords: Face micro-expression recognition · Weber Local Descriptor · Deep Belief Nets · Feature extraction

1 Introduction

Facial micro-expression feature extraction and classification is crucial in the human-computer interaction [1]. Because of its high dimension nature and increasingly larger volume, big image data is categorized as the "biggest big data", making accurate expression recognition a challenging problem. As a typical deep learning net model, Deep Belief Net (DBN) dig out more effective features in higher layers derived from lower layers by simulating the brain tissue structure [2]. Though it can greatly reduce the big dimension of image data, Deep Belief Net, however, is still not sufficient for the accurate representation for local micro-expression structure features.

We argue that a two-stage feature extraction strategy can achieve both speedup and high accuracy. The former stage focuses on the accuracy by the operator advised in this paper to denote, label and extract detailed local texture features; the latter stage

Supported by the National Natural Science Foundation of China (Grant No. 61572345).

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_68

emphasizes the speedup by reducing the calculation of each neurons for the given feature weights. In realizing the strategy, we claim the following contributions: (1) A notion called Double Weber Local Descriptor with considering more space gradient change is devised to accurately abstract more detailed local feature. (2) An efficient cascaded feature extraction method for facial micro-expression data consists of extracting initial local texture feature by the operator and more global feature by Deep Belief Net. The experiment results show that our method works more accurately than the related algorithms.

2 Related Work

The facial micro-expression contains abundant texture features with invariance advantage of shift, angle and rotation. Texture extracting algorithms are proposed to obtain the local features. Guo put forward completely local binary pattern (CLBP) [3] and local binary pattern based on Fisher criterion (FCL-CLBP) [4]. Furthermore, wavelet is superior to local binary operators in the representation capability in time and space domain. For Gabor wavelet, Liu introduced Gabor multi-orientation fused features [5] and Zhong presented the eigenvalue extraction method based on local Gabor feature [6]. Although both of them solved the problems of insufficient and redundant extraction for local micro-feature, they have no strong robustness to noise and high light. Therefore Weber Local Descriptor (WLD) was raised based on the advantage of strong performance in the distinguishing description and the robustness to the noise [7]. This paper extends the descriptor and design Double Weber Local Descriptor, which optimized gradient calculation method of the central point by not only reflecting the gradient changing in horizon and vertical direction, but also focusing on the variation of diagonal direction.

3 Preliminary of Basic Theory

3.1 The Deficiency of Deep Belief Networks

Compared with the traditional neural network, it effectively reduces the time complexity by layer-by-layer training instead of all the layers simultaneously. Meanwhile, the inputs for each layer within RBM would be reconstructed and its weight would be continuously tuned. In spite of these, it usually ignores the two-dimension structure of the images, so it is hard to obtain the more detailed micro-feature, which is crucial to micro-expression recognition. Besides this, an input image with the noise, such as highlight, would lead the net to obtain adverse characters.

3.2 Weber Local Descriptor

Weber Local Descriptor (WLD) consists of two parts. One is called the Differential Excitation ξ, the ratio of Weber formula, which is used to describe the ratio between the point and its surrounding pixels' changing, as showed in formula (1). The other is the

Gradient Orientation θ of the central pixel, which reflects the ratio between the changes in horizon level and vertical level. It displays the space distribution of grey changing within local window, as showed in formula (2).

$$\xi(x_c) = \gamma_s^0 = arctan(\frac{f_{00}(x_s)}{f_{01}(x_s)}) = arctan(\sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c}) \tag{1}$$

$$\theta(x_c) = \gamma_s^1 = arctan(\frac{f_{10}(x_s)}{f_{11}(x_s)}) = arctan(\sum_{i=0}^{p-1} \frac{x_5 - x_1}{x_7 - x_3}) \tag{2}$$

4 Micro-expression Recognition Algorithm with Deep Belief Set Based on Double Weber Local Descriptor

4.1 Double Weber Local Descriptor

In formula (2), four neighbor pixels of the central point can be taken into account in original gradient calculation. It only demonstrates the gradient changes in the horizontal and the vertical direction, which can not fully embody the grey changes in the space distribution. For example, three different local texture patterns are showed in Fig. 1. When calculated by traditional formula, their ξ and θ are all zero, so they cannot be distinguished.

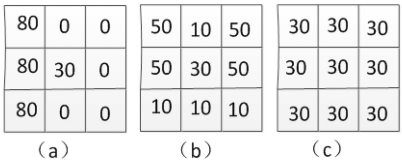


Fig. 1. Examples of local texture

To avoid it, the new gradient calculation method is showed in formula (3).

$$\theta(x_c) = tan^{-1} \frac{2(x_5 - x_1) + x_4 - x_2 + x_6 - x_0)}{2(x_7 - x_3) + x_0 - x_2 + x_6 - x_4)} \tag{3}$$

In this formula, eight neighbor pixels of the central point would be taken into account in case of loss of more space detail. Besides it, these pixels are used twice to eliminate the error, which is usually caused by different appearing frequency of pixels. That is why we call it Double Weber Local Descriptor (DWLD). By formula (3), their θ are different, which means they can be distinguished. Its calculation diagram is displayed is Fig. 2.

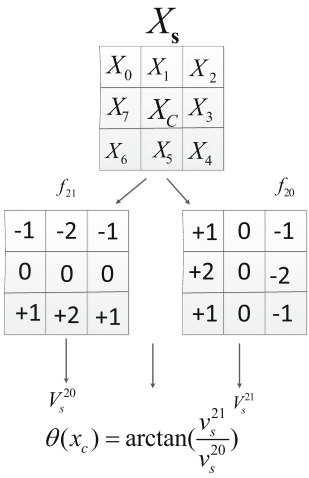
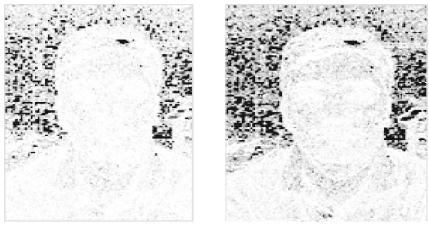


Fig. 2. Schematic diagram of DWLD gradient direction

The comparison experiment by DWLD and WLD is conducted and showed in Fig. 3. From it, the clearer outline of the face by DWLD indicates that more distinguished feature detail would be achieved and more accurate information would be extracted.



a. θ of WLD b. θ of DWLD

Fig. 3. The comparison of experiment results by WLD and DWLD in gradient calculation

4.2 The New Proposed Algorithm

From above, it is concluded that Double Weber Local Descriptor (DWLD) focuses on more detailed local character of the gradient and the edges, while Deep Belief Nets (DBN) emphasizes more global feature and less computation. But when DWLD is only used in micro-expression recognition, it can not effectively represent the overall structure feature. Therefore, two-stage feature extraction algorithm is proposed. By DWLD, initial feature description can be obtained and then introduced to DBN. In second stage, by pre-training and fine-tuning in DBN, the more advanced feature can

be achieved to accomplish classification of face micro-expression. The new proposed algorithm is as following.

Step 1: Set the facial expression database, and locate, crop and normalize these images.
Step 2: Group preprocessing images into the training samples and test samples, and extract the DWLD feature.
Step 3: Input the pixel vectors of training samples to the visual layer of DBN and initialize the parameters. By the top-down unsupervised greedy learning, the weights are updated.
Step 4: After the training of RBM, according to the input data and the lost function of the reconstruction data, the net parameters are retuned by Back Propagation net.
Step 5: Judge whether the difference between the initial vectors and the reconstruction vectors with the optimized weight, is less than the preset value. If it is, it goes to step 6, otherwise return to step 4.
Step 6: Put the testing samples into DBN model for classifying and output the results.

5 Experiment Results and Analysis

The paper selects JAFFE database as the experimental samples, and it has a total of 213 images. We would compare the new proposed micro-expression recognition algorithm (DWLD+DBN) with Deep Belief Nets based on Weber Local Descriptor (WLD+DBN). Test results in JAFFE are showed in Table 1 when the number of the hidden nodes in DBN is 300 and the hidden layers are 1, 2 or 3 respectively. The results are showed in Table 2 when the number of hidden layers is 1 and the number of nodes in the hidden layer is 50, 100, 300 and 500 respectively.

Table 1. DBN, DBN+WLD and DBN+DWLD recognition results in different number of hidden layers (%)

	1	2	3
DBN	86.91	83.38	40.95
WLD+DBN	90.23	87.54	44.36
DWLD+DBN	92.66	88.12	45.11

Table 2. DBN, DBN+WLD and DBN+DWLD recognition results with different nodes (%)

	50	100	300	500
DBN	81.42	84.76	86.91	71.42
WLD+DBN	83.37	85.21	90.23	79.56
DWLD+DBN	84.69	87.34	92.66.	81.33

In the best performance of DWLD+DBN, the recognition results of seven expressions are shown in Table 3, where the average recognition rate is 92.66%, The recognition rate of four facial micro-expression is over 90%, and the rate of neutral facial expression is nearly 100%.

Table 3. Recognition rate of face micro-expression by DWLD+DBN in JAFFE database

Expression	Anger	Happiness	Sadness	Surprise	Disgust	Fear	Neutral
Anger	93.33	0.00	0.00	0.00	3.27	3.40	0.00
Happiness	0.00	88.17	2.72	2.05	0.00	0.00	7.06
Sadness	0.00	2.07	97.90	0.00	0.00	0.00	1.03
Surprise	0.00	2.11	0.00	95.87	0.00	2.02	0.00
Disgust	3.25	0.00	3.65	0.00	84.83	8.27	0.00
Fear	0.00	0.00	3.02	2.75	2.54	88.56	2.91
Neutral	0.00	0.03	0.00	0.00	0.00	0.00	99.97

Furthermore, we compare the new proposed algorithm with local binary pattern (LBP), Local Difference Pattern (LDP), Gabor wavelet transform and PCA feature extraction. The experimental results are presented in Fig. 4. The recognition algorithm proposed in this paper has increased recognition rate on the average.

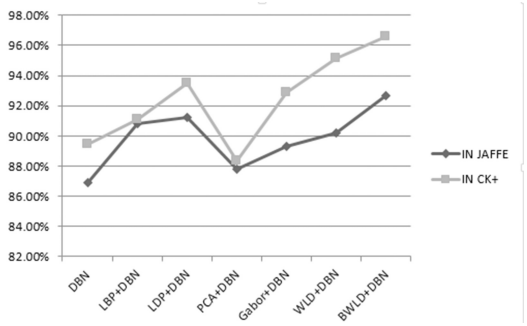


Fig. 4. The recognition rate of DBN combined with different feature extraction

Table 4 shows the comparison of the time used in training and recognition by DBN and DWLD+DBN. From it, it can be concluded that the new method has improved the efficiency.

Table 4. The time used in training and recognition by different methods

Nodes in hidden layers	Method	Training time (s)	Recognition time (s)	Total time (s)
300	DBN	25.38	2.08	27.46
	DWLD+DBN	22.46	1.65	24.11
500	DBN	51.79	2.47	54.26
	DWLD+DBN	49.08	1.94	51.02

6 Conclusion

This paper suggests that the two-stage recognition strategy is able to achieve both efficiency and accuracy. Compared with Weber Local Descriptor, Double Weber Local Descriptor could cover the deficiency of Deep Belief Net, which lack the extraction of the texture structure features. Moreover, it is then introduced into the Net to reduce the amount of learning for the redundant features and speed up the efficiency.

References

1. Zilu, Y., Jingwen, L., Youwei, Z.: Facial expression recognition based on SLLE with expression weighted distances. Pattern Recogn. Artif. Intell. **23**(2), 278–283 (2010)
2. Jung, H., Lees, S., Park, S., et al.: Development of deep learning for emotion recognition system. In: 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV 2015), pp. 1–4. IEEE, Mokpo (2015)
3. Zhenhua, G., Lei, Z.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. **19**(6), 1657–1663 (2010)
4. Guo, Y.: Discriminative features for texture description. Pattern Recogn. **45**(10), 3834–3843 (2012)
5. Liu, S.-S.: Facial expression recognition method based on Gabor multi-orientation features fusion and block histogram. Acta Automatica Sinica **37**(12), 1455–1463 (2011)
6. Sizhi, Z.: Research on Facial Expression Recognition. School of Information Science Technology, Shanghai (2015)
7. Chen, J.: WLD: a robust local image descriptor. IEEE Trans. Pattern Anal. Mach. Intell. **32** (9), 1705–1720 (2010)

Design for Network File Forensics System Based on Approximate Matching

Aonan Zhai[1,2], Fei Xu[1,2], Haiqing Pan[1,2], Junzheng Shi[1,2], and Gang Xiong[1,2(✉)]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{zhaiaonan, xufei, panhaiqing, shijunzheng, xionggang}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. Network forensics is a comparatively new field of forensics science. The growing popularity of the Internet means that computing has become network-centric and data is now available outside of disk-based digital evidence. To collect certain network data for forensics, real-time network file packet inspection becomes a hot topic as it is needed in many applications such as virus detection, intrusion and attack forensics. Most of the traditional techniques use exact matches on keyword and/or white/black MD5 lists to have an efficient inspection. However, it is well-known that exact matches may not be effective to identify similar files such as the same videos with small changes, e.g. titles, posted by different users or metamorphic viruses (mutated computer viruses). Approximate matching is known to be more robust to identify similar files and has been proven to be effective in digital forensics. In this paper, we design a network forensics system by recording objective network files for future analysis. We try to confirm that by using an appropriate approximate matching approach, it is feasible and effective to inspect real-time traffic in order to identify similar files. Our experiments with real data show that our solution achieves good usability in practical.

Keywords: Approximate matching · Network traffic · File detection · Network forensics

1 Introduction

Nowadays, computer networks are vulnerable to cyber-attacks from both inside and outside. Furthermore, the threats of the cyber-attacks such as personal information disclosure, DDoS (Distributed Denial of Service) attacks and APT (Advanced Persistent Threat) attack are occurring continuously [1–10]. Therefore, network traffic forensics is of great significance to anti-virus, data leak preventions and so on.

In order to secure their networks, companies install intrusion detection systems (IDS) which usually use a keywords list or a black MD5 list to identify virus, malicious software and certain internal files. However, this method can only detect exact matches. In network level, it is a common observation that files captured from network flow are

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_69

often incomplete due to packets loss or processing error. Middle box can't completely recover the original file due to a burst of network traffic or a DdoS attack. And there is a lot of human factors that could compromise the integrity of captured files. Traditional cryptographic hash cannot cope with this situation. Approximate matching is a rather new working field but has been proven to be useful for similar input detection.

In this work, we apply fuzzy hashing to network packet inspection in order to detect similar files in real time and therefore record certain files for forensics. The solution can be easily applied and maintained. The main contribution of our work is the framework design for network file forensics system based on approximate matching.

2 Background and Related Work

Approximate matching is a rather new field and probably had a breakthrough in 2006 with an algorithm called context triggered piecewise hashing proposed by Kornblum. Since then, a few more algorithms were presented. Currently there are two main SPH implementations: ssdeep [2] and sdhash [1].

Ssdeep is a program for computing context triggered piecewise hashes (CTPH). The key idea of ssdeep is to break up file into pieces by finding trigger values using a rolling hash function. Every hash value for the separated block is transformed to a character and the generated file digest is a string. This string is used to provide the similarity percentage between 0 and 100 when compared with another file.

Sdhash was proposed in 2010 by Roussev [1]. It attempts to pick characteristic features for each object that are unlikely to appear by chance in other objects based on results from an empirical study. There are two general classes of problems where sdhash can provide significant benefits, namely fragment identification and version correlation.

Also there are some typical network forensics system proposed aiming at different situation.

VAST: A Unified Platform for Interactive Network Forensics [6], is a distributed platform for high-performance network forensics and incident response that provides both continuous ingestion of voluminous event streams and interactive query performance. VAST leverages a native implementation of the actor model to scale both intra-machine across available CPU cores and inter-machine over a cluster of commodity systems.

Designing a Data Warehouse for Cyber Crimes [7] explores designing a dimensional model for a data warehouse that can be used in analyzing cyber crime data. They also present some interesting queries and the types of cyber crime analysis that can be performed based on the data warehouse. However their work is based on our conceptual analysis of literature.

Data Warehousing Based Computer Forensics Investigation Framework [8] proposed the design of an efficient computer forensics investigation framework. The proposed framework improves the investigation efficiency using Data Warehouse (DW) concept, which provides a selective evidence identification, collection and analysis.

3 Implementation Detail

3.1 Forensics System Framework

The upcoming subsection provides an overview of our solution. The forensics framework is shown in Fig. 1. By comparing the similarity between network traffic and target files, we can record the trace of certain network attacks.

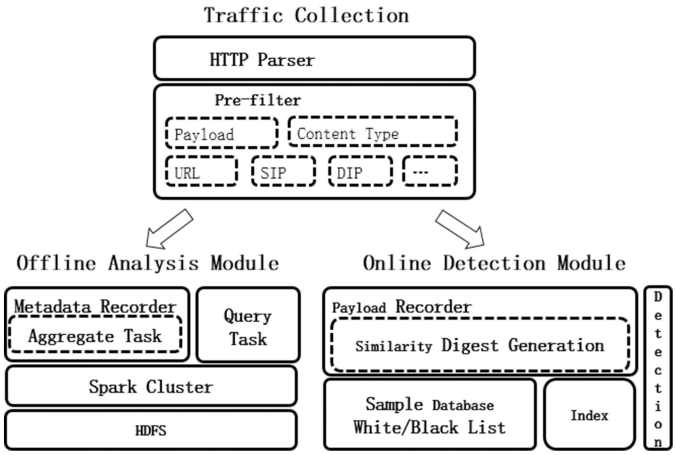


Fig. 1. System framework

3.2 Traffic Collection

We aim at detecting similar files or fragments in network traffic and record matched files for forensics in the future. To achieve this goal, first we need to capture network files. In this paper, only network files transmitted by HTTP protocol is concerned. We extract HTTP packets based on PF_RING, and filter out desired traffic based on meta data (for example, we focus on HTTP method including POST and GET, others are ignored), finally we restore network files which are used for similarity digest generation.

3.3 Online Detection Module

In online detection module, we need to determine whether the current transmitted file is similar to target files. We use approximate matching algorithm to generate network file similarity digests, compare the digests against the target database and record files with metadata if matched in the target database. In order to speed up comparison, we need to build index for the database. Target files database is comprised of malicious or confidential files typically and can be generated offline ahead of time. Therefore, the

performance of online process is not influenced by database generation time. In this paper, we choose sdhash as the basic approximate matching algorithm due to its outstanding performance in precision and recall compared to other algorithms.

We aim at detecting network files in real time, but the time complexity of traditional all-against-all comparison is unsatisfactory. Similarity digests cannot be ordered, so that common data structure like B-tree is unable to be used as index. As far as we know, one efficient solution is to insert all target file features into one Bloom filter.

For every selected feature of each network file, we check whether the feature is in the Bloom filter, if a certain number of consecutive features are not in the index filter or the total features in the filter not accounts for a certain proportion, the file is viewed as not matched

3.4 Offline Analysis Module

Meta data with respect to file traffic is recorded for offline analysis in online detection module, including IP, port, bytes and packets transferred etc. A json file composed of records is created every hour and Spark cluster is used to aggregate connection attributes based on IP address. In this way, we can obtain the network behavior of certain IP address within a particular time interval. Besides, query jobs to find the trace of transmitted files similar to specified files can also be done taking advantage of Spark.

4 Experiment Results

We analyze the network file traffic and obtain the experiment dataset from captured files. Then we measure the system throughput based on practical data.

4.1 Network File Traffic Analysis

In this paper, we collect network files from one ISP level network gateway, namely China Science and Technology Network. We focus on HTTP traffic only and there are 1,152,874 files captured for 24 h. 9.3% of the files are not complete compared to the downloaded version from the recorded URL due to some kind of reason.

We analyzed the captured files and the distribution figure of file counts according to incomplete file size proportion for the top five content type is displayed below in Fig. 2. Vertical axis represents for the file count proportion of incomplete files and horizontal axis represents for captured file size proportion compared to its downloaded version, note that the proportion is discretized by 10%. From the figure we can see quite a part of captured files are incomplete in real traffic, therefore it's of great significance to cope with this situation in our framework by using approximate matching.

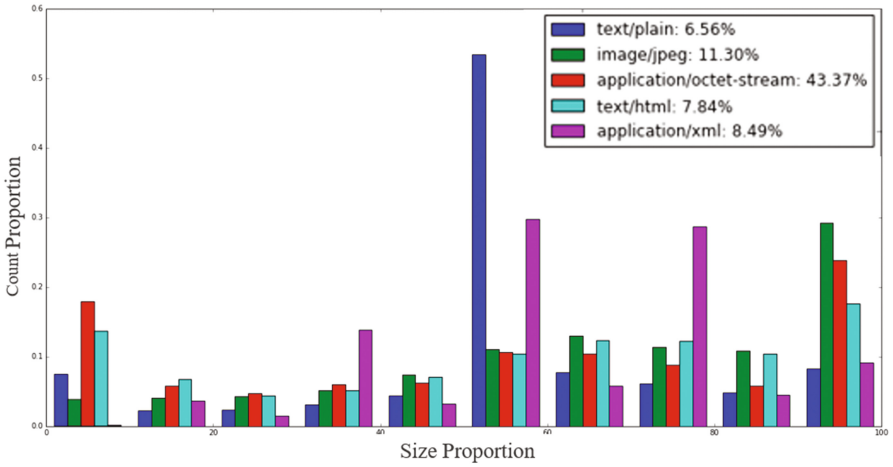


Fig. 2. Analysis of file counts according to incomplete file size proportion.

4.2 Online Detection Module

In online detection module, we build the index for a collected dataset including 4508 files, the files total size is 2.78 GB and the memory consumption is about 36 MB. Compared to brute force comparison method, the index lookup gains about one thousand times speed up. Besides time effectiveness, index also has a significant filter precision, it helps a lot to reduce the number of network files that need to be checked in detail. In our experiment, the built index can filter more than 90% dissimilar files and almost none of the similar files is filtered out.

4.3 Offline Analysis Performance

To analyze the recorded traffic meta data and file digests for future forensics, we set up a Spark cluster which is consisted of one master and four slaves, every worker possesses 16 cores and 256 GB memory. The json file created per hour contains about ten million records, and taking advantage of Spark, corresponding aggregation computation jobs can be done within 10 min. Furthermore, we can accomplish query jobs about the similarity comparison between target file and recorded traffic in parallel.

4.4 System Throughput

To determine the throughput we measured result includes the time for generating file digests and looking up in the Bloom filter. The total process time for a dataset containing 2.14 GB files was 46.603 s. Overall, the throughput is 2140 MB/46 = 46.52 MB/s. The test was performed on a 2 GHz Intel Core i7 CPU, single threaded. However, the approach allows for easy parallelism without any synchronization. We only need to run multiple threads to hash the packets of each file separately and compare them to the Bloom filter.

5 Conclusion

In this paper, we considered the challenge of similar file identification on network traffic using approximate matching (a.k.a. fuzzy hashing) and design a network forensics system to record suspicious network files. To detect files in high speed network in real time, how to deal with concurrency and incomplete file capture is of great importance, we come up with a detection framework to cope with this situation and evaluate its applicability. The performance of the algorithm in terms of time and space is also evaluated, and the result shows a good usability in practical.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61602472, No. U1636217), and the National Key Research and Development Program of China (NO. 2016YFB0801200).

References

1. Roussev, V.: Data fingerprinting with similarity digests. In: IFIP International Conference on Digital Forensics. Springer, Heidelberg (2010)
2. Kornblum, J.: Identifying almost identical files using context triggered piecewise hashing. Digit. Invest. **3**, 91–97 (2006)
3. Breitinger, F., Ibrahim, B.: File detection on network traffic using approximate matching. J. Digit. Forensics Secur. Law JDFSL **9**(2), 23 (2014)
4. Breitinger, F., Harald, B.: Properties of a similarity preserving hash function and their realization in sdhash. In: Information Security for South Africa (ISSA 2012). IEEE (2012)
5. Roussev, V.: An evaluation of forensic similarity hashes. Digit. Invest. **8**, S34–S41 (2011)
6. Vallentin, M., Paxson, V., Sommer, R.: VAST: a unified platform for interactive network forensics. In: Proceedings of the 13th USENIX Symposium on Networked Systems Design and Implementation, pp. 345–362 (2016)
7. Halboob, W., Mahmod, R., Abulaish, M., Abbas, H., Saleem, K.: Data warehousing based computer forensics investigation framework. In: 2015 12th International Conference on Information Technology - New Generations, pp. 163–168 (2015)
8. Song, I.Y., Maguire, J.D., Lee, K.J., Choi, N., Hu, X.H., Chen, P.: Designing a data warehouse for cyber crimes. J. Digit. Forensics Secur. Law **1**(3), 5–22 (2006)
9. Pilli, E.S., Joshi, R.C., Niyogi, R.: Network forensic frameworks: survey and research challenges. Int. J. Digit. Forensics Incident Response Arch. **7**, 14–27 (2010)
10. Roussev, V.: An evaluation of forensic similarity hashes. Digit. Invest. **8**(Suppl. 1), 34–41 (2011). The Proceedings of the Eleventh Annual DFRWS Conference

The Comparative Analysis of the Repeat Regions from the Assembled Contigs

Jaehee Jung(✉)

General Education, Hongik University, Seoul, South Korea
jhjung@hongik.ac.kr

Abstract. With the active sequencing studies, along with the advancement of the next-generation sequencing (NGS) technology, rapid progress has been made in genome analysis of various species. For the completion and analysis of the whole genome map, it is necessary to assemble the read results from NGS data for map completion. When reference models are available for assembly, similar sequences can be used for mapping assembly, whereas the de novo assembly method is applied when there are no models available. At this time, if the number of repeats in repeat regions is unclear, it would be challenging to assemble the whole genome map. Thus, the aim of this study was to conduct comparative analyses of the repeat regions using assemblies from various assembler tools automatically yielding repeat regions, and to carry out effective assembly analysis including repeat regions.

Keywords: Sequence assembly · Repeat region · NGS · RepeatMasker · RepeatModeler

1 Introduction

After the emergence of the next-generation sequencing (NGS) technology, which produces short reads unlike the Sanger sequencing method, DNA sequence analysis has made a remarkable progress. Nonetheless, assembling short reads and combining them into long scaffolds remains a significant challenge. When a reference model species is available, the whole genome sequences can be assembled through the alignment of the read results using the analytical tools referring to the similarities with the reference. Nevertheless, there are few reference models in most cases. If no reference genome is available, de novo assembly is required. Most de novo assembly tools are based on the De Bruijn graph (DBG) method. The De Bruijn graph includes the Velvet [1], SOAPdeNOVO [2], and the ABySS [3] packages, which have two problems, unlike the assembly method using the reference. First, performance of the assembly can depend on how the k-mer value is set. When a too high K-mer is set, it hampers accurate assembly, whereas too small a value makes it hard to assemble short reads. Second, it is difficult to find regions of repetitive sequences. If the reads are assembled under the assumption that the K length is a vertex, it is difficult to figure out how many repeats are included. Therefore, it is necessary to identify the repetitive sequences that are commonly found after mapping the assembled sequences from different de novo assembler tools using a database of repeat regions. In other words, this approach may

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_70

lay a solid foundation for the development of an algorithm to automatically determine repeat regions via assembly of correct repeat regions regardless of the method used for assembly.

2 Related Works

Assemblers for compilation of the read results from NGS data and the repeat search tool for finding repeat regions will be discussed.

2.1 Sequence Assembly Software

2.1.1 Mimicking Intelligent Read Assembly (MIRA) [4]

It assembles various input data just as the Solexa (Illumina) sequencing or Ion Torrent does and merges short reads with long reads to build optimal contigs for reads.

2.1.2 St. Petersburg Genome Assembler (SPAdes) [5]

This is an assembly tool kit composed of Illumina or Ion Torrent pipelines, in which data are mainly processed. There are four pipelines, among which two pipelines are inspection steps for errors of Illumina reads and IonTorrent reads. The other pipeline is the key core module that carries out repeated short-read assembly by setting k and by considering the types and lengths of the reads automatically. The existing DBG draws a graph with a certain preset K. When the K size was small, tangle was easily made, but when the k value was relatively large, it was hard to attain connection for assembly. To complement these drawbacks, this pipeline removes mismatch information between the read parts or unnecessary information between the connection of the 2 par gemone, aimed at constructing a multisized de Bruijn graph. Finally, the last pipeline is the MismatchCorrector step, which improves the short-indel rate including the mismatched region using the Burrows-Wheeler Aligner [6] (BWA) tool. BWA is a low-divergent sequence mapping tool using a large genome sequence as a reference.

2.1.3 Velvet

The Velvet method, one of the most popular methods among De Bruijn Graph methods, can be applied to relatively smaller genomes. Reads are split into a number of k-mers, and the overlapping sequences are aligned per number of k-1 numbers. The aligned contigs are regarded as one vertex; if the outgoing edges and the incoming edges form the same contig, the two vertices are assembled into one. By merging the vertex, the assembled contigs are established. Performance of the De Bruijn Graph method is dependent on the k-mer value, where a small k value can result in many contigs, but a high k value can make the assembly relatively difficult. This method was proposed at the early stages, and is currently used mostly as a filtering process in many tools; it is rarely used as an independent assembly method.

2.2 Repeat Analysis Software

2.2.1 RepeatMasker

RepeatMasker [7] is a tool designed to identify repeat regions in genomic data and to mask them. A repeat database reference is required; for this purpose, Repbases or Dfam libraries have been mainly used. Repbase [8] is a repetitive-sequence database derived from another eukaryotic species. The main function of RepeatMakser is to identify repeat regions using CrossMatch, RMBlast, WUBlast, and HMMER by means of the repeat database such as Dfam Repbase. Although it is advantageous that repeat regions are easily found by matching data with the reference genome, it is difficult to find the reference genome for analysis of an unknown sequence, and success is dependent on quality of the reference, i.e., on the accuracy of the repeat region in the database (Table 1).

Table 1. The repeat database from RepeatModeler

Repeat category	RepeatModeler	
	RECON	RepeatScout
SINEs	0	0
LINEs	19	3
LTR elements	65	5
Unclassified	434	771
Small RNA	0	0
Simple repeats	0	1
RC	1	0
DNA/Maverick	1	0
buffer	0	1
Total	520	781

2.2.2 RepeatModeler

A database which can be established by the RepeatModeler [9] is necessary to find repeat regions using RepeatMasker. The RepeatModeler is a repeat family modeling from the two different packages - RECON [10] and RepeatScout [11]. These two programmings extract the repeat elementary and identify family relationship from the sequence.

A database for mitochondria, not for a eukaryotic species, was constructed using RepeatModeler according to the goal of the study. A format for the desired species was determined for RepeatModeler. RECON and RepeatScout software tools were run in sequence with the predetermined format to find the repeat regions, yielding a database for RepeatModeler. Through analysis of the Repeat database from the RepeatModeler, a total of 520 Repeat databases were completed in the first run RECON. After that, RepeatScout was run four times repeatedly, yielding a total of 781 Repeat databases.

3 Approach

3.1 Dataset

The tested dataset corresponds to a photosynthetic cryptophyte alga that contains various pigments. Although cryptophyte algae are important as primary producers in a water ecosystem, there have been only a few phylogenetic and ecological studies in Korea. The cryptophyte alga was selected as a dataset in this study in order to complete its genome map and to provide fundamental data for research on genome evolution. Data from the cryptophyte alga *Rhodomonas salina* was obtained through actual collection. Analysis of NGS results and of already reported repeat regions in NCBI enabled us to automatically find repeat regions in other cryptophyte algae and allowed for annotation of the regions. The already revealed data on *Rhodomonas salina* can be used to reconstruct the genetic map as shown in Fig. 1. The genetic map can be drawn for the already known parts by means of GenBank data of NCBI using OGenomeDRAW [12]. As shown in the map, the regions matching gene information were marked on the map with the relevant information, whereas the regions without any match (indicated by a solid line) are Variable Number of Tandem Repeat (VNTR) regions. Thus, the process was designed to automatically find these regions using NGS data. According to the information about *Rhodomonas salina* (accession number: NC_002572) in GenBank, most repeat regions were tandem repeats, which were reported as simple repeats.

3.2 Method

Here, the aim was to analyze repeat regions using below flow chart (Fig. 2). The yellow box stands for the dataset and empty boxes are process for each steps. NGS read results on *Rhodomonas salina* were used as in Sect. 3.1. Reads were assembled using three independent assemblers. To comparatively analyze repeat regions of the mitochondrial genome of *Rhodomonas salina*, mitochondrial data in fasta format were retrieved from NCBI to establish a repeat library using RepeatModeler. The established repeat library was used to find repeat regions by means of three assemblers.

3.3 Result

Table 2 shows the results of analysis of sequences assembled by three assembly tools (MIRA, SPAdes, and Velvet). N50 refers to the contig length at the position corresponding to the half of the total assembled contigs from the genome sequencing. In other words, assembled contigs of various lengths were sorted in the descending order, and all the contig lengths were summed up from the longest to the shortest, where the contig length with the median value was regarded as N50. N50 length is sometimes used to evaluate performance of an assembler, where a higher N50 value means a relatively larger contig size. The number of contigs means the number of assembled contigs, which can be used to estimate how many contigs were used for assembly.

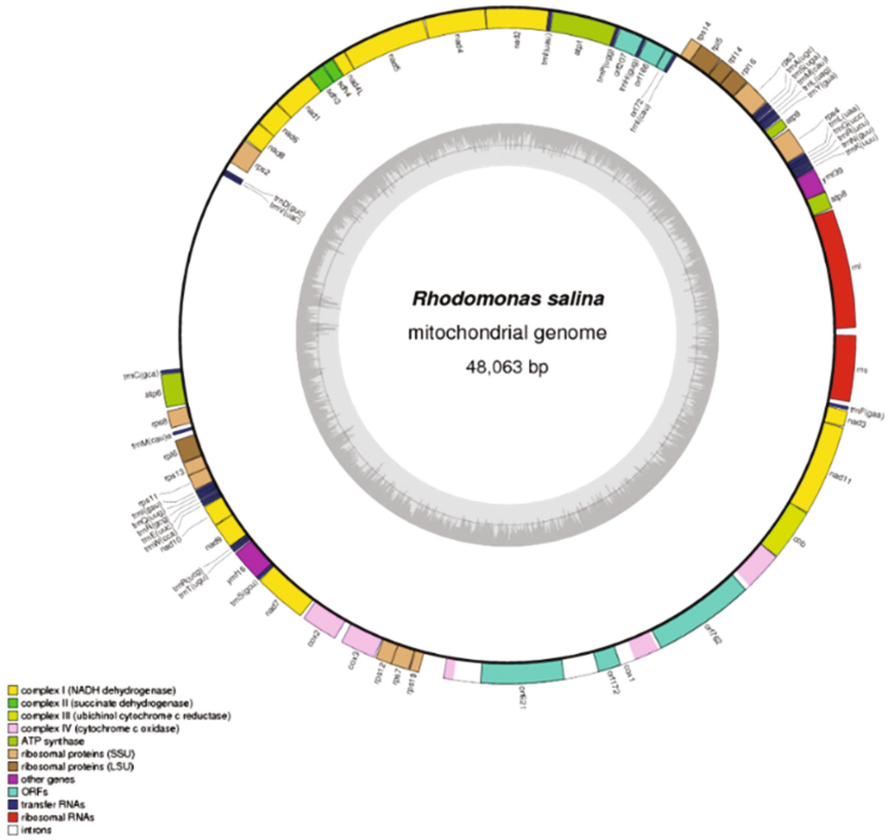


Fig. 1. Map of *Rhodomonas salina* (accession number: NC_002572)

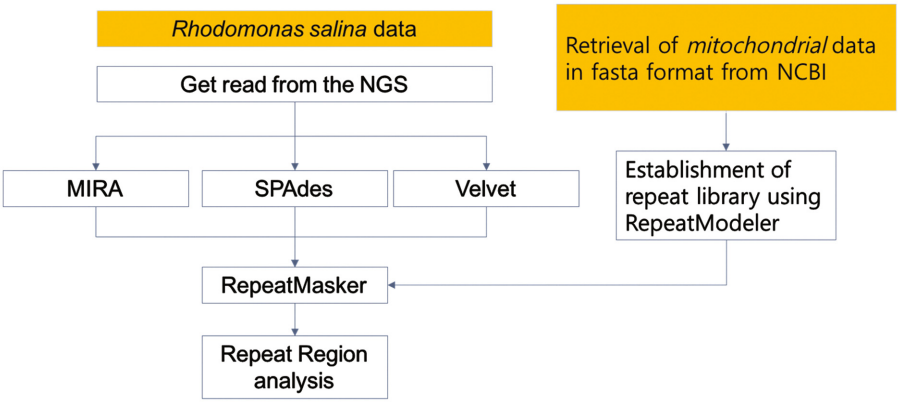


Fig. 2. The flow chart of the analysis of repeat regions

Table 2. The repeat database from RepeatModeler

	Number of contigs	N50
MIRA	1285	6729
SPAdes (K = 21)	229049	234
Velvet	421956	131

Results from the three assemblers are summarized in Table 3. MIRA produced the longest contig, and SPAdes increased the number of contigs by setting a K value smaller than default. Velvet had a higher contig number and the smallest N50 value. In general, performance of an assembler on a genome that has available references is highly dependent on the k value and assembly type of the software.

Table 3. Finding the repeat categories from the different assemblers software

Repeat category	Number of elements			Length occupied		
	MIRA	SPAdes	Velvet	MIRA	SPAdes	Velvet
SINEs	0	0	0	0	0	0
LINEs	1	1	2	64 bp	64 bp	107 bp
LTR elements	6	7	28	1312 bp	1406 bp	2469 bp
Unclassified	241	389	1842	71016 bp	88058 bp	165355 bp
Small RNA	0	0	0	0	0	0
Simple repeats	294	4838	5618	12334 bp	191533 bp	236889 bp
Low complexity	26	890	1058	1226 bp	42058 bp	63952 bp

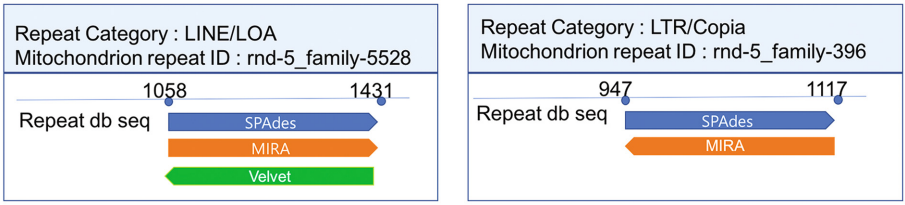


Fig. 3. The examples of the commonly founded repeats

The contig number and N50 number of assembly results for the cryptophyte alga were dependent on the characteristics of each assembly tool. Nonetheless, matching the assembly results to the mitochondrial database from RepeatModeler resulted in data shown in Table 3. Figure 3 shows examples of repeats that were commonly found in data from Table 3. The left panel on the top of Fig. 3 indicates that sequence #5528 from RepeatModeler was commonly mapped to Long interspersed nuclear elements (*LINEs*) among the repeat categories. Moreover, the direction of arrows stands for the mapping direction. That is, SPAdes and MIRA matches the forward direction, but Velvet aligns the backward one. Sequences assembled by SPAdes and MIRA were

mapped to the positive strand in the database, while the sequence from Velvet was aligned with the negative strand. The right panel of Fig. 3 shows that the sequences assembled by MIRA and SPAdes were mapped to entry #396 in the database among Long terminal repeats (*LTRs*). In the assembled sequences from Velvet, the corresponding item was partially aligned as fragments. For *Unclassified* and *Simple sequences*, there was a commonly found sequence library, while some sequences had different starting and ending positions when aligned. Thus, when contigs were assembled from NGS data by different assemblers, there were repeat library sequences common for the different assemblers, leading to identification of repeat regions.

4 Conclusion

NGS yields fragmented reads as results, and these read results need to be assembled to complete a genome map, which is the core of genome analysis. Although substantial portions of eukaryotic genomes have been found to be repeat regions, assembly methods such as de novo assembly have limitations in terms of finding repeat regions. In this study, we were able to identify repeat regions that were commonly mapped to contigs by different assembly methods. These regions can be considered candidate groups for repeat regions. In the future, automatic filtering components should be added through comparison with already known repeat regions. This approach will lay the foundation for reliable assembly of repeat regions without experimentation.

Acknowledgments. This work was supported by the Hongik University new faculty research support fund.

References

1. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**, 821–829 (2008)
2. Luo, R., et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**, 18 (2012)
3. Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I.: ABySS: a parallel assembler for short read sequence data. Genome Res. **19**(6), 1117–1123 (2009)
4. Chevreux, B., et al.: Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. **14**, 1147–1159 (2004)
5. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Pevzner, P.A.: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**(5), 455–477 (2012)
6. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754–1760 (2009)
7. Tempel, S.: Using and understanding RepeatMasker. Methods Mol. Biol. **859**, 29–51 (2012)
8. Jurka, J.: Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. **9**, 418–420 (2000)
9. Smit, A.F.A., Hubley, R.: RepeatModeler Open-1.0 (2008–2015). http://www.repeatmasker.org

10. Bao, Z., Eddy, S.R.: Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. **12**, 1269–1276 (2002)
11. Price, A.L., Jones, N.C., Pevzner, P.A.: De novo identification of repeat families in large genomes. In: Proceedings of the 13 Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2005) (2005, to appear)
12. http://ogdraw.mpimp-golm.mpg.de/

Senior Tourism and Information and Communication Technologies

Blanka Klimova(✉)

University of Hradec Kralove, Hradec Králové, Czech Republic
blanka.klimova@uhk.cz

Abstract. Current demographic trends show that there is a growing number of older population groups, which might result in serious social and economic problems in future. Therefore national governments want to prevent them by implementing different strategies which could contribute to the maintenance and enhancement of quality of life of older people. One of the approaches to handle this issue is also traveling which can be enhanced by information and communication technologies (ICT). The purpose of this article is to explore the use of ICT by senior travelers. This is done by literature search of available studies on the research topic in the world's databases Web of Science, Scopus, ScienceDirect, and Springer. The findings show that senior tourists are now more technologically savvy than they used to be two decades ago. In addition, their use of ICT reflect their confidence and independence in travelling. Nevertheless, since senior tourism is a new developing branch of tourism, much more research has to be done in this field, including the research on the use of ICT by these older travellers.

Keywords: Information and communication technologies · Seniors · Tourism · Trends · Types

1 Introduction

At present senior tourism is rapidly expanding due to the current demographic changes which result in the increase of older population groups worldwide. In 2000 the percentage of older individuals aged 65+ years reached 12.4%. In 2030, this number should rise to 19% and in 2050 to 22% [1]. In Europe this population group aged 65+ represent 18% of the 503 million Europeans, which should almost double by 2060 [2]. Therefore there is constant effort to maintain these older people active as long as possible, improve the quality of their life [3–5] and enable them to stay socially and economically independent [6]. One of the approaches to handle this issue is also traveling which can be enhanced by information and communication technologies (ICT). Since senior tourism is a quite new branch of tourism and it has not been clearly and comprehensibly defined yet, much research needs to be done in this area [7]. This is confirmed by the fact that there exist different names for this type of tourism. Apart from the term senior tourism, mature age tourism or third-age tourism are interchangeably used. Currently, senior tourism is considered to be part of the so-called

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_71

accessible tourism since the disability is related to older people. Accessible tourism was for the first time comprehensibly defined by Darcy and Dickson in 2009 as follows [8]:

Accessible tourism enables people with access requirements, including mobility, vision, hearing and cognitive dimensions of access, to function independently and with equity and dignity through the delivery of universally designed tourism products, services and environments. This definition is inclusive of all people including those travelling with children in prams, people with disabilities and seniors.

Since senior tourism is a new developing branch in tourism, the use of ICT in it has not been researched much either. Therefore the purpose of this article is to explore the use of ICT by senior citizens with respect to tourism.

2 Methods

The methods used for this article include a literature search of available studies on the research topic in the world's acknowledged databases such as Web of Science, Scopus, ScienceDirect, and Springer. The search was based on the key words: *accessible tourism* AND *information and communication technologies, mature age* AND *information and communication technologies, tourism* AND *information and communication technologies, third-age tourism* AND *information and communication technologies, senior tourism* AND *information and communication technologies*, *senior tourism* AND *information and communication technologies* in the period of 2000 till present. In addition, methods of comparison and evaluation of the findings from the selected studies were applied. Most of the studies, however, started to occur after 2010, when it became obvious that there would be a rise in the older population groups and these *newly retired baby boomers* would have slightly different needs in comparison with the previous generation, for example, as far as the use of ICT is concerned.

3 Seniors and Their Use of ICT in Tourism

Generally, there are three main preconditions which support the rising use of ICT by older people. These include:

- an increase in the number of older people worldwide;
- ICT as a tool for providing older people with the promise of greater independence;
- the generation of "baby boomers" approaching retirement being relatively comfortable using ICT; they will bring many technology-related skills into their retirement years [9, 10].

The present older generation is much better at using ICT than it used to be two decades ago [11, 12]. Nowadays, more than 80% of older people in the developed countries have access to computers [13]. Apart from e-mail messaging and writing documents, [13, 14] they start exploiting other tools such as sharing photos or calendar information, [15] video chats, [16] or web blogs [17]. Overall, younger older people use ICT more frequently and confidently than their older counterparts.

In senior tourism, altogether three age range groups of seniors can be detected due to slightly different needs of these three groups (55–64, 65–74, 75 and more). The first group includes more active seniors with higher incomes and desire to travel to exotic countries. As the age increases, these aspects, including the use of ICT, gradually decrease in the other two age groups [18].

Although the youngest group of senior tourists is digitally competent, still the word-of-mouth information is the most frequent source for their travel arrangements [7]. The EU report on Preferences of Europeans towards Tourism [19] states that 44% of the senior travellers take advantage of the word-of-mouth information for their travel arrangements, which is then followed by 30% of the seniors who use the Internet and 29% of the senior tourists trust personal experience.

The most recent Finnish study by Pesonen, Komppula, and Riihinen [20] on the use of ICT by senior travelers has revealed that there could be three different types of senior travelers resulting from their use of online travel services. Figure 1 illustrates these three types of senior travelers and provides their description.

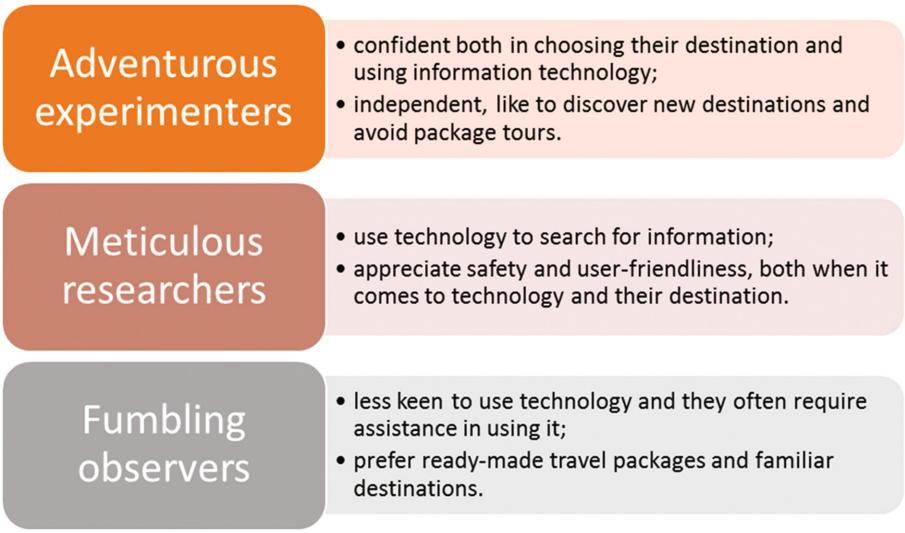


Fig. 1. Typology of senior travelers resulting from their use of online travel services, author' own processing based on the data from [20]

Thus, the findings on the typology of senior travelers based on their active use of ICT indicate that the higher adherence to the use of ICT by senior travelers means their higher independence when it comes to travel arrangements and travelling itself. In fact, with the increasing number of seniors, the market offer in senior tourism will be definitely changing and since this group of people is still heterogeneous, the offer of travel providers will have to reflect individual senior travelers' needs, including those who will be less technologically competent. Therefore travel providers will have to

offer more personalized services for these different groups of senior travelers. However, they should always care about some common specifications which the tourism product should possess (Fig. 2) [21].

Fig. 2. Specifications of a senior traveler product (author's own processing)

4 Conclusion

The findings show that senior tourists are now more technologically savvy than they used to be two decades ago, especially the adventurous travellers are eager to use smart tourism digits during their travels. In addition, the results indicate that the use of ICT reflect seniors' confidence and independence in travelling. Nevertheless, since senior tourism is a new developing branch of tourism, much more research has to be done in this field, including the research on the use of ICT by these older travellers.

Acknowledgments. This work was supported by IGS project 2017 run at the Faculty of Informatics and Management of the University of Hradec Kralove in the Czech Republic.

References

1. Transgenerational Design Matters (2009). http://transgenerational.org/aging/demographics.htm
2. Patterson, I.: Growing Older. Tourism and Leisure Behaviour of Older Adults. Cabi, Cambridge (2006)
3. Klimova, B., Kuca, K.: Alzheimer's disease: potential preventive, non-invasive, intervention strategies in lowering the risk of cognitive decline – a review study. J. Appl. Biomed. **13**(4), 257–261 (2015)
4. Klimova, B., Maresova, P., Kuca, K.: Non-pharmacological approaches to the prevention and treatment of Alzheimer's disease with respect to the rising treatment costs. Curr. Alzheimer Res. **13**(11), 1249–1258 (2016)
5. Klimova, B., Maresova, P., Valis, M., Hort, J., Kuca, K.: Alzheimer's disease and language impairments: social intervention and medical treatment. Clin. Interv. Aging **10**, 1401–1408 (2015)
6. Klimova, B., Maresova, P.: Elderly People and Their Attitude Towards Mobile Phones and Their Applications – A Review Study. LNEE, vol. 393, pp. 31–36 (2016)
7. Alen, E., Dominguez, T., Losada, N.: New opportunities for the tourism market: senior tourism and accessible tourism (2012)
8. Darcy, S., Dickson, T.: A whole-of-life approach to tourism: the case for accessible tourism experiences. J. Hospitality Tourism Manag. **16**(1), 32–44 (2009)
9. Virginia Assistive Technology System: Assistive technology and aging. A handbook for Virginians who are aging and their caregivers. http://www.vda.virginia.gov/pdfdocs/Assistive%20Technology%20&%20Aging%20-%20All.pdf
10. Klimova, B.: Use of the Internet as a prevention tool against cognitive decline in normal aging. Clin. Interv. Aging **11**, 1231–1237 (2016)
11. Klimova, B., Simonova, I., Poulova, P., Truhlarova, Z., Kuca, K.: Older people and their attitude to the use of information and communication technologies – a review study with special focus on the Czech Republic (Older people and their attitude to ICT). Educ. Gerontol. **42**(5), 361–369 (2016)
12. Wu, Y.H., Damnee, S., Kerherve, H., Ware, C., Rigaud, A.S.: Bridging the digital divide in older adults: a study from an initiative to inform older adults about new technologies. Clin. Interv. Aging **10**, 193–201 (2015)
13. Heart, T., Kalderon, E.: Older adults: are they ready to adopt health-related ICT? Int. J. Med. Inform. **82**, e209–e231 (2013)
14. Sayago, S., Blat, J.: Telling the story of older people e-mailing: an ethnographical study. Int. J. Hum. Comput. Stud. **68**, 105–120 (2010)
15. Brush, A.B., Inkpen, K.M., Tee, K.: SPARCS: exploring sharing suggestions to enhance family connectedness. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, San Diego, pp. 629–638 (2008)
16. Ames, M.G., Kaye, J., Spasojevic, M.: Making love in the network closet: the benefits and work of family video chat. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, Savannah, Georgia, USA, pp. 145–154 (2010)
17. Thurlow, C., Lengel, L., Tomic, A.: Computers-Mediated Communication. Social Interaction and the Internet. Sage, London (2009)
18. Slavik, L.: Senior tourism. Diploma Thesis, Praha, UK (2012)
19. Pesonen, J., Komppula, R., Riihinen, A.: Typology of senior travellers as users of tourism information technology. Inf. Technol. Tourism **15**(3), 233–252 (2015)

20. European Commission: Preferences of Europeans towards Tourism (2015). http://ec.europa.eu/geninfo/query/resultaction.jsp?swlang=en&QueryText=senior+travellers&sbtSearch=Search
21. CBI Product factsheet (2016). https://www.cbi.eu/sites/default/files/market_information/researches/product-factsheet-europe-senior-travel-2015.pdf

Wearable and Portable Monitoring Devices for Older People

Blanka Klimova(✉) and Petra Maresova

Faculty Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, Czech Republic
{blanka.klimova, petra.maresova}@uhk.cz

Abstract. Currently, there is a growing number of older people worldwide. This demographic change results in serious social and economic problems. Therefore governments, especially in developed countries, attempt to intervene in this process and help to enhance quality of life of older population groups with different means. One of the approaches in this respect is the use of mobile and wireless technologies in healthcare. The purpose of this article is to discuss the use of wearable and portable monitoring devices for older people in three areas of healthcare: fall detection, dementia care and low access to healthcare. The findings revealed that there is a lack of clinical studies examining the use of wearable devices in healthcare for older people. Therefore more research should be performed because the benefits of the use of wearable devices, such as their unobtrusiveness, sensitivity, or reliability, can contribute to the enhancement of quality of life of older people.

Keywords: Wearable and portable monitoring devices · Older people · Fall detection · Dementia care · Low-access · Benefits

1 Introduction

At present there is a steady increase of older population groups. In 2013 there were around 45 million people at the age of 65+ living worldwide. By 2020 this number of older people should reach 98 million [1]. In the developed countries of Europe, the proportion of older people at the age of 65+ is estimated to grow from 18.2% in 2013 to 28.1% in 2050 [2, 3]. These demographic changes obviously bring about serious problems such as important social and economic issues [4–6]. Thus, there is an attempt to increase quality of life of this older population groups in order to support them to stay independent as long as possible. This is also the goal of the world's governments. They are working on the strategic plans which would meet this aim [7]. One of the approaches which the governments try to implement in this process of the improvement of quality of life of older people is the use of information and communication technologies (ICT) for the maintenance of their health problems. They have good prerequisites for it because the present generation of seniors is relatively digitally aware than they used to be ten years ago [8]. Furthermore, they are comfortable with the exploitation of the so-called eHealth technologies, i.e., mobile and wireless devices which can help them improve their health. Especially, wearable technology is

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_72

becoming increasingly popular nowadays. It includes electronics that can be worn on the body, either as an accessory or as part of material used in clothing. One of the major features of wearable technology is its ability to connect to the Internet, enabling data to be exchanged between a network and the device [9]. As Arean, Ly, and Andersson [10] claim, wearable sensors have the potential to collect a warehouse full of physiological, social, emotional, and behavioral data in real time with limited burden on the client. These data can enable important decisions about treatment options and monitoring response. The sensors are objects whose purpose is to detect events or changes in its environment, and then provide a corresponding output. They may provide various types of output, but typically use electrical or optical signals [11, 12].

The purpose of this article is to discuss the use of wearable and portable monitoring devices for older people in three areas: fall detection, dementia care and low access to healthcare.

2 Methods

The authors used a method of literature review of available sources exploring research studies focused on Wearable and portable monitoring systems for older people in the acknowledged databases and a method of comparison and evaluation of their findings. This review was done by searching databases such as Web of Science, Science Direct, and Springer from 2010 until September 2016 for the following key words: *elderly people* AND *wearable and portable monitoring devices*. In addition, other relevant studies were reviewed on the basis of the reference lists of the research articles from the searched databases. The selection period starts with the year of 2010 since this is the year when older adults started to be more digitally literate and able to use technologies on a daily basis [13], as well as to trust in the effects of wearable devices for the improvement of their health.

3 Use of Wearable and Portable Monitoring Devices for Older People

Wearable technology covers a broad area of devices. With its use becoming more common in the healthcare sector the issue concerning privacy becomes more crucial. New devices can help physicians monitor patients' vital signs; sleep patterns and heart rhythms remotely transforming the face of medicine as we know it. These developments in technology will help detect early signs of diseases and aid in diagnosing medical conditions. Essentially these devices are mini computers that send and receive data which can be used for further analysis. The data that these incredibly powerful devices collect can be stolen, which raises concerns for data safety and encryption. Furthermore, there is a valid concern that companies will use large amounts of personal healthcare data for marketing and insurance purposes [14]. As they grow more sophisticated, wearable health devices will help users track and even diagnose various conditions and potentially advise a course of action or, more simply, remind users to take medications or contact medical professionals as necessary. In the process, these

health and fitness devices and applications could eventually become "lifestyle remotes" that help consumers control or automate many other systems around them, regardless of whether they are in their homes, offices, cars, or the like. [15] The wearable and portable devices are mainly used for monitoring mobility, physical activity, and rehabilitation progress. [16] At present, they are widely used for monitoring older people in the following three area:

- Fall detection
- Dementia care
- Low access to healthcare

3.1 Fall Detection

Fear of falling has been shown to be associated with negative consequences such as avoidance of activities, less physical activity, falling, depression, decreased social contact and lower quality of life. Over one third of older people at the age of 65+ fall every year. 10% of these falls usually result in a serious injury. Therefore there is an attempt to prevent and detect these falls [17].

The fall detectors can be divided into two types: context-aware systems and wearable devices. [18] The context –aware systems use sensors such as cameras, floor sensors, or microphones to detect falls in the environment, while the wearable devices are miniature electronic sensor-based devices that are worn by the bearer under, with or on top of clothing. These are usually accelerometers which collect data during the falls.

As the research studies illustrate, especially the floor sensors [19–22] and wearable cameras [23–27] have been widely tested. The tests with the floor sensors are usually conducted with computer generated subjects. These tests are based on a series of predefined simulated movements which are generated to simulate an elderly person living alone, or cohabiting with 1 or 2 family members. In particular, ADLs, a fall from bed after waking up, a fall after getting up from a chair, and a fall when walking or standing are simulated. Then several scenarios are designed to provide a successful recovery, remaining unconscious, or unable to stand. The simulator can also produce a number of simple and complex scenarios involving 1 or more people walking. Similarly, the wearable cameras, usually with some volunteers, attempt to simulate unpredicted falls, both indoor and outdoor surroundings.

3.2 Dementia Care

Nowadays, there are about 44 million people living with dementia and by 2050 this number is expected to triple due to increasing life expectancies and the aging population worldwide. [28] People with dementia experience a considerable loss of memory, orientation problems, impaired communication skills, depression, behavioral changes and confusion. [29] Wearable monitoring devices can at least help them with some of these problems such as disorientation and finding their way back home. Currently, there are several tracking devices which help both patients and their

caregivers to manage this situation. [30] These include, for example, iTraq, which is a global location devices. Its use is based on cellular towers which can enable to identify person's location. It is a small device in the form of a credit card and its location is available through a mobile application. Another monitoring device for people with dementia is Mindme, which uses GPS to provide location, updates and an alarm if a person moves out of a preset location. It is again a small device in the size of an electronic care key. Similarly, PocketFinder provides a GPS location of the lost person. It is also small, water-proof and have one of the longest battery lives. It can provide updates "at the touch of a button" on a computer or mobile device, through email and text notifications. It also stores the person's monitoringTracking history up to 60 days and provides an unlimited number of "geo-fences" that alert you when the GPS leaves a specified area and provides up to 60 days of tracking history.

In addition, healthcare providers can obtain with the help of wearable devices information from caregivers who look after the people with dementia. [31] For example, Matthews et al. [31] designed a wearable and wireless camera system to capture in situ the challenges of dementia caregiving over many hours per day at home. It goes where the user goes to capture whatever the user sees, hears, says, and does, with the user retaining control of when and what data are collected. The system consists of an outward-facing CMOS camera with a fisheye lens, a MEMS microphone that records ambient sound, and the electronics, battery, and garment necessary to support their function. The camera is mounted on a custom-printed circuit board enclosed within a plastic housing that is clipped onto the frames of overglasses, to ensure data capture at eye level and to isolate the circuit from the user's head. A cable connects the camera housing to a computer that is encased in a second housing and, in turn, embedded within the upper, inner pocket of a multi-layered "video vest".

3.3 Low Access to Healthcare

Wearable and portable devices can also help older people in remote areas whose access to healthcare is thus limited. These wireless monitoring sensors can lower the costs of transportation of these people to their doctor or hospital. Furthermore, these wireless sensors enable that these people living in the remote rural areas do not have to be present at the doctor for their diagnostic. The doctor can use the data collected through these wireless sensors attached to their bodies and then provide them with real-time diagnosis advices which are important to their recovery. These monitoring devices can play an important role in case of emergency since sensors can autonomously send data about the patient health [32].

4 Conclusion

This study aimed to emphasize the importance of the use of wearable and portable monitoring devices for older people in healthcare, especially in its three areas: fall detection, dementia care and low access to healthcare. The findings revealed that there was a lack of clinical studies examining the use of wearable devices in healthcare for

older people. Therefore more research should be performed because the benefits of the use of wearable devices can contribute to the enhancement of quality of life of older people.

Acknowledgments. This paper is published thanks to the support of SPEV project 2017 run at the Faculty of Informatics and Management of the University of Hradec Kralove. The authors also thank Jan Freiberg for his help with data collection.

References

1. AoA (2016). http://www.aoa.acl.gov/aging_statistics/index.aspx
2. Bujnowska-Fedak, M.M., Grata-Borkowska, U.: Use of telemedicine-based care for the aging and elderly: promises and pitfalls. Smart Homecare Technol. TeleHealth **3**, 91–105 (2015)
3. Erostat: Databse Population (2016). http://ec.europa.eu/eurostat/data/database
4. Klimova, B., Maresova, P., Kuca, K.: Non-pharmacological approaches to the prevention and treatment of Alzheimer's disease with respect to the rising treatment costs. Current Alzheimer Research (2016, in press)
5. Klimova, B., Maresova, P., Valis, M., Hort, J., Kuca, K.: Alzheimer's disease and language impairments: social intervention and medical treatment. Clin. Interv. Aging **10**, 1401–1408 (2015)
6. Maresova, P., Mohelska, H., Dolejs, J., Kuca, K.: Socio-economic aspects of Alzheimer's disease. Curr. Alzheimer Res. **12**, 903–911 (2014)
7. National Strategy for an Ageing Australia (2001). http://www.ifa-fiv.org/wp-content/uploads/2012/11/062_Australia-2001-National-Strategy-for-an-Ageing-Australia.pdf
8. Virginia Assistive Technology System: Assistive technology and aging. A Handbook for Virginians Who Are Aging and Their Caregivers (2016). http://www.vda.virginia.gov/pdfdocs/Assistive%20Technology%20&%20Aging%20-%20All.pdf
9. Wearable Technology (2016). http://www.investopedia.com/terms/w/wearable-technology.asp#ixzz4LLX0aRe7
10. Arean, P.A., Ly, K.H., Andersson, G.: Mobile technology for mental assessment. Dialogues Clin. Neurisci. **18**, 163–169 (2016)
11. Kretschmar, M., Welsby, S.: Capacitive and Inductive Displacement Sensors, Sensor Technology Handbook. Newnes, Burlington (2005)
12. Semantic Sensor Network XG Final Report (2016). http://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/
13. Klimova, B., Simonova, I., Poulova, P., Truhlarova, Z., Kuca, K.: Older people and their attitude to the use of information and communication technologies – a review study with special focus on the Czech Republic (older people and their attitude to ICT). Educ. Gerontol. **42**, 361–369 (2016)
14. Dave, P.: Privacy and security in an age of wearable devices. Wearable Devices (2016). http://www.wearabledevices.com/2016/01/06/privacy-security-age-wearable-devices/
15. Thierer, A.: The internet of things and wearable technology. Mercatus Working Paper (2014). http://mercatus.org/sites/default/files/Thierer-Wearable-Tech.pdf
16. Kang, H.G., Mahonez, D., Hoenig, H., Hirth, V.A., Bonato, P., Hajjar, I., et al.: In situ monitoring of health in older adults: technologies and issues. J. Am. Geriatr. Soc. **58**, 1579–1586 (2010)

17. Chaudhuri, S., Thompson, H., Demiris, G.: Fall detection devices and their use with older adults: a systematic review. J. Geriatr. Phys. Ther. **37**, 178–196 (2014)
18. Igual, R., Medrano, C., Plaza, I.: Challenges, issues and trends in fall detection systems. Biomed. Eng. Online **12**, 66 (2010)
19. Litvak, D., Zigel, Y., Gannot, I.: Fall detection of elderly through floor vibrations and sound. In: Conference Proceedings IEEE Engineering in Medicine and Biology Society, pp. 4632–4635 (2008)
20. Ariani, A., Redmond, S.J., Chang, D., Lovell, N.H.: Software simulation of unobtrusive falls detection at night-time using passive infrared and pressure mat sensors. In: Conference Proceedings IEEE Engineering in Medicine and Biology Society, pp. 2115–2118. (2010)
21. Zigel, Y., Litvak, D., Gannot, I.: A method for automatic fall detection of elderly people using floor vibrations and sound–proof of concept on human mimicking doll falls. IEEE Trans. Biomed. Eng. **56**, 2858–2867 (2010)
22. Zhang, Z., Kapoor, U., Narayanan, M., Lovell, N.H., Redmond, S.J.: Design of an unobtrusive wireless sensor network for nighttime falls detection. In: Conference Proceedings IEEE Engineering in Medicine and Biology Society, pp. 5275–5278 (2011)
23. Bourennane, W., Charlon, Y., Bettahar, F., Campo, E., Esteve, D.: Homecare monitoring system: a technical proposal for the safety of the elderly experimented in an Alzheimer's care unit. IRBM **34**, 92–100 (2013)
24. Charlon, Y., Bourennane, W., Bettahar, F., Campo, E.: Activity monitoring system for elderly in a context of smart home. IRBM **34**, 60–63 (2013)
25. Doukas, C.N., Maglogiannis, I.: Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components. IEEE Trans. Inf Technol. Biomed. **15**, 277–289 (2011)
26. Gietzelt, M., Spehr, J., Ehmen, Y., et al.: GAL@Home: a feasibility study of sensor-based in-home fall detection. Z. Gerontol. Geriatr. **45**, 716–721 (2012)
27. Tasoulis, S.K., Doukas, C.N., Maglogiannis, I., Plagianakos, V.P.: Statistical data mining of streaming motion data for fall detection in assistive environments. In: Conference Proceedings IEEE Engineering in Medicine and Biology Society, pp. 3720–3723 (2011)
28. Langa, K.M., Foster, N.L., Larson, E.B.: Mixed dementia: emerging concepts and therapeutic implications. JAMA **292**, 2901–2908 (2004)
29. Klimova, B., Kuca, K.: Speech and language impairments in dementia – a mini review. J. Appl. Biomed. **14**, 97–103 (2016)
30. Dementia and elderly GPS tracking devices (2015). http://www.aplaceformom.com/blog/4-29-15-dementia-and-elderly-gps-tracking-devices/
31. Matthews, J.T., Lingler, J.H., Campbell, G.B., Hunsaker, A.E., Hu, L., Pires, B.R., et al.: Usability of a wearable camera system for dementia family caregivers. J. Healthc. Eng. **6**, 213–238 (2015)
32. Abidoye, A.P., Azeez, N.A., Adesina, A.O., Agbele, K.K., Nyongesa, H.O.: Using wearable sensors for remote healthcare monitoring systém. J. Sens. Technol. **1**, 22–28 (2011)

A Novel Anomaly Detection Method in Wireless Network Using Multi-level Classifier Ensembles

Bayu Adhi Tama and Kyung-Hyune Rhee(✉)

Laboratory of Information Security and Internet Applications,
Department of IT Convergence and Application Engineering,
Pukyong National University, Busan, South Korea
bayuat@pukyong.ac.kr, khrhee@pknu.ac.kr

Abstract. Anomaly detection is very crucial in an intrusion detection task since it has capability to discover new types of attacks. The major challenges of anomaly detection are how to maximize the accuracy while maintaining low positive rate. In this paper, we propose new approach on anomaly detection using multi-level classifier ensembles. We employ an ensemble learner as a base classifier of ensemble rather than a single classifier algorithm. We run several experiments to choose the best combination of two-level classifier ensemble model. From our experimental result, it is revealed that the performance of our proposed approach yields satisfactory results over classical classifier ensembles and single classifiers.

Keywords: Two-level ensembles · Classifier ensembles · Anomaly detection · Wireless network

1 Introduction

Classifier ensembles has been active research in the last two decades. They have been applied to many real-world applications such as computer security, credit and financial risk analysis, medicine, and recommender system. When a single classification algorithm cannot guarantee us to have a good performance, the combination of multiple classifiers might give a better solution. The key success factor of classifier ensembles is splitting the original dataset into several subsets, classifier algorithm is trained using each subset, and then taking voting combination for final prediction. Diversity can be obtained either from random sampling of feature or instance.

Classifier ensembles have been considered to solve underlying problems in anomaly detection such as maximizing detection rate and reducing false positive rate [1]. Technically speaking, anomaly detection is binary classification problem and the classical classifier ensembles still might not produce good results due to poor ensemble design [2]. Diversity is one aspect in the classifier ensemble design, so that it is believed to improve classifier performance. However, so far it has not been proven yet either empirically or theoretically. In contrary, in practical applications, classifier ensembles have shown improved performance and robustness to noisy data.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_73

Concerning to this, we focus on designing a classifier ensemble using other ensemble as a base classifier. By employing ensemble in ensemble, we intend to maximize the diversity of an ensemble using both feature subsets and bootstrapping samples from the training set. The contributions of this paper can be summarized as follows: (i) we introduce a new approach of anomaly detection in wireless network using two-level classifier ensembles; (ii) we show that the proposed model yields the promising result for anomaly detection; and (iii) we thoroughly compare and discuss a number of classifier models based on two-level classifier ensembles, classical ensemble model, and single classifier.

The remaining part of this paper is structured as follows. Section 2 briefly describes existing approaches for anomaly detection. Section 3 details the proposed method, whilst Sect. 4 presents result and discussion. Finally, some concluding remarks are drawn in Sect. 5.

2 Literature Review

Earlier approach for anomaly detection using classifier ensemble is proposed by [3]. Ensemble of neural network, support vector machine, and multivariate regression splines are fused using majority voting for final class prediction. KDD Cup 99 dataset is used for classifier modeling and testing. The classifier performance is measured using accuracy metric. However, because of many redundant records are exist in KDD Cup 99 dataset, the performance result of the classifier is biased. To overcome this problem, a new dataset, namely NSL-KDD is proposed [4].

Moreover, GPRS dataset is proposed since the number of available dataset specific to wireless networks is quite limited [5]. It is deployed based upon the intrusion detection on the IEEE 802.11 environment. Three traditional classifier, i.e. multilayer perceptron (MLP), radial basis function (RBF), and Bayes network (BN) are used for evaluating the proposed dataset. More recently, a classifier design based on rotation forest [6] for anomaly detection in wireless network is proposed by [7]. They evaluate and benchmark the performance of rotation forest with different number of classifiers as a base classifier.

3 Material and Method

3.1 Dataset

We use the same dataset as it is discussed in the previous works [5, 7]. Dataset comprises two distinct network topologies, i.e. WEP/WPA and WPA2. Either WEP/WPA or WPA2 dataset possesses the same 15 attributes and 1 class label. In this experiment, we consider full training WPE/WPA set which consists of 2 classes, i.e. normal class (6000 instances) and attack class (3600 instances). The full training WPA2 set contains 4500 instances of normal class and 3000 instances of attack class.

3.2 Proposed Method

In this paper, we propose a new approach of anomaly detection in wireless network using two-level classifier ensembles. Unlike in the traditional ensemble models, which comprise a simple prediction model, i.e. support vector machine (SVM), neural network (NN), and decision tree (DT), our proposed approach takes the advantage of ensemble learner as a base classifier of the first-level ensemble. We chose rotation forest [6] and boosting [8] for the first-level and the second-level classifier ensemble, respectively. Furthermore, we also considered conjunctive rule (CR) as a base classifier of the second-level ensemble as indicated by previous work [7].

Actually we could employ other classification ensembles depending on the construction techniques are applied. However, we would intend to maximize the diversity of two classifier ensembles. Rotation forest uses different attribute subsets, whilst boosting uses different training subsets for induction strategies. By integrating these two ensemble approaches we might have the improved performance and construct a robust classifier at once. Figure 1 depicts the proposed method of constructing two-level classifier ensembles for anomaly detection in wireless network.

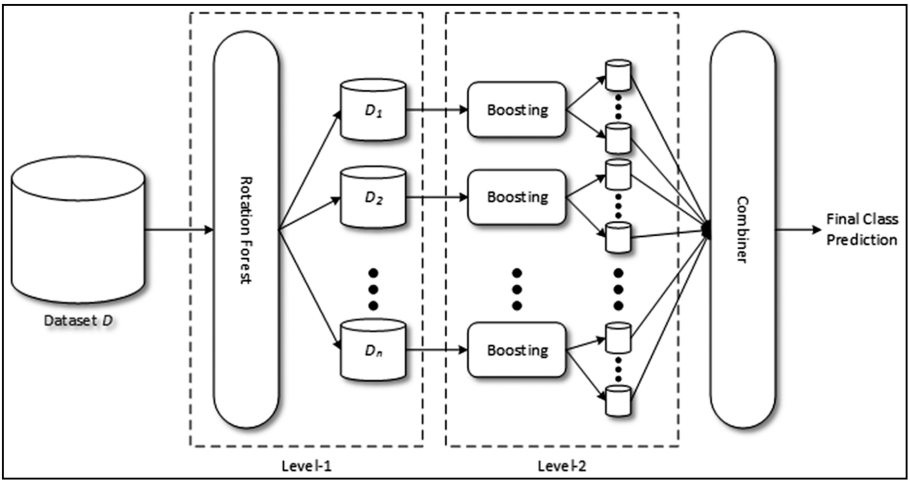


Fig. 1. Construction of two-level classifier ensemble

A two-level classifier ensemble is constructed based on the combination of rotation forest and boosting. Level-1 ensemble generates feature set D into n feature subsets. Afterwards, each feature subset is split into k bootstrap samples in the level-2 ensemble. Final class prediction is made by taking majority voting from a total number of $n \times k$ classifiers.

3.3 Experimental Design

We compare our proposed approach with two traditional ensembles, i.e. boosting of CR (Boost-CR) and rotation forest of CR (RoF-CR). In addition, several single

classifiers, i.e. CR, MLP, RBF, and BN are also considered for further benchmark. 10-fold cross validation (10 *cv*) is used for validation method, whilst accuracy, false positive rate (FPR), precision (detection rate) are used as performance measures.

We further asses the performance differences of the aforementioned classifiers using statistical significance test. Friedman test [9] is employed to prove that the classifier differences are significant. Furthermore, it is meaningful to conduct a post-hoc test after Friedman using Nemenyi test [10–12].

4 Results and Discussion

In this section, we discuss the average performance of seven classifiers in term of accuracy, FPR, and precision metrics. Figures 2 and 3 present the performance results for WEP/WPA and WPA2 dataset, respectively. First of all, we describe the classifier performance for WEP/WPA dataset. It is obvious that our proposed approach outperforms other classifiers significantly in terms of accuracy, FPR, and precision metric. Nevertheless, the proposed approach only outperforms other classifiers in term of accuracy for WPA2 dataset. Moreover, using the same dataset, Bayesian network (BN) is the best performer in terms of FPR and precision metric.

Consecutively, we conduct statistical significant test to further assess the significant differences among classifiers. We only include the accuracy metric for performance assessment. Friedman test ranks the classifiers, with the best classifier receiving rank 1, and the worst classifier receiving rank 7 (number of classifiers) [9, 11]. Nemenyi test

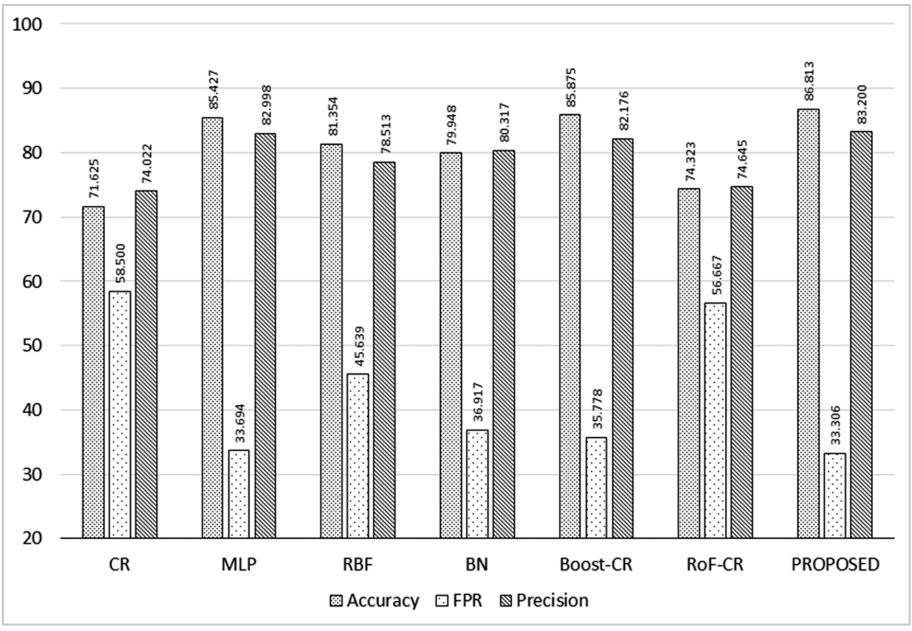


Fig. 2. Performance average of classifiers for WEP/WPA dataset

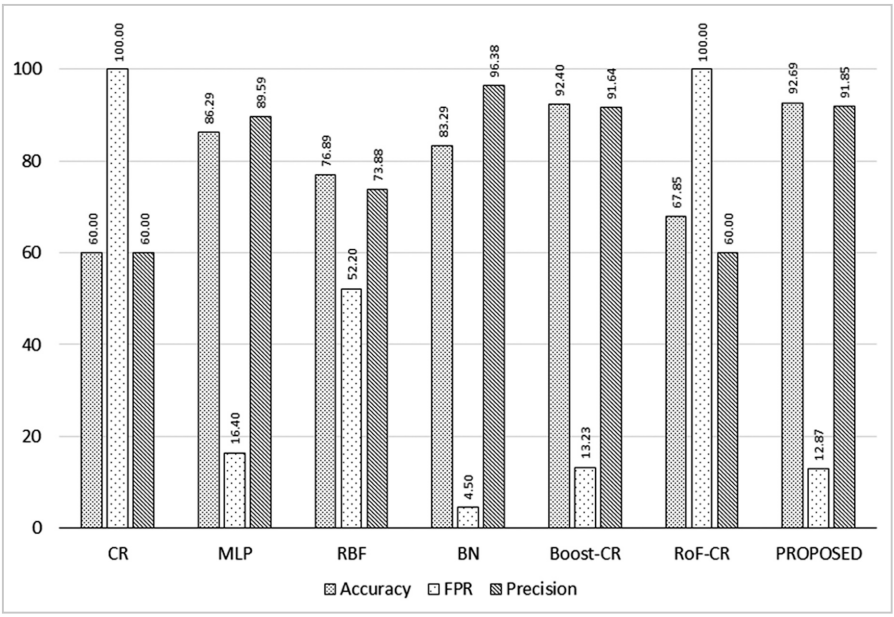


Fig. 3. Performance average of classifiers for WPA2 dataset

indicates that two classifier are significantly different if the corresponding average ranks differ by at least the critical difference (CD).

In our case, the number of elements is equal to the number folds (n = 10), the number of classifier (k = 7), significant level $\alpha = 0.05$, degree of freedom for Nemenyi $(df = k - 1)$, and degree of freedom for Nemenyi $(df = (k - 1)(n - 1))$. The result of significant test using Friedman and Nemenyi test are shown in Table 1. Based on the Friedman test, it can be said that there is highly significant difference $(p < 0.01)$ among the seven classifiers.

Table 1. The results of significant test using Friedman and Nemenyi test

Dataset	Friedman			Nemenyi	
	χ_F^2	df	$p-$value	CD	df
WEP/WPA	55.056	6	4.517E–10	2.9423	54
WPA2	57.161	6	1.695E–10	2.9423	54

Moreover, to determine which classifiers are significantly different, it is necessary to compute the average rankings of the accuracy and then compare which differences are greater than $CD = 2.9423$. From Fig. 4, it can be seen that the proposed classifier is statistically better than any other classifiers. However, with regards to the experimental result of the proposed classifier applied on WEP/WPA dataset, its accuracy is not

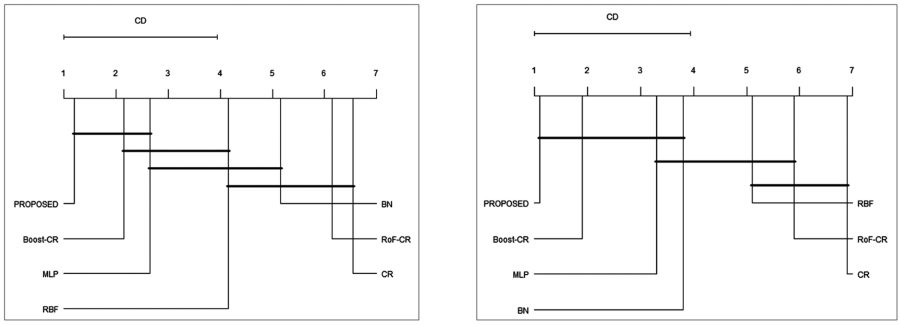


Fig. 4. Plot of critical difference of Nemenyi test for WEP/WPA (left) and WPA2 (right)

significantly different with two classifiers, i.e. Boost-CR and MLP. Performance accuracy of the proposed classifier applied on WPA2 dataset is also similar, which it is not significantly different compared to three classifiers, i.e. Boost-CR, MLP, and BN (as denoted by bold line).

5 Conclusion

We propose a new approach for anomaly detection in wireless network. The proposed classifier is based on two-level ensembles, which an ensemble learner is induced as a base classifier in another ensemble. Based on the experimental results, we prove that the proposed method outperforms both traditional ensembles and single classifiers. For future work, we intend to investigate the proposed classifier's performance applied on other intrusion datasets.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) No. NRF-2014R1A2A1A11052981 and Korean Government Scholarship Program (KGSP) for Graduate 2013–2018.

References

1. Tama, B., Rhee, K.: Performance analysis of multiple classifier system in DoS attack detection. In: Information Security Applications, vol. 9503. Springer, Heidelberg (2016)
2. Tama, B., Rhee, K.: A combination of PSO-based feature selection and tree-based classifiers ensemble for intrusion detection systems. In: Advances in Computer Science and Ubiquitous Computing, pp. 489–495. Springer, Singapore (2015)
3. Mukkamala, S., Sung, A., Abraham, A.: Intrusion detection using an ensemble of intelligent paradigms. J. Netw. Comput. Appl. **28**(2), 167–182 (2005)
4. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.: A detailed analysis of the KDD CUP 99 data set. In: The Second IEEE Symposium on Computational Intelligence for Security and Defence Applications (2009)

5. Vilela, D., Ferreira, E., Shinoda, A., de Souza Araujo, N., de Oliveira, R., Nascimento, V.: A dataset for evaluating intrusion detection systems in IEEE 802.11 wireless networks. In: IEEE Colombian Conference on Communications and Computing (COLCOM), pp. 1–5 (2014)
6. Rodriguez, J., Kuncheva, L., Alonso, C.: Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1619–1630 (2006)
7. Tama, B., Rhee, K.: classifier ensemble design with rotation forest to enhance attack detection of IDS in wireless network. In: 11th Asia Joint Conference on Information Security (AsiaJCIS), Fukuoka, pp. 87–91 (2016)
8. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**(2), 337–407 (2000)
9. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Ann. Math. Stat. **11**(1), 86–92 (1940)
10. Nemenyi, P.: Distribution-free multiple comparisons. Biometrics **18**(2), 263 (1962)
11. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
12. Japkowicz, N., Shah, M.: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, Cambridge (2011)

Learning Reaction Analysis Engine for Interactive Digital Textbook Platform

Kwang Sik Chung(✉)

Department of Computer Science, Korea National Open University, Seoul, Korea
kchung0825@knou.ac.kr

Abstract. As we use smartphones as daily use as televisions and cars, smartphones are always used for daily life. And information technology and sensor technology matures enough to detect user's movement, direction, shapes and bio-signals. With sensor technology and mobile learning device (smartphone, tablet PC) with sensors, it is possible to support and satisfy instant personal learning activity tracking and learner's learning emotional state for learning environment and learning contents. Lastly, on distance learning environment or e-learning environment, learning interactions and feedbacks between learners and teachers is almost impossible, since there is online communication and learning contents which can not deliver or transmit learning emotion of learners. But, in order to deliver personalized learning contents and modify learning environment according to a learner, learning emotional state from learning contents is very important. Learning environment state means analyzes learning environment information consists of learner's property, learner's environment and learner's activity from learner's e-portfolio and learner's educational devices. And besides learning environment states, real-time learning emotional state (learning emotion state) from learning contents is necessary.

In this paper, we design Learning Reaction Analysis Engine for interactive digital textbook platform. Proposed interactive digital textbook platform has Learning Reaction Analysis Engine that could detect learner's learning emotional state that means emotional learning interests and learning concentration state, and Learning Reaction Analysis Engine can analyze learner's learning emotional state from learning activity state, learning device execution state and learners' bio-signals from smartphone, tablet PC, or smart watch. In order to analyze learning emotional state information, we construct learning emotional model that defines learning emotional information, and classification and inference rules of learning emotional information. Learning emotional state inference rules can be applied into Automated Tutoring Engine and Personalized Learning Contents Modification Engine. Learning Reaction Analysis Engine decides learning strategy for type of learning contents frameworks, learning sequence, difficulty change of learning contents and learning contents modification for current learning contents according to a learner who is accessing to the learning contents.

Keywords: u-learning · Automated tutoring engine · Personalization of learning contents · Adaptation of learning contents · Educational devices

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_74

1 Introduction

Rapid development of communication technologies and sensor technologies widely affect various ITC services fields and multimedia contents interactions formations. Especially e-learning and distance learning delivery styles can be newly affected and altered and learning interactions between learners and learning contents. At the previous age of e-learning environment, a learner can communicate with desktop computer, and finish his/her learning activities by himself/herself. These kinds of learning interactions demand selfcontrols of learners and make loneliness of growing or poor learning results. These realities reveals lack of emotional issues in distance learning but also emotional demands of distance learners [1, 2].

But, With learning device sensors (smart phone or Tablet PC, etc.), sensors can initiate interactions with learning contents, and activate learners' learning actions, that are impossible to track learner's learning emotional state on desktop computer or learning devices without sensors. At the first time, a smart phone has limited sensors and can not fully support learners nor track various learning actions and interactions with learning contents. Nowadays, smart phones have several sensors and smart watches can collect bio-signals of a learner so that learners' learning emotional state can be inferred and learners' attitudes toward learning contents and learning environments can be modified according to learners' emotional state. [3] defines "affective computing is emotion-related, from emotions or to influence the emotions of calculation" and calls affective computing trying to make computers as human observation, understanding and the ability to generate a variety of emotions.

Those varieties learners' learning emotional state, that can be produced on learning devices, various learning contents and learning environments, should be tracked by smart learning system. And learning emotional state should be used to establish learning strategy and stimulate learners who can feel boring on learning contents or lose learning interests. But, as the more learners access to distance learning contents with m-learning environment, the more academic affair lecturers and teachers should do for the learners and personalized learning responses can be a big burden to lecturers and teachers. Learning Reaction Analysis Engine is computer system that are designed to infer learners' learning emotional state and make useful information for automated tutoring engine's learning strategy establish. For Learning Reaction Analysis Engine, learners' emotional state model should be established and several sensors of a smart phone and a smart play an important role. We propose Learning Reaction Analysis Engine that tracks and analyses learner's learning emotional state with learner's activity (learning activity, learning interaction), and decides and deliveries appropriate learning contents for each learner.

We propose Learning Reaction Analysis Engine that considers learner activity (learner's learning activity) and learning emotional state (activity level of learning time learning contents reaction: emotional reaction). And Learning Reaction Analysis Engine analyze learners' emotional state. And infer learners' learning interests and learning concentrate level. In order to analyze learners' emotional state. And infer learners' learning interests and learning concentrate level, we propose learners' emotional state model and define learners' emotional state information. The remainder of paper is

organized as follows. In Sect. 2, we review previous affective computing research for e-learning or distance learning. The proposed learners' emotional state model and learner activity (learner's learning activity and learning emotional state (activity level of learning time learning contents reaction: emotional reaction) are described in Sect. 3. We introduce the architecture and function modules of Learning Reaction Analysis Engine and Learning Reaction Analysis Engine are defined and described in Sect. 4. Finally, we conclude in Sect. 5.

2 Related Works

In [4], emotion state of a learner is collected and analysed with keystroke data of a subject. The relation and between subject theme and keystroke range is analysed with touch log. In [5], problems or gaps in technology usage in the learning environment is identified and affective solutions for technology integration into programming courses at the University levels is suggest. In [6], affective computing platform was proposed and was used basic platform for learners' emotional system. In [7], in order to design a non-simultaneous distance instruction system with affective computing, integrates interactive agent was proposed with the curricular instruction of affective design and the questionnaire for user interaction satisfaction, observation, and interviews were used. The respondents' survey results showed high-level satisfaction regarding interaction with the affective learning system. In [8], conceptual architecture of agent-based affective tutoring system was presented and provided the effectiveness of selected teaching approach on student's emotional state, behaviour, and learning progress.

But previous works have focused on one or two factors of learning emotional state information. And only one or two sensors are used for tracking for learner's learning state. The paper tries to collect various learning emotional state information with various sensors.

3 Architecture of Learning Reaction Analysis System and Learners' Emotional State Information

Proposed Learning Reaction Analysis Engine tracks and analyzes learner's learning emotional state and decides learner's learning interest level and concentration level so that appropriate learning strategy for a learner can be delivered to Automated Tutoring Engine. Proposed Learning Reaction Analysis Engine considers learner's personal learning emotional state and learner's learning activity patterns. Learner's personal learning emotional model is emotional attitude toward learning contents and means learning interest, learning difficulty and learning concentration. According to Learner's personal learning emotional model, learning interest level, learning difficulty level and learning concentration level are decided by Learning Reaction Analysis Engine. Proposed Learning Reaction Analysis Engine delivers learners' learning emotional state to Automated Tutoring Engine (Fig. 1).

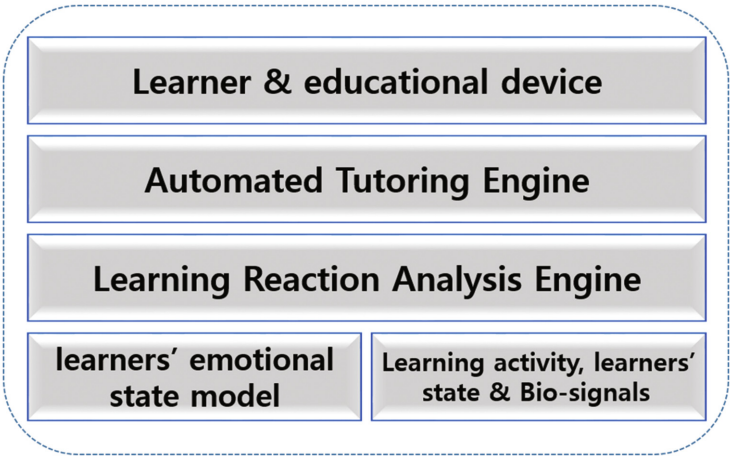


Fig. 1. Architecture of Learning Reaction Analysis System

Proposed Learning Reaction Analysis System consists of Learning Reaction Analysis Engine, Learners' Emotional State Model, Learners' state database, Bio-Signals DB and Learning Activity database.

- Learner and Educational Devices: source of learners' learning activities and emotional information with learning agents
- Learning Reaction Analysis Engine: collect learners' learning emotional state information, and infer and decision of learners' learning contents
- Automated Tutoring Engine: collection and analysis of learning environment information
- Learning Activity: activity level of learning time
- Learners' Emotional State Model: model of learning interests, learning contents reaction
- Bio-signals: learner's bio-signals for learning contents

Learning interests means activity level of learning time. Learning contents reaction means emotional reaction (boring, fun, difficult, etc.) from learning contents.

Learning Reaction Analysis Engine needs various learner's learning emotional information that can be classified into three groups. Three groups are learner's various actions and emotions that are learner's activity, learner's smartphone attitude, Bio signals and Learner's attitude. Learner's activity learner's smartphone attitude, Bio signals and Learner's attitude are not stored in learners e-portfolio DB and easily changeable.

Learner's learning emotional information is formatted and transferred in format of XML. learner's learning emotional information is like as bellows;

- Learner's activity
  - learning activity: activity level of learning time
  - learning interaction: initiative level of interaction with learning contents
- Learner's smartphone attitude

  - Lean angle:
  - Touch frequency: number of smart phone touch
  - Sector stay time: time to concentrate on learning contents
  - shaking frequency: time to make irrelevant learning
- Bio signals
  - Heart bit
  - Temperature
  - Hands movement and shaking
- Learner's attitude:
  - Another APP launching times: number of other Apps and launching time
  - face portion: learner's concentration level on learning contents
- Learners' Emotional State
  - learning interest: activity level of learning time
  - learning concentration: emotional reaction (boring, fun, difficult, etc.) from learning contents

4 Information Flow of Learning Reaction Analysis Engine

We assume that learners' learning device has various sensors and communication utility, and learners' smart watch can collect and detect bio-signals. Lastly, there is no middleware to convert and manage noise of bio-signals, we assume (Fig. 2).

① Learners log on a smart learning portal server (web server) with learning mobile devices or smart watch that have several sensors and seamless communication facility. At the same time, Learners log on Learning Management Server (LMS) with single-sign-on protocol. And LMS delivers learners' previous performance data and learner's property from e-portfolio DB to Automated Tutoring Engine.

②③ Learning Reaction Analysis Engine receives and checks APPs launching state on the smart phone and smart phone lean angle, face portion, touch frequency and moving and shaking frequency.

④ Learning Reaction Analysis Engine infers checks APPs launching state on the smart phone and smart phone lean angle, face portion, touch frequency and moving and shaking frequency and constructs Learning Concentration Analysis results.

⑤ After the inference, Learning Reaction Analysis Engine sends Learning Concentration State to Automated Tutoring Engine.

⑥ Automated Tutoring Engine constructs Learning Strategy for the learner, according to the Learning Concentration State.

⑦ Learning Reaction Analysis Engine sends the learner's Learning Concentration State for the subject to e-portfolio database. Learner's Learning Concentration State is result from Learner's Learning Emotional State Model that formulates the learning emotional level and kinds.

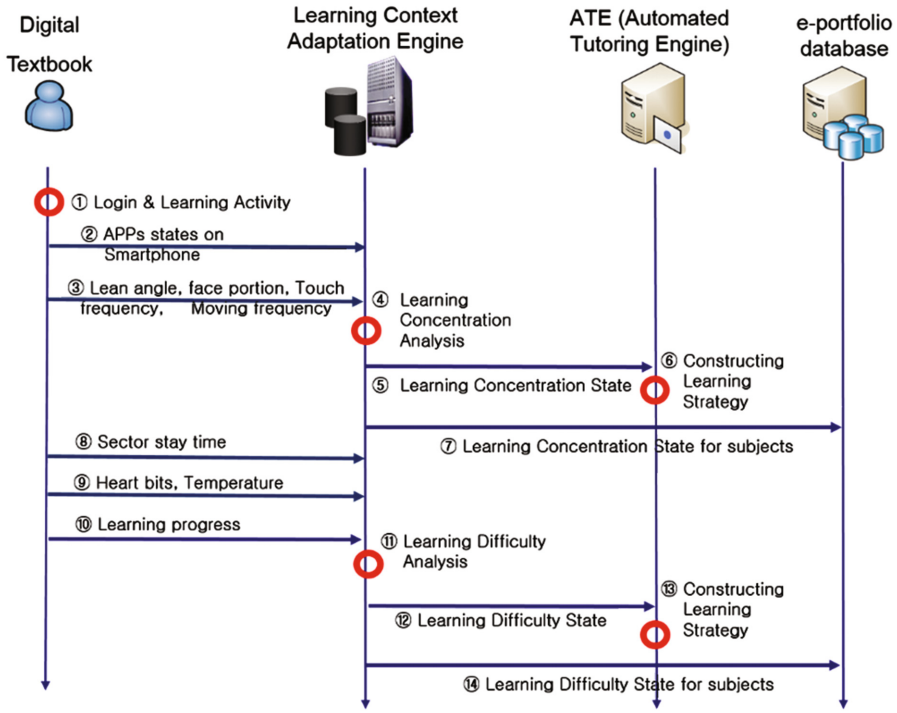


Fig. 2. Information flow of Learning Reaction Analysis Engine

⑧⑨⑩ Learning Reaction Analysis Engine receives and checks learner's Sector stay time on the smart phone, Heart bits, Temperature of smart phone and learner's Learning progress.

⑪ Learning Reaction Analysis Engine infers learner's Sector stay time on the smart phone, Heart bits, Temperature of smart phone and learner's Learning progress and constructs Learning Difficulty Analysis results.

⑫ After the inference, Learning Reaction Analysis Engine sends Learning Difficulty State to Automated Tutoring Engine.

⑬ Automated Tutoring Engine constructs Learning Strategy for the learner, according to the Learning Difficulty State.

⑭ Learning Reaction Analysis Engine sends the learner's Learning Difficulty State for the subject to e-portfolio database. Learner's Learning Difficulty State is result from Learner's Learning Emotional State Model that formulates the learning emotional level and kinds.

5 Conclusion

We propose Learning Reaction Analysis Engine that can collect and analyze learning emotional state for a learner. Especially proposed Learning Reaction Analysis Engine considers learners' attitude toward educational devices characteristics as like lean angle, face portions of a learner with an education devices. Thus education devices have to be equipped with sensors. It means that learning emotional state information can be delivered to Learning Reaction Analysis Engine and be instantly analyzed by Learning Reaction Analysis Engine. In order to extract exact learning emotional state, learning emotional state model is construct and defined. Learning emotional sate information consists of learner's activity, learner's smartphone attitude, learner's Bio signals and learner's earner's attitude. Learning emotional state consists of learning interest level and learning concentration level. After analysis of learning emotional state information, intelligent learning contents adaptation engine defines and infers learning emotional state according to learning emotional state models. And learning contents adaptation engine defines learning emotional state to Automated Tutoring Engine. Automated Tutoring Engine can construct personalized learning strategy according to learner's learning emotional state.

We have a plan to design XML rule for learning emotional state model.

Acknowledgement. This work was supported by 2016 Korea National Open University Research Fund.

References

1. Liu, J., Tong, J., Han, J., Chen, S., Yang, F.: Affective computing applications in distance education. In: Proceedings of the International Conference on Education Technology and Information System, pp. 927–930 (2013)
2. Shen, L., Callaghan, V., Shen, R.: Affective e-Learning in residential and pervasive computing environments. Inf. Syst. Front. **10**(4), 461–472 (2008)
3. Brave, S., Nass, C.: Emotion in human–computer interaction. Hum. Comput. Interact. **53**, 81–96 (2003)
4. Lee, P.-M., Tsui, W.-H., Hsiao, T.-C.: The influence of emotion on keyboard typing: an experimental study using auditory stimuli. PLoS ONE **10**(6), e0129056 (2015)
5. Prasad, A., Farik, M.: Integration of innovative technologies and affective teaching and learning in programming courses. Int. J. Sci. Technol. Res. **4**(10), 313–317 (2015)
6. El-Abbasy, K., Angelopoulou, A., Towell, T.: Affective computing to enhance e-Learning in segregated societies. In: Proceedings of 2015 Imperial College Computing Student Workshop, pp. 13–20 (2015)
7. Lin, H.-C.K., Su, S.-H., Chao, C.-J., Hsieh, C.-Y., Tsai, S.-C.: Construction of multi-mode affective learning system: taking affective design as an example. Educ. Technol. Soc. **19**(2), 132–147 (2016)
8. Petrovica, S., Pudane, M.: Simulation of affective student-tutor interaction for affective tutoring systems: design of knowledge structure. Int. J. Educ. Learn. Syst. **1**, 99–108 (2016)

Rule-Based Topic Trend Analysis by Using Data Mining Techniques

Yunwan Jeon[1], Chanho Cho[1], Jongwoo Seo[1], Kyunglag Kwon[1], Hansaem Park[2], and In-Jeong Chung[1(✉)]

[1] Department of Computer and Information Science, Korea University, 2511 Sejong-ro, Sejong-si, Republic of Korea
{juwl23, sail777, sjw007s, helpnara, chung}@korea.ac.kr
[2] Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
saem@kaist.ac.kr

Abstract. Many users in social web environments share and publish user-generated contents such as tastes, opinions, and ideas in the form of text and multimedia data. Various research studies have been conducted on the analysis of such social data, which can be used for discovering users' thoughts on specific topics. But, there are still challenging tasks to find out the meaningful patterns from the social data due to rapidly increasing amount of data. In this paper, we therefore propose a rule-based topic trend analysis by using On-Line-Analytical Processing (OLAP) and Association Rule Mining (ARM) to detect information such as previously unknown or abnormal events or situations. For the verification of the proposed method, we conduct experiments to demonstrate that the method is feasible to perform rule-based topic trend analysis.

Keywords: On-line analytical processing · Association rule mining · Decision support system · Topic trend analysis · Social data

1 Introduction

Users' activities on Social Network Services (SNSs) have reached unpredictable levels with the explosive propagation of smart mobile devices and development of web 2.0. Also, SNSs enable people to participate in on-line, social activities, and shatters the barrier for on-line users to generate and consume user-generated contents in anywhere or anytime. Users share and publish various information including product reviews, interests, and opinions with other people via SNSs. To date, a variety of research studies have shown that the huge amount of social data is very useful in many ways such as predicting the stock market, supporting decisions for government, accessing political polarization in the public, etc. But, the tremendous volume and rapidly increasing amount of data still make it difficult to discover meaningful patterns from the data. Therefore, there is a need to scrutinize such social data and define suitable knowledge extraction approaches for supporting decision making. In this paper, we utilize one of

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_75

data warehouse and data mining techniques on social data to analyze topic trend and extract meaningful knowledge (i.e., association rules).

The proposed method consists of three main steps, including data collection, data extraction, and data analysis. First, we collect social data from SNSs and then conduct natural language processes to reduce noisy data. Second, we apply Latent Dirichlet Allocation (LDA) algorithm to detect candidate topics in social data, and create documents and terms matrices. Next, we construct multidimensional data cube model to apply On-Line Analytical Processing (OLAP). In this step, we use a star schema, consisting of a large central table (i.e., fact table) and a set of auxiliary tables (i.e., dimension tables), and manually define concept hierarchies to perform diverse OLAP operations for the detection of topic trends from a variety of social data. Lastly, we extract topic-based meaningful rules from the analyzed data by using Association Rule Mining (ARM). For the feasibility of proposed method, we conduct experiments with real social data (e.g., Twitter). The experimental results show how trends of topics on social media are analyzed by OLAP, and how relevant keywords are extracted by ARM.

2 Background and Related Work

Data mining is the process of extracting and analyzing data for users to investigate the correlations or patterns in large dataset. Among various data mining techniques to analyze and find the correlations between a set of items or useful pattern from data, we use a hybrid approach, which is a combination of two data mining techniques such as On-Line Analytical Processing (OLAP) [1] and Association Rule Mining (ARM) [2].

The OLAP is a data warehouse technology for analyzing much of historical data stored in data warehouse [1]. It allows users to manipulate a set of indicators according to different dimensions with one or more hierarchies. Moreover, the OLAP is one of data analysis techniques with functionalities such as summarization, consolidation, aggregation, and the ability to view information from multiple angles [1, 3, 4]. Consequently, the OLAP enables users to gain deeper insights of data for the better understanding of various aspects of their corporate data through a fast, consistent, interactive access to a wide variety of possible views of the data [5].

The ARM is one of the most frequently used techniques to deduct unknown rules or knowledge in various fields. In the ARM, associated operators allow users to navigate an intuitive insights on different levels of hierarchies. Each association rule is represented in the form of "$X \rightarrow Y$", where X and Y can be defined as a set of attributes or keywords. The left and right hand sides indicate a condition part ("if") and rule part ("then"), respectively. To be specific, we use the Apriori algorithm to find frequent item sets and association rules from transactional databases. We exploit this algorithm to identify the frequent, individual items in dataset which is related to particular topics. This algorithm generates association rules between items based on two criteria, namely support and confidence [6, 7]. The support for an association rule is the fractions of transactions including both X and Y in the databases. And the confidence for the rule is the probability of transactions including X which also contains Y in the database, reflecting accurate prediction of the rules. As both minimum values for support and confidence will thus affect the quality of association rule mining, it is very important to

ensure that user-specified minimum support and confidence are appropriate to generate meaningful rules [8]. Generally, the support threshold is set to 30% and the confidence threshold is over 80% [9]. If minimum support threshold is too low, the number of frequent set of keywords and the number of rules generated will be increased.

3 Topic Trend Analysis Based on OLAP and ARM

Figure 1 shows the overall architecture of method, consisting of three steps: (1) data collection and preprocessing, (2) topic extraction, and (3) rule-based topic trend analysis.

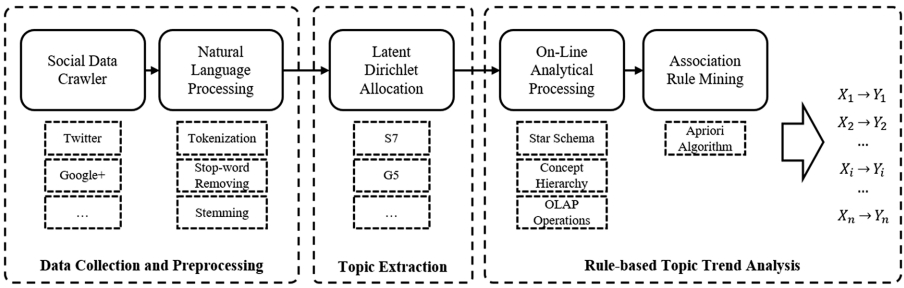


Fig. 1. The overall architecture of proposed method

3.1 Data Collection and Preprocessing

We first collect social data from Twitter. It contains a large number of unnecessary information for detecting topic trends. For example, a message contain unnecessary words such as RT, @WSJD, type of URL. Since these terms are insignificant to extract topic trends on users' contents, we apply filtering methods such as tokenization, stop-word removing, and stemming.

3.2 Topic Extraction

After collecting and preprocessing social data, we apply LDA method to the preprocessed data to detect the topics from users' contents. LDA is very popular method for extracting topics. In LDA, we assume that topics are specified before any document has been generated. Thus, for any document in the corpus, the generative process contains two stages. First, a topic distribution vector modeled by a Dirichlet random variable has been chosen randomly to determine which topics are the most likely appear in a document. Then, for each terms that are to appear in the documents, a single topic from the topic distribution vector is randomly selected. From the analysis by LDA, it enables users to organize and summarize electronic archives on a scale that would be impossible using human annotation [10]. Through applying LDA in social data, a variety of topics are extracted on each particular theme such as iPhone, S7, and G5, as shown in Table 1.

Table 1. Examples of the extracted candidate topics for each theme

Theme	Media	Top 10 topic words
iPhone	Twitter	SE, Launch, Apple, Now, Galaxy, Htc, Cable, Huawei, 6s, Amazon
S7		Galaxy, Samsung, Iphone, Edge, Note, Htc, Shipping, Huawei, Android, Tablet
G5		LG, Launching, Module, B&O, Preorder, Sandragon, S7, Battery, Android, Galaxy

With extracted results on social data, we construct multidimensional data cube model to detect the trends of topics in the next step.

3.3 Rule-Based Topic Trend Analysis

To perform a topic trend analysis, we first use OLAP and its operations based on multidimensional data cube model [3, 5]. In this paper, we use a star schema to construct the multidimensional data cube model.

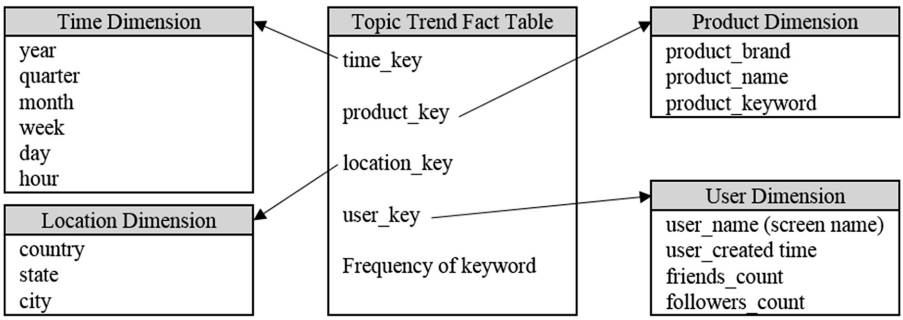


Fig. 2. Star schema for multidimensional data cube in the proposed method

Figure 2 show the star schema, including a fact table and four dimension tables. The fact table is composed of the names of facts, or measure values and foreign keys to each of the related dimension tables. In the fact table, we use frequencies of each keyword for aggregation criteria (i.e., measure values). In general, facts are numeric values as quantities by which users want to analyze relationships between dimensions. Each dimension table includes perspectives or entities to describe what to analyze. In this paper, we add four dimension tables: *time_key* for temporal trend analysis, *location_key* for a spatial analysis of particular keywords on social media, and *product_key* and *user_key* for which keywords are spatio-temporally grown from each user. The data warehouse needs a subject-oriented schema which makes it possible to analyze on-line data along with a variety of viewpoints.

Next, we manually define concept hierarchies by discretizing or grouping values for a given dimension or attribute. It is a sequence of mappings from a set of low-level

concepts to the higher level concepts [5]. Concept hierarchies help users handle data at varying level of abstraction. The attributes of each dimension are organized in a partial order. The partial order for the *location* dimension such as city, region, and country is 'city<region<country,' indicating the hierarchical structure from low-level (city) to higher level (country). In addition, a concept hierarchy for the *time* dimension contains attributes such as hour, day, day of week, month, quarter, and year. The partial order for *time* dimension is 'hour<day<{day of week, month, quarter}<year'. We also construct *product* and *user* hierarchies. For example, the *Product* hierarchy contains various attributes such as product brand, product name and related keywords. The constructed concept hierarchies for each dimension are stored at the multidimensional data cube model. As time goes by, the low-level node data in each hierarchy is summarized and aggregated into upper level nodes. For instance, Los Angeles (LA), San Diego (SD), etc. in the lower level node of location hierarchy are aggregated into California (CA), upper node (region) in location hierarchy.

With the constructed multidimensional data cube model, we perform OLAP operations such as roll-up, drill-down, slice and dice to detect useful information. In the model, data is generally organized into multiple dimension along with the concept hierarchies, thus providing flexibility to view data from a variety of users' perspectives. We then employ ARM on analyzed results of OLAP operations and identify which keywords and rules related to the topic are extracted on user-generated contents. In this proposed method, the keyword refers to a term related to particular topics while frequently occurred set of keywords refer to a group of terms. The association rules generated in this step are represented in the form of "X→Y," indicating that the keyword Y is likely to appear if keyword X is contained in the user-generated content about the particular topics. Each extracted rule is used to make a right decision with consideration of detected co-occurred keywords, which represent trendy issues in real time. With the extracted rules, decision makers can establish product marketing strategies, and predict consumers' demands about their products.

4 Performance Evaluation

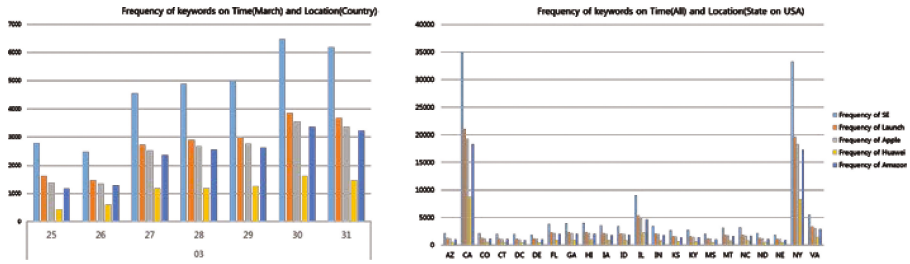
For experiment, we collect a bulk of data from Twitter. It provides a variety of users' information and their contents. In this paper, we collect some of them, including *text*, *retweet_count*, *created_id*, *location*, *time*, etc. by using Twitter API. We then filter text messages with no hashtag, and ones which do not contain any smartphone-related keywords such as iPhone, Galaxy, G5, etc. The social data is collected from March 25, 2016 to April 30, 2016. The collected dataset consists of approximately 500,000 tweets.

4.1 Analysis of Topic Trend by Using OLAP Operations

We show how topic trends on social media are analyzed by various OLAP operations such as roll-up, drill-down, etc., and how ARM extracts rules from the topics. In this paper, we employ the Excel Pivot table with MySQL to construct the multidimensional

data cube model. All of data collected from social media is stored in database, i.e., MySQL, and then analyze a huge amount of data to detect the topic trend.

In OLAP, multidimensional analysis is provided by selecting dimensions within the viewpoint of users. We show that topic trend is analyzed for changes in topics mentioned from user-generated contents along with *time* dimension. Figure 3(a) shows that the frequency of extracted topics is explosively increased in end of April than beginning of April. It implies that iPhone SE is firstly launched on starting from end of March to country selected as a first releasing country such as USA, Canada, Japan, and France. Thus, we are able to detect the trend of particular topics mentioned by many users from social media. The results are very useful for supporting the right decision making in multiple choices to decision makers. In addition, we also know that the usage of social media relatively grows when particular events occur on real world. In the same fashion manner, we can induce the spatial topic trend with analysis of measure values based on *location* dimension. From the result, we can get an insight that the higher usage of particular social data on particular location such as United States, United Kingdom, Japan, India, and Canada than any other country. We also infer knowledge on usage trends of other continent in the world. For example, iPhone is more popular in North America than any other continent, thus aiding decision makers to establish effective marketing strategies.



(a) Analysis of measure values based on *time* (March)

(b) Analysis of measure values based on *time* (All) and *location* (State on USA)

Fig. 3. Results with dimensions of *time* and *location*

For deeper understanding of topic trends on the North America, we then employ drill-down operations in OLAP. It represents more general viewpoint from *location* hierarchy. For example, there is a need to detect unknown information by analyzing specific viewpoint into more general one. Figure 3(b) shows the frequency of keywords on time and location. From the result, two states such as CA and New York (NY) have the higher measure values than other states, indicating the higher usages of social media of users. Also, particular topic (iPhone) is more popular in these states than others.

By using a multidimensional data analysis, we represent a process for detecting spatial-temporal topic trends based on a variety of scenario that gives users intuitive information. The results show that multidimensional analysis with OLAP provides useful and interesting insights for decision makers.

4.2 Analysis of Extracted Topic Rules with ARM

For the extraction of latent, unknown rules or patterns, we use the dataset generated from two states, CA and NY, to recognize which terms are simultaneously occurred and correlated to each other. In the experiments, we employed R program, which is one of the most frequently used program for analyzing the tremendous volume of data, to perform ARM efficiently. As mentioned in the related work, two values (support, confidence) are set to 40% and 80%, respectively, to detect association rules from given dataset.

To show how ARM is used in the proposed method, we provide exemplary results. At the top of Table 2, support values of the top five keywords for each topic are shown. In the table, a topic (Huawei) is eliminated due to low support and confidence values than initially defined values. Such removed topics are considered as candidate of frequent keywords. These frequent keywords are then merged together to generate two-keyword sets by grouping in the next iteration, as shown in the bottom of Table 2. We create the association rules by applying two-keyword sets into user-generated data represented in particular location (CA, NY). As described in Table 3, we obtain five association rules with higher values of initially defined support and confidence values as useful, meaningful rules. For example, the term (Case), which represent a kind of accessory for iPhone, appears when the two terms (e.g., SE and Launch) simultaneously appear in user-generated contents.

Table 2. Support values of the top five keywords and two-keyword sets

	Keyword(s)	Support (%)
Keywords	SE	73
	Apple	52
	Launch	48
	Amazon	41
	Huawei	38
Two-keyword sets	SE, Launch	68
	SE, Apple	61
	SE, Amazon	57
	Apple, Launch	54
	Apple, Amazon	46
	Launch, Amazon	41

Table 3. Valuable association rules in the form of IF-THEN with support and confidence values

	IF	Then	Support (%)	Confidence (%)
1	SE, Launch	Cases	68	96
2	Apple, Launch	SE	54	91
3	Apple, SE	Galaxy, Huawei	61	88
4	Amazon, Apple	Cable	46	85
5	Launch, Amazon	Now	41	84

5 Conclusions

We present a rule-based topic trend analysis method by using OLAP and ARM. In the method, we collect a real dataset from social media. We then extract a variety of topics about a particular theme such as iPhone, S7, and G5 from text data by using LDA. Next, trend analysis is performed based on OLAP for spatial-temporal analysis of extracted topics on social data. In addition, ARM is employed for which keywords and rules related to topics are extracted with computations of support and confidence values. To verify the proposed method, we conduct experiments with real dataset from Twitter. The experiment results demonstrate that the proposed method is feasible to analyze topic trends and detect association rules from social data.

The main contributions of this paper are presented as follows. First, we consider social trends of topics in SNSs by using OLAP and ARM. Second, we provide possibility for analyzing various social data which have different insights of users on similar topics. Lastly, we show the feasibility of combination of two different data mining techniques. However, there still exist some challenges for better understanding of topic trends. Since we use frequencies of each topic as measure values in the fact table, we need to develop the proposed method to handle other types of measure values such as relative ratio of topics. Moreover, we will devise an improved method to deal with unstructured data such as image, video, etc., which cannot be represented by numerical values.

Acknowledgment. This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (grand number NRF-2016R1A2B1010975).

References

1. Chaudhuri, S., Dayal, U.: An overview of data warehousing and OLAP technology. ACM Sigmod Rec. **26**(1), 65–74 (1997)
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Rec. **22**(2), 207–216 (1993)
3. Park, D., et al.: NetCube: a comprehensive network traffic analysis model based on multidimensional OLAP data cube. Int. J. Netw. Manag. **23**(2), 101–118 (2013)
4. Sohn, J.-S., Chung, I.-J.: Dynamic FOAF management method for social networks in the social web environment. J. Supercomput. **66**(2), 633–648 (2013)
5. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques (2011). Elsevier
6. Sarwar, B., et al.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce. ACM (2000)
7. Inokuchi, A., Washio, T., Motoda, H.: An apriori-based algorithm for mining frequent substructures from graph data. In: Principles of Data Mining and Knowledge Discovery, pp. 13–23. Springer, Heidelberg (2000)
8. Lim, A.H., Lee, C.-S., Raman, M.: Hybrid genetic algorithm and association rules for mining workflow best practices. Expert Syst. Appl. **39**(12), 10544–10551 (2012)
9. Lee, C., et al.: A hybrid OLAP-association rule mining based quality management system for extracting defect patterns in the garment industry. Expert Syst. Appl. **40**(7), 2435–2446 (2013)
10. Ma, B., et al.: Semantic search for public opinions on urban affairs: a probabilistic topic modeling-based approach. Inf. Process. Manag. **52**(3), 430–445 (2015)

A Case Study of Hierarchical Safety Analysis for Eliciting Traceable Safety Requirements

Daehui Jeong, Anit Thapaliya, and Gihwon Kwon(✉)

Department of Computer Science, Kyonggi University, Suwon, South Korea
{jdhpenl00, anitthapaliya, khkwon}@kgu.ac.kr

Abstract. In this paper, we present the hierarchical safety analysis for eliciting traceable safety requirements. The proposed technique was used to the case study of railway system as an example. In this work, FMEA and HAZOP analysis are used as safety analysis technique in order to illustrate hierarchical safety analysis showing traceability.

Keywords: Hierarchy · Safety analysis · FMEA · HAZOP · Safety requirements

1 Introduction

With the development of IT technologies, the portion of software in the system has increased. Because of this tendency, the safety of software became important and required safety related activities.

Safety analysis is a part of safety activities that includes hazard identification and risk assessment to drive safety requirements. In other words, safety analysis identifies hazards that required risk assessment performed before the requirements analysis, and the identified hazard is mitigate to derive the safety requirements.

Clear requirements lead to successful development of software. Hence, lead to successful development of safety-oriented software by clarifying requirements through safety requirements from safety analysis [1]. The following are widely used safety analysis techniques [2].

- Preliminary Hazard Analysis (PHA)
- System Hazard Analysis (SHA)
- Fault Tree Analysis (FTA)
- Failure Mode and Effects Analysis (FMEA)
- HAZard and OPerabilityStudy (HAZOP)

These safety analysis techniques are selectively applied on the targeted system depending on the nature and the stage of software development. However, it is difficult to express the safety requirements traceability when the safety analysis is composed of various techniques because each technique consists of different scope, and purpose. In this paper, we present a hierarchical safety analysis using the characteristics of the hierarchical structure of the system development to support traceability analysis. We

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_76

apply this method in the development of railway control software being made up by model train.

This paper is structured as mentioned further: In Sect. 2, we proposed hierarchical safety analysis. In Sect. 3, we applied hierarchical safety analysis to the case study of railway system as an example. In addition, the results of this hierarchical safety analysis are presented. Section 4 concludes by examining how the work described supports the hierarchical safety analysis and identifying possible directions for future work.

2 Hierarchical Safety Analysis

Safety analysis is the main safety activities, generating safety requirements to control risks to the acceptable level. In order to develop safe software, safety analysis is required. In safety analysis, it is difficult to show traceability, because several techniques are used depending upon the nature and stage of software development. Likewise, it is difficult to show the traceability due to different scope, propose and style of each techniques. In this paper, we present hierarchical safety analysis to support the traceable safety analysis.

2.1 Hierarchical Structure

Safety is a property of system that means analysis of failure in subsystems and components must always be considered to their direct contribution to system level effects. The system consists of several subsystems, and the subsystem in turn consists of another subsystem or component. The hierarchical structure of these systems also affects the flow of hazards as shown in Fig. 1. If a component fault occurs, it affects the subsystem, which is a fault of the subsystem. Likewise, fault in the subsystem affects the system, which is the faults in system and finally it leads to accidents. Thus, the flow of hazards also shows a hierarchical structure.

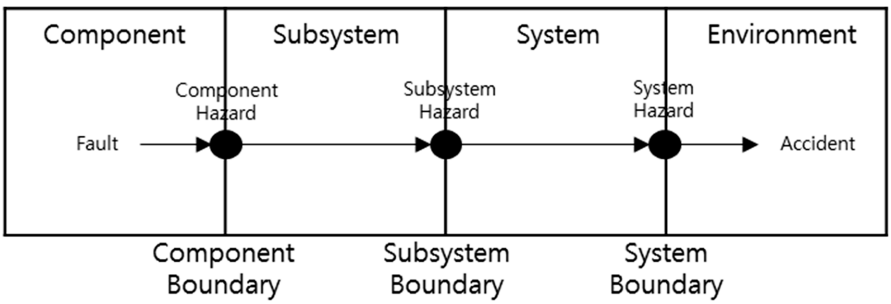


Fig. 1. Hazard flows

Hierarchical safety analysis uses the hierarchical structure of the flow of system and risk sources. We performed safety analysis for each boundary of the system and have analysed the traceability of the results.

2.2 HAZOP and FMEA

In this methodology, HAZOP and FMEA are used for hierarchical safety analysis. HAZOP and FMEA are often used in combination to complement each other's strengths and weakness [3]. Analysis starts hierarchically from the lowest component. Initially, the fault type identification is required for each component. Identification of the failure modes in FMEA uses parameters and guidewords from HAZOP. In this case, the parameter and guidewords were used as defined in HAZOP-KR for the railway [4]. Once the failure modes identified, the cause and effects in FMEA are estimated. Here, the influence effect is the cause of the next layer subsystem. The analysis method for subsystems and systems are similar as described above.

2.3 Traceability Analysis

This section proposes a hierarchical safety analysis. After identifying the failure modes, causes, and effects for each level to Sect. 2.2, this section analyses the results, to determine the traceability of the risk sources and further examines, controls and eliminates it.

Traceability analysis begins by identifying the traceability of risk sources. Based on the results in Sect. 2.2, component faults, subsystem faults, and system faults are analysed. Here, the faults of the subsystem are the effects of the faults on the components, and the faults of the system are the effects of the faults of the subsystem. Finally, we analysed the environmental impact of system faults.

Once the traceability of the risk source is identified, each risk sources are reviewed, controlled and removed. For the purpose of review, measures are used to mitigate risks and the risk is assessed by measuring RPN (Risk Priority Number) before and after mitigation measures. The RPN is measured by two factors, severity and occurrence,

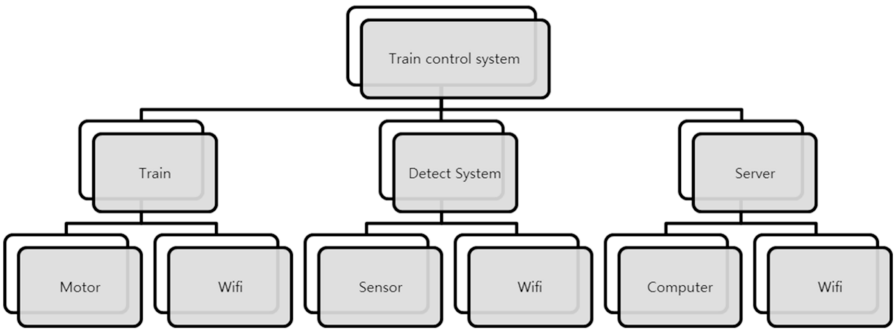


Fig. 2. System structure of case study

and the tolerance level is obtained. If the RPN does not meet the acceptable level, an additional mitigation measure should be established and the RPN should be measured again until it reaches acceptable levels.

3 A Case Study

In this research, hierarchical safety analysis is introduced to the development of train control software. It's a software for controlling the direction for two train as shown in Fig. 2 above. Usually, the sensor detects the train which is transmitted to the server using the wifi from the detect system. Then, finally compute the information control the train. In this paper, because of paper space we have shown the analysis results of one-subsystem and components. Tables 1, 2 and 3 are analysis results of each component, results of subsystem analysis, and analysis of the system respectively. Table 4 shows the results of the traceability analysis, and Table 5 shows the RPN before and after mitigation measures respectively.

Table 1. Component level

Component	Failure modes		Causes	Effects
	Parameters	Guidewords		
Motor	Data	No	Speed change fail	Motor control fail
		Other	Speed change order error	Motor control wrong
Wifi	Interface	No	Connection fail	Speed receive fail
			Disconnection	Speed receive fail
	Data	Part of	Receive part of speed	Change wrong speed
		Early	Receive early speed	Change early speed
		Late	Receive late speed	Change late speed

Table 2. Subsystem level

Subsystem	Failure modes		Causes	Effects
	Parameters	Guidewords		
Train	Interface	No	Speed receive fail	Train control fail
	Action	No	Motor control fail	Train control fail
		Early	Change early speed	Train control early
		Late	Change late speed	Train control late
		Other	Motor control wrong	Train control wrong
			Change wrong speed	Train control wrong

Table 3. System level

System	Failure modes		Causes	Effects
	Parameters	Guidewords		
Train control system	Interface	No	Train control fail	Train collision
		Other	Train control wrong	Train collision
			Train control early	Train collision
			Train control late	Train collision

Table 4. Traceability analysis

No.	Component	Subsystem	System	Effects
1	Speed change fail	Motor control fail	Train control fail	Train collision
2	Speed change order error	Motor control wrong	Train control wrong	Train collision
3	Connection fail	Speed receive fail	Train control fail	Train collision
4	Disconnection	Speed receive fail	Train control fail	Train collision
5	Receive part of speed	Change wrong speed	Train control wrong	Train collision
6	Receive early speed	Change early speed	Train control early	Train collision
7	Receive late speed	Change Late speed	Train control late	Train collision

Table 5. Measures and RPN

No.	RPN before measures			Measures	RPN after measures		
	Severity	Frequency	Tolerance		Severity	Frequency	Tolerance
1	A	3	Intolerable	M1: Compare changing value	A	5	Tolerable
2	A	3	Intolerable	M2: Error exception	A	5	Tolerable
3	A	3	Intolerable	M3: Heartbeat	A	5	Tolerable
4	A	3	Intolerable	M4: Heartbeat	A	5	Tolerable
5	A	3	Intolerable	M5: Speed range Selection	A	5	Tolerable
6	A	3	Intolerable	M6: Heartbeat	A	5	Tolerable
7	A	3	Intolerable	M7: Heartbeat	A	5	Tolerable

4 Conclusions

In this research, we proposed an analysis of hierarchical safety to extract the traceability requirements and implemented the working examples of train control system. As a result, the following safety requirements was derived:

- SR1: The system compares and checks the speed change.
- SR2: Command error exception handling should implement in the system.

- SR3: The system needs to set the range of speed values.
- SR4: The system should exchange Heartbeat and always check communication.

Safety requirements are derived from mitigation measures of safety analysis. If the contents of the mitigation measures are different, each refers a new safety requirement. SR1 was extracted from M1, SR2 from M2, SR3 from M5, and SR4 from M3, M4, M6, and M7.

Safety requirements are also traceable. Since the safety requirements are extracted from the mitigation measures, traceability can be confirmed through Tables 4 and 5. Figure 3 shows the traceability of SR1.

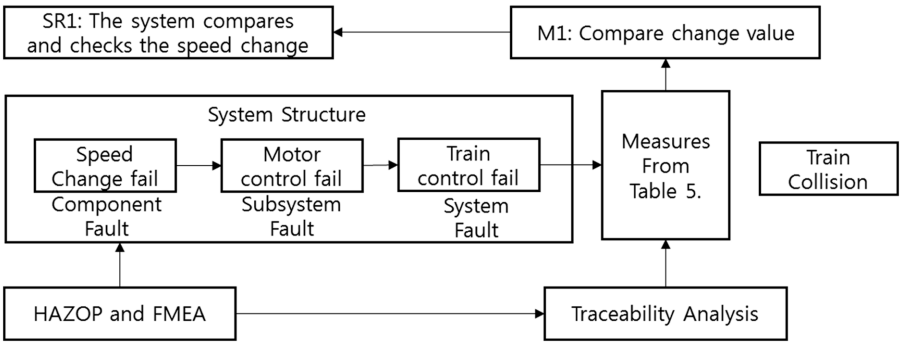


Fig. 3. Traceability of SR1

The derived safety requirements are reflected into subsequent design, implementation and testing. In the conventional development without safety analysis, when the sensor is defective, the train collision problem occurs, whereas after safety analysis, safety requirements are reflected and the problem is controlled.

The contribution of this paper is twofold. First, the traceability of safety requirements is clearly shown. Secondly, the problematic parts of the system can be understood by understanding the flow from defects to the occurrence of an accident.

Future research directions will extend the experience of this study, and a model for analysing software incidents and techniques for analysing theoretical aspects of this method.

Acknowledgement. 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-R0992-16-1014) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

1. Na, K.: the impacts of requirement uncertainty and standardization on software project performance: a comparison of Korea and USA. J. Inf. Technol. Appl. Manag. **11**(2), 15–27 (2004)
2. Ericson II, C.A.: Hazard Analysis Techniques for System Safety. Wiley-Interscience, Hoboken (2005)
3. Trammell, S.R., Davis, B.J.: Using a modified HAZOP/FMEA methodology for assessing system risk. In: Proceedings of the 2nd International Workshop on Engineering Management for Applied Technology, EMAT 2001, pp. 47–53 (2001)
4. Hwang, J., Jo, H., Han, C., Cho, W., Ahn, J., Ha, D.: A study on the HAZOP-KR for hazard analysis of train control systems. In: The Korean Society for Railways Symposium, vol. 13, no. 4, pp. 369–403 (2010)

Smart Home Technology and Energy Savings with a Special Focus on Europe

Blanka Klimova, Petra Maresova, and Ondrej Krejcar(✉)

Faculty Informatics and Management, University of Hradec Kralove,
Rokitanskeho 62, Hradec Kralove, Czech Republic
{blanka.klimova, petra.maresova, ondrej.krejcar}@uhk.cz

Abstract. With the rapid expansion of new technologies in all spheres of human life, technological devices have inevitably penetrated in people's homes, where in most cases are targeted at facilitating people's life in order to improve people's quality of living. The purpose of this article is to address current issues with respect to smart home technology and energy savings, especially in European countries. The specification is based on the basis of available studies between 2010 and 2015. The authors used a method of literature review of available sources exploring research studies focused on smart home technology with respect to energy savings in the acknowledged databases Web of Science, Elsevier, Science Direct, and Springer. The findings show that the number of studies exploring the issue of smart home technology and energy savings is gradually rising since consumers are getting more aware of potential savings, respectively of waste of energy.

Keywords: Smart home technologies · Quality of living · Energy savings · Benefits · Limitations

1 Introduction

With the rapid expansion of new technologies in all spheres of human life, technological devices have inevitably penetrated in people's homes, where in most cases are targeted at facilitating people's life in order to improve people's quality of living. At present these labour-saving devices at home are called smart home technology.

Smart home technology can be defined as the integration of technology and services through home networking with the purpose to enhance quality of living [1]. Generally, it is the use and control of home devices remotely or automatically. Therefore, sometimes it is simply called home automation. Some home automation devices can be used on their own; they do not communicate such as a programmable light switch, while others are part of the internet of things and are networked for remote control and data transfer.

Smart home technology can be classified into two main types: wiring and wireless system. In the wiring system the equipment is connected into the main power supply directly and the data are transmitted to the appliances to activate or deactivate them. There are many types of wires that people may want to install in-wall. Many home automations are connected through the wiring system such as new wire (twisted pair,

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_77

optical fiber), Powerline, or Busline. The most common and quality technology include X10, HomePlug, Consumer Electronics Bus (CEBus), or European Installation Bus. In the wireless system two elements are required: a sender and a receiver. Many new devices exploit wireless technology to communicate with other devices such as microwaves, Infrared (IR), radio frequency (RF), Wi-Fi, Bluetooth, or IEEE 802.11. An example of wireless communication system for smart home is Z-wave, which is a reliable and affordable wireless home automation solution [2].

Every home now has some type of smart home technology [3, 4], which can be divided into hardware devices and software application. Hardware devices can include sensors such as cameras and thermometers, controllers, actuators (to do things), and communication systems. Remote control can range from a simple remote control to a smartphone with Bluetooth, to a computer on the other side of the world connected by internet. Home automation systems are available which consist of a suite of products designed to work together. These typically connected through Wi-Fi or power line communication to a hub which is then accessed with a software application. Popular applications include thermostats, security systems, blinds, lighting, and door locks.

Statista's Digital Market Outlook [5] presents the five largest markets for smart home technology, which include the United States, Japan, Germany, China and the United Kingdom, with the U.S. leading the pack both in terms of revenue and household adoption. There are currently 4.6 million smart homes in the United States, a number that is expected to increase fivefold by 2020. Worldwide it is assumed that market with smart home devices is expected to grow from 40 million US dollars in 2012 to 26 billion US dollars in 2019 [6].

Important benefit and motivation for the development of smart technology is an effective approach to the use of electrical energy. Currently, the most popular solutions in this area are [7]:

- use of renewable energy sources (such as wind and solar photovoltaic) to decrease carbon emission;
- use of energy storage devices such as batteries or flywheels in order to enable the enquiry to follow the generation;
- concept and implementation of smart grids, which enable effective integration of renewables and new large scale electric devices and reduces the demand;
- implementation of smart autonomous agents such as smart meters or intelligent software agents that manage, schedule and control the electric consumption of loads, minimizes its inefficient usage and maximizes consumer's savings.

The purpose of this article is to address current issues with respect to smart home technology and energy savings, especially in European countries. The specification is based on the basis of available studies between 2010 and 2015.

2 Methods

The authors used a method of literature review of available sources exploring research studies in the period from 2010 to 2015, and a method of comparison and evaluation of their findings. In addition, other relevant studies were reviewed on the basis of the

reference lists of the research articles from the searched databases. For a more detailed specification of the research studies, the authors conducted a search in Web of Science for the following keywords: smart AND home AND energy. Altogether 1,595 studies were generated. Afterwards, the authors of this review study refined the search and only focused on the research articles (450), reviews (21), books (631), patents (488), meetings (10), and others. In Europe 385 documents originated, in the USA it was 200, and the rest came from other continents, particularly from Asia.

3 Findings and Their Discussion

Altogether five research studies were identified. Two studies were Italian, one Czech, Irish and British. The studies were included provided that they were written in English and described the energy savings in Europe by using smart technology. As the findings from the studies described show, apart from the development and evaluation of the energy saving smart devices, there is an increased interest in the development of methodologies which could help to select the best smart devices that contribute to energy savings. In addition, there is considerable effort to improve consumers' attitude and awareness of energy savings. Studies have shown that receiving a regular and effective feedback on energy consumption behavior allows individuals to change their behavior. Based on the assessments of the energy-saving potential in changing people's behavior, carried out in Switzerland, Germany, United Kingdom, Netherlands and USA, it can be said that people's behavior in relation to energy consumption affects the number of macro-level and personal factors. Macro-level factors include technological progress, economic development level, demographic, institutional and cultural country factors, while personal factors include personal human qualities, attitudes, beliefs, norms, motivation, skills, knowledge, habits and routines. [13] Moreover, as the study by Ehrhardt-Martinez, Donnelly and Laitne [14] emphasizes, it is also important to provide feedback on energy usage to users since research [15] confirms that displaying real-time information on electricity usage may challenge user's behavior on this issue and contribute to energy savings of up to 30%. The data that should be provided to the users should include the direct and indirect feedbacks. The direct feedbacks involve real-time plus feedback (real-time information about the level of energy used by the appliance) and real-time feedback (real-time premise level information). The indirect feedbacks comprise daily or weekly feedback, which is household-specific information and advice on a daily or weekly basis, estimated feedback (typically web-based energy audits with information supplied on an on-going basis), and enhanced billing (household-specific information and advice) (Table 1).

The findings also indicate that there is a public interest in the use of smart meters which help people understand and quantify the energy they use. In comparison with traditional meters, these smart meters can show accurate energy use data in real-time.

Table 1. Home smart technologies and energy savings

Study	Objective	Methods	Results
Building heating technology in Smart Home using PI System management tools [8]	The aim is to describe a functionality of technologies of the BHT part and the BAT part, and an interconnection with a superior system in the form of the PI System	Comparison and evaluation of possibilities of the PI Coresight and the PI ProcessBook in dependency on performing the real-time fast statistical analysis of the measured temperature data, dynamic programming or dynamic time warping (DTW)	The PI ProcessBook is very useful. On the other hand, in terms of the possibility of easy access to the acquired data and measured temperature values via some Web-based application or via a smart phone, the PI Coresight is sufficient and better than the use of the PI ProcessBook
Energy performance of an all solid state electrochromic prototype for smart window applications [9]	The aim is to present performance data of an home-made all solid state EC prototype based on amorphous tungsten trioxide	Measuring indicators related to energetic performance, specifically, the optical transmittance and reflectance coefficients, the solar heat gain coefficient, the thermal transmittance and, in addition, the lifetime	The prototype meets (and in some cases exceeds) most of these specifications. Some desirable improvements concern higher visible transmittances in the bleached state and enhanced reflectance modulation in the NIR spectral range
Smart meters and energy savings in Italy: Determining the effectiveness of persuasive communication in dwellings [10]	The aim is to evaluate a smart monitoring system which improves awareness of energy behavior in homes, enabling better management via the visualization of consumption	The system was tested on 31 Italian families selected among volunteers all over Italy, participating in the first trial phase from October 2012 to November 2013	The study concludes that energy related persuasive communication is effective in reducing electricity consumption in dwellings on average −18% and up to −57%
Energy saving potential of heat insulation solar glass: Key results from laboratory and in-situ testing [11]	The aim is to test efficiency of HISG (heat insulation solar glass) multi-functional glazing technology to mitigate energy consumption of buildings	Testing of parameters in laboratory settings such as shading coefficient, UV, IR and visible light intensity investigated through the tests conducted in real operating conditions	HISG provides 38 and 48% energy saving in heating and cooling season

(*continued*)

Table 1. (*continued*)

Study	Objective	Methods	Results
Reducing household electricity demand through smart metering: The role of improved information about energy saving [12]	The aim is to answer two research questions: First, does feedback improve a household's stock of information about potential energy reducing behaviours? And second, do improvements in such information help explain the demand reductions associated with the introduction of smart metering and time-of-use tariffs	Randomised controlled smart metering trial (Ireland)	Although treated households also increased their self-reported energy-reducing information, such improvements are not correlated with demand reductions in the short-run

Acknowledgments. This paper is published thanks to the support of the internal projects of University of Hradec Kralove: Economic and Managerial Aspects of Processes in Biomedicine.

References

1. Robles, R.J., Kim, T.: Applications, systems and methods in smart home technology: a review. Int. J. Adv. Sci. Technol. **15**, 37–48 (2010)
2. Sripan, M., Lin, X., Ponchan, P., Ketcham, M.: Research and thinking of smart home technology. In: International Conference on Systems and Electronic Engineering (ICSEE 2012), Thailand, pp. 61–63 (2012)
3. Smart Homes and Building Association. http://www.shaba.eu/. Accessed 2 Apr 2016
4. Prospero, M., Wolpin, S.: Best smart home gadgets of 2016. http://www.tomsguide.com/us/best-smart-home-gadgets,review-2008.html. Accessed 15 Apr 2016
5. Statista: Smart home (2015). https://www.statista.com/outlook/279/smart-home. Accessed 15 Apr 2016
6. International Energy Agency (IEA): Energy efficiency market report. International Energy Agency (IEA), Paris, France (2013)
7. Mehdi, G., Roshchin, M.: Electricity consumption constraints for smart-home automation: an overview of models and applications. Energy Procedia **83**, 60–68 (2015)
8. Vanus, J., Vojcinak, P., Martinek, R., et al.: Building heating technology in Smart Home using PI System management tools. Perspect. Sci. **7**, 114–121 (2016)
9. Piccoloa, A., Simone, F.: Energy performance of an all solid state electrochromic prototype for smart window applications. Energy Procedia **78**, 110–115 (2015)
10. D'Oca, S., Corgnati, S.P., Buso, T.: Smart meters and energy savings in Italy: determining the effectiveness of persuasive communication in dwellings. Energy Res. Soc. Sci. **3**, 131–142 (2014)

11. Cucea, E., Cucea, P.M., Younga, C.H.H.: Energy saving potential of heat insulation solar glass: key results from laboratory and in-situ testing. Energy **97**, 369–380 (2016)
12. Carrolla, J., Lyonsa, S., Denny, E.: Reducing household electricity demand through smart metering: the role of improved information about energy saving. Energy Econ. **45**, 234–243 (2014)
13. Simanavicienea, Z., Volochovica, A., Vilke, R., et al.: Research review of energy savings changing people's behavior: a case of foreign country. Procedia Soc. Behav. Sci. **191**, 1996–2001 (2015)
14. Ehrhardt-Martinez, K., Donnelly, K.A., Laitne, J.: Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities. American Council for an Energy-Efficient Economy, Washington, DC, USA (2010)
15. Darby, S.: The effectiveness of feedback on a review for defra of the literature on metering, billing and direct displays. Environmental Change Institute, University of Oxford, Oxford, UK (2006)

Senior Citizens' Views of Using Medical Technologies – Case Study in Central Europe

Petra Maresova and Jaroslav Kacetl(✉)

Faculty Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove, Czech Republic
{blanka.klimova, petra.maresova}@uhk.cz

Abstract. The aim of this paper is to specify the seniors' attitudes towards modern technology as a tool for improving the quality of life in case of chronic disorders. The research is conducted in the Czech Republic. This study employs the analysis of both primary and secondary data. Primary data was collected through a questionnaire survey among senior citizens living in the Czech Republic. 170 questionnaires were distributed, 112 of which were returned and processed. The source of secondary data was the Czech Statistical Office. The case study done in the Czech Republic showed that local patients' attitude to modern technologies is rather positive. 86% of respondents use personal computers and 32% wield smart phones. Patients tend to be hesitant only concerning the utilization of smart household appliances, which would be welcome by 44% of the respondents and only 26% of them agreed strongly.

Keywords: Senior citizen · Medical technologies · Survey · Central Europe - Czech Republic

1 Introduction

Population ageing has been seen for several years as a hot issue of the future. In the course of several decades to come the developed countries are going to witness an unprecedented growth of older population. The issue of ageing population is becoming extremely vital [1, 2] and it is connected with different type of chronic disease such as mental disorders or depression. Depression is nowadays the fourth common cause of morbidity and by the year of 2020 it should seize the first position [3]. Therefore, there is persistent effort of searching for alternative non-pharmacological treatments which are less costly and improve quality of life. Rapidly developing modern technologies can provide better care in many areas. For instance, a new term telemedicine, meaning distance technologies for patient monitoring, represents these new possibilities.

Telemedicine includes a variety of distance technologies for patient monitoring. It is a part of e-healthcare. Distant medical monitoring saves both financial and human resources. It also shortens emergency response times, which is extremely vital especially in cases when a lot of patients are treated without sending a doctor by means of communication technologies in their homes. Li [4], who focus in their article on cooperation with patients, maintain that a lot of patients are willing to cooperate in distant monitoring of cardio-respiratory signals transmitted from their homes to

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_78

centralized interactive voice response systems. The authors continue to say that sensors are the basis and communication networks are pillars supporting various telemedicine applications and facilitate providing intelligent home environment to individuals4 Mobile health monitoring systems may help both patients and doctors by providing an instant access to medical information, in particular in case of emergency. According to clinical assessments, attached sensors do not hinder the patient in their ordinary life. The system's battery time of approximately 6 h is enough to cover the active part of a subject's normal day [5]. The fact that telemedicine is a prospective area can be supported by data from the year of 2010. Wootton et al. [6] found in their April 2010 literature survey 20 studies in which the percentage of cases when telemedicine enabled patients to avoid travelling to their doctors. The total number of such patients in these studies was 5,199. "The mean percentage avoided travel reported in the 12 store-and-forward studies was 43%. In the 7 real-time studies and in a single study with a hybrid technique, 70% of the patients avoided travel". In the area of mental illness technology brings online treatment applications [7] for running online treatment trainings, online interactive psychotherapeutic courses, online cognitive behavioral therapy sessions, online motivational home reminders, or monitoring of the symptoms of illness [8]. As Topolovec-Vranic et al. [9] showed, self-monitoring treatment approaches for mental illnes seem to be more accessible for patients since they can exploit them from anywhere and at any time. Overall, among widespread devices belong mobile phones with an emergency button, lost item trackers, wander door alarms, GPS tracking units, GPS watches, and many others. New trends encompass training senior citizens to use computers and other multimedia devices as well as the development of healthcare facilities alleviating senior citizens' pain and fear, or improving their movement.

In order to facilitate further development in this area, it is important to ensure that patients, specifically senior citizens not only have basic skills needed to use modern technologies, but also show positive approach to using them. All this as a precondition for the application of technology in the treatment of chronic diseases, including mental. The aim of this contribution is therefore to determine senior citizens' ICT skills and find out their attitudes to using communication and medical technologies. The research was done as a study in a Central European country – namely the Czech Republic.

2 Methods

For research purposes, there were used quantitative methods, including the analyses of both primary and secondary data. The basic method used was a survey among seniors in the Czech Republic, Central Europe. All respondents were at least 60 years of age. The questionnaires were distributed in the period between 19th January 2015 and 19th March 2015. The researchers distributed 170 questionnaires. They collected and could use 112 of them. The rate of questionnaire return was 66%. The questionnaire was divided into three parts. First, identification questions about the sex, age, education and job of respondents were asked. The second part aimed to specify the respondents' knowledge of information technologies. The final part focused on senior citizens' attitudes towards medical technologies.

3 Results

3.1 Characteristics of Respondents

The University of Hradec Kralove (UHK), the Czech Republic carried out a questionnaire survey called 'The Potential of Information Technologies for Supporting Senior Citizens' Quality of Life'. Its aim was to describe the approach of senior citizens to modern technologies.

The survey was divided into three parts and was aimed at the 60-plus age group. The first part consisted of identification questions. Another one focused on senior citizens' knowledge of information technologies (IT) they use. The final part collect data about senior citizens' willingness to use both IT monitoring their health and interconnected smart electronic appliances. There were 59% of female respondents and 41% of male ones. The ratio can be considered almost balanced. Most respondents belonged to the 60–64 and 65–69 age groups. These two groups combined amounted to 67% of all respondents. There were only 23% of financially active respondents. The others draw their state pension and consequently do not belong among economically active population.

3.2 Senior Citizens' Utilization of Modern Technologies

The collected data indicate that 86% of respondents use personal computers and only 32% wield smart phones. This result differs from the data obtained by the Czech Statistical Office, which suggest lower utilization of both PCs and mobile technologies. It may mean that there was further development between 2010 (when the Czech Statistical Office published their survey) and 2015 (when The UHK survey was carried out). If the real situation is better reflected by the UHK survey, it should be possible to use telemedicine on a much larger scale. Similarly, 85% of respondents answered that they had and used home internet access. As few as 15% stated that they could not access the Internet from home.

On the other hand, as regards tablet computers and e-book readers, there is still a relatively small potential of their exploitation among senior citizens. Consequently, another question was asked, namely whether there are any barriers to even more significant usage of information technologies. Surprisingly, 63 respondents agreed that nothing hinders them from using information technologies. Other options, e.g. fear from new things, lack of knowledge, lack of courses, lack of time, safety concerns, were chosen by about ten respondents each.

3.3 Attitudes of Senior Citizens to Using Monitoring and Smart Devices

Another aim of the UHK survey was to analyse senior citizens' attitudes to using interconnected monitoring and smart devices. The collected data should determine whether senior citizens are open to using new technologies, which may lead to discerning a suitable direction of further development in using these devices. Figure 1 depicts whether or not the respondents would allow home camera systems monitor their

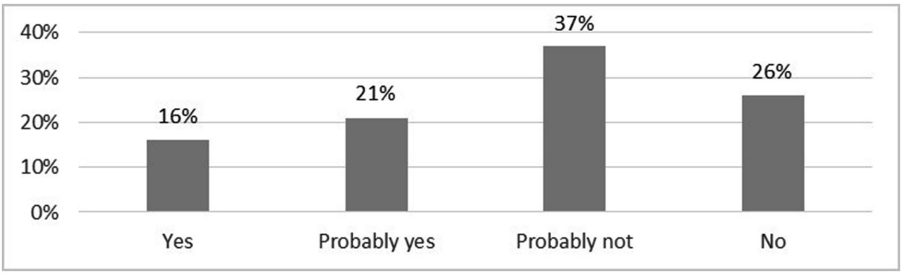


Fig. 1. Senior citizens' attitudes (willingness) to using interconnected monitoring and smart devices

movement at home. Surprisingly, most of them would not mind being monitored at home. Only 16% of respondents expressed their utter disagreement. These results reflect senior citizens' trust in these devices and relatively low fear of any abuse or misuse of information gathered by these devices by other people.

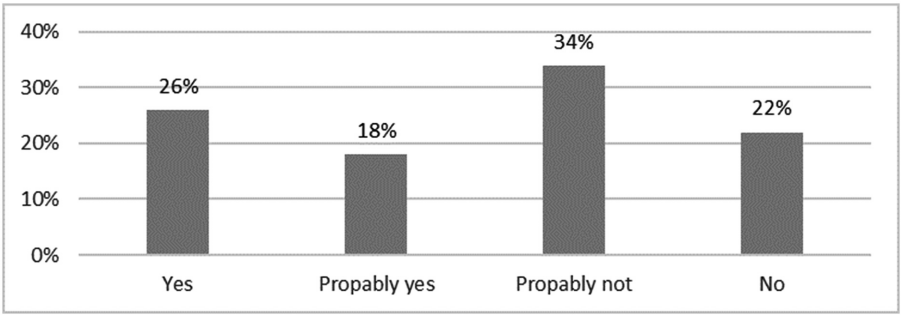


Fig. 2. Attitudes to smart household appliances

When asked if they would be willing to wear sensors (blood pressure monitors, ECG monitors, glucose monitoring devices, body temperature sensors) on their bodies, 80% of respondents answered they would. As many as 40% responded they would certainly wear these sensors, the remaining 40% would assent to wearing them when persuaded about involved benefits. Telemedicine is much more popular than it may seem. As few as 8% of respondents expressed their utter disagreement with using wearable sensors. It is an optimistic result. The last explored attitude was the one to smart household appliances. This area is not well-known among senior citizens. The respondents' answers were therefore quite reserved. Smart appliances would be welcome by 44% of respondents although only 26% of respondents agreed strongly (Fig. 2).

All in all, the survey of how Czech Republic's senior citizens use information technologies proved that the interest in new technologies is quite high and this age group members' attitudes to using telemedicine technologies are satisfactory.

4 Discussion

The results of this survey in Central Europe correspond with findings of several other authors [12, 14–16] who maintain that patients' attitude to using new technologies monitoring their health is rather positive. The collected data indicate that 86% of respondents use personal computers and only 32% wield smart phones. This result differs from the data obtained by the Czech Statistical Office, which suggest lower utilization of both PCs and mobile technologies. It may mean that there was further development between 2010 (when the Czech Statistical Office published their survey) and 2015 (when The UHK survey was carried out). Patients tend to be hesitant only in case of smart household appliances, which would be welcome by 44% of respondents and only 26% of respondents agreed strongly. Also Li [4] stated that patients are willing to cooperate in home telemonitoring of cardiorespiratory signals that are sent into centralized interactive voice-activated systems.

Sensors are the basis and communication networks act like pillars supporting various telemedicine apps and they facilitate providing smart home environment to individuals.

5 Conclusion

The growth of the number of senior people using modern technologies is currently a worldwide trend. The renowned analytical agency Gartner identified the top 10 strategic technology trends for 2016, including the device mesh, ambient user experience, information of everything, advanced machine learning, autonomous agents and things, mesh app and service architecture, internet of things architecture and platforms. [17] The issues of safety, ethical concerns as well as legislative aspects go hand in hand with these trends. It therefore seems vital to look into problems like privacy protection and protection of personal data, which is an important challenge.

Acknowledgments. This paper is published thanks to the support of the internal projects of University of Hradec Kralove. In addition, the authors thank Bohdana Kasparova for her help with the research.

References

1. Maresova, P., Mohelska, H., Kuca, K.: Social and family load of Alzheimer's disease. Appl. Econ. **48**, 1936–1948 (2015). doi:10.1080/00036846.2015.1111986
2. Maresova, P., Mohelska, H., Dolejs, J., Kuca, K.: Socio-economic aspects of Alzheimer's disease. Curr. Alzheimer Res. **12**(9), 903–911 (2015)
3. WHO: Depression. Key facts (2016). http://www.who.int/mediacentre/factsheets/fs369/en/ 27 February 2016)
4. Li, K.F.: Smart home technology for telemedicine and emergency management. J. Amb. Intell. Humaniz. Comput. **4**(5), 535–546 (2012). doi:10.1007/s12652-012-0129-8

5. Reiss, A., Stricker, D.: Aerobic activity monitoring: towards a long-term approach. Univ. Access Inf. Soc. **13**(1), 101–114 (2014). doi:10.1007/s10209-013-0292-5
6. Wootton, R., Bahaadinbeigy, K., Hailey, D., Taniguchi, E.: Estimating travel reduction associated with the use of telemedicine by patients and healthcare professionals: proposal for quantitative synthesis in a systematic review. BMC Health Serv. Res. **11**, 185 (2011). doi:10.1186/1472-6963-11-185
7. White, G., Caine, K., Connelly, K., et al.: Designing consumer health technologies for the treatment of patients with depression: a health practitioner's perspective. Interact. J. Med. Res. **3**(1), e2 (2014)
8. Kramer, L.T., Owen, R.R., Wilson, C., Thrush, C.R.: Relationship between self-report and clinician-rated impairment in depressed outpatients. Commun. Ment. Health J. **39**(4), 299–307 (2003)
9. Topolovec-Vranic, J., Zhang, S., Wong, H., et al.: Canadian brain injury and violence research team. PLoS One **10**(11), e0141699 (2015). doi:10.1371/journal.pone.0141699
10. Rychtaříková, J.: Successful aging 21st century. Demography **44**(1), 43–46 (2002)
11. CSO (Czech Statistical Office): Homepage on the internet (2014). www.csu.cz
12. Seto, E., Leonard, K.J., Masino, C., Cafazzo, J.A., Barnsley, J., Ross, H.J.: Attitudes of heart failure patients and healthcare providers towards mobile phone-based remote monitoring. Med. Internet Res. **12**(4), e55 (2010)
13. McGillicuddy, J.D., Weiland, A.K., Frenzel, R.M.: Patient attitudes toward mobile phone-based health monitoring: questionnaire study among kidney transplant recipients. J. Med. Internet Res. **15**(1), e6 (2013)
14. Mudano, A.S., Gary, L.C., Oliveira, A.L., et al.: Using tablet computers compared to interactive voice response to improve subject recruitment in osteoporosis pragmatic clinical trials: feasibility, satisfaction, and sample size. Patient Prefer. Adherence. **7**, 517–523 (2013). doi:10.2147/PPA.S44551
15. Liu, F., Zou, Y., Huang, Q., et al.: Electronic health records and improved nursing management of chronic obstructive pulmonary disease. Patient Prefer. Adherence **9**, 495–500 (2015). doi:10.2147/PPA.S76562
16. Lang, A.R., Martin, J.L., Sharples, S., et al.: Medical device design for adolescent adherence and developmental goals: a case study of a cystic fibrosis physiotherapy device. Patient Prefer. Adherence **8**, 301–309 (2014). doi:10.2147/PPA.S59423
17. Gratner: Homepage on the internet. Strategic-technology-trends (2015). https://www.gartner.com/doc/3143618/top–strategic-technology-trends. Accessed 14 Dec 2015

Audiences Counting in Cinema by Detecting Occupied Chairs

Zhitong Su[1], Jun Lan[1], Wei Song[1,2(✉)], Simon Fong[3], and Yifei Tian[1]

[1] Department of Digital Media Technology, North China University of Technology, Beijing, China
{suzhitong, sw}@ncut.edu.cn, junlanl234@163.com, tianyifei0000@sina.com
[2] Key Laboratory on Integration and Analysis of Large-Scale Stream Data, North China University of Technology, Beijing, China
[3] Department of Computer and Information Science, University of Macau, Macau, China
ccfong@umac.mo

Abstract. Human counting in cinema is easily influenced by varied illumination, so as to become a complicated problem. This paper develops an audience counting system in cinema by detecting occupied chairs in captured images. Firstly, we initialize chair regions in a background image manually. Then, the differences between the background and current images are detected as foreground regions. Such rough segmentation results always contain noise because of environmental illumination changing. Thus, a contour difference detection algorithm is applied to refine the audience detection results. Next, if both foreground and contour differences in a chair region are larger than a threshold, this chair is recognized to be occupied by an audience. Finally, the audience number is estimated by counting the occupied chairs.

Keywords: Audiences counting · Foreground segmentation · Contour detection

1 Introduction

Box office receipt is a measurable indicator to evaluate film's favorable rate. However, deception means exist in counting the receipts to attract cinema attendance. Counting audiences in the cinema is a reliable way to estimate attendance rate. Traditionally, this work is always done manually, that cost huge workload. To realize automatic cinema attendance estimation, this paper proposes an occupied chairs detection method in video images.

Foreground segmentation is the first step of people and moving objects detection [1]. The typical segmentation methods based on pixel difference detection are easily influenced by varied environmental illumination [2]. The feature-based object recognition methods detect and track the objects of special shape, which have highly reliability and anti-interference for environmental illumination changing situation [3]. To enhance

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_79

audience counting accuracy, this paper detects occupied chair based on foreground detection and contour detection.

In our method, a background image is captured by a fixed camera, when the cinema has no audience. Before a movie is played, an image is captured with lots of people. Our application aims to count the number of the audiences seating in chairs. Thus, only the chair regions are necessary to be analyzed. If the pixel and counter differences between these two images are larger than a threshold, this chair is considered to be occupied. The audience number is estimated by counting the occupied chair.

Our proposed method does not only solve deception problem, but also count the attendance rate automatically. This office receipts counting work provides reliable information for big data of cinema favorable analysis.

The remainder of this paper is organized as follows, Sect. 2 overviews related work. Section 3 discusses the method of audiences counting. Section 4 describes the experiments using the proposed method. Section 5 concludes this paper.

2 Related Works

Face and body detection methods in color images are widely researched in human counting. Zou et al. [4] counted human number by detecting faces in RGB (Red, Green, Blue) and HSV (Hue, Saturation, Value) spaces. After the pixels in skin color were segmented, the face regions were eliminated based on face contour features. Xu et al. [6] proposed a head-shoulder component detection method with a Kalman filter to track people in the video sequences. In the foreground segmentation process, a mixture of Gaussians model was applied to extract the foreground pixels roughly. Because the head and shoulder components were less varied, the people were recognized based on these features. However, such color-based segmentation methods were easily influenced by illumination changing and shadow effect [5].

Based on a background model, Tong et al. [7] extracted foreground connected regions and classified human upper body using a support vector machine. To remove the noise caused by the illumination changing, an energy function was applied to compute the region contour so as to increase the accuracy of human segmentation results. By integrating foreground segmentation and contour detection algorithm, this paper proposes an occupied chair method for counting audiences in cinema.

3 The Audience Counting System

In order to count the audiences number in cinema, we propose an audience counting system described in Fig. 1. A background image of the cinema is captured and initialized when nobody exists. The current image is captured by a fixed camera for monitor the cinema scene before the movie starts. The system aims to detect the occupied chairs in the current images, which are assumed as audiences on them.

Firstly, the chairs are manually labeled in the background image to obtain the regions where the audiences seat. Next, the contour pixels of the background and current images are detected. To detect contour pixels, a simple kernel-based contour

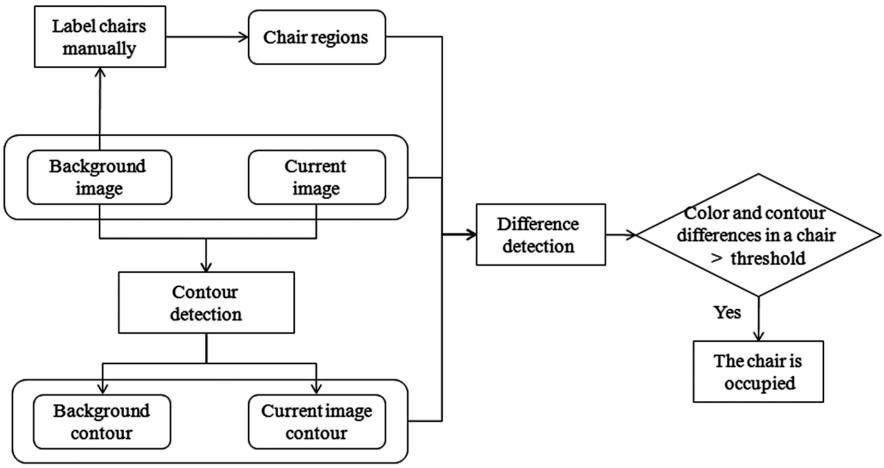


Fig. 1. Framework of audiences counting

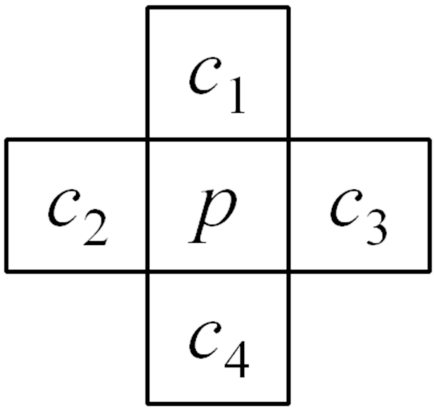


Fig. 2. The kernel for contour detection

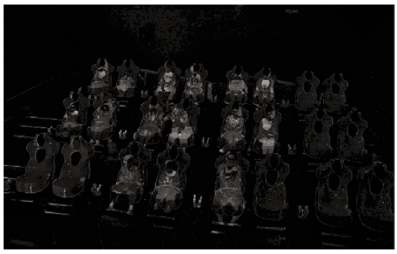
detection algorithm is applied. Figure 2 shows a kernel clique, including a center pixel p and its four neighboring pixels c_n. The changing gradient g_p in pixel p is formulated as Eq. (1).

$$g_p = \sum_{n=1}^{4} |p - c_n|/4 \tag{1}$$

Then, the contour and color differences are extracted from RGB and color maps of the background and current images. If both color and contour differences in a chair region are larger than a threshold, this chair is determined to be occupied by an audience. Finally, audiences' number is computed by counting the occupied chairs.

(a) The background image

(b) A current image



(c) The color difference map

(d) The contour difference map

(e) The detection result of occupied chairs

Fig. 3. An audience detection result

4 Experiments

In this section, we analyze the performance of the proposed system, as shown in Fig. 3. The experiments were implemented on a 3.20 GHz Intel® Core™ Quad CPU computer with a GeForce GT 770 graphics card, 4 GB RAM. In our cinema, more than 5,000 images of 5184 × 3456 were captured and tested. The audience detection accuracy is around 99.37%.

The initialized background and a current image were shown in Fig. 3(a) and (b) respectively. The color difference was shown as the gray map in Fig. 3(c). After computing the contour maps in the background and the current images, the contour difference map was generated as in Fig. 3(d). If the pixel count of contour differences in a chair region was larger than a threshold, this region was marked in red. If the sum of color differences in a chair region was larger than a threshold, this region was marked in green. Thus, the chair was determined to be occupied, when both red and green marks were labeled on it. As shown in Fig. 3(e), tow bottom-left chairs were affected by high lighting projection, so the color difference satisfied the detection condition. But the contour difference was not larger than the threshold so that these chairs were recognized as empty chairs.

5 Conclusions

This paper proposed an audiences counting system in cinema by detecting occupied chairs. Using the color and contour difference detection for background image and current image, the system was able to detect the occupied chairs in the situation of varied illumination. The experimental results confirmed that the proposed method was implemented effectively for counting box office receipts. In the future, we will develop some interaction interfaces for interactive movie using the proposed audience detection method.

Acknowledgment. This research was supported by the National Natural Science Foundation of China (61503005), by Beijing Natural Science Foundation (4162022), and by High Innovation Program of Beijing (2015000026833ZK04).

References

1. Chen, S., Fern, A., Todorovic, S.: Person count localization in videos from noisy foreground and detections. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1364–1372 (2015)
2. Rachmawati, E., Khodra, M.L., Supriana, I.: Edge based approach in object boundary detection on multiclass fruit images. In: 4th International Conference on Information and Communication Technology (2016)
3. Li, B., Zhang, J., Zhang, Z., Xu, Y., et al.: A people counting method based on head detection and tracking. In: Smart Computing, pp. 136–141 (2014)

4. Zou, L.H., Liu, Y.C.: A new algorithm of counting human based on segmentation of human faces in color image. In: International Conference on Computational Intelligence and Security, pp. 505–509 (2009)
5. Hafiz, F., Shafie, A.A., Khalifa, O., et al.: Foreground segmentation-based human detection with shadow removal. In: International Conference on Computer and Communication Engineering, pp. 1–6 (2010)
6. Xu, H., Lv, P., Meng, L.: A people counting system based on head-shoulder detection and tracking in surveillance video. In: International Conference on Computer Design and Applications, pp. 394–398 (2010)
7. Tong, R., Xie, D., Tang, M.: Upper body human detection and segmentation in low contrast video. IEEE Trans. Circ. Syst. Video Technol. **23**, 1502–1509 (2013)

Efficient Distributed Index Structure and Encrypted Query Processing Scheme for Cloud Computing Environment

Yeonwoo Jang, Hyunjo Lee, and Jae-Woo Chang(✉)

Department of Computer Engineering, Chonbuk National University, Jeonju, Republic of Korea
{kp7050, o2near, jwchang}@jbnu.ac.kr

Abstract. To the best of our knowledge, there has been no research on index structure for the encrypted data. In addition, the existing query processing schemes over the encrypted data can support limited types of queries [1–3]. To solve the problems, in this paper, we propose a distributed index structure and a query processing scheme for the encrypted data. The proposed distributed index structure guarantees data privacy preservation and performance improvement for the various types of queries. In addition, the proposed query processing scheme provide both high query performance and 100% accuracy while preserving the data privacy. Finally, we show from our performance analysis that our proposed index structure and query processing scheme are suitable for protecting the data privacy of the mobile sensitive data.

Keywords: Data encryption · Encrypted query processing · Distributed index structure · Mobile data privacy

1 Introduction

To protect sensitive data in outsourced database environment, it is necessary to develop not only an efficient distributed index structure, but also a query processing scheme over the encrypted data. First, HBase [4], which was developed as a data management model for a large amount of data in Apache Hadoop project, is a typical distributed index structure. Because HBase stores data in distributed manner using the cluster system and provides MapReduce programming model, it can support efficient data analysis and query processing. However, HBase is not adequate as a distributed index structure for large sensitive data because it cannot support efficient data management for the encrypted data. Second, CryptDB [5] is a typical query processing scheme over the encrypted data. CryptDB encrypts data in a column-wise way by considering various query types, i.e., exact matching, range query and so on. However, CryptDB cannot support operations among data with different columns because it uses the different types of encryption schemes depending on attribute types.

To solve the problems, in this paper, we propose a distributed index structure and a query processing scheme for the encrypted mobile data. The proposed distributed index structure uses a bitmap to preserve the privacy of sensitive data. The index structure

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_80

consists of a prefix-based upper index structure and a signature-based lower index structure. The prefix-based upper index structure can relieve data skewness whereas the signature-based lower one can improve query processing performance. Meanwhile, the proposed query processing scheme basically supports exact matching, range and partial matching queries. In addition, our scheme can support not only arithmetic and comparison operations among data with different columns.

The rest of the paper is organized as follows. We explain the proposed distributed index structure and our query processing algorithm over the encrypted data in Sects. 2 and 3, respectively. In Sect. 4, we compare the performance of our query processing scheme with that of the existing schemes. Finally, we conclude this paper with future work in Sect. 5.

2 Proposed Distributed Index Structure for Encrypted Data

Figure 1 shows the overall architecture of the proposed distributed index for the encrypted mobile data. The proposed distributed index consists of three components: (i) bitmap key generation (ii) prefix-tree based upper index (iii) signature based lower index.

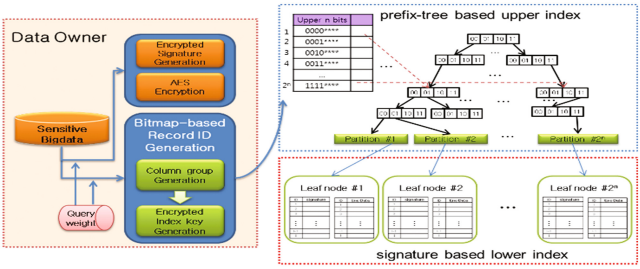


Fig. 1. Overall architecture of the proposed distributed index structure.

A. Bitmap key generation

To enhance the performance for query processing over the encrypted data, it is necessary to store data in a distributed manner. For this, we generate the bitmap keys of records by using Random Projection tree (RP-tree) [6] which is a typical data partitioning scheme. By using RP-tree, we can protect data partitioning information from an attacker. This is because RP-tree performs clustering by projecting data to an arbitrary partition axis. However, data may be skewed in a node when clustering data using the RP-tree. To solve the problem, we can choose column pairs by performing a correlation analysis which is a statistical method to analyze the correlations among data in different data groups. The weak correlation means that the data are scattered. To relieve the data skewness, we generate a RP-tree with a pair of columns whose correlation is weak. For this, we calculate a sample correlation coefficient by using Eq. (1). The IDs of the leaf nodes in the RP-tree are converted into bit patterns. By concatenating the bit patterns of each RP-tree, we generate a bitmap keys.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i + \overline{X})^2 \sum_{i=1}^{n}(Y_i + \overline{Y})^2}} \quad (1)$$

B. **Prefix-tree based upper index**

We can support a partial matching query on a column by using the prefix-tree based upper index. We define four query types based on the position of a wildcard; *back*, *front*, *center*, and *split* patterns. The wildcard can be generated by replacing into '*' a bit being excluded in the designated column of the query. When processing a range query using the prefix-tree, the query performance can be deteriorated. This is because the existing prefix-tree performs the range query with two steps; searching for the starting position of the range and expanding the range with a linked-list. To solve the problem, we build a hash table with the upper k-bits of the bitmap key. We calculate the cost of the hash table by considering the memory usage and the data retrieval time, as shown in Eq. 2.

$$\text{Cost}_{Hash} = w_{SearchTime} * \frac{O(1)}{O(\log_2 n)} + w_{MemoryUsage} * HashTableSize \quad (2)$$

C. **Signature based lower index**

A data node in the signature based lower index contains a data <encrypted signature, AES [7] encrypted data>. Because the size of the encrypted signature is much smaller than that of the AES encrypted data, we can store the encrypted signature into a cache, leading to the better retrieval performance by reducing the cost of disk I/O. Thus, we can determine the encrypted signatures to be cached by considering both their frequency and their recency. For this, we define a validity of recency (V_{Rec}) and a value of frequency (V_{Fre}). The V_{Rec} can be computed by measuring how much time is elapsed since a recent query has accessed the encrypted signatures. The V_{Fre} can be computed by the number of accesses to the encrypted signature in a cache for a specific period of time. So the cache ratio using both V_{Rec} and V_{Fre} is calculated by using Eq. 3. Here, t_{cur} is the current time and *timeInterval* means the elapsed time from the recent query time to t_{cur}. *Count* is the number of queries issued from t = 0 to t_{cur}. And w_r and w_f mean the weights of recency and frequency, respectively.

$$\begin{aligned} V_{Rec}(i, t_{cur}) &= V_{Rec}(i, t_{cur-1}) + t_{cur} * \sqrt{\frac{t_{cur} - timeInterval}{2}} \\ V_{Fre}(i, t_{cur}) &= \frac{Count(i, t_{cur}) - Count(i, t_{cur-1})}{timeInterval} \\ w_{cache}(i) &= w_r * V_{rec}(i, time_{current}) + w_f * V_{frequency}(i, time_{current}) \end{aligned} \quad (3)$$

3 Proposed Query Processing Scheme over the Encrypted Data

Figure 2 shows the overall architecture of the proposed query processing scheme over the encrypted data. First, when an authorized user (AU) requests a service, a data owner (DO) not only generates an upper index key using the RP-tree, but also creates the encrypted signature for a given query. Here, we support simple SQL-like query for AU. Second, a service provider (SP) selects the data nodes corresponding to the upper index key for processing the query. Then, both the encrypted signature and the query are sent to the data nodes being selected. Third, the nodes retrieve the candidate results by performing the bit operations of the encrypted signatures. Fourth, the SP returns the encrypted candidate result to the DO. Fifth, the DO decrypts it and obtains the final result by filtering out unnecessary candidates. Finally, the DO sends the final result to the AU.

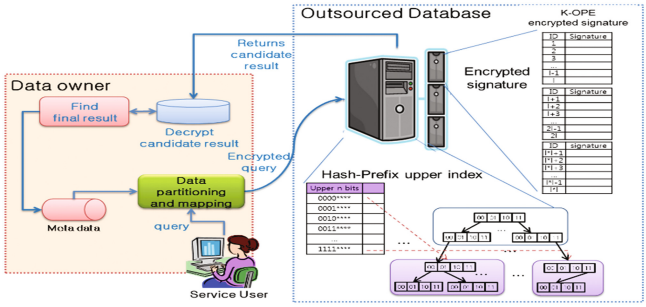


Fig. 2. Overall architecture of the proposed query processing scheme over the encrypted data.

A. Exact-match and partial-match query processing

The exact-match query processing algorithm is performed as follows. First, the algorithm generates a b-bit key for a given query by using the grid-tree index and creates an encrypted signature to retrieve the encrypted data in a data node. Second, it searches the hash table by using the upper n-bits of the key and generates the inverted list of partitions of the RP-tree. Third, the algorithm finds data nodes corresponding to each partition, in order to process the query by using the lower $(b - n)$ bits of the key. Fourth, it determines whether or not the candidate nodes are included in the search area. Finally, it sequentially searches the signatures being stored in the data nodes for filtering out unnecessary candidates. To the best of our knowledge, there are no existing works to support a partial-match query on the encrypted data. Thus, we propose a partial-match query processing algorithm. First, the algorithm analyzes a query to find out a query pattern. Second, it stores the indexes of the prefix sub-trees corresponding to the query pattern. Third, it assigns them to GPU threads and retrieves candidate results. Finally, it merges the candidate results and returns the final result to the user.

B. **Range query processing**

The range query processing algorithm is performed as follows. First, the algorithm finds the search range of the prefix-tree for a given query. Second, it selects the prefix sub-trees corresponding to the query and assigns them to threads to retrieve the data. Third, the exact-match query is performed on the assigned sub-trees by GPU threads. Finally, the algorithm determines whether or not the candidates retrieved from each thread are included in the query region. The final result is returned to the user.

C. **Operations among data with different columns**

In data analytics, it is necessary to obtain meaningful information by analyzing data from multiple columns. However, CryptDB cannot support a query on multi-columns because it uses different types of encryption schemes depending on its attribute types. To solve the problem, we propose a grid-based query processing algorithm over the encrypted data. When a query contains k number of columns, the algorithm generates both a column-wise grid-tree mapping index and a *k*-dimensional grid index. Then, the algorithm obtains the result by projecting the query onto the *k*-dimensional grid index. To find the query range based on the projected query, the algorithm determines a location relation between a cell and the query plane by calculating the distance between them. For this, the algorithm first applies a center point of the grid cell $p = (p_1, p_2, \ldots, p_k)$ to the *k*-dimensional query $ax_1 + bx_2 + \ldots + kx_k > 0$. If $ap_1 + bp_2 + \ldots + kp_k$ is greater than 0, the gird cell is selected as a candidate because the cell can include records satisfying the query. Otherwise, it calculates a distance h between p and v by using Eq. (4). In addition, it calculates a distance d between p and a vertex of the grid by using Eq. (5). If $h < d$, the query plane is meant to overlaps the grid cell. The grid cells satisfying this condition become the candidate result.

$$h = \frac{|v \cdot p|}{|v|} = \frac{|a(x_1 - p_1) + b(x_2 - p_2) + \ldots + k(x_k - p_k)|}{\left|\sqrt{a^2 + b^2 + \ldots + k^2}\right|} \tag{4}$$

$$d = \sqrt{\frac{k}{(2 \times m)^2}} = \frac{1}{2 \times m} \times \sqrt{k} \tag{5}$$

4 Performance Analysis

In this section, we present the experimental analysis of our encrypted query processing algorithm. For performance analysis, we compare our encrypted query processing algorithm with the existing CryptDB [5], in terms of the query processing time. For our experiments, we use an expanded Census dataset whose size is 2 GB. The dataset includes sensitive data such as name, married or single, number of children, sex, age, level of education, job, majority, salary, and income/expense by property. We did our performance analysis on the Window 7 Enterprise K system with Intel Core2 Quad 2.4 GHz and 2 GB memory.

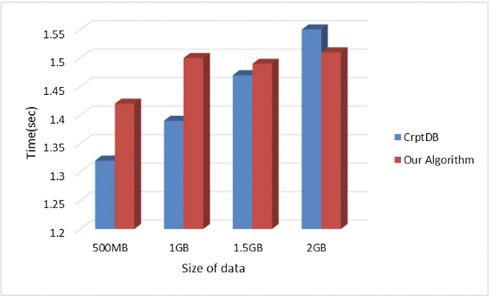


Fig. 3. Exact-match query processing time with varying the size of data.

Figure 3 shows the exact-match query processing time of our algorithm and that of the existing CryptDB over the encrypted data. Our algorithm shows 1.42, 1.5, 1.49, and 1.51 s with varying the size of the data from 500 MB to 2 GB, whereas CryptDB shows 1.32, 1.39, 1.47 and 1.55 s. Our algorithm shows better performance when the data size becomes large. This is because our algorithm can select less number of candidates by using our prefix-tree based index structure.

Figure 4 shows the range query processing time for both our algorithm and the existing CryptDB, with varying the size of the query range from 0.0001% to 0.0005%. Here the size of the query range is computed as the ratio of the query range size over the data size. When the query range size is equal to or smaller than 0.0002%, our algorithm is 1.1 times faster than CryptDB. However, when the query range size is greater than 0.0002%, our algorithm is about 15% slower than CryptDB. This is because the number of data with the same signature value is increased in our lower index as the query range size increases. So the data transmission cost becomes greater than CryptDB when the query range size is larger than 0.0002%.

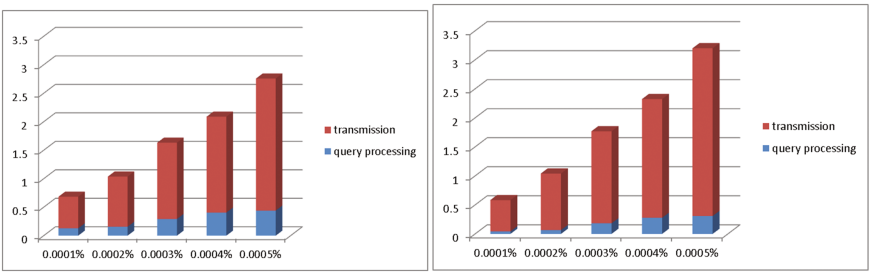


Fig. 4. Range query processing time with varying query range size

5 Conclusion and Future Work

We proposed a bitmap-based distributed index structure that guarantees data privacy preservation and a query processing scheme that can process a query over the encrypted data without data decryption. Our query processing scheme provides both high query processing performance and high query result accuracy while preserving data privacy. We showed from our performance analysis that our proposed index structure and query processing scheme are suitable for protecting large-scale sensitive data from attacker in data outsourcing environment.

As the future work, we will expand our work to handle elaborate types of queries, such as k-NN queries and skyline queries, over the encrypted data.

Acknowledgments. This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. R0113-16-0005, Development of a Unified Data Engineering Technology for Largescale Transaction Processing and Realtime Complex Analytics). Also, this work was partly supported by the Human Resource Training Program for Regional Innovation and Creativity through the Ministry of Education and National Research Foundation of Korea (NRF-2016H1C1A1065816).

References

1. Corena, J., Ohtsuki, T.: Secure and fast aggregation of financial data in cloud based expense tracking applications. J. Netw. Syst. Manag. **20**(4), 534–560 (2012)
2. Hu, H., Xu, J., Ren, C., Choi, B.: Processing private queries over untrusted data cloud through privacy homomorphism. In: ICDE, pp. 601–612 (2011)
3. Yao, B. Li, F., Xiao, X.: Secure nearest neighbor revisited. In: ICDE, pp. 733–744 (2013)
4. http://hbase.apache.org/
5. Popa, R.A., Redfield, C., Zeldovich, N., Balakrishnan, H.: CryptDB: protecting confidentiality with encrypted query processing. In: Proceedings of the 23rd ACM Symposium on Operating Systems Principles, SOSP 2011, pp. 85–100 (2011)
6. Dasgupta, S., Freund, Y.: Random projection trees and low dimensional manifolds. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing, pp. 537–546 (2008)
7. Advanced Encryption Standard (AES): NIST Federal Information Processing Standards Publication 197 (2001)

A Data Encryption Scheme Using Periodic Functions for Efficient Query Processing on Encrypted Data

Hyunjo Lee, Youngho Song, and Jae-Woo Chang(✉)

Department of Computer Engineering, Chonbuk National University, Jeonju, Republic of Korea
{o2near, songyoungho, jwchang}@jbnu.ac.kr

Abstract. Due to advancement in cloud computing technology, an order-preserving encryption schemes, called Programmable Order-Preserving Secure Index (POPIS), has been proposed. This scheme hides the original data while keeping the order of the encrypted values the same as that of the original data. So the service provider can perform query processing without decryption. However, because the encrypted data in POPIS is sorted by certain column values, it is weak to both order matching attacks and count attacks. To solve this problem, we propose a data encryption scheme using periodic functions. Our scheme generates encryption signatures based on data groups and periodic functions. With this, we can preserve the order of each data group and also can guarantee the data privacy. Finally, we show from the performance analysis that the proposed scheme is better in terms of the degree of privacy protection than the existing data encryption scheme.

Keywords: Wireless order-preserving encryption scheme · Data privacy protection · Database outsourcing

1 Introduction

With the development of cloud computing technology, researches on the order-preserving encryption scheme (OPES) has been proposed to enhance the query processing efficiency on the encrypted databases [1–5]. OPES hides the original data while keeping the order of the encrypted values the same as that of the original data. So the service provider can perform query processing without decryption. One of the most recent works is programmable order-preserving index scheme (POPIS) [6]. POPIS divides the original data into groups based on the range of the value. Then, for each data group, it encrypts the data with an encryption function that is monotone increasing. Here, it adds a random noise to hide the exact value of the original data. However, POPIS is weak to both order matching attacks and data counting attacks because the encrypted data are sorted by a given column.

To solve the problem, we propose a group order preserving data encryption scheme, using the periodic function. Our scheme called GOPES consists of three parts: a data grouping algorithm supporting k-anonymity, a data transformation algorithm

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_81

based on the periodic function, and a signature generation algorithm. First, with the data grouping algorithm, GOPES is able to preserve the order of data groups which contain more than k number of data. So GOPES can protect the encrypted data from the order matching attacks. Secondly, with the data transformation algorithm, GOPES modifies the frequency of data by assigning the same transformation value to the data whose original values fall into the same encryption group. So GOPES can preserve the encrypted data from the data counting attacks. Finally, with the signature generation algorithm, GOPES can perform both an exact matching query and a range query efficiently on the encrypted database by using bit operations.

The rest of the paper is organized as follows. First, Sect. 2 describes our GOPES. Secondly, we do the performance analysis by comparing our GOPES with POPIS in terms of both the query processing time and the data privacy protection. Finally, we conclude the paper with a brief summary and the future work.

2 Data Encryption Scheme Using the Periodic Function

2.1 Overall Architecture of the Proposed Encryption Scheme

Figure 1 shows the architecture of our group order-preserving encryption scheme (GOPES) based on periodic functions. GOPES consists of four steps; data group based transformation, function segment based transformation, periodic function based transformation and signature generation.

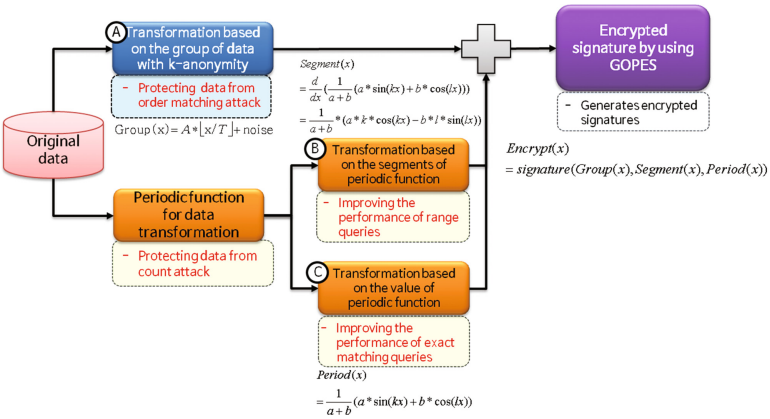


Fig. 1. An overall structure of the proposed encryption scheme

At first, in the data group based transformation step, GOPES makes groups of the original data by considering k-anonymity. Then it transforms the data based on their groups. Because our GOPES preserves the order of data groups, instead of the order of the encrypted data, it can protect the encrypted data from the order matching attack. Secondly, in the function segment based transformation step, GOPES transforms the

data based on the segments which is calculated by differentiating a periodic function. Because the data in the same segment have the same encryption value, GOPES can modify the frequency of the encrypted data. So it protects the encrypted data from the data count attack. Thirdly, in the periodic function based transformation step, GOPES modifies the data based on the periodic function in order to improve the performance of the exact matching query. Because GOPES shuffles the order of the encrypted data, it reduces a probability that the original data can be disclosed from the order matching attack. Finally, in the signature generation step, GOPES generates the encrypted signatures by concatenating group transformation, segment transformation, and function transformation values. Here, a signature means an abstract value of the data. So GOPES can modify 1-to-1 relationship between the original data and encrypted data into 1-to-many relationship. As a result, it is difficult for the attacker to obtain the original data from the encrypted one.

2.2 A Data Group Based Transformation Algorithm Supporting K-Anonymity

To protect the encrypted data from the order matching attacks, we propose a data group based transformation algorithm by considering k-anonymity. Our data group based transformation algorithm consists of two steps; (i) it makes groups of the data by clustering the k number of similar data. (ii) it transforms the data of each group into a random value. Because the range of the selected random value is set based on the order of the groups, it can preserve the order of data groups. By using our data group based transformation, GOPES has two main advantages. First, the probability of the disclosing data from the attacker is reduced to 1/k because each group contains the k number of random values whose orders are not the same as those of the original data. Second, because GOPES does not preserve the order of the original data, it is difficult to differentiate the data distributions of GOPES from those of the original data. Equation (1) shows the data grouping function. Here, u(x) means a unit function which outputs a result value 0 if x < 0. Otherwise, a result value is 1. And f(x) means a unit step function based on u(x) and Ti means the maximum value of the ith data group.

$$u(x) = \begin{cases} 0, & where\ x < 0, \\ 1, & where\ x < 1 \end{cases} \qquad (1)$$
$$f(x) = A \times (u(x) + u(x - T_i) + u(x - T_{i+1}) + \cdots)$$

Equation (2) shows a regulation function of Eq. (1). Here, x means the original value, T means a size of the original data for each group, and A means a size of the encrypted data for each group.

$$Group(x) = A \times \lfloor x/T \rfloor + random\_noise \qquad (2)$$

2.3 A Function Segment Based Transformation Algorithm

To protect the encrypted data from the data count attacks, we propose a function segment based transformation algorithm. A segment means an area generated by dividing an interval of a periodic function. If a set of the original data is in the same segment, GOPES encrypts the data into the same value. So GOPES can modify the frequency of the encrypted data. To generate segments, we choose a periodic function p (x) which repeats its values in regular intervals w. Here, the interval w, also called basic period, is a positive constant value. Considerations for selecting p(x) are as follows. (i) To improve the privacy of the encrypted data, we select the interval w which has no common multiple with the size of the original data (T). For example, if T is a rational number, we set w into an irrational number such as π. (ii) In a specific range of the original data, their encryption values can be skewed. So we select a periodic function which can solve this problem. For example, if we use p(x) = 2x − i * w (where i is an integer). For the original value x in a range [9.0, 10.0], its encrypted value is in [512.0, 1024.0]. However, x in [0.0, 1.0] is encrypted to a value in a range [512.0, 1024.0]. So the size of the encrypted data for x in [9.0, 10.0] is 256 times greater than that of the encrypted data for x in [0.0, 1.0]. As a result, an attacker can estimate the transformation function accurately. (iii) To make difficult to estimate the function, we select the function whose pattern in its interval is not simple. For example, if we select sin(x) or cos(x) as p(x), an attacker can easily estimate the function due to its simplicity.

Based on these considerations, we choose a trigonometric function as p(x). Equation 3 shows our p(x). Here, the range of p(x) is $-1 \leq p(x) \leq 1$, and the interval of p(x) is a the least common multiple of k and l. Equation 4 shows the segment generation function which is calculated by differentiating Eq. 4.

$$Periodic(x) = \frac{1}{a+b} \times (a \times \sin(k \times x) + b \times \cos(l \times x)) \tag{3}$$

$$\begin{aligned} \{x|0 \leq x \leq w\} &= \bigcup_{i=0}^{n} \{y|y \in Group_i\}(where,\ x \in R) \\ Segment(x) &= \begin{cases} \{x|0 \leq x < x_0\},\ where\ 0 < x_0 < x_i, \\ \{x|x_i \leq x < x_{i+1}\},\ where\ 0 < x_{i+1} < x_n, \\ \{x|x_i \leq x < x_n\},\ otherwise \end{cases} \\ &(where, \{x_i|0 \leq x_i < w\},\ Periodic(x_i) = 0,\ 0 \leq i < n) \end{aligned} \tag{4}$$

3 Performance Evaluation

We present performance results of our GOPES by comparing it with the existing POPIS [6] in terms of the data privacy and the query processing time. To measure the query processing time, two query types are used; an exact matching query and a range query. For our experimental environment, we use Microsoft Visual Studio 2010 compiler for Windows 7 Enterprise K OS running on Intel Core 2 Quad Q6600

2.4 GHz CPU with 2 GB Memory using. For the experiment, we generate a set of data based on the US census dataset. The size of the generated data is 2 GB. Similar to the US census dataset, our dataset includes sensitive attributes such as age, incomes, jobs and so on. So we encrypt those column data with POPIS and our GOPES.

3.1 Experimental Results in Terms of Data Privacy

Table 1 shows the probability of exposing data from the order matching attack. In POPIS, the ratio of the correctly matched pairs to total number of pairs is 100% and the average number of the indistinguishable pairs is 0. This is because the order of the encrypted data is the same as that of the original data. Thus, POPIS is weak to the order matching attack. However, in our GOPES, the ratio of the correctly matched pairs to total number of pairs is reduced to 31.16%. The reason is that, by using the group transformation, our GOPES preserves the order of each data group, but not each data. In addition, GOPES shuffles the order of the encrypted data based on the function transformation. And the average number of the indistinguishable pairs is increased to 2.56. This is because, by using our segment transformation, GOPES sets the same encryption value for the original data in the same segment. As a result, GOPES outperforms the POPIS in terms of the data privacy against the order matching attack.

Table 1. Probability of leaking data by order matching attack

Encryption scheme	Ratio of the correctly matched pairs to total number of pairs	Average number of the indistinguishable pairs	Probability of exposing data from order matching attack
POPIS	100%	0	100%
GOPES	31.16%	2.56	12.18%

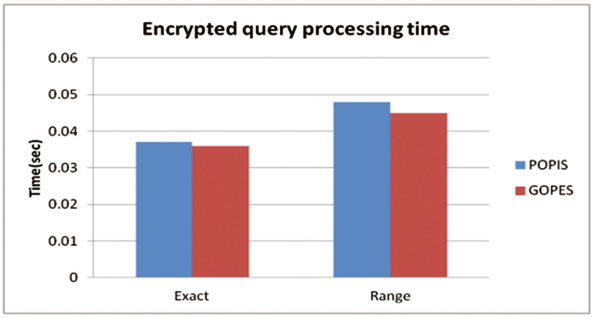


Fig. 2. Encrypted query processing time

3.2 Experimental Results in Terms of the Query Processing Time

Figure 2 shows an encrypted query processing time of both POPIS and GOPES. In case of the exact matching, the query processing time of POPIS and our GOPES are 0.037 and 0.036. In case of the range query, the query processing time of POPIS and our GOPES are 0.048 and 0.045. For two types of queries, GOPES shows slightly better performance than POPIS. This is because GOPES can reduce the computational cost by using bit operations.

4 Conclusions and Future Work

To protect the sensitive users' data, it is important to encrypt the database prior to outsourcing it to the service provider. However, with the most data encryption schemes, it is inefficient to process queries on the encrypted databases. To solve this problem, we propose a group order preserving data encryption scheme using the periodic function (GOPES). The proposed GOPES generates encryption signatures based on data groups and periodic functions. By using the group transformation and the segment transformation, GOPES is robust to both the order matching attacks and counting attacks. Through the performance results, we have shown that our scheme is much more efficient than the existing POPIS in terms of data privacy.

As a future work, we plan to apply our encryption scheme to the outsourced databases environment.

Acknowledgments. This work was partly supported by the Human Resource Training Program for Regional Innovation and Creativity through the Ministry of Education and National Research Foundation of Korea (NRF-2016H1C1A1065816). Also, this work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number 2016R1D1A3B03935298).

References

1. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Order preserving encryption for numeric data. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 563–574. ACM (2004)
2. Wong, W.K., Cheung, D.W.L., Kao, B., Mamoulis, N.: Secure kNN computation on encrypted databases. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, pp. 139–152. ACM (2009)
3. Boldyreva, A., Chenette, N., O'Neill, A.: Order-preserving encryption revisited: improved security analysis and alternative solutions. In: Annual Cryptology Conference, pp. 578–595. Springer, Heidelberg (2011)
4. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y: Order preserving encryption for numeric data. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 563–574. ACM (2004)

5. Liu, D., Wang, S.: Programmable order-preserving secure index for encrypted database query. In: 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 502–509. IEEE (2012)
6. Chung, S.S., Ozsoyoglu, G.: Anti-tamper databases: processing aggregate queries over encrypted databases. In: Proceedings of the 22nd International Conference on Data Engineering Workshops, p. 98. IEEE (2006)

Software Product Line Lifecycle Management-Integration Engineering and Management Process

Jeong Ah Kim[1] and Jin Seok Yang[2(✉)]

[1] Catholic Kwandong University, 579 BeonGil 24, GangNeung GangWonDo 25601, Korea
clara@cku.ac.kr
[2] SPID Consulting Inc., 93 St. 27, 4th Floor SunNeungRo, KangNam, Seoul, Korea
edeward@gmail.com

Abstract. In this paper, we describe the software product line process centered for small and first introduced company. Most software product line process focused on domain engineering and application engineering since variability management and product configuration technique are main differences from other software development methodologies. When the organization considers the adoption of new methodology, manager consider the whole lifecycle management. In this paper, we define whole lifecycle for software product line engineering and explain the experiences of real pilot project. Our suggested model can be used as software product line engineering transfer model.

Keywords: Software product line · Process · Lifecycle management · Adoption process · Technology transfer model

1 Introduction

In software product line engineering, many methodologies can improve both the productivity and quality of software. Typical examples are the Feature-Oriented Reuse Method (FORM) from the POSTECH software engineering lab [1, 2], Product Line Practice (PLP) from SEI [3], Family-Oriented Abstraction, Specification, and Transformation (FAST) from Lucent Technologies [4], Algebraic Hierarchical Equation for Application Design (AHEAD) from the University of Texas, Generative Programming (GP) from the University of Ilmenau [5], and ProdUct Line Software Engineering (PuLSE) from Fraunhofer IESE in Germany.

However, these companies are still reporting difficulties in applying software product line engineering to actual businesses. Nonetheless, a recent survey on corporate management and working-level executives of Korean companies reveals that many of them wish to apply software product line engineering and see that relevant training is promoted. A lack of technical knowledge and materials for software product line engineering is becoming a stumbling block for companies that wish to adopt it.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_82

Hence, resistance against adopting software product line engineering technology must be reduced. In order to do this, it is necessary to discover cases of successful adoption of software product line engineering, and to share them with the software companies. In particular, a methodology for supporting whole lifecycle is needed that defines not only the engineering methodology but also the decision methods for adoption and range, operation methods for organization, and the product line.

This suggests a work process classified into four phases: SPL adoption planning, domain engineering, application engineering, and SPL operation. The concept is easy to understand as an independent technical module because the process is modularized by the technologies used in the software development process. This paper also suggests new ways to configure the concept by using relevant technical modules according to various environments and methods.

This paper is organized as followings. Section 2 briefly explains the most popular SPL process. Section 3 defines the conceptual model to define the technology transfer strategy in SPL area. Technology transfer process is described in Sect. 4 and case study is written in Sect. 5 with contributions. Conclusion is described in Sect. 6.

2 Analyzing Current Status of SPL Process

Among the existing methodologies, FORM, PuLSE, and PLP are the most advanced methodology. Although actual usage is low, a feature-model-based ISO standard [7–9] has also been established. The feature model has been used as a de facto standard of the tool used to analyze commonality and variability. PuLSE can easily apply a method that extracts the assets that comprise a software product line from existing products through reengineering. This study aims to investigate FORM, PLP, and ISO standards, and to provide introductory strategies for companies that attempt to take proactive or extractive approaches.

FORM is a software reuse methodology developed by Professor Kang at the Gyo-Chul of Pohang University of Science and Technology in 1998. It is an expansion of the 1990 Feature-Oriented Domain Analysis (FODA) from architecture and the component design perspectives. FORM is a feature-oriented product line methodology. It is a systematic method in which commonality and variability of the product line are derived by the features to develop the architecture and components that can be reused in a product line. The range is determined by marketing and product planning. By using analysis results, it focuses on developing reusable product line assets.

SPL practice V5.0 by SEI requires that the product line include core asset development and product development using this core asset. These two activities must be done together with technical and organizational management. Which activity should start first is not important. If there is an existing product, extracting a normal asset from this product must take place first in order to fill up the storage for the core assets needed in the product line. If there is no product available, the core assets must be developed considering future reuse.

3 FORM-Based Technology Transfer Model

3.1 Conceptual Model

SPL technology transfer reference model (tech transfer model) collects the information needed for technology transfer from an organization and development environment of the receiving company. Then it selects a management process and engineering method for SPL technology that are most appropriate to the respective company, and processes the transfer. Through the technology transfer reference model, the entry barrier to new technologies for the companies that need this technology can be lowered. This reduces the effort and time required to absorb the technologies. Technical advisory organizations can also easily identify what types of technology are core ones and are necessary for the companies.

This paper establishes a technology transfer model based on the FORM methodology that effectively defines a feature-based proactive and extractive method. The FORM-based technology transfer model is a process with four phases: SPL adoption planning, domain engineering, application engineering, and SPL operation. This model is used by software analysts, architects, or developers who want to build reusable software assets on the basis of FORM methodology. In other words, the adoption planning phase and the operation phase have been combined into the SPL process and used in total.

3.2 Management Process Integrated Model

The project management part of the technology transfer model suggested in this research uses a Software Process (SP) model's [11] project management area and supporting area. By using an SP model, the quality of the software is controlled through the process management of the software's development. Based on the SP model, the project management and supporting work outputs that are to be applied to the respective company are defined.

3.2.1 Project Management and Supporting Area

Once the approach method is selected, the management processes of SPL asset development and operation are established, and detailed activities and works are defined. The SP model is used to determine the management process. There may be differences in the application levels of the SP model depending on the circumstances, but the quality management, configuration management, and requirement management must be included (Fig. 1).

3.2.2 Project Management and Supporting Work Outputs

Based on the output required in the SP model, the outputs appropriate to the company's characteristics can be defined. The outputs for the Work Breakdown Structure (WBS), work history, and issues need to be defined.

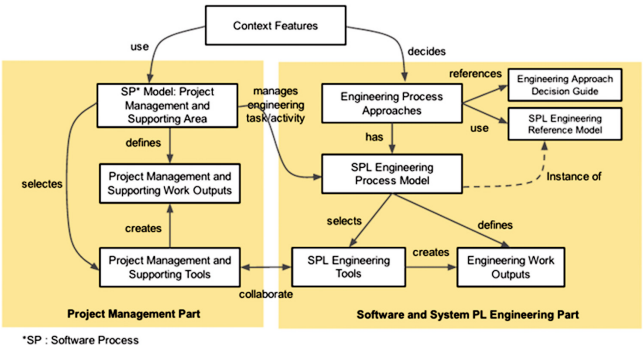


Fig. 1. Components and relationship between process of product, technology management, and engineering method [10]

3.2.3 Project Management and Supporting Tools

The tools that support project management can be chosen according to the items managed in the SP model, and the outputs can be correspondingly defined. The candidates for the tools that can be utilized for each item are as follows:

- Project management and control/requirement management: Redmine (Impasse)
- Quality management tools: PMD, CppCheck, NSIQ Collector, Hudson (Jenkins), JUnit, CppUnit, Emma, gcov
- Configuration management: CVS, SVN, Git.

3.3 Process Integration Model

The core part of the SPL engineering of the technology transfer model suggested by this paper allows the company to develop SPL-based software with the company characteristics in mind by referring to an SPL engineering reference model and the engineering approach decision guide for the approach. Then, a corporate SPL development model can be proposed.

3.3.1 SPL Engineering Process Approach

After the completion of a risk analysis or a product line evaluation, the phased development approach can be determined by considering the organizational structure, conditions, and domain characteristics. There are three types of approaches (Proactive Approach, Reactive Approach, Extractive Approach). All of them can be applied to every stage, or different approaches can also be applied in different phases:

3.3.2 SPL Engineering Reference Model

The SPL engineering reference model defines the engineering phases needed for software development using SPL technology at the highest abstract level. ISO/IEC

26550 is an international standard for this, and a further defined methodology can be seen in a development model defined in FORM. The SPL development process suggested in these reference models normally goes through the following steps: ① to ③ are the asset development process, and ④ to ⑥ belong to the product development process.

① Domain and PL requirement analysis
② Design of PL architecture
③ PL component development
④ Analysis and configuration of product requirements
⑤ Application of product architecture
⑥ Generation, combination, and adaptation of product component.

3.3.3 Engineering Approach Decision Guide

This provides guidelines for deciding the approach of SPL technology transfer for each engineering phase, and it is defined as the reference model for the development process. Through a number of questions suggested in the guide, the SPL technology transfer approach can be decided for each domain engineering stage. In every phase, the same approach can be applied, or several different approaches can also be mixed and used.

4 Technology Transfer Reference Model Based on Integration Process

To easily support SPL technology transfer, the integrated process integrates the management and the support processes with the development process. The integrated process suggested by this paper consists of (1) adaptation planning phase, (2) domain engineering phase, (3) application engineering phase, and (4) SPL operation phase.

The SPL adaptation planning stage analyzes the condition of the company, determines possible SPL applications, and chooses the right application method. The domain engineering phase is the stage where the assets that can be reused between the products within the product line are developed. During the application engineering phase, individual products are developed on the basis of the reusable assets obtained as a result of the domain engineering. The SPL operation phase forms an SPL operation group to enable close cooperation in the executing organization while considering the company's development atmosphere and characteristics.

Assets with high possibility for reuse are identified from the entire SPL development cycle, and the consistency and integrity between the assets and the products are maintained through asset tracking management. The following Table 1 lists the actions needed in each phase and their expected outcomes:

Table 1. FORM-based Technology transfer process

Phase	Task	Work products
SPL adoption planning phase	Adoption decision making	Risk analysis/evaluation report
		Candidates for product line
		Product family evaluation
		Interview/questionnaires
		Interview/questionnaires analysis report
		Feasibility report for SPL
		SPL adoption plan
Domain engineering	Domain analysis	Domain specification
		Feature model
		Domain requirement specification
		Feature-requirement traceability matrix
	Domain architecture	Conceptual architecture
		Process architecture
		Deployment architecture
		Architecture traceability matrix
		Feature-architecture traceability matrix
	Domain component development	Object design model
		Component specification
		Feature-component traceability matrix
Application engineering	Product requirement decision	Product requirement specification
	Product architecture decision	Product architecture decision table
	Product component development	Product architecture specification
		Product component
	Product configuration	Product software
		Product testing report
SPL operation	Organizational policy decision	Organizational architecture
	Asset and product management policy decision	Product line baseline
		Asset repository
		Product baseline
		Change management strategy

5 Case Study of SPL Technology Transfer Model

In order to understand the application effects and issues of the SPL technology-reference-model-based integration process, a case study was conducted on a domestic medium-sized enterprise. The company was Donga** Co., Ltd., which develops control systems for automobile air conditioners. As a second-level vendor in the automobile industry, Donga had issues with the quality and maintenance improvements for the automobile air-conditioner controlling software delivered according to various requirements of the customers.

The company has been delivering their air conditioner control systems to two large first-level vendors in the automobile industry. They also developed air conditioner control systems needed in various models for each vendor. These conditions were sufficient for Donga to establish product lines per vendor/control type. Although they obtained CMM-I level-2 certification, they were having difficulties in adopting SPL. One of the biggest reasons was not being able to find the right way to adopt the SPL process in preparation for level-3 certification. By utilizing the technology-reference-model-based integration process, the training, process development, and execution of a lead project were done over five months, and finally the company was able to establish its own SPL process.

In order to verify the effects of the technology transfer according to the suggested model, an evaluation using a BAPO model was conducted, and the ratings before and after the transfer were measured. Considering the duration of the pilot task, the evaluation of the organization was not completed. The result shows the business area improved by one grade to level 2, the architecture aspects also improved by one grade to level 2, and the process aspect reached the third level. However, the application engineering aspect reached only the second level because the maturity of the storage was still low. Although these three aspects showed improvement, the company was hesitant to adopt SPL. However, because of this maturity in their processes, Donga is now able to apply SPL more easily.

6 Conclusion

FORM is a representative methodology of SPL and exists in the form of multiple theses and methodological application guidelines. However, there is no phased application guide for the developers who are not familiar with the concept and model of SPL methodology. There is also no process establishment guide available for managers. A survey reveals that many companies expressed willingness to adopt software product line engineering technology. However, the lack of organizational standards and support for the technology transfer process needed to introduce a new technology are the stumbling blocks to adoption of SPL.

The SPL methodologies suggested to date can be classified into two phases: domain engineering and application engineering. Domain engineering is a process of developing reusable assets between the products in a product line that displays a set of products with similar characteristics. Application engineering refers to the activity of

developing individual products based on reusable assets obtained as a result of domain engineering.

This paper defines an integrated process that includes domain engineering, application engineering, all supporting activities needed for project planning and management, and SPL platform maintenance management. The suggested integration process becomes an important factor to encourage companies that are preparing for process quality certifications to adopt SPL.

A case study shows that a company with a CMM-I level-2 certification found SPL to be an effective way to prepare for level-3 certification. By adopting SPL, the competence of the organization improved. In other words, by enabling the organization's systematic software development and organizational asset management, productivity and quality can be improved to another level. Therefore, a company's adoption of SPL can be made easier by integrating the management of standards that the organization needs with the development process.

References

1. Kang, K.C. et al.: Feature-oriented domain analysis (FODA) feasibility study. Technical report (CMU/SEI-90-TR-21), Carnegi-Mellon University Software Engineering Institute (1990)
2. Kang, et al.: FORM: a feature-oriented reuse method with domain- specific reference architectures. Ann. Softw. Eng. **5**, 143–168 (1998)
3. SEI: A Framework for Software Product Line Practice, Version 5.0. http://www.sei.cmu.edu/productlines/frame_report/index.html
4. Ardis, M., Daley, N., Hoffman, D., Siy, H., Weiss, D.: Software product lines: a case study. Softw. Pract. Exp. **30**(7), 825–847 (2000)
5. Czarnecki, K., Eisenecker, U.W.: Generative Programming: Methods, Tools, and Applications. ACM Press/Addison-Wesley Publishing Co., New York (2000)
6. Bayer, J., et al.: PuLSE: a methodology to develop software product lines. In: Proceedings of the 1999 Symposium on Software Reusability (1999)
7. ISO/IEC 26551:2012: Software and systems engineering — reference model for product line engineering and management
8. ISO/IEC 26550:2013: Software and systems engineering — reference model for product line engineering and management
9. ISO/IEC 26555:2013: Software and systems engineering — tools and methods for product line technical management
10. Kim, J.A., Yang, J.S.: Strategy for improving the domestic supply of SPL. Technical report of NIPA (2013)
11. Cho, Y.-H., Lee, S.J.: Research of SW process (SP) quality certification for public SW development of effective project management. In: Proceedings of the Korea Project Management Conference, pp. 168–174 (2012)

OLSR Improvement with Link Live Time for FANETs

Yan Jiao, Wenyu Li, and Inwhee Joe(✉)

Mobile and Network Convergence Laboratory,
Department of Computer Science, Hanyang University,
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, South Korea
{kubohao007, fishl989, iwjoe}@hanyang.ac.kr

Abstract. In recent years, flying Ad-hoc networks (FANETs), which consist of small unmanned aerial vehicles (UAVs), is being used in the increasing of civilian and military applications. Due to the high mobility of the UAVs nodes, the link between the UAVs may frequently be disrupted. Hence, the existing routing protocols are inability to perform in FANETs. Motivated by this, we propose a new routing protocol named UAV-OLSR for FANETs in this paper. This protocol is based on the well-known protocol called optimized link state routing protocol (OLSR). We focus in our protocol on the lifetime of a communication link between the UAVs nodes and named link live time (LLT). We propose a new multipoint relay (MPR) selection algorithm where the UAVs node with maximum LLT is selected as the MPR. Our emulation results show that UAV-OLSR protocol outperforms OLSR in the packet loss rate, total time delay, the average time delay, and traffic received.

Keywords: FANET · Link live time · MPR · UAV-OLSR · OLSR

1 Introduction

Unmanned aerial vehicles (UAVs) or Drones are small aircraft equipped with the sensors, cameras and communication equipment, which can fly autonomously or can be operated remotely without carrying any human personnel. Because of their flexibility, versatility, easy installation and rather lower operation expense, they are increasingly used for civilian and military applications, such as traffic monitoring, managing wildfire, search and destroy operations, disaster monitoring and so on. Recently, due to the limitation performance of single UAVs operation, the conception of flying ad-hoc networks (FANETs) is proposed.

FANETs, is a group of UAVs communicate with each other with no need to access point, but at least one of them must be connected to a ground base or satellite [1]. By this definition, it can be classified as a special case of mobile ad-hoc network (MANETs). However, the topology of these networks is more dynamic than MANETs. According to [2], UAVs has a speed of 30–46 km/h, and this situation result in FANETs with high mobility character. Hence, the routing protocol design is one of the most challenging issues. Because, the existing conventional routing protocol of MANETs is not suited for such high mobility.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_83

In recent years, the specific routing protocols for FANETs are proposed in some literature [3–8]. In a certain aspect, they focused on the UAVs' high mobility for protocol design. Nevertheless, they ignored a communication link may be disrupted owing to UAVs' move. It is pointless to establish a route for two nodes without available communication link. In addition, as demonstrated in [8], optimized link state routing protocol (OLSR) performance outperforms Ad-hoc On-demand Distance Vector (ADOV) and Dynamic Source Routing (DSR) in FANETs. For this reason, we propose a new protocol named UAV-OLSR is based on OLSR protocol. We focus on the lifetime of communication link between two UAVs, then propose a measure metric called link live time (LLT) for the lifetime of the communication link. The highlighting function of the proposed protocol is the MPR selection algorithm where a UAVs node with the maximum LLT is selected as the MPR. The best of our knowledge, it is the first time to introduce the LLT in routing protocol design for FANETs.

The rest of the paper is organized as follows: In Sect. 2, we describe the FANETs model and present the functional architecture of the UAV-OLSR protocol. Then, we specify the modification of the *Hello* messages and *TC* (topology control) messages, demonstrate the mathematical foundation of LLT and present LLT based MPR selection algorithm in Sect. 3. In Sect. 4, we present the NS3 simulation results. Finally, we summarized the main conclusion and future work in Sect. 5.

2 System Model and Functional Architecture of UAV-OSLR

In this paper, we assume the a case of FANETs where each UAVs node can act as a router are grouped with no central infrastructure and 80211a/b/g access point act as connecting to ground base [1]. The MPR UAVs is a subset of selected UAVs used to perform routing advertisement. That is, MPR UAVs act as forwarders to transport the data from the source UAVs to the destination UAVs.

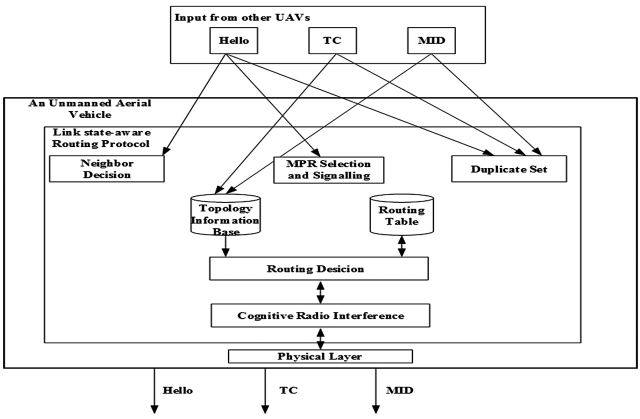


Fig. 1. The functional structure of UAV-OLSR

There are two main messages of UAV-OLSR protocol: *Hello* message and *TC* message discover and disseminate routing UAVs node information. Figure 1 depicts the functional structure of UAV-OLSR.

3 The Proposed UAV-OLSR Protocol

3.1 The Modified *Hello* Message and *TC* Message

Figures 2 and 3 illustrate the structure of the modified *Hello* message and *TC* message respectively. Comparing with the original *Hello* message, modified *Hello* message additionally contains UAVs node 3D position information (includes latitude, longitude and altitude), UAVs node movement speed, link live time. The 3D position of a UAV node can be obtained accurately via GPS and movement speed can be known by itself. The role of a *TC* message is that transfers topology control information as soon as possible. Hence, it must know the all LLTs of neighbors. Hence, *Link Live Time Vector* is designed for saving LLTs.

Fig. 2. Format of modified *Hello* message

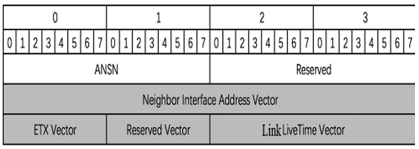


Fig. 3. Format of modified *TC* message

3.2 LLT Based MPR's Selection

We assume $l_{ij}(i \neq j)$ is a communication link between node N_i (represent UAV i) and node N_j (represent UAV j). The concept of LLT is defined as below

Definition 1. When a node N_i moves out of the communication range of a node N_j in an interval I_{ij}, such interval I_{ij} is named the LLT of l_{ij}.

We assume N_i receives the *Hello* messages from N_j located in its communication rage at t_1 and t_2 respectively. L_1^j and L_2^j denote the location information of N_j at t_1 and t_2. The average speed of N_j and N_i is expressed in Eq. (1), respectively

$$V_j = \frac{\left|L_2^j - L_1^j\right|}{|t_2 - t_1|} \quad V_i = \frac{\left|L_2^i - L_1^i\right|}{|t_2 - t_1|} \tag{1}$$

We denote V_{ij} represent N_j relative speed to N_i, and V_{ij} can be expressed as Eq. (2)

$$V_{ij} = V_i - V_j \tag{2}$$

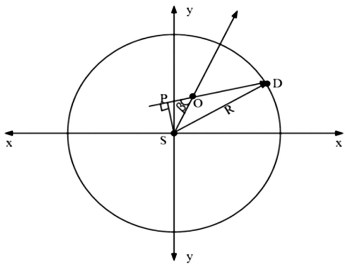


Fig. 4. The geometry of link l_{ij}

The geometry of link l_{ij} is shown in Fig. 4, where R denotes maximum transmission radius of the node, O denotes N_j location at t_2, S denotes N_i location at t_2. Therefore, LLT of a link l_{ij}, t_{ij} is expressed as Eq. (3)

$$t_{ij} = \frac{|OD|}{V_{ij}} \tag{3}$$

According to trigonometric relations, $|OD|$ can be expressed as Eq. (4)

$$|OD| = \sqrt{|OD|^2 - (|SO| \cdot \sin\beta)^2} - |SO| \cdot \cos\beta \tag{4}$$

Combining (3), we can obtain t_{ij} as expressed as Eq. (5)

$$t_{ij} = \frac{\sqrt{|OD|^2 - (|SO| \cdot \sin\beta)^2} - |SO| \cdot \cos\beta}{V_{ij}} \tag{5}$$

Hence, the LLT based MPR selection algorithm as described as follows
Input: Neighbor information through Hello and TC message
Output: Select a neighbor node with maximum LLT as the MPR nodes
Steps:

1. For $j = 1$ to $n, j \neq i$,
2. Obtain location Coordinates of neighbor node
3. Let transmission range be R
4. Calculate distance from current location to D which the edge of transmission range, according to trigonometric relations, $|OD| = \sqrt{|OD|^2 - (|SO| \cdot \sin\beta)^2} - |SO| \cdot \cos\beta$

5. Let V_i, V_j are the average speed of the node and neighbor
6. Calculate LLT between the node N_i and its neighbor

$$t_{ij} = \frac{\sqrt{|OD|^2 - (|SO| \cdot \sin\beta)^2} - |SO| \cdot \cos\beta}{V_{ij}}$$

7. LLT set holds the t_{ij}
8. End Loop
9. Select a node N_j with maximum t_{ij} as the MPR of the node N_i
10. MPR set holds the selected the MPR node N_j for the node N_i

4 Simulation

To demonstrate the performance of the UAV-OLSR protocol present above, we compared it to the original OLSR protocol. For our simulation, we use NS3, which is a discrete-event network simulator includes a rich set of detailed models for ad-hoc networks. There are 20 UAVs are placed in a 1000 × 1000 × 250 m area, data rate is 2048 Mbps, the power transmission is 4 mw. The simulation period is 1200 s and the UAVs are moving according to Gauss-Markov mobility model [9] with a random speed of (10–30) m/s. All UAVs are configured to run the OLSR protocol in the first scenario and then UAV-OLSR protocol in the second scenario. We compare UAV-OLSR and OLSR protocols.

Figure 5 compares the two protocols in terms of packet loss rate for the different packet size. As shown in the figure, our proposed protocol gave better packet loss rate (PLR) than the original OLSR. Especially, the PLR of UAV-OLSR is significantly better than OLSR with transmitted packet size increase. Since the MPRs is a set of nodes with maximum LLT. Thereby, UAV-OLSR is most of avoiding packet loss owing to the communication link invalid by node movement.

Figure 6 shows the performance comparison results for total time delay (TTD) between two protocols. The figure shows that UAV-OLSR has less TTD than the original OLSR. The TTD for both protocols is high at the beginning of the simulation. Since the control traffic is high before the MPR set selection at the beginning. Namely, the large number of nodes used for flooding the control messages in the initial simulation. When a certain number of MPRs are selected, the control message transmission will be restricted only in MPRs. Thus, the TTD also be reduced.

We introduce the average time delay (ATD), which is defined as the time of per unit packet transmission, comparison results are shown as Fig. 7. It reflects a fact that the size of the transmitted packet increase needs more transmission time. However, the UAV-OLSR needs less time than OLSR when transmitted packets are the same size.

Figure 8 compares the traffic received using the UAV-OLSR and OLSR. It can be seen that UAV-OLSR precedes OLSR in all simulation time. The reason is that

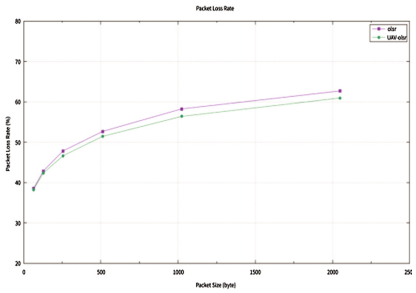


Fig. 5. Comparison between UAV-OLSR and OLSR protocol for packet loss rate.

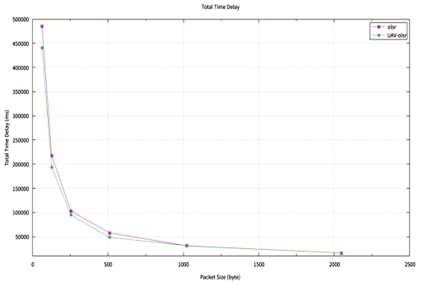


Fig. 6. Comparison between UAV-OLSR and OLSR protocol for total time delay.

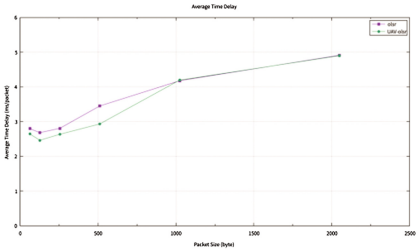


Fig. 7. Comparison between UAV-OLSR and OLSR protocol for average time delay.

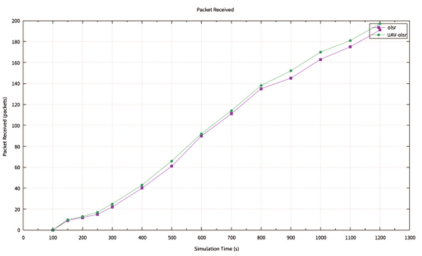


Fig. 8. Comparison between UAV-OLSR and OLSR protocol for traffic recevied.

UAV-OLSR selects the nodes with maximum LLT as MPRs, consequently the number of lost packets by the communication failure is minimized. There is no such function in OSLR protocol.

5 Conclusion and Future Works

In this paper, we proposed a link live time based Optimized Link State Routing named UAV-OLSR for FANETs. Our protocol is capable of avoiding the OLSR protocol performance will be under the influence of the communication links between the two nodes failed owing to UAVs high mobility. Furthermore, we compared the performance of UAV-OLSR and the OLSR for packet loss rate, end-to-end time delay, the average time delay and throughput using NS3 simulator. The simulation results show that UAV-OLSR achieves better performance than OLSR in four aspects: packet loss rate, total time delay, the average time delay and traffic received. Since the UAV-OLSR performance avoids the worst communication situation where the communication link is disrupted. However, it needs to continue to improve the performance of traffic received in UAV-OLSR. We will focus on this problem in future works.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT & Future Planning) (No. 2016R1A2B4013118).

References

1. Bekmezci, I., Sahingoz, O.K., Temel, S.: Flying ad-hoc networks (FAMETs): a survey. Ad Hoc Netw. **11**, 1254–1270 (2012). http://dx.doi.org/10.1016/j.adhoc.2012.12.004
2. Clapper, J., Young, J., Cartwright, J., Grimes, J.: Unmanned systems roadmap 2007–2032. Technical report, Department of Defense (2007)
3. Forsmann, J.H., Hiromoto, R.E., Svoboda, J.: A time-slotted on-demand routing protocol for mobile ad hoc unmanned vehicle systems. In: SPIE 6561 (2007)
4. Shirani, R., et al.: The performance of greedy geographic forwarding in unmanned aeronautical ad-hoc networks. In: 2011 Ninth Annual Communication Networks and Services Research Conference, CNSR 2011, pp. 161–166. IEEE Press, Canada (2011)
5. Alshabtat, A.I., Dong, L.: Low latency routing algorithm for unmanned aerial vehicles ad-hoc networks. Int. J. Electr. Comput. Eng. **5**, 1–7 (2011)
6. Lin, L., Sun, Q., Li, J., Yang, F.: A novel geographic position mobility oriented routing strategy for UAVs. J. Comput. Inf. Syst. **8**, 709–716 (2012)
7. Tareque, M.H., Hossain, M.S., Atiquzzaman, M.: On the routing in flying ad hoc netoworks. In: Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 13–16. IEEE Press, Poland (2015)
8. Rosati, S., et al.: Dynamic routing for flying ad hoc networks. IEEE Trans. Veh. Technol. **65**, 1690–1700 (2016)
9. Biomo, J.-D.M.M., Kunz, T., St-Hiaire, M., Zhou, Y.: Unmanned aerial *ad hoc* networks: simulation-based evaluation of entity mobility models' impact on routing performance. Aerospace **2**, 392–422 (2015). doi:10.3390/arospeace20303927

An Efficient Partition-Based Filtering for Similarity Joins on MapReduce Framework

Miyoung Jang, Archana B. Lokhande, Naeun Baek, and Jae-Woo Chang(✉)

Department of Computer Engineering, Chonbuk National University, 567 Baekjae-daero, Jeonju, Republic of Korea
{brilliant, archana, nannanil00, jwchang}@jbnu.ac.kr

Abstract. Similarity join is an important operation in MapReduce framework to find pairs of similar objects like images, video and time series. Since MapReduce basics do not support efficient join processing, the duplicate reduction of candidates and load-balancing among partitions are the major challenges. Recently, many partition based similarity join algorithms have been proposed to solve such problems. However, the existing algorithms still have limitations for supporting efficient join processing over large-scale data set. In this paper, we proposed a similarity join algorithm with an efficient filtering technique on MapReduce to overcome the limitations of traditional partitioning method in two ways: (1) the number of outputs records generated by the filtering matrix reduces duplicates and (2) the estimated join cost generated by using a partition matrix leads to a better load-balance among reducers. Moreover, we have conducted experimental evaluations using sequential data to show the speed-up and scale-up of proposed method.

Keywords: MapReduce-based similarity join algorithm · Join matrix · Load balancing · Parallel join processing

1 Introduction and Background

Recently, the amount of data is rapidly increasing with the popularity of the Social Networking Services (SNS) and the development of mobile technology. The increasing volume of high dimensional data triggers new challenges towards how to efficiently analyze the big data. Analytical join queries become important due to their applicability for decision making applications. For such data-intensive applications, the MapReduce [1] has attracted much interest as a new paradigm of data processing framework. The MapReduce introduced by Google is used to perform large-scale data processing in a distributed manner. The MapReduce adopts a flexible computation model with a simple interface consisting of map and reduce functions, which can be customized by application developers. The MapReduce has been proved to be a suitable, error-tolerant framework for parallel join algorithms. However, the MapReduce does not consider the join operation of multiple datasets, which is an important operation in the advanced data analysis. Because the size of data sets used for join processing are huge, efficient

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_84

analytical query processing algorithms in MapReduce are vital to provide a fast response time.

Among the analytical join queries, similarity joins have been considered as key operations in many data analysis tasks [2–4]. The goal of similarity join is to find all pairs of records that have scores greater than a predefined similarity threshold (θ) under a given similarity function (or distances does not exceed a given threshold). There are two key challenges to be addressed in order to process the similarity join on large-scale data. First, data should be partitioned and processed in a distributed manner. Hence, workload-aware data partitioning techniques are required, which ensure the balance of not only the input data but also the output of each machine. Second, a sophisticated design of a filtering technique is required because the number of comparisons is drastically growing as data size and dimension increase.

Recently, Sarma et el. [5] proposed the ClusterJoin to perform a similarity join of large-scale data sets on MapReduce. The ClusterJoin differentiate itself by proposing a general framework to compute metric distances, whereas most of the existing similarity join algorithms focus on string similarity measures. For this, they proposed a filter that can prune away candidate pairs by using a bisector-based data clustering. Furthermore, the authors proposed a dynamic load balancing scheme that is adaptive to data distribution and skewness, by using a hash function. However, because of the characteristics of the hash function, there is a high possibility that the data skewness is not fully solved when the density is high. Also, the bisector-based clustering algorithm using randomly sampled data may cause high data duplication among clusters.

To resolve these problems, we propose a new similarity join algorithm on MapReduce framework. First, we propose a join data filtering technique by using a join matrix. This can reduce data duplication among join partitions. Second, we propose a load balancing method that is based on the join matrix with an estimated join costs.

The rest of this paper is organized as follows. In Sect. 2, we present our partition-method for a join matrix and propose a new similarity join algorithm using the matrix Sect. 3 is devoted to experimental results. Finally, we conclude our work with future direction in Sect. 4.

2 Similarity Join Algorithm with a New Filtering Scheme on MapReduce

To solve the problems of the existing similarity join algorithms, we propose a new similarity join algorithm to evenly distribute data into clusters and to perform a join in a parallel way. To this end, we first design a filter matrix that efficiently estimates join costs of tuples. This can reduce unnecessary data computation and communication costs by sending only similar data to the same reducer when perform a join. Secondly, we propose a load-balancing method that evenly distributes data to MapReduce jobs by considering data distribution.

2.1 Preprocessing Phase: Join Matrix Generation

In our similarity join algorithm, the join Matrix is utilized to estimate the size of each join groups and to evenly distribute the corresponding join groups for each Map or Reducer. For two dataset S and T, where the $S = \{s_1, s_2, s_3, \ldots, s_N\}$ and $T = \{t_1, t_2, t_3, \ldots, t_M\}$, the row ID is assigned as the ascending order of the values in S data. It also stores $C_s(\alpha)$, which is the cardinality of data with value α. This is also applied to the column with T data set. And, for the i-th row and j-th column, a cell M(i, j) stores the cross product of the SxT, which indicates the estimated join cost. Equation 1 explains how to compute the join cost for a cell M(i, j) in our join matrix.

$$C_s(\alpha) = \{cardinaltiy(S_x) | S_x.value = \alpha,\ S_x \in S\}$$
$$C_T(\beta) = \{cardinaltiy(T_x) | T_x.value = \beta,\ T_x \in T\}$$
$$\text{For all } i > 1, j > 1,$$

A matrix cell M(i, j) is computed as

$$C_s(\alpha) \times C_T(\beta) \tag{1}$$

Matrix generation is performed through single MapReduce phase. In the Map phase, all data tuples are loaded and transformed in the form of <(origin, value), 1>. In the reduce phase, the data is assigned to the reducers based on the key. As a result, the cardinality of data based on the value is computed and stored in HDFS.

Since we have k reducers, the optimal workload ratio for each reducer is 1/k. However, it is notable that the resulting matrix m has huge workloads and many duplicates for estimation costs. Therefore, to balance the workload among k reducers, we further process the filtering of m with similarity join threshold θ before send it to the similarity join MapReduce phase.

2.2 Similarity Join Phase

To optimize the load-balancing among reducers, we proposed a filtering technique on the join matrix. The filtering algorithm (Fig. 1) is performed as follows. For the given similarity threshold θ, the matrix cell M(i, j) is set to TRUE if the i-th tuple value from R and j-th tuple value from S satisfy the join condition (lines 1–4), and FALSE otherwise (line 5). Finally, the algorithm returns the filtered matrix m_f whose elements are candidates for the similarity join.

In the second MapReduce phase, a join processing is actually performed by using the result of our partition matrix. In the Map phase, all the data are transformed in the form of <partition ID, data> and are sent to the Reducers. The corresponding data with the same partition ID is sent to the same reducer to perform a similarity join. The result of Reduce phase is then returned to a user as a final similarity join result.

The running example of our similarity join algorithm is described in Fig. 2. Given two input sets R and S, and the user defined similarity threshold t, the algorithm starts partitioning of the join datasets (Fig. 2(a, b)). In the pre-processing phase, each S and T data is mapped in the form of <value, (origin, ID)> and then merged by the value to

```
Algorithm 1.Filtering
Input    : Matrix m, similarity threshold(θ)
Output : Filtered Matrix m_f
/*Find join candidates on Matrix m*/
1:   For each s_i = m.getS()
2:      Each t_j= m.getT()
3:      If (|s_i − t_j| ≤ θ) Then
4:         m_f.m(s_i,t_j) = TRUE
5:      Else m_f.m(s_i,t_j) = FALSE
6:   Return m_f
```

Fig. 1. Filtering algorithm

(A) S data map

Input tuple	Output Tuple
S_1 A=2	(S1 2)
S_2A=3	(S2 3)
S_3A=6	(S3 6)
S_4A=5	(S4 5)
S_5A=7	(S5 7)
S_6A=4	(S6 4)
S_7A=5	(S7 5)
S_8A=3	(S8 3)
S_9A=2	(S9 2)
S_{10}A=1	(S10 1)
-	-
-	-
-	-

(B) S data map

Input tuple	output tuple
T_1A=2	(T1 2)
T_2A=4	(T 2 4)
T_3A=7	(T3 7)
T_4A=3	(T4 3)
T_5A=2	(T 5 2)
T_6A=2	(T 6 2)
T_7A=1	(T 7 2)
T_8A=3	(T 8 3)
T_9A=6	(T 9 6)
T_{10}A=4	(T10 4)
-	-
-	-
-	-

(C) Merged info

S	
1	2
2	4
3	7
4	3
5	2
6	2
7	1
8	3
9	6
10	4
T	
1	2
2	3
3	6
4	5
5	7
6	4
7	5
8	3
9	2
10	1

(D) Partition info

Value	Partition ID
S.A = 1	P(1)
S.A = 2	P(1)
S.A = 3	P(1, 2)
S.A = 4	P(1, 2)
S.A = 5	P(2)
S.A = 6	P(2, 3)
S.A = 7	P(2, 3)
S.A = 8	P(3)
S.A = 9	P(3)
S.A = 10	P(3)
T.A = 1	P(1)
T.A =2	P(1)
T.A =3	P(1)
T.A =4	P(2)
T.A =5	P(2)
T.A =6	P(2)
T.A =7	P(3)
T.A =8	P(3)
T.A =9	P(3)
T.A =10	P(3)

(E) Reduce

Reducer 1	
Input tuple	(S1 2), (S2 3), (S8 3), … (T1 2), (T4 3), (T5 2), …
Output tuple	(S1, T1), (S2, T1), (S8, T1), …

Reducer 2	
Input tuple	(S2 3), (S3 6), (S4 5), … (T2 4), (T9 6), (T12 5), …
Output tuple	(S2, T2), (S3, T9), (S4, T9), …

Reducer 3	
Input tuple	(S3 6), (S5 7), (S11 9), … (T3 7), (T13 8), (T17 10), …
Output tuple	(S3, T3), (S5, T3), (S11, T13), …

Fig. 2. An example of join processing algorithm

analyze data histogram (Fig. 2(c)). This result is used to generate a join matrix, which will generate a join partition based on the join candidates. We divide the data into 3 join partitions and generate partition info table (Fig. 2(d)). In the join processing phase, the algorithm read dataset S and T, and map the data based on the join partition information. By this means, all join candidates are assigned to the same reducer. Finally, each reducer compares its input tuples and finds pairs of S and T whose distance is less than the similarity threshold t (Fig. 2(e)). The merged join result is finally returned to the user.

3 Performance Evaluation

We conduct experiments on 4-machine clusters with Hadoop 0.20.2. Each machine has 2.9 GHz Quad-core processor, 4 GB RAM, OS Ubuntu 12.4. Table 1 shows the parameter settings for the experiments. For the dataset, we use a synthetic data generated by using Generate Spatio Temproal Data (GSTD) [6].

Table 1. Experimental parameters

Parameters	Range (default in bold)
Number of data	$5 \times 10^3, 10 \times 10^3, 20 \times 10^3, 30 \times 10^3, 40 \times 10^3, 50 \times 10^3$
Similarity threshold (theta)	0.02, ***0.04***, 0.06, 0.08, 0.1

Figure 3 shows the query processing time of the similarity join algorithms with varying data size. When the number of tuples increases, the query processing time of ClusterJoin is drastically increased, while our algorithm remains steady. When the data size is 50,000, our proposed algorithm required about 820 s to process data, while ClusterJoin took almost 11,000. This reveals our proposed algorithm achieves significant improvement compared to the ClusterJoin algorithm. The main reason is that our join matrix can greatly reduce the number of duplicated join candidates for each partition. Whereas the hash-based clustering does not guarantee the optimal data distribution among clusters.

To evaluate the effectiveness of join algorithms, we perform the join with different theta value, i.e. 0.02 to 0.1 with 0.02 interval. As shown in Fig. 4, the execution time of ClusterJoin is significantly increased with the increasing theta value whereas our algorithm shows moderate performance. When the similarity threshold is 0.02, our proposed algorithm requires 80 s to execute join operation, whereas ClusterJoin required 148 s. This is because in the map phase our powerful filter is capable to eliminate more duplicates, than ClusterJoin. As the similarity threshold increases, there is high possibility that the number of duplicated data among clusters is greatly increased. In case of ClusterJoin, this trend leads to a radical performance deterioration because the hash-based partitioning shows worse performance with densely populated data. On the other hand, our filtering technique using the join matrix determines the

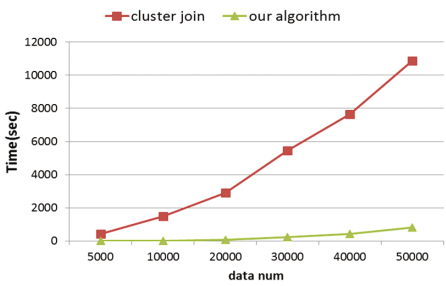


Fig. 3. Performance with varying data size

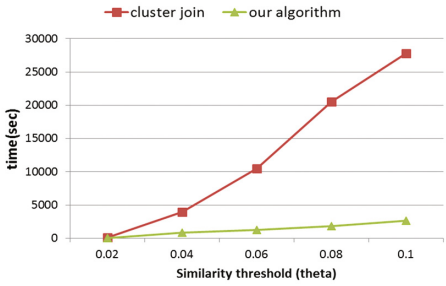


Fig. 4. Performance with varying theta value

join partitions where the duplicated data is minimized, which can greatly improve overall processing performance.

4 Conclusion

Handling load balance and pruning duplicates are essential for efficient join algorithms in MapReduce. In this paper, a similarity join algorithm is successfully executed using Matrix-based filtering schemes in the MapReduce framework. Our algorithm efficiently finds the similarity join candidates by using a join Matrix. Using the matrix we can successfully reduce the duplicates and efficiently estimate the input-output load for the reducers. Hence, only datasets whose values exceed the similarity threshold are sent to the MapReduce, thus minimized the similarity join computation. The performance of our algorithm is evaluated and the results confirm that our proposed algorithm outperforms the existing ClusterJoin algorithm.

My future work aim to extend the proposed similarity join algorithm to support join on multi-dimensional data using MapReduce framework.

Acknowledgments. This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. R0113-16-0005, Development of a Unified Data Engineering Technology for Largescale Transaction Processing and Realtime Complex Analytics). Also, 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-2016H1C1A1065816).

References

1. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI), pp. 1–13 (2004)
2. Blanas, S., Patel, J.M., Ercegovac, V., Rao, J.: A comparison of join algorithms for log processing in MapReduce. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, pp. 975–986. ACM (2010)
3. Metwally, A., Faloutsos, C.: V-smart-join: a scalable MapReduce framework for all-pair similarity joins of multisets and vectors. Proc. VLDB Endowment **5**(8), 704–715 (2012)
4. Okcan, A., Riedewald, M.: Processing theta-joins using MapReduce, In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pp. 949–960. ACM (2011)
5. Sharma, A.D., He, Y., Chaudhuri, S.: Clusterjoin: a similarity joins framework using map-reduce. Proc. VLDB Endowment **7**(12), 1059–1070 (2014)
6. Theodoridis, Y., Silva, J.R.O., Nascimento, M.A.: On the generation of spatiotemporal datasets. In: Proceedings of SSTD, vol. 1651, pp. 147–164 (1999)

A Study of Algorithm Replacement Mechanism for Environment Adaptive Load Balancer

JongMin Lee[1] and Young B. Park[2(✉)]

[1] Department of Computer Engineering, Dankook University Juk-jeon, Yongin, Republic of Korea
jongmin07ll@daum.net
[2] Department of Computer Science and Engineering, Dankook University Juk-jeon, Yongin, Republic of Korea
ybpark@dankook.ac.kr

Abstract. Recently the use of computing technologies for distribution systems such as cloud or grid is increasing for effective management of IT operation and investment cost reduction. One of the important tasks in distributed systems is to prevent the server from overloading. Although a load balancer has been introduced to solve this problem, there is a limitation that an existing load balancer must use only one initially set algorithm. To solve these limitations, this paper proposes a load balancing method that dynamically replaces the load balancer algorithm according to the server status and environment. The proposed method is that the load balancer collects the server status information in real time, analyzes the result of the server status, and replaces it with an appropriate algorithm if necessary. We experimented with several algorithms and proved that this proposed load balancing is more effective than the existing load balancing.

Keywords: Adaptive · Dynamic · Load balancing

1 Introduction

In recent years, the development of information technology resulted in constructing distributed system environments for handling large-scale data. In a distributed system environment, when a large amount of data is processed, traffic is concentrated on one server, which may cause a problem of performance degradation. In a distributed system environment, performance degradation is solved by distributing traffic through a load balancer [1, 2]. The load balancer improves server performance by appropriately distributing or reallocating tasks in consideration of system resources and load conditions. The load-balancing algorithm used at this time varies depending on the monitored elements. These algorithms are divided into static load balancing and dynamic load balancing. The difference between the two type of algorithm is whether they consider the status of the server in real time or not [3, 4]. If appropriate algorithms are applied according to the situation, the minimum response time and usage rate for nodes can be improved. But general load balancers have a limitation that the initial algorithm, once set, can not be replace unless the user replace it. So, those existing load balancers have

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_85

a problem that it can not cope with various node conditions caused by environment change. In this paper, we propose an adaptive load balancing mechanism, which automatically changes the algorithms used in the static and dynamic load balancing methods according to the status and environment of the system. By using the proposed method, the load balancer can flexibly cope with various loads while minimizing overhead occurrence.

The rest of this paper is organized as follow. We discuss background of load balancing in Sect. 2. Then Sect. 3 describes how the proposed load balancer works and how it is implemented. Section 4 shows and analyzes the results of the proposed load balancing experiment. Finally, Sect. 5 shows the conclusion.

2 Background

Load balancing allow the workload to be distributed evenly over before task is unbalanced and a node is overloaded [5]. To do load balancing there are load balancing algorithms suitable for required the monitoring data and node environment. For more reliable load recovery, the load balancing algorithm requests more monitoring information. Also, the scope of the load balancing algorithm varies widely [6]. However, there is a limitation in solving various load condition in one algorithm. The solution is adaptive load balancing. Adaptive load balancing use some policy that is used for load balancing based on load conditions. In order to resolve the load conditions adaptive load balancing replaces policies such as information policy, the trigger policy, transfer policy and location policy [7]. JaegeolPark [8] studied the adaptive load balancing method based on the RFID middleware standard architecture. A. Omar Portillo-Dominguez [9] studied the adaptive load balancing strategy based on garbage collection. However, the adaptive load balancing also requires many data and increases the complicity of the implementation. In this paper, we propose adaptive load balancing mechanism to resolve the problem. The mechanism can load balancing by changing appropriate load balancing algorithm.

3 Suggested Adaptive Load Balancer

This chapter introduces the configuration and operation of the load balancer applying the proposed adaptive load balancing mechanism. Each server has an agent to provide the necessary information to the load balancing. The structure of the proposed adaptive load balancing mechanism consists of Monitor, Datacenter, and AlgorithmSelector. Monitor collects server status data from servers. The data includes physical specifications as well as the server system, process information. The data is stored in Server info database. Some data stored in the database is used to gather statistics of the collected information. Also, some data of server status information is grouped according to the elements required by the pre-defined rule. The grouped data will be used whether or not to use a dynamic load balancing algorithm. And the data stored in the Server info data base judge whether or not to change with other load balancing algorithm. The DataCenter analyzes grouped data. Also, DataCenter is statistics on the

data accumulated in real time from the server information database are provided. The analysis data and statistic data judge whether to change with other load balancing algorithm. Elements grouped by the monitor are analyzed. Also statistics on the data. AlgorithmSelector determines whether the selected load balancing algorithm is an optimal. A.S receives the DataCenter results and the information about server specifications from the Server info database. Then, A.S matches result of DataCenter and Pre-Define Rule of the DataCenter to determine the optimal algorithm. Once an optimal algorithm is determined, the AlgorithmSelector informs the LB of the information about the selected algorithm. The following Fig. 1 shows the structure of the proposed adaptive load balancer.

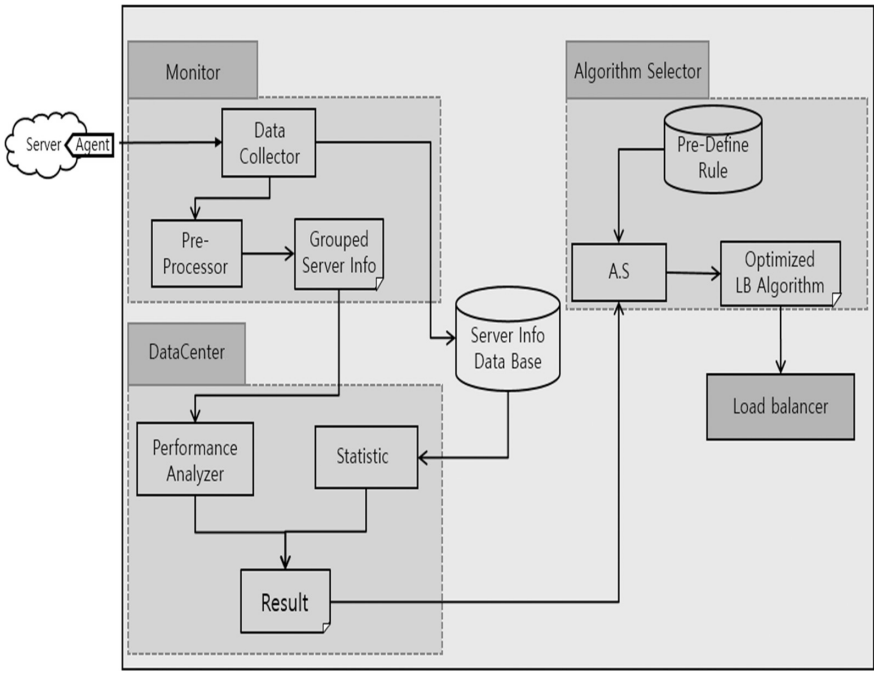


Fig. 1. Adaptive load balancer

4 Experiment

We conducted a load test to evaluate the performance of the proposed adaptive load balancing mechanism. The experiment consisted of a basic load balancer, an adaptive load balancer, and three Apache servers. The experiment is to measure the transaction and response time of the servers when using the existing load balancer and the proposed load balancer. The load balancing algorithms used for the experiments are round robin algorithm, least connection algorithm and weighting algorithm. Though there is a slight difference in the transactions between Tables 1 and 2, when you compare the

Table 1. Load balancer

Parameter name	Parameter value		
	Server1	Server2	Server3
Transection (s)	*1681.54*	*1587.64*	*1583.52*
Response time (ms)	0.51	0.55	0.53

Table 2. Adaptive load balancer

Parameter name	Parameter value		
	Server1	Server2	Server3
Transection (s)	*16543*	*1590.64*	*1589.53*
Response time (ms)	0.49	0.54	0.53

response times, you can see that the performance has improved. This indicates that the proposed load balancing scheme can not only improve the performance but also respond optimally to various load environments by replacing algorithms.

5 Conclusion

In this paper, we proposed a method to deal with server load, with the purpose that the load balancer replaces the algorithm with an appropriate one by considering the status of the server. The existing load balancer has a limitation that it can not cope with various node conditions of environment change because it can use only the initial set algorithm. However, the proposed load balancer can cope with various environment changes by applying various algorithms, and the test results show that the proposed load balancing algorithm is effective. With the proposed adaptive load balancer, we could not apply a little bit more of various algorithms. In the future, it is expected that a load balancer can be more efficient if it can predict the possible load scenarios and bring the algorithms necessary for load balancing from other servers.

Acknowledgments. This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the SW master's course of a hiring contract program (H0116-16-1015) supervised by the IITP (Institute for Information & communications Technology Promotion). This research was supported by NextGeneration Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2012M3C4A7033348). This research was supported by The Leading Human Resource Training Program of Regional Neo industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (No. NRF-2016H1D5A1909989).

References

1. Deshmukh, S.C., Deshmukh, S.S.: A survey: load balancing for distributed file system. Int. J. Comput. Appl. **111**, 0975–8887 (2015)
2. Kahanwal, B., Singh, T.P.: The distributed computing paradigms P2P, grid, cluster, cloud, and jungle. Int. J. Latest Res. Sci. Technol. **1**, 183–187 (2012). arXiv preprint
3. Al Nuaimi, K., et al.: A survey of load balancing in cloud computing: Challenges and algorithms. In: 2012 Second Symposium on Network Cloud Computing and Applications (NCCA), pp. 137–142. IEEE (2012)
4. Alakeel, A.M.: A guide to dynamic load balancing in distributed computer systems. Int. J. Comput. Sci. Inf. Secur. **10**(6), 153–160 (2010)
5. An, I.: Scalable, adaptive load sharing for distributed systems. IEEE Parallel Distrib. Technol. **1**(3), 62–70 (1993)
6. Mesbahi, M., Rahmani, A.M.: Load balancing in cloud computing: a state of the art survey. Int. J. Mod. Educ. Comput. Sci. **8**(3), 64–78 (2016)
7. Shivaratri, N.G., Krueger, P., Singhal, M.: Load distributing for locally distributed systems. Computer **25**(12), 33–44 (1992)
8. Park, J.-G., Chae, H.-S.: An adaptive load balancing method for RFID middlewares based on the Standard Architecture. KIPS Trans. Part D **15**(1), 73–86 (2008)
9. Portillo-Dominguez, A.O., et al.: TRINI: an adaptive load balancing strategy based on garbage collection for clustered Java systems. Sof. Pract. Experience, **46**(12), 1705–1733 (2016)

Face Recognition Based on Enhanced CSLBP

Chen Li[1,2(✉)], Shuai Zhao[1], Ke Xiao[1,2], and Yanjie Wang[1]

[1] College of Computer Science,
North China University of Technology, Beijing, China
lichen@ncut.edu.cn
[2] Beijing Key Laboratory on Integration and Analysis of Large-Scale
Stream Data, Beijing, China

Abstract. Face image with illumination variation usually contains redundant data that will seriously reduce the recognition rate. To combat the influence of illumination variation and extracting illumination-robust feature, a novel feature extraction method is proposed. The novel method is based on the combination of Center-Symmetric Local Binary Pattern (CS-LBP) and the fusion of the vertical and horizontal component images derived from wavelet decomposition. Numerous experiments have been done on the Extended Yale B to verify its effectiveness. The experimental results show that by applying the proposed method, redundant data caused by severe illumination variation can be filtered, while useful texture information can be reserved and enhanced. Compared with CSLBP, it significantly improves the face recognition performance under severe illumination variation.

Keywords: Wavelet decomposition · Image fusion · CSLBP · Feature extraction

1 Introduction

Face recognition has been a research hotspot ranging from the field of biometric recognition to modern security application due to its' distinct advantage compared with other biometric characteristics. Abundant face recognition approaches have been proposed, such as Center Symmetric Local Binary Pattern (CSLBP) [1–3], Local Directional Binary Pattern (LDBP) [4] and so on. Most of them can achieve satisfactory result if the face images are captured under well-constrained circumstances. However, in real scenario, there are many unconstrained factors such as illumination variation, expression variation, pose variation or occlusion that can weaken the face recognition performance. To combat these influence factors, further explores are urgent. This paper focuses on study novel feature extraction method that can be more robust to illumination variation.

Besides the rich information for identification, natural face image with illumination variation also contains redundant data that can lead to appearance variation of the image. To combat the variation and lower the impact of the redundant data, various pre-processing algorithms such as histogram equalization [5] and wavelet transform [6] have been applied. However, some useful information for recognition can also be suppressed during the preprocessing. Another category of method to combat illumination

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_86

variation is feature based method. Researchers try to propose feature which is more robust to illumination variation, such as Local Binary Pattern (LBP) [7], CSLBP, Gabor wavelet [8, 9] and so on. Among these, CSLBP has been proven to be efficient at describing the texture features of image. However its robustness to illumination variation is also finite.

In order to depress the redundant data caused by illumination variation while extracting illumination-robust feature, combination of the two methods mentioned above can be a promising way. Hence, by combining the CSLBP with wavelet-based image fusion method, a new enhanced CSLBP method is proposed in this paper. This paper is organized as follows: in Sect. 2, the proposed feature extraction method is detailed. In Sect. 3, abundant face recognition experiments are conducted to evaluate the proposed method. The conclusions are drawn in Sect. 4.

2 Technical Approach

2.1 Center Symmetric Local Binary Pattern

The proposed novel algorithm is based on CSLBP which is an effective texture feature descriptor. CSLBP introduces central symmetry principle to encode the image and describe the local texture feature by comparing the center-symmetric gray values between four pairs of pixels. CSLBP has been proven to be more robust to the flat area of image than traditional LBP. Besides, due to the central symmetry principle, CSLBP has lower computational complexity. The formula of CSLBP can be shown as follows:

$$CSLBP_{(R,N)}(x,y) = \sum_{i=0}^{L/2-1} p(f_i - f_{i+L/2})2^i \quad p(x) = \begin{cases} 1, & x \geq T \\ 0, & x < T \end{cases} \tag{1}$$

Where (R, N) denotes a circle neighborhood, f_i,$f_{i+L/2}$ are center symmetric pixels pairs, the threshold T is used to determine the flatness of the local texture region.

When $L = 8$, the calculation process of CSLBP feature is shown in Fig. 1.

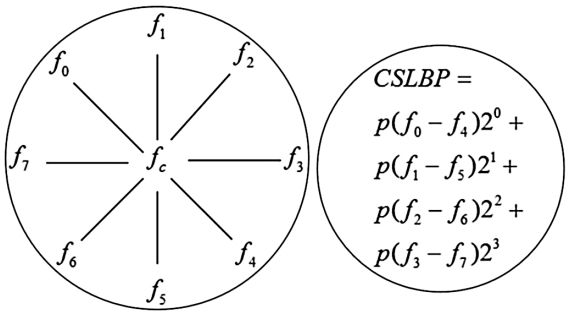


Fig. 1. Calculation process of CSLBP descriptor

2.2 Enhance Center Symmetric Local Binary Pattern

To achieve more robust illumination-insensitive features, a novel feature extraction method based on the combination of CSLBP and fused component image derived from wavelet decomposition is proposed. This section details the specific method. First of all, verdict the grayscale distribution of the image. The nonlinear enhancement transform is adopted to the images with extreme gray scale distribution before wavelet transform. Then the wavelet transform is applied to the sample images J^*. The equations are shown as follows:

$$\begin{aligned} J(i,l) &= \sum_{k,m} d(k-2i)d(m-2l)J^*(k,m)\ S(i,l) = \sum_{k,m} g(k-2i)d(m-2l)J^*(k,m) \\ C(i,l) &= \sum_{k,m} d(k-2i)g(m-2l)J^*(k,m)\ D(i,l) = \sum_{k,m} g(k-2i)g(m-2l)J^*(k,m) \end{aligned} \tag{2}$$

where d denotes the low-pass filter, g denotes the high-pass filter, J, S, C and D represent approximate component, horizontal component, vertical component and diagonal component respectively. The decomposition results are shown in Fig. 2.

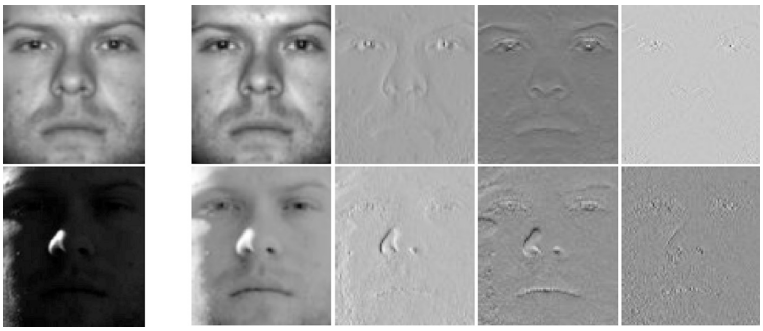


Fig. 2. Wavelet decomposition

Thirdly, according to certain fusion rules, image fusion based on wavelet decomposition is conducted on the component images with more detailed information. and the fused image is obtained by wavelet inverse transform of the combined coefficients. Finally, the fusion result is combined with CSLBP to obtain the Enhance Center Symmetric Local Binary Pattern (ECSLBP). The specific steps are as follows.

Step 1. Analyse the gray scale distribution of the samples, and conduct nonlinear extension to the images with drastic gray scale change.
Step 2. Decompose the sample image by wavelet to obtain approximate, vertical, horizontal and diagonal component.
Step 3. Fuse the vertical and horizontal component as two separate images. To obtain more abundant detailed features, choose the coefficient with higher absolute

value as the fusion coefficient for the high frequency component and the weighted average coefficient as the fusion coefficient of the low frequency component.
Step 4. Conduct CSLBP on the fusion image to obtain the feature vectors.

By combining CSLBP and the fused component images derived from wavelet decomposition, redundant data caused by severe illumination variation can be filtered while useful texture information can be reserved, enhanced and extracted.

3 Analysis of Experimental Data and Result

In this paper, Extended Yale B databases is selected to verify the effectiveness of the proposed algorithm. This database was created by the Computing Visual and Control Center of Yale University. Face images in this database have severe illumination changes and are suitable for testing the illumination robustness of algorithm. 640 face images captured from 10 different persons are selected for our experiment. It consist of 5 subsets. Among them, subset 1 has 7 images, subset 2 and subset 3 have 12 images, subset 4 has 14 images, and subset 5 has 19 images. Some examples are shown in Fig. 3.

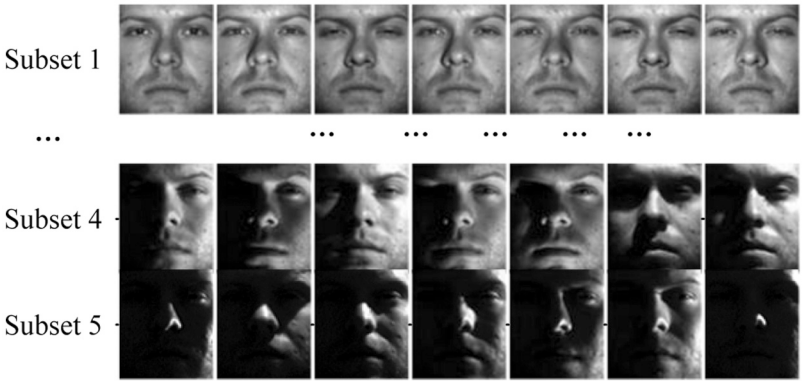


Fig. 3. Images of the subsets divided from the Extended Yale B

In this experiment, subset 1 under normal illumination circumstance is used as the training sample, and the other four subsets with severe light effects are used as the test samples respectively. The proposed new method is compared with the original CSLBP method to illustrate its effectiveness. Considering the recognition results is also affected by the classification algorithms, for both feature descriptors, this paper discusses two different classifiers to further demonstrate the effectiveness of the proposed method.

In the first experiment, Label Consistent K-SVD (LCKSVD) [10] is employed as classifier. The experimental results of the proposed method compared with the original CSLBP are shown in Fig. 4. As shown, the proposed ECSLBP achieve equal recognition rate on subset 2, which is 100%. Besides, it achieves much higher recognition

rate in subset 3, 4 and 5. Especially when the recognition rate of CSLBP has dropped to 36.42%, our proposed method can still achieve almost 79.28% recognition rate in subset 4. Similarly, when the recognition rate of CSLBP has dropped to 27.37%, our proposed method achieve recognition rate of 67.89% in subset 5.

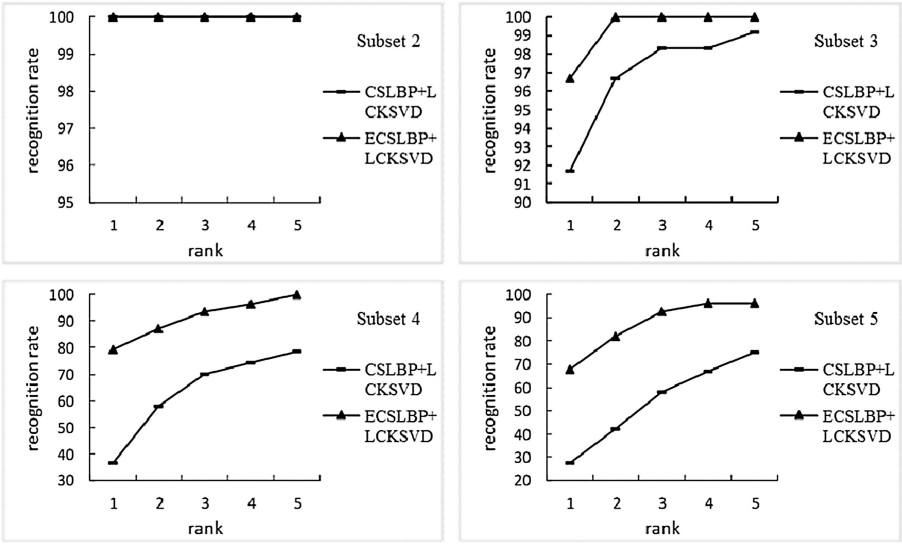


Fig. 4. The experimental results of each subset

In the second experiment, Euclidean distance is selected as the classifier. The experimental results are shown in Table 1.

Table 1. Recognition rates of different methods with various illuminations (%).

Methods	SubSet2	SubSet3	SubSet4	SubSet5
CSLBP + ED	100.00	95.00	39.28	21.58
ECSLBP + ED	100.00	98.33	85.71	77.89

As shown above, the recognition rate of the proposed ECSLBP and CSLBP both decline as the illumination condition becomes severe, however the proposed method invariably shows a better performance than CSLBP. Besides, even though adopting different classification algorithm achieves different recognition rate, but for both classification algorithm, the proposed new method achieve higher recognition rate than CSLBP. This means that the proposed method is more robust to illumination variation.

4 Conclusion

In this paper, we propose an enhanced CSLBP to combat the effect of illumination variation on face recognition performance. By combining CSLBP with the fused component image derived from wavelet decomposition, redundant data caused by severe illumination variation can be filtered while texture information beneficial for face recognition are reserved, enhanced and extracted. Abundant experiments are conducted on Extended Yale B database to verify the effectiveness of the proposed method. Regardless of which classifier are adopted, compared with CSLBP, our proposed method makes significant improvement in recognition rate for all the subsets of Extended Yale B.

Acknowledgment. This research is supported by National Natural Foundation of China (Grant No. 61503005), Beijing Natural Science Foundation (4162022), High Innovation Program of Beijing (2015000026833ZK04), Science and Technology Innovation Service Ability Construction (2017).

References

1. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest region with local binary pattern. Pattern Recogn. **42**(3), 425–436 (2009)
2. Changming, L., Jianyun, L., Lin, L.: Three-level face features for face recognition based on center-symmetric local binary pattern. In: IEEE International Conference on Computer Science and Automation Engineering, pp. 394–398 (2011)
3. Zhongshi, H., Jianyun, L., Lei, Y.: Face recognition method based on multi-channel Gabor filtering and center-symmetric local binary pattern. Comput. Sci. **37**(5), 261–264 (2010). (in Chinese)
4. Y.W., G.H.: Expression recognition algorithm based on local directional binary pattern. J. Comput. Inf. Syst. 3221–3228 (2014)
5. Ramirez-Gutierrez, K., Cruz-Perez, D., Olivares-Mercado, J., et al.: A face recognition algorithm using Eigenphases and histogram equalization. Int. J. Comput. **5**(1), 34–41 (2011)
6. Fengxiang, W.: Face recognition based on wavelet transform and regional directional weighted local binary pattern. J. Multimedia **9**(8), 1017–1023 (2014)
7. Guangcheng, Z., Xiangsheng, H., Stan, Z.L., Yangsheng, W., Xihong, W.: Boosting local binary pattern (LBP)-based face recognition. In: Conference on Advances in Biometric Person Authentication, pp. 179–186 (2004)
8. Ashraf, A.B., Simon, L., Tsuhan, C.: Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1335–1341 (2010)
9. Weifeng, L., Zengfu, W.: Facial expression recognition based on fusion of multiple Gabor features. In: 18th International Conference on Pattern Recognition, pp. 536–539 (2006)
10. Zhuolin, J., Zhe, L., Larry, S.D.: Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In: Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1697–1704 (2011)

A Supporting Environment for Formal Analysis of Cryptographic Protocols

Jingchen Yan, Kazunori Wagatsuma, Hongbiao Gao, and Jingde Cheng(✉)

Department of Information and Computer Sciences, Saitama University, Saitama, Japan
{jingchenyan, wagatsuma, gaohongbiao, cheng}@aise.ics.saitama-u.ac.jp

Abstract. Formal analysis of cryptographic protocols is to find out flaws in the protocols by various formal methods. Some supporting tools for formal analysis of cryptographic protocols have been proposed and applied, but the tools failed to support the whole processes of formal analysis automatically. Therefore, a supporting environment which can support formal analysis automatically is needed for analysts. This paper presents the first supporting environment for formal analysis of cryptographic protocols.

Keywords: Cryptographic protocols · Formal analysis · Supporting environment

1 Introduction

Formal analysis of cryptographic protocols is used to find out flaws in the protocols [5, 6, 10]. Formal analysis methods mainly include model checking [1, 2], theorem proving [1, 2] and formal analysis method with reasoning [7].

Model checking is proving method that properties of the cryptographic protocol and specification present fatal action of attacks must be converted into a formalism acceptance, then verify the attacks succeed or not [1, 2, 9]. Theorem proving is also based on proving that security targets are constructed into formal theorems, then analysts should prove whether the formal theorem holds or not by using inference rules and model theorem [1, 2, 13]. Formal analysis method with reasoning is a method that actions of attack are deduced by forward reasoning from formalized cryptographic protocol as premises. It consists three processes which are formalization, forward reasoning and analysis [18].

Various supporting tools such as Scyther [11], ProVerif [3] for model checking, CafeOBJ [13], Isabelle [15] for theorem proving and FreeEnCal [8] for formal analysis method with reasoning have been developed and applied. However, these supporting tools can only support a part of formal analysis methods, but cannot support the whole processes of formal analysis. Therefore, it's necessary to develop a supporting environment which can support analysts to perform the whole processes of formal analysis automatically.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_87

This paper presents the first supporting environment for formal analysis of cryptographic protocols. The rest of this paper is organized as follows. Section 2 shows difficulties in formal analysis of cryptographic protocols. Section 3 presents requirements analysis and function definition for the supporting environment. Section 4 explains the usage of the supporting environment. And concluding remarks is given in Sect. 5.

2 Difficulties in Formal Analysis of Cryptographic Protocols

Although various supporting tools have been proposed and applied, there are still some difficulties when analysts perform formal analysis of cryptographic protocols.

Firstly, formalization process and its inverse process still need to perform manually [16, 17]. when analysts use supporting tools to perform formal analysis, formalization process should be performed at first. In particular, properties of the cryptographic protocol should be converted into formalism acceptances or formal theorems. However, the incompleteness of informal properties makes the formalization process difficult [14], which leads to the time-consuming and error-prone problems of manual formalization.

Similarly, the inverse process of formalization also needs to perform manually. In formal analysis method with reasoning, it's difficult for analysts to transform the deduced logic formulas into their natural language representation.

In addition, if analysts want to use different supporting tools to analyze a cryptographic protocol, they must master each language used by those tools, for example, Scyther's input language is loosely based on a C/Java-like syntax [12] but ProVerif's input language is a variant of the applied pi calculus [4], which increases the difficulty of using them.

Secondly, in formal analysis method with reasoning, analysis of deduced formulas by forward reasoning is also a time-consuming complex task [18]. In the analysis process, analysts need to find possible successful attacks from millions of formulas. It's difficult for analysts to deal with massive amount of deduced formulas to find possible successful attacks.

Thus, to solve the automatic formalization problem and its inverse process of various supporting tools and provide convenient for analysts to analyze the deduced logic formulas, in other words, to support the whole processes of formal analysis for cryptographic protocols automatically, a supporting environment is necessary and urgent.

3 Requirement Analysis and Function Definition

To develop the supporting environment for formal analysis of cryptographic protocols, we enumerated following requirements for the supporting environment.

R1: The supporting environment should support the formalization process. R1.1: The supporting environment should provide the function of transforming informal properties of the cryptographic protocol into the language of formal analysis tool of the

selected tool. R1.2: The supporting environment should provide the function of transforming logical formulas into the language of formal analysis tool.

R2: The supporting environment should support the inverse process of formalization process. R2.1: The supporting environment should provide the function of transforming language of formal analysis tool into logic formulas. R2.2: The supporting environment should provide the function of transforming logic formulas into informal properties.

R3: The supporting environment should support the process of formal analysis. R3.1: The supporting environment should support the automatic tool for forward reasoning method. R3.2: The supporting environment should support the automatic tools for model checking method. R3.3: The supporting environment should support the automatic tools for theorem proving method.

R4: The supporting environment should support analysts to analyze the result of forward reasoning. R4.1: The supporting environment should provide the function of filtering the logical formulas. R4.2: The supporting environment should provide a function that sorts out the filtered logical formulas according to the possibility of successful attacks.

R5: The supporting environment should systematically store the inputting properties of cryptographic protocols. R5.1: The supporting environment should store the properties for each user. R5.2: The supporting environment should store the properties of each cryptographic protocols. R5.3: The supporting environment should store the properties for each creation times.

R6: The supporting environment should make possible for analysts to manage the properties of cryptographic protocols. R6.1: The supporting environment should make possible for analysts to search the properties of cryptographic protocols. R6.2: The supporting environment should make possible for analysts to modify the properties of cryptographic protocols. R6.3: The supporting environment should make possible for analysts to delete specification of cryptographic protocols.

R7: The supporting environment should systematically store the result of formal analysis. R7.1: The supporting environment should store the result of formal analysis for each user. R7.2: The supporting environment should store the result of formal analysis for each cryptographic protocols. R7.3: The supporting environment should store the result of formal analysis for each time of formal analysis. R7.4: The supporting environment should store and manage the result of formal analysis for each divided cases in tasks of forward reasoning.

R8: The supporting environment should make possible for analysts to manage the result of formal analysis. R8.1: The supporting environment should make possible for analysts to search the result of formal analysis by each formal method. R8.2: The supporting environment should make possible for analysts that they can delete the result of formal analysis.

It is a hard task for analysts to perform formalization, forward reasoning or proving, and analyzing the result of formal analysis. So R1, R2, R3, R4 are the basic requirements of the supporting environment for formal analysis method.

It is possible that analysts analyze many cryptographic protocols. As the result, it takes lots of time to find target properties or results from many properties or results.

Therefore, the supporting environment should satisfy R5, R7 to save the properties and results of each cryptographic protocol.

It can be assumed that analysts improve a target cryptographic protocol from the result of formal analysis. Therefore, the supporting environment should satisfy R6 to manage the properties of cryptographic protocols.

It can be assumed that analysts manage the result of the formal analysis of target cryptographic protocol. Therefore, R8 should be satisfied by the supporting environment.

To satisfy the requirements above, a list of functions the supporting environment is as follows.

F1: Formalization function (satisfied R1) F1.1: Transform informal properties into logic formulas. F1.2: Transform logic formulas into language of formal analysis tool.

F2: Inverse formalization function (satisfied R2) F2.1: Transform language of formal analysis tool into logic formulas. F2.2: Transform logic formulas into informal properties.

F3: Formal analysis function (satisfied R3) F3.1: Forward reasoning. F3.2: Proving by model checking. F3.3: Proving by theorem proving.

F4: Filter function (satisfied R4) F4.1: Filter logic formulas. F4.2: Sort out the filtered logic formulas.

F5: Making directory function (satisfied R5, R7) F5.1: Make directories of the properties and results of cryptographic protocols for each user. F5.2: Make directories of the properties and results of cryptographic protocols for each cryptographic protocol. F5.3: Make directories of the properties and results of cryptographic protocols for each creation time.

F6: Properties managing function (satisfied R6) F6.1: Search the properties of cryptographic protocols. F6.2: Modify the properties of cryptographic protocols. F6.3: Delete the properties of cryptographic protocols.

F7: Result managing function (satisfied R8) F7.1: Search the result of cryptographic protocols. F7.2: Delete the result of cryptographic protocols.

4 Usage of the Supporting Environment

Overview of the supporting environment is represented as Fig. 1. Details process are that analysts input user name, cryptographic protocol name and properties of the cryptographic protocol as a target of formal analysis. Then select the formal analysis tool. After that, the properties will be transformed into LF (logical formulas) and LAT (language of formal analysis tool) which correspond to the formal analysis tool.

If analysts would like to use the formal analysis method with proving, only formalization and proving tasks be needed. For example, when we select ProVerif, the function of formalization will put informal properties of the cryptographic protocol into ProVerif's input language, a variant of the applied pi calculus which support types [4]. Then Proverif can automatically verify whether flaws exist or not. Because the output produced by ProVerif is rather Verbatim, we don't have to do transformation work.

If analysts would like to use the formal analysis method with reasoning, the properties will be transformed into LF and LAT corresponding to FreeEnCal [8] and then forward reasoning is performed by using the result of formalization. After that, the

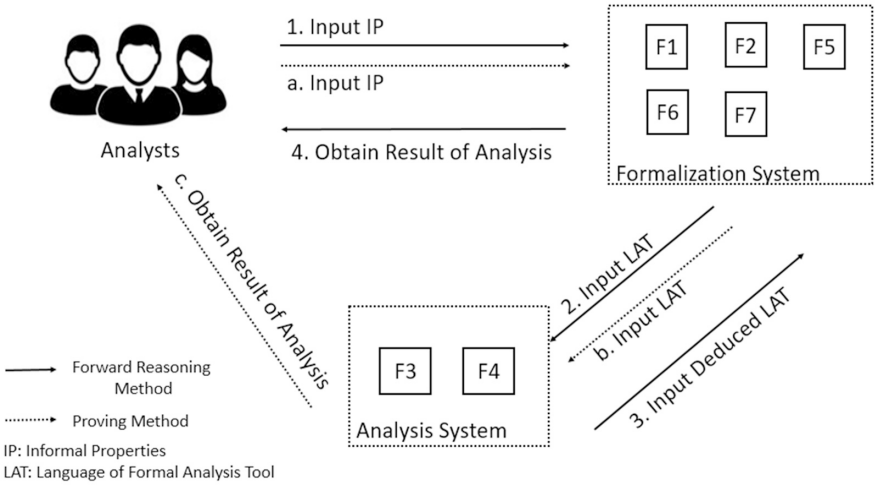


Fig. 1. This figure shows the overview of supporting environment. Procedure a, b, c show the process when analysts select proving method. Procedure 1, 2, 3, 4 show the process when analysts select forward reasoning method. Formalization system is corresponding to the formalization process and analysis system is to perform forward reasoning or proving. F1–F7 denotes functions mentioned in Sect. 3.

results of forward reasoning presented by LAT will be filtered and sorted. Then the sorted results will be transformed into LF. At last, target LF will be transformed into informal language for analysts to analyze whether there are possible successful attacks or not. If the possible successful attacks are difficult for analysts to confirm, they can use the proving method of the supporting environment to verify whether the possible successful attacks succeed or not. By the two rounds of analysis, analysts can find all flaws in the cryptographic protocols without enumerating the attacks before formal analysis.

Analysts can manage the input properties and results of each performing of formal analysis. They can read each property and result of cryptographic protocols by searching the user name or cryptographic name or creation time. Certainly, they can also search, modify or delete them.

5 Concluding Remarks

In this paper, we have proposed the first supporting environment for formal analysis of cryptographic protocols. The supporting environment integrates various supporting tools and it can support analysts to perform formal analysis of cryptographic protocols through the whole processes. Currently, we are developing the environment and evaluating it through some case studies.

References

1. Avalle, M., Alfredo, P., Riccardo, S.: Formal verification of security protocol implementations: a survey. Formal Aspects Comput. **26**(1), 99 (2014)
2. Bau, J., Mitchell, J.C.: Security modeling and analysis. IEEE Secur. Priv. **9**(3), 18–25 (2011)
3. Blanchet, B.: An efficient cryptographic protocol verifier based on prolog rules. In: Proceedings of the 14th IEEE Computer Security Foundations Workshop, pp. 82–96. IEEE, (2001)
4. Blanchet, B., Smyth, B., Cheval, V.: ProVerif 1.96: Automatic Cryptographic Protocol Verifier, User Manual and Tutorial (2016)
5. Meadows, C.A.: Formal verification of cryptographic protocols: a survey. In: International Conference on the Theory and Application of Cryptology. Springer, Heidelberg (1994)
6. Meadows, C.A.: Formal methods for cryptographic protocol analysis: emerging issues and trends. IEEE J. Sel. Areas Commun. **21**(1), 44–54 (2003)
7. Cheng, J., Miura, J.: Deontic relevant logic as the logical basis for specifying, verifying, and reasoning about information security and information assurance. In: Proceedings of the 1st International Conference on Availability, Reliability, and Security, pp. 601–608. IEEE-CS (2006)
8. Cheng, J., Nara, S., Goto, Y.: FreeEnCal: a forward reasoning engine with general-purpose. In: Proceedings of the 11th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Artificial Intelligence, vol. 4693, pp. 444–452. Springer-Verlag, Heidelberg (2007)
9. Clarke, E.M., Grumberg, O., Peled, D.: Model Checking. MIT Press, Cambridge (1999)
10. Cortier, V., Steve, K., Bogdan, W.: A survey of symbolic methods in computational analysis of cryptographic systems. J. Autom. Reasoning **46**(3–4), 225–259 (2011)
11. Cremers, C.: On the protocol composition logic PCL. In: Proceedings of the 2008 ACM Symposium on Information, Computer, and Communications Security, pp. 66–76. ACM (2008)
12. Cremers, C.: Scyther User Manual (2014)
13. Futatsugi, K., Diaconescu, R.: CafeOBJ report. World Scientific (1998)
14. Heidorn, G.E.: Automatic programming through natural language dialogue: a survey. In: Readings in Artificial Intelligence and Software Engineering, pp. 203–214 (1986)
15. Paulson, C.: The inductive approach to verifying cryptographic protocols. J. Comput. Secur. **6**, 85–128 (1998)
16. Wagatsuma, K., Shogo, A., Goto, Y., Cheng, J.: Formalization for formal analysis of cryptographic protocols with reasoning approach, In: Future Information Technology, Lecture Notes in Electrical Engineering, vol. 309, pp. 211–218. Springer, Heidelberg (2014)
17. Wagatsuma, K., Harada, T., Anze, S., Goto, Y., Cheng J.: A supporting tool for spiral model of cryptographic protocol design with reasoning-based formal analysis, In: Advanced Multimedia and Ubiquitous Engineering - Future Information Technology, LNEE, vol. 354, pp. 25–32. Springer, Heidelberg (2015)
18. Yan, J., Wagatsuma, K., Gao, H., Cheng, J.: A formal analysis method with reasoning for cryptographic protocols. In: Proceedings of the 12th International Conference on Computational Intelligence and Security, pp. 566–570. IEEE Computer Society (2016)

Modeling and Simulation of LoRa in OPNET

Andrew D. Jun[1], Seokjoon Hong[1], Wooyeob Lee[1], Kyungrak Lee[1], Inwhee Joe[1(✉)], Kyeseon Lee[2], and Tae-Joon Park[2]

[1] Department of Computer Science, Hanyang University, Seoul, Korea
{dosungjun, daniel379, matiasl2, esilote82, iwjoe}@hanyang.ac.kr
[2] Electronics and Telecommunications Research Institute, Daejeon, Korea
{seonny, tjpark}@etri.ac.kr

Abstract. LPWA networks are getting attention as a solution to support massive number of IoT devices. LoRa is one such low-power based long-range technology. In this paper, we discuss methods to build a LoRa simulation environment. We first discuss the characteristics of the LoRa PHY and MAC layers, defined in the LoRa and LoRaWAN specifications. Then, we show how LoRa PHY and MAC functions can be realized in the OPNET simulation environment. For LoRa PHY implementation, we adopted LoRa modulation curve based on BER vs Eb/No. We also implemented various process models depends on LoRa node model for LoRa MAC functions. We conclude with future directions for performance enhancement.

Keywords: IoT · LPWA · LoRa · OPNET · Modeling · Simulation

1 LoRa Functionality

1.1 PHY Layer

LoRa PHY, based on CSS (Chirp Spread Spectrum) modulation [1–4], is capable of delivering low power, long range communications. CSS facilitates a few orthogonal (logical) channels over a common physical spectrum with different SF (Spreading Factor) values. The choice of an SF value determines the communication range and the data rate. There is a tradeoff between the communication range and the data rate - communication ranges become longer with lower data rates.

For the evaluation or implementation of a modulation technique, BER (Bit Error Rate) vs. Eb/No (Energy per bit to Noise density ratio) relationships are required. With no such data at hand (either from Semtech data sheets or from field experiments), we calculate BER-to-Eb/No relationships as follows.

For PER (Packet Error Rate) = 1% and N = 100 bits transmitted (payload 80 bits, code rate 4/5), BER is calculated as 0.01% from $PER = 1 - (1 - BER)^N$. For all the SNR values from Table 1, Eb/No values are calculated as ~6 dB from Eq. (1). Here, BW (bandwidth) is 125 kHz, and Rb (Bit Rate) is as given in Table 1.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_88

Table 1. LoRa demodulator SNR [5, 6]

Spreading factor	Data rate (bps)	Demodulator SNR
7	5470	−7.5 dB
8	3125	−10.0 dB
9	1760	−12.5 dB
10	980	−15.0 dB
11	440	−17.5 dB
12	250	−20.0 dB

$$E_b/N_0\,(\mathrm{dB}) = SNR\,(\mathrm{dB}) + 10 * \log(\frac{BW}{R_b})\,(\mathrm{dB}) \tag{1}$$

According to Semtech data sheet sx1276 [5], LoRa has 3 different cases of interference rejection conditions for signal strengths and receiver sensitivities: co-channel, co-channel with different SF, and adjacent channel.

- *Co-channel:* when the target (T) and the interferer (I) packets are received simultaneously with the same SF, frequency, and bandwidth. The rejection condition is
  - signal strength (T) < 3 dB + sensitivity, or
  - signal strength (T) < 6 dB + signal strength (I).
- *Co-channel with different SF:* when the target (T) and the interferer (I) packets are received simultaneously with the same frequency, but different SF. The rejection condition is
  - signal strength (T) < 6 dB + sensitivity, or
  - signal strength (T) + THc < signal strength (I). (THc: rejection threshold in Table 2).
- *Adjacent channel:* when the target (T) and the interferer (I) packets are received simultaneously with different SF and frequency. The rejection condition is
  - signal strength (T) < 3 dB + sensitivity, or
  - signal strength (T) + THa < signal strength (I). (THa: rejection threshold in Table 2).

Table 2. Rejection threshold for co-channel with different SF

SF	Co-channel (with different SF) rejection threshold (dB)	Adjacent channel rejection threshold (dB)
7	6	60
8	9.5	
9	12	
10	14.4	
11	17	
12	19.5	70

1.2 MAC Layer

- *Class A*

Class A aims for minimal power consumption for both signal transmission and reception. As such, Class A end-devices wake up, from the dormant state, for transmission only when there is some data to send out. Channel access in Class A is based on ALOHA-type random access: i.e. contention-based transmission. Successful transmissions and receptions of uplink data can be confirmed by the receiving end (gateway) via ACK message replies. Collisions or other errors can be detected with NACK messages or no reception of any replies.

After transmission, Class A end-devices open up two short "Receive Windows" to receive data from the network/server (downlink traffic). RX1 (the first receive window) utilizes the same frequency and data rate as those of the preceding uplink transmission. Here, the delay times of RECEIVE_DELAY1 and RECEIVE_DELAY2 are pre-defined in the network. The frequency and data rate of RX2 (the second receive window) is pre-configured. If required, this value can be changed through a MAC command between the end-device and the network. Note that RX2 opens up only if no data can be successfully received in RX1.

- *Class B*

Departing from the random access techniques of Class A, downlink communication of Class B is based on periodic placement of "Receive Windows." As explained in Fig. 1, end-devices are synchronized on the "system" time through the Beacon signals transmitted periodically by the gateway. Once a Beacon signal is received, the timing of short "Receive Windows" (Ping Slots) can be calculated based on a mutually-agreed formula between the end-device and the network. Class A end-devices can be converted to Class B via application-level decisions if preferred.

The timings of Ping Slots are calculated in a pseudo-random manner so that Ping Slots for different end-devices do not overlap from one another. For each Beacon period, the end-device and the server calculate the "pingOffset" as follows.

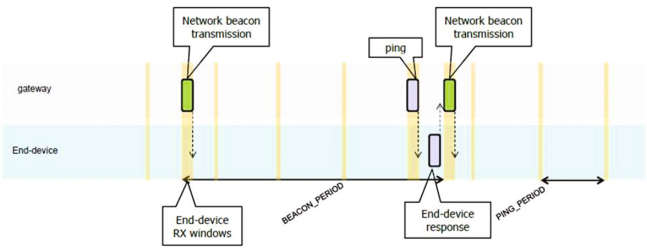


Fig. 1. Beacon window and ping slot (Class B)

- *Class C*

When end-devices have sufficient power sources, they do not need to limit the size of "receive windows" for the purpose of power conservation. In this case, end-devices employ the Class C mode for continuous reception of signals. Similar to Class A, there exist two types of "Receive Windows" in Class C: RX1 and RX2 (Fig. 2). Class C does not close the RX2 window until next transmission. Class C end-devices spend most of time for reception of downlink data in RX2 window.

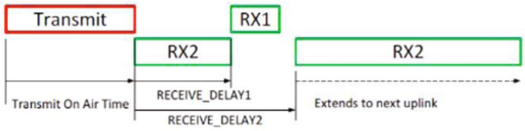


Fig. 2. Receive window timing for Class C end-device

2 LoRa Implementation in OPNET

In this paper, we implemented LoRa PHY and MAC functions by using OPNET network simulator [7]. OPNET is a famous simulator which offers graphical support as well as various network model components and functions. In OPNET, object-oriented programming technique is used to create the mapping from graphical design to the implementation of real system.

2.1 PHY Layer

OPNET does not provide mechanisms to literally fulfil PHY layer functionalities. Instead, characteristics of PHY are assumed as implicit, and only the consequences and effects of the functionality are explicitly expressed.

The effect of modulation, for instance, is revealed as a BER-to-Eb/No graph mentioned in Sect. 1.1. Other PHY characteristics, such as noise figure and interference level, are configured as attributes in the fourteen stages in the "Radio Transceiver Pipeline" (the first six for transmission and the following eight for reception):

(1) *Receiver Group:* to determine which receivers are capable of receiving the signal.
(2) *Transmission Delay:* to determine the time required for transmission.
(3) *Link Closure:* to check if the signal can be reached to the receiver.
(4) *Channel Match:* to classify the packet to VALID, NOISE and IGNORE.
(5) *Tx Antenna Gain:* to calculate the gain of Tx antenna.
(6) *Propagation Delay:* to calculate the propagation delay.
(7) *Rx Antenna Gain:* to calculate the gain of Rx antenna.
(8) *Received Power:* to calculate the power of the signal received.
(9) *Interference Noise:* to calculate the interference level of other signals to the target signal.

(10) *Background Noise:* to calculate the noise level.
(11) *Signal to Noise Ratio:* to calculate SNR of received packet.
(12) *Bit Error Rate:* to calculate bit error probability based on SNR.
(13) *Error Allocation:* to determine which bit contains an error.
(14) *Error Correction:* to determine if the received packet is valid after.

- *Implementation of Modulation*

Figure 3 shows the LoRa modulation curve based on the methodologies discussed in Sect. 1.1 to calculate BER-to-Eb/No relationships.

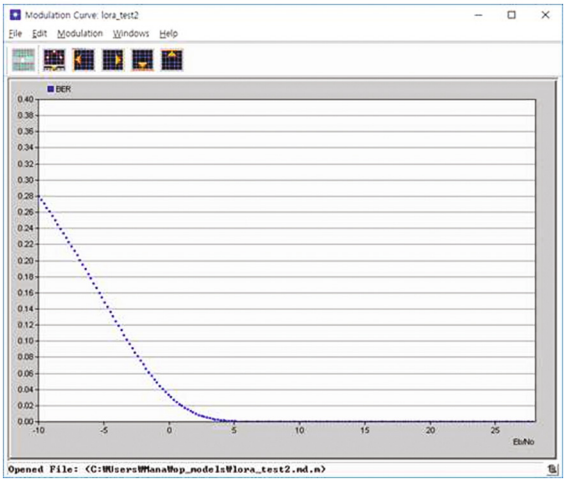


Fig. 3. LoRa modulation curve.

- *Implementation of Interference*

In Stage 4 of the Radio Transceiver Pipeline, received packets are classified as IGNORE, NOISE or VALID state, based on the values of frequency, bandwidth, data rate, SF, and modulation.

– *IGNORE:* if the received signal frequency does not overlap the frequency of interest.
– *NOISE:* if the frequency of the received signal partially overlaps the frequency of interest, or if fully overlaps, but with different SF.
– *VALID:* if both the frequency and the SF are the same as those of interest.

The basic model of the "Channel Match" stage is modified to reflect the LoRa characteristics defined in SX1272/73 data sheets [6]. The range of adjacent channels is defined to be 1.5 times wider than the bandwidth of interest.

The power level of VALID packets is summed up as interference noise in Stage 9 if the frequency falls into that of interest. After validity check and interference noise and background noise summations, the 3 cases of rejection conditions (co-channel, co-channel with different SF, and adjacent channel) are exercised as described in Sect. 1.1.

- *Implementation of Path Loss*

Although the base model for path loss in OPNET is the "Free Space" model, we have adopted the field-measured values presented in [8].

2.2 MAC Layer

- *Network Model*

A LoRa network consists of three base elements: End-Node, Gateway, and Network Server. End-devices and gateways communicate through LoRa channels in wireless while any communications medium, including wireline, can be assumed in between gateways and the network server [9].

- *Node Models*

LoRa end-device model consists of PHY Rx/Tx modules, LoRa MAC module, and tx_gen (application) module. Simulation parameters associated with end-devices, such as Name, *AppEUI*, *DevEUI, Class Type, etc.*, are defined as the attributes of the End-device Model.

The LoRa Gateway node model contains 3 sets of LoRa PHY Rx/Tx modules (for 922.1 MHz, 922.5 MHz, and 922.5 MHz, respectively) to communicate with end-devices, and Ethernet Rx/Tx modules to communicate with the network server [5]. Also included is a "Gateway MAC," which conveys uplink and downlink packets between the network server and end-devices, and controls the periodic transmission of beacons. LoRa gateways are uniquely identified via GwEUI attribute.

The LoRa Network Server node model include "Wireline Rx/Tx" modules for connection with gateways, and the "Server MAC" module for LoRaMAC-level communications with end-devices.

- *Process Models*

Three LoRa process models are produced for one-to-one mapping with the three LoRa Node Models. Note that lora_dispatch is the top-level process to be fetched first, and one of the child processes (lora_end_device, lora_gateway, and lora_network_server) is launched depending upon which node model is being instantiated. Sub-level child processes are also defined for further processing of deeper-level functionalities.

When lora_dispatch, the root process, provides services to an end-device, the lora_end_device process is called out. As displayed in Fig. 4, sub-level processes (aloha_proc or csma_proc) are then dispatched for transmission and reception of packets.

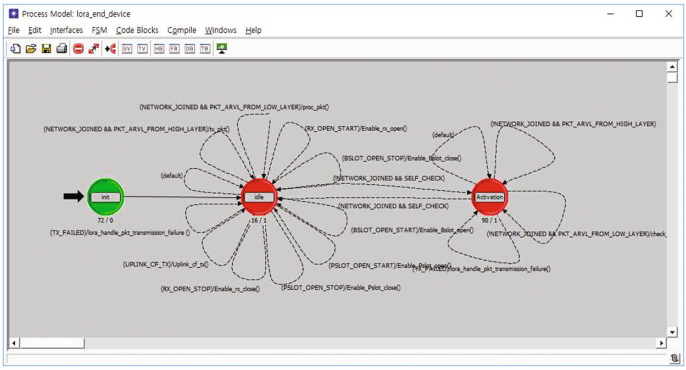


Fig. 4. lora_end_device process model

The lora_end_device process is executed in three states: init, idle, and activation.

- *"init" state:* When initialized, various parameters and attributes are filled with default set-up values.
- *"idle" state:* The process remains in the idle state until an event takes place (except that the transition from the init state automatically makes a change to the activation state). When a packet arrives from the higher layer, this process launches a child process for packet transmission. When a packet is received from outside, an appropriate child process is spawned for reception of the packet. Packet transmissions and receptions follow all the rules and regulations of PHY and MAC as described in Sects. 1.1 and 1.2, respectively, as one of Class A, Class B or Class C end-device. After proper handling of an event, the process always comes back to the idle state.
- *"activation" state:* When initialized, each end-device has to be registered in the network server through the exchanges of "join request" and "join accept" messages with the network server. Other messages are ignored until activated.

LoRa gateway nodes launch the lora_gateway process (Fig. 5) with two states:

- *"init" state:* All the initialization gets done, including the launch of "beacon_proc" for the preparation of periodic beacon transmissions.
- *"idle" state:* The process waits for a packet reception event either from the LoRa side or from the wireline side. Once received properly from one side, the packet gets transmitted to the other side.

The lora_network_server process has three states as follow.

- *"init" state:* In this state, all the attributes are initialized with default values.
- *"idle" state:* The process waits for an event in this state.
- *"proc_pkt" state:* This takes care of received packets from end-devices via gateways. For "Join Request", "Join Accept" is replied back upon successful activation. ACK messages are sent back for data packets. Replies are queued in the system

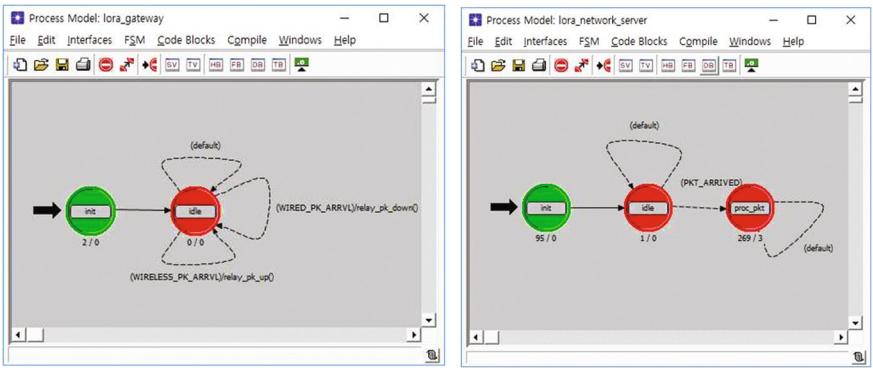


Fig. 5. lora_gateway and lora_network_server process models

until end-devices open up "receive windows." Ping messages are sent out to Class B end-nodes via "tx_ping_proc" in synchronization with the beacon periods.

3 Concluding Remarks

In this paper, we have presented the characteristics and algorithms of LoRa PHY and MAC, followed by the OPNET implementations. Basic OPNET features have been adopted and modified to simulate LoRa PHY and MAC. As a network simulation environment, OPNET does not provide functions and tools for the actual implementation of PHY. Rather, the net effects of PHY are built into the OPNET "Radio Transceiver Pipeline," projecting the real results from the data sheet and the field. Different uplink and downlink MAC behaviors are implemented for Class A, Class B and Class C.

For future work, we are very interested in developing advanced features in IoT networks for good performance and functionality without losing the practicality of battery power conservation and low implementation complexity. LoRa MAC can be extended to incorporate LBT and transmission prioritization techniques. Of special interest is "reservation-based" uplink transmission mechanisms. "Soft traffic segregation" methodology, where opportunistic random access traffic is laid in the background with lower priority, is being considered for high utilization of bandwidth.

Acknowledgement. This work was supported by Institute for Information and Communications Technology Promotion (No. R7117-17-0197, Development of Core Technologies for LPWA IoT Network) and partly by National Research Foundation of Korea (No. 2016R1A2B4013118), both funded by Ministry of Science, ICT and Future Planning.

References

1. Springer, A., Gugler, W., Ruemer, M., Reindl, L., Ruppel, C.C.W., Weigel, R.: Spread spectrum communications using chirp signals. In: Information Systems for Enhanced Public Safety and Security, EUROCOMM 2000, IEEE/AFCEA17, pp. 166–170, May 2000
2. Sarni, Y., Sadoun, R., Belouchrani, A.: On the application of chirp modulation in spread spectrum communication systems. In: Sixth International Symposium on Signal Processing and its Applications, vol. 2, pp. 501–504, 13–16 August 2001
3. IEEE P802.l5.4a/D4 (Amendment of IEEE Std 802.15.4). Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LRWPANs), July 2006
4. Wang, X., Li, X., Fei, M.: Performance evaluation of wireless networks based on chirp spread spectrum. In: 11th IEEE International Conference on Communication Technology, pp. 58–61, 10–12 November 2008
5. SX1272/73 Datasheet. Semtech Corporation, March 2015
6. On the Coverage of LPWANs: Range Evaluation and Channel Attenuation Model for LoRa Technology. University of Oulu (2015)
7. OPNET. http://www.opnet.com
8. LoRaWAN Specification, the LoRa Alliance, July 2016
9. LoRaWAN Regional Parameters, the LoRa Alliance, October 2016

Dynamic Analysis Bypassing Malware Detection Method Utilizing Malicious Behavior Visualization and Similarity

Jihun Kim and Jonghee M. Youn(✉)

Department of Computer Engineering, Yeungnam University, Gyeongsan, Gyeongbuk 38541, South Korea
fl352l@ynu.ac.kr, youn@yu.ac.kr

Abstract. Malware attacks have been posing various security threats such as data losses, personal information and financial information, system damage, and IT infrastructure destruction. To prevent these security threats in advance, many anti-malware programmers and malware analyzers have been analyzing malware. But methods of attacks are diversifying and it makes it harder for analyzer to analyze malware. For instance, bypass dynamic analysis malwares such as time-trigger are much more difficult to analyze than general malware because its function is executed at a particular time. In this paper, we proposed that automatic analysis of bypass dynamic analysis malware such as time-trigger. First, for our proposal, we utilizes BFS (Breadth-First Search) algorithm to track malicious behaviors flows from the beginning to the end. And such flows of malicious behaviors were visualized into graph. Furthermore, we calculated malware similarity using SSIM (Structural Similarity Image Metric) based on malicious graph.

Keywords: Malware analysis · Malware similarity · Visualization

1 Introduction

Recently, following the rapid development of hacking techniques and computer networks, diverse kinds of malware have been exponentially increasing. Furthermore, malware's attack forms have been diversified drastically and malware variants made by falsifying existing malware are also showing explosive rates of increase. Especially, bypass dynamic analysis malwares make it harder for analyzer to analyze because such as time triggers that initiates malicious behavior when a certain time has come or when the user takes a certain action. So these malware have to use static technologies such as reverse engineering because these types of malware cannot be analyzed through

This work was supported by the MSIP(Ministry of Science, ICT and Future Planning, Korea, under the ITRC (Information Technology Research Center) support program (IITP-2016-R2718-16-0035) supervised by the IITP (National IT Industry Promotion Agency), the Basic Science Research Program through the NRF funded by the Ministry of Education (NRF-2015R1C1A1A02037561) and the 2016 Yeungnam University Research Grant.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_89

dynamic analysis. In the paper, a method that utilizes the advantages of static analysis to detect malware that bypasses dynamic analysis is presented. We used the method applies breadth-first searches to track malicious behaviors' flows from the beginning to the end. And then, such flows of malicious behaviors were visualized into graphs. Each of the graphs are used in calculation similarity. The present paper is composed as follows. In Sect. 2, studies of malware that bypasses dynamic analysis and malware similarity related studies that were conducted earlier are introduced, in Sect. 3, the method proposed in the present study is presented, in Sect. 4, the proposed method is experimented for verification, and finally, in Sect. 5, conclusions are drawn.

2 Related Work

Dynamic analysis is a method that actually executes malware to analyze malicious behavior with observed changes in the system [1]. However, this analysis technique also has limitations in analysis because those types of malware that is activated when the user has taken a certain behavior or those types of malware such as time triggers that act at a certain time do not conduct any malicious behavior unless certain requirements have been satisfied. Recent malware makers make malware that bypasses these two analysis technique to make the detection of malicious behaviors by anti-malware programs difficult. Therefore, related researchers have been conducting studies to use static analysis and dynamic analysis together in order to detect malware using the common attributes of malware [2] or to detect malware involving bypassing technique [3]. Although the present study is based on static analysis, it conducts analysis similar to dynamic analysis because it tracks the flows of reverse engineered codes at the time points of beginning and ending and extracts common attributes of malware such as time triggers to figure out the similarity thereby presenting a new classification technique for those types of malware that bypasses dynamic analysis.

To present and respond to the threat of malware, diverse methods of malware analysis are being studied and studies of similarity can be said to be a representative study subject. Similarity analysis is not only utilized for the analysis of large quantities of malware that is explosively increasing in scale and the maximization of the efficiency of automatic analysis systems but also utilized as a measure of fast decision making. Malware similarity analyses have been conducted with diverse methodologies. Representative ones include those that compiled the characteristics of malware, that is, Opcodes or Strings and classified the differences to measure similarity levels [4], those that analysed file DNA based behavior patterns to compare similarity levels [5], and those that used N-gram to compare similarity levels [6]. In addition, algorithms to figure out similarity levels are also used in measuring malware similarity. Representative ones are cosine similarity measurement methods that numerically calculate the distance between two object to measure similarity [7] and those similarity measurement methods that utilize Jaccard coefficients [8].

3 Proposed Method

3.1 Composition of Behavior Graphs

Visualization of malicious behaviors draws graphs consisting of edges and vertexes. Here, the vertexes are fixed coordinates that indicate the types of APIs and the colors of edges vary with branches in algorithms.

Figure 1 is intended to improve the understanding of the method proposed in the present study and visualization by explaining them together with pseudo codes.

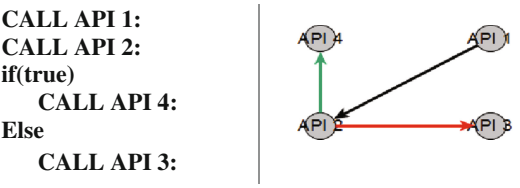


Fig. 1. Visualization of malicious behaviors

Vertexes are drawn with fixed coordinates such as API 1–API 4. As can be seen in the pseudo codes, API 1 is called first followed by API 2. Since this part is simple call relationships without logical operations for truth and falsehood, a black edge is drawn for this part. After API 2 is called, API 3 or API 4 are called. If the conditional statement is true, API 4 will be called and if the conditional statement is false, API 3 will be called. As such, the results of logical operations according to truth and falsehood are indicated by green and red edges in the present study. Finally, an edge in green that indicates truth is drawn from API 2 to API 4 and an edge in red that indicates falsehood is drawn from API 2 to API 3.

3.2 Analysis of the Flows of Malicious Behaviors Utilizing BFS Algorithm

From the malware reverse assembled using IDA [9], character strings including prefixes such as sub_ and loc_ that indicate sub routines or instructions are used at BFS algorithm search vertexes. That is, if a branch point including a prefix appears after a compare instruction such as cmp and test or a logical operation instruction such as xor appeared, the branch point will be enQueued in the Queue. When a branch point enQueued according to a BFS algorithm has been deQueued, the deQueued vertex is searched. If any prefix such as sub_ or loc_ is found in the process of searching, the enQueueing is repeated and this process is repeated until there is no vertex left. Under pseudo codes explain BFS algorithm which is applied our proposal.

```
Target BFS(Target function)
{
 if(isExitCompare() == true)
   if(isExitAddress() == true)
    enQueue(function);
 return deQueue();
}
```

3.3 API Collection and Categorization

The entire flows of reverse engineered malware behaviors were collected utilizing BFS algorithms. However, the vertexes used in visualization in the present study are used as APIs. In this section, APIs included in sub routines or instructions collected.

Since the types of APIs used in Windows are extremely diverse, the APIs should be categorized. In the present study, individual APIs classified into upper categories referring to MSDN [10] (Table 1).

Table 1. 26 abstraction categories presented in behavior graph based malware variant detection

FILE_CREATE	FILE_OPEN	FILE_READ	LIBRARY
SERVICE_OPEN	SERVICE_ACCESS	SERVICE_CLOSE	STRING
NETWORK-AND-SOCKET_OPEN	NETWORK-AND-SOCKET_ACCESS	NETWORK-AND-SOCKET_CLOSE	MUTEX
PROCESS_OPEN	PROCESS_ACCESS	PROCESS_READ	DEBUGGING
MEMORY_READ	MEMORY_ACCESS	FILE_ACCESS	SHELLL-AND-CONSOLE
REGEDIT_OPEN	REGEDIT_READ	REGEDIT_ACCESS	WINDOWS-GUI-AND-BITMAP
TIME-AND-THREAD	SYSTEM-INFORMATION		

3.4 Calculation of Similarity Utilizing SSIM (Structural Similarity Image Metric)

In the present study, a method of measuring similarity based on malware behavior graph images is presented. Here, SSIM (Structural Similarity Image Metric) [11] is utilized as an algorithm for measurement of similarity. SSIM is a method of calculating similarity between two images and is calculated by windows between two images. The resultant values of SSIM closer to 1 indicate higher levels of similarity. Before experimenting the method presented in the present study, the similarity of test images was verified.

4 Experiment and Result

4.1 Detection of Those Pieces of Malware that Bypass Dynamic Analysis

First, to detect those pieces of malware that bypass dynamic analysis, those APIs that are mainly used by the relevant pieces of malware should be detected and those pieces of malware that use the relevant APIs should be classified. To this end, among more than 200 pieces of malware, those in the categories thread and time that are mainly used by time triggers were mainly checked.

Figure 2 shows the behaviors of Backdoor/W32.Agent.25088.AF drawn into a graph. Whereas this graph does not use functions included in categories thread and time, Fig. 3 that visualized the behaviors of Backdoor/W32.Way.700928 uses functions included in categories thread and time and the fact that edges are drawn between the two categories indicates that mutual relationships exist between the functions in the categories. As shown in Fig. 4, malware can use functions in the category thread or time for other purposes than the purpose of dynamic analysis bypassing. However, edges are drawn between the two categories because there are causal relationships for

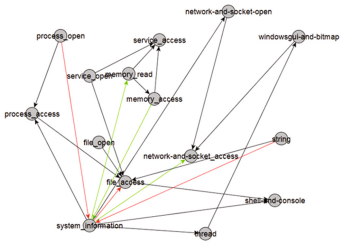


Fig. 2. Backdoor/W32.Agent.25088.AF behavior visualization

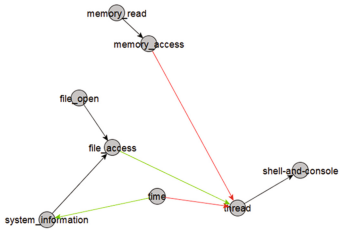


Fig. 3. Backdoor/W32.Way.00928 behavior visualization

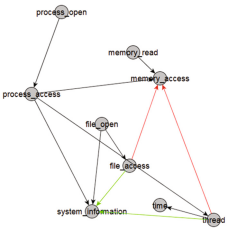


Fig. 4. Trojan/W32.Dialer.11296 behavior visualization

calls of the functions for time triggers between the categories time and thread and the edges that connect the two categories should have not been drawn if the foregoing had not been the case.

4.2 Similarity Analysis and the Results

In the present study, those pieces of malware that bypass dynamic analysis were detected and similarity between those pieces of malware was calculated. To this end, those pieces of malware that included edges drawn between thread and time were extracted and as a result three pieces of malware were extracted from among 200 pieces of malware. In the present study, similarity was calculated utilizing SSIM algorithms. Since SSIMs are calculated based on windows between two images, individual images were fixed with 950 * 950 sized PNG extensions and white backgrounds were used to enhance the similarity of graphs.

Using Table 2 and time and thread functions, at least 70% similarity could be identified among those pieces of malware that bypass dynamic analysis and this proves that even different types of malware conduct similar behaviors based on the commonality of functions used by them.

Table 2. Dynamic analysis bypassing malware similarity analysis

Malware	Backdoor/W32.Way.700928	Trojan/W32.Dialer.11296	Trojan/W32.Dialer.26973
Backdoor/W32.Way.700928	1	0.7183939	0.7745499
Trojan/W32.Dialer.11296	0.7183939	1	0.7627291
Trojan/W32.Dialer.26973	0.7745499	0.7627291	1

5 Conclusion

In the present study, a method of detecting those pieces of malware that bypass dynamic analysis utilizing static analysis and measuring similarity among such pieces of malware was proposed. Although static analysis is a major analysis method, it enables observing the flow of behaviors even without executing malware because it uses BFS algorithms to observe the flow of behaviors. In addition, this method derives common features from those pieces of malware that conduct malicious behaviors when a certain time has been reached such as time triggers so that those pieces of malware that make dynamic analysis can be efficiently detected. Furthermore, a new automated analysis classification method was presented that visualizes the behaviors of those pieces of malware that bypass dynamic analysis to calculate similarity.

References

1. Willems, C., Holz, T., Freiling, F.: Toward automated dynamic malware analysis using CWSandbox. IEEE Secur. Priv. **5**(2), 32–39 (2007)
2. Wu, D.J., Mao, C.H.: Malware detection based on mining API calls. In: Proceedings of the ACM Symposium on Applied Computing, SAC 2010, pp. 1020–1025 (2010)
3. Daryabar, F., Dehghantanha, A., Udzir, N.I.: Investigation of bypassing malware defences and malware detections. In: 7th International Conference on Information Assurance and Security (IAS) (2011)
4. Massacci, F., Wallach, D., Zannone, N.: Idea: opcode-sequence-based malware detection. In: Engineering Secure Software and Systems. LNCS, vol. 5965, pp. 35–43 (2010)
5. Jang, E.-G., Lee, S.J., Le, J.I.: A study on similarity comparison for file DNA-based metamorphic malware detection. J. Korea Soc. Comput. Inf. **19**(1), 85–94 (2014)
6. Santos, I., Penya, Y., Devesa, J., Bringas, P.: N-Grams-based file signatures for malware detection. In: International Conference on Enterprise Information Systems (ICEIS), vol. AIDSS, pp. 317–320 (2009)
7. Karnik, A., Goswami, S., Guha, R.: Detecting obfuscated viruses using cosine similarity analysis. In: The Proceeding of IEEE Symposium First International Conference on Modelling and Simulation (2007)
8. Jang, J., Brumley, D., Venkataraman, S.: BitShred feature hashing malware for scalable triage and semantic analysis. In: 18th ACM Conference on Computer and Communications Security, CCS 2011, pp. 309–320 (2011)
9. IDA. https://www.hex-rays.com/products/ida/
10. MSDN. https://msdn.microsoft.com/en-us/
11. 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 (2001). doi:10.1109/TIP.2003.819861. ISSN 1057–7149

Mongolian Internet Consumers' Attitude Towards Web Advertising

Bulganmaa Togookhuu, Junxing Zhang(✉), and Wuyungerile Li

School of Computer Science, Inner Mongolia University,
Inner Mongolia, Hohhot, China
2493308520@qq.com, junxing@imu.edu.cn

Abstract. Web advertising is being given various opportunities to individuals and business enterprises. Web advertising has following advantages such as distributed directly announcing or data and marketing events without barriers of time and of location to the customers.

Aim of this study is to emphasis the facing problems, interrelation and attitudes of internet customers of Mongolia at websites of advertisements. This study based on survey of customers and 500 customers of internet participated at this survey. As a result of this study, information feature of web advertising and customers' attitudes at website of advertisement are positive and impact the deepest influence. As a studying their real consumption and customers' attitudes at web advertising, communication of customers and service providers will be improve and can exchange safety information each other.

Keywords: Web advertising · Mongolia internet customers' attitudes · Statistic significant

1 Introduction

Mackenzie et al. researcher defined that attitudes at advertisement is "preliminary attitudes for positive or negative reacting at interaction of any advertisement". The deepest relationship has between positive and negative attitudes at the advertisement and customers' assessment for the advertisement. Ducoffe [1] processed forecast structure and designs for customers' attitudes at advertisement and their value, and they proved that attitudes at web advertising directly relevant from value of the advertisement. Bracket and Carr [2] proved further confirmation on Ducoffe designs and additionally entered factor of demography in frame of the issue. Gender of population or their degree of education is only impacting at attitudes of advertisement. We have changed factors of designs for attitudes of advertisement according to our study and as a result of the change we could prove as exciting customers' desire for using web advertising, there sources real consumption because of increasing information feature. Pervious conditions for attitudes of web advertising are like followings.

Desire Using Web Advertising: Individual good or bad attitudes at certain objects are his/her preliminary attitudes. In other words, desire using is probability censorship of customers' purchase for service or goods. In order to define a real measure of frequency

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_90

and it's time for new consumption, accepted a model of technology [3] is focused on real consumption. However, it is impossible to real measure for a desire for certain characteristics of an individual. Therefore there should be studied how to the deepest relationship between desire using web advertising and satisfaction.

Information Feature: Advertisement has the main role for distributing information to customers. Information feature of a website of any organizations directly distributes about concepts of information of the organization, its products or services to their customers. Therefore information of website should be clear or useful and distribute to customers on time and these requirements can source favorable impression to the customers. Brackett et al. [2] considered that if announcing quality of advertisement is good, it will be more favorable impression for customers or their need and it will positively impact their purchase.

Pleasure Condition: Used words or phrases, its stylistics and its picture for posting any advertising should be induced favorable impression for customers and it will positively impact to the customers. Munusamy et al. [4] considered that the experience of advertising could be a pleasure upon exposure or in recollection. Tsang et al. [5] reported that the favorable/pleasure factor is considerably important in effective Web advertising.

Economic Conditions: The condition of economics reflects the viewpoints that advertising accelerates the acceptance of new products, fosters full employment, lowers the average cost of production, promotes a healthy competition between producers to all consumers' benefit, and raises the average standard of living. If the customers have positive attitudes at web advertising, it will impact in the improvement of economic condition in the country.

The negative characteristics or depression condition: to be negative feelings or non-respectable attitudes of advertising mainly impact in the customers' attitudes at advertising.

Factors of Demography: Two variables of demography such as education and age directly impact attitudes at web advertising according to the study of Bracket and Carr [2].

Age: Youths please more favorable attention at common advertising. Although advertising should be for middle-aged people, they assess more carefully at advertising. Therefore the age is being the main factor for predicting attitudes of internet customers. Considering these factors, youths may positively assess at web advertising.

Gender: There is the study about different of men and women attitudes at web advertising. Generally, men assess more positive than women at web advertising.

Education: People with lower education or standard of living assess more favorable attention at web advertising. Different of education and sex directly impact attitudes at the website of advertising. However, there is no enough evidence or information about this concept. Therefore the experiment or study should be done about factors of demography impacting at attitudes.

2 Related Work

According to Abdul et al. [6], they indicated that the level and quality of information provided in an advertisement is an important factor that impacts customers' perception of the company and the company's products. Therefore, the company should ensure that the information delivered to the public through online advertising is sufficiently informative. Economic conditions reflect market circumstances in the environment with respect to inflation, interest rate and price movement. Consumers tend to have positive attitudes towards advertising if they believe that advertising can aid the economic condition of the marketplace. The emergence of globalization has created more dimensions for researchers to explore the role of advertising in products and services performance including consumers' attitude [7]. According to Arnett [8], young adults have the power to make decisions. Advertising is a vital medium for them to keep up-to-date with latest products or services.

Chan [9] commented that youth market is important to the marketer because of their purchasing power including those of their parents.

Beard [10] found that college students believe strongly that advertising can cause people to buy things that they should otherwise not buy.

3 Hypotheses and Model of the Study

We have formulated the structure of the model for this study based on theories and studies have been written about attitudes at web advertising. We have shown about its trends, their interrelation of real consumption and factors impacting attitudes at web advertising in the model. We have started following hypothesizes according to each interrelation (Fig. 1).

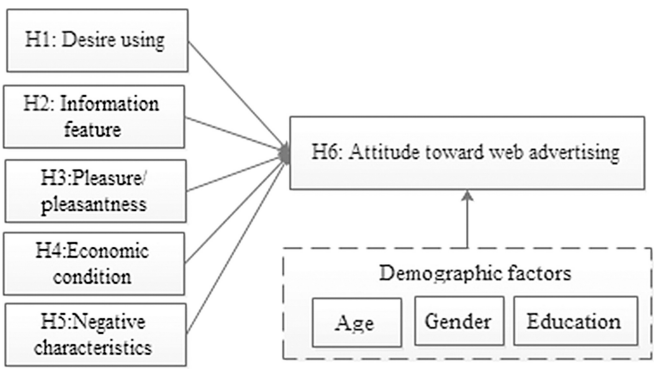


Fig. 1. Interrelation model of factors and attitudes at web advertising

Factors in the model: we have formulated following predictions in order to prove interrelation of the variables.

H1(+): in case, there is more desire using at web advertising, it will positively impact at their customers' attitudes.

H2(+): in case, there is more information feature at web advertising, it will positively impact at their customers' attitudes.

H3(+): in case, there is more pleasantness for a condition of service and products, it will positively impact at their customers' attitudes.
H4(+): in case, there is the more flexible condition of economics at web advertising, it will positively impact at their customers' attitudes.
H5(−): in case, there is the more negative feature at web advertising, it will negatively impact at their customers' attitudes.
H6(+): in case, the customers' attitudes at web advertising are more positive, real consumption will be more increase.

4 Research Methodology

Research design is to specify the methods and procedures for data collecting and data analysing. Then, quantitative research and descriptive research are being used in this study [6]. Quantitative research is to quantify the data as well as the findings or conclusion evidence. In other words, quantitative is used of numerical data in data collection technique and data analysis procedure. In this study, the purpose of confirming the quantitative research is to test the hypothesis from the questionnaires. Descriptive research is widely used research design to gather the information about the present existing conditions by the use of questionnaire, personal interviews and observation.

Aim of this study is to emphasis how to assess internet customers of Mongolia at websites of advertising detecting the facing problems and factors or conditions for implementing successfully web advertising.

4.1 Questionnaire Design and Data Collection Method

In our survey, total 500 internet customers participated and only 217 questions had right or full answers, so we have used these 217 answers in order to make the conclusion of this study. We organized the action for data of the study by the survey measure for the customers. Survey is more flexible than other methods and easy for analyzing data of study.

The survey consists of A and B sections. In A section, there are general questions about the participants in the survey and in B section, there are variables (questions) expressing their attitudes at web advertisings. Each variable consists of answers of 2–5 questions. And each answer of the question has been expressed by Likers measure of 1–5 points. Including: 1 - Strongly disagrees, 5 - Strongly agree at the highest level.

4.2 Data Analysis and Its Results

We have made data analyzes using Statistical Package for Social Sciences (SPSS) program. In order to use this program, we have used following methods; statistical method (common methods), reliability analyzes, correlation analyzes and linear regression analyzes. We showed answers of questions for demography factors.

There are 53% men and 38.1% women from total 217 people of the survey and men mainly assess at website of advertising than women. 36.8% from total participants in the survey are regular customers continuously using internet in recent seven years. In the survey, participated more people with 19–29 ages (40%) and less people with above age than 41 age (5%). Students are 42.6% and people with bachelor degree or higher education is 39.9% of total participants in the survey. In other words, in the survey, students, who have no permanent income or job, have mainly been participated. As a result of this indication, youths are 60% of Mongolian population and they have actively participated in the online survey. Number of internet customers' assessing at websites of advertising in a month is 11 times (27.7%).

4.3 Averages of Factors and Reliability Analysis of Data

We accumulated average indication of answers of the questions for defining the factors using Likers measure of 1–5 points. In case, an average indication of any factors is higher, it is expressed that the customers' proposals meeting with the feature. As a reliability analysis of data is expressed whether high interrelation of the factors or use this indication in the analysis. To be Cronbach's α is expressed correctly contain questions in the survey and is considered to possible analyze in the factors. Experts defining; if α is more than 0.6, it can be concluded reliability data and can analyze data of the survey or to formulate based on correct questions.

Table 1. Averages and Cronbach's (α) on the factors

Factors	Mean	Std. deviation	Item (N)	Cronbach's (α)
Desire using	3.41	2.73	4	0.63
Information feature	3.31	3.044	3	0.68
Pleasure/ Pleasantness	3.39	1.83		0.33
Economic condition	3.52	2.32	3	0.61
Negative characteristics	3.52	2.45	3	0.62
	3.48	2.30	3	0.49

We showed averages of all variables, Std. Deviation, numbers of the item and Cronbach's Alpha in Table 1. Because of more means than three of an average of all variables, its difficulties = 3.48, rick = 3.52, economic condition = 3.52, pleasantness = 3.39, information feature = 3.31 as well as using desire = 3.41 for the independent variables, recognizing questions of the survey will be examined further. In this study, we showed the independent or the dependent variables using Cronbach's Alpha. Cronbach stated that a value higher than 0.50 was a satisfactory level of good internal consistency, Therefore, the reliability analysis for all variables in this study (except = diff) are still acceptable.

It can be showed matrix relationship of variables as an indication of Table 2. A two-tail test at 0.05 possible levels indicates that there are positive relationships between the independent variables and the dependent variable.

UW → AWA: using desire is positive statistically significant at attitudes of web advertising.

INF → AWA: information feature is being positively impacted at web advertising.

Table 2. Pearson correlation coefficient matrixes

	INF	PLE	Eco	UW	Risk	Diff	ATT
INF	1						
PLE	.333**	1					
Eco	.100	.415**	1				
UW	−.039	−.215**	.050	1			
Risk	−.196**	−.249**	−.023	.065	1		
Diff	−.045	.184**	.342**	.235**	−.005	1	
ATT	.398**	.474**	.671**	.383**	.263**	.615**	1

PLE → AWA: a favorable condition of website impacts positive attitudes at web advertising according to the result of this study. Therefore, there should be used proper words or phrases for posting advertising and paid on the graphic density of pictures.

ECO → AWA: economic condition of customers is almost independent significant at the website of advertising.

Risk → AWA: the facing rick is positive significant to attitudes at the website of advertising; therefore there should be prevented conditions of risk.

Diff → AWA: difficulties are no positive significant to attitudes at the website of advertising.

According to showing results of regression analysis, we have analyzed regression analysis on the independent (UW, INF, PLE, ECO, RISK, DIFF) variables segmenting ages of demography by the dependent variables and we have considered that relationship of the factors is mutual significant because of less Sig(P) $\leq$ 0.05 of all variables. And we have analyzed regression analysis segmenting level of education, only information feature has been confirmed but all variables of other factors have not been confirmed. As a result of the indication, people with high of educational level positively assess at web advertising and their real consumption is more. Total participants in the survey to segment by sex, facing rick is positive statistic relationship to attitudes at web advertising but other factors are more than P $\geq$ 0.05 and there is being considered no statistic significant.

5 Conclusion

Therefore positive impacting at the website of advertising has been confirmed according to this study. As result of this study, age is one of impacting factors that was confirmed to be positive attitudes at the website of advertising. Also, information feature had the highest indication and it should be mainly paid attention to this issue. Pleasantness and negative feature had the worst attitudes and also it should be mainly improved this direction. Internet customers of Mongolia have a higher desire for becoming real consumptions using web advertising. Therefore there should be paid attention to issues such as mainly improve contains or positive features of advertising and the consumers' attitudes using this condition.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (Grant No.: 61261019), the Inner Mongolia Autonomous Region Natural Science Foundation (Grant No.: 113113 and Grant No.: 2014BS0603), and the Program of Higher-level talents of Inner Mongolia University.

References

1. Ducoffe, R.H.: Advertising value and advertising on the web. J. Advertising Res. **36**(5), 21–35 (1996)
2. Brackett, L.K., et al.: Cyberspace advertising vs. other media: consumer vs. mature student attitudes. J. Advertising Res. **41**(5), 23–32 (2001)
3. Davis, F.D., et al.: Perceived usefulness, perceived ease of use and user acceptance of information technology. MIS Q. **13**, 319–340 (1989)
4. Munusamy, J., et al.: Attitude towards advertising among students at private higher learning institutions in Selangor. J. UNITAR **3**(1), 31–55 (2007)
5. Tsang, M., et al.: Consumer attitude toward mobile advertising: an empirical study. Int. J. Electron. Commer. **8**(3), 65–78 (2004)
6. Azeem, A., et al.: Perception towards internet advertising: a study with reference to three different demographic groups. Glob. Bus. Manag. Res. Int. J. **4**(1), 28–45 (2012)
7. Kanso, A., et al.: Multinational corporations and the challenge of global advertising: what do US headquarters consider important in making media-selection decisions? Int. Mark. Rev. **24**(5), 563–590 (2007)
8. Arnett, J.J.: Emerging adulthood: a theory of development from the late teens through the twenties. Am. Psychol. **55**, 469–480 (2000)
9. George, T., et al.: Persistent impacts of West Nile virus on North American bird populations. Proc. Natl. Acad. Sci. U.S.A. **112**(46), 14290–14294 (2015)
10. Beard, F.K.: College student attitudes toward advertising's ethical, economic, and social consequences. J. Bus. Ethics **48**, 217–228 (2003)

A Study on the Quantitative Design and Documentation for the VR Based Training Contents Realism

Gyungchang Lee[1,2], Kyoil Chung[3], and Cheong Youn[2(✉)]

[1] Ptolemy System Corp., Daejeon, Republic of Korea
gchang.lee@gmail.com
[2] Department of Computer Science and Engineering, Chungnam National University, Daejeon, Republic of Korea
cyoun@cnu.ac.kr
[3] Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea
kyoil@etri.re.kr

Abstract. The serious game which is utilized for educational training has been operated using I/O (Input/Output) devices like keyboard, mouse and monitor. Recently, serious game is advancing towards VR (Virtual Reality) based training system by providing interaction functions between virtual visualization environment and real action of trainee by wearable devices, such as motion recognition system and HMD (Head Mounted Display). For acceptance test of VR based training system, hardware testing is possible by using test metrics which can be described on its specification. But VR based training contents as a software including virtual visualization environment has many difficulties for testing due to lack of quantitative test metrics about realism which means similarity between real environment and its virtualized environment. In this study, it is suggested quantitative design and documentation methodology of VR based training contents using test models derived from ISO/IEC 25010:2001 SQuaRE (System and software Quality Requirements and Evaluation). The suggested quantitative design document has been experimented through the development of a VR based plant safety training system and has verified its usefulness as training contents test metrics for testing realism which is an emotional factor.

Keywords: Virtual reality · Training system · 3D contents · Serious game

1 Introduction

The training system has purpose of operator training, such as mastering the control skill of target system, preventing potential malfunction, and handling an accident, etc. A development of "America's Army" in 2002 was presented as a successful case of applying a serious game to training system by enhanced graphic visualization [1]. Recently, interaction between training contents of serious game and real action of trainee became possible from using wearable devices instead of keyboard and

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_91

mouse [2]. As a result, recent serious game is advancing to VR based training system offering higher training fidelity. The VR (Virtual Reality) based training system consists of hardware such as HMD (Head Mounted Display), motion recognition system, etc. and software that is training contents providing visual experience similar to real environment [3]. The training contents is built from integration of 3D modeling to be generated by computer graphic and function software that is implemented by coding. In acceptance test to deliver VR based training system to customer, hardware testing is possible according to quantitative performance requirements specified in development specification. But software in other words "training contents" is very difficult to test because it requires verification of realism that is similarity between real environment and its computer graphic visualization. Realism is difficult to test because it is an emotional factor that is difficult to describe in the development specification with quantitative design.

This research shows quantitative design and documentation methodology of VR based training contents using test model derived from ISO/IEC 25010:2011 SQuaRE (System and software Quality Requirements and Evaluation).

2 Previous Researches

Viknashvaran et al. [4] distinguished game, serious game and training simulator at earlier research, and defined serious game as "applications of interactive technology that extend far beyond the traditional video-game market, including: training, policy exploration, analytics, visualizations, simulation, education, and health and therapy."

The VR based training system has similar features with a training simulator but it should be classified as an advanced serious game because the most important feature of it is computer generated graphic contents. Hardware like sensors for trainee's motion recognition is using for improving training fidelity instead of keyboard/mouse or joystick on traditional serious game. By this reason, the development process for VR based training system should be considered to be improved the game contents development process rather than training simulator development process. VR based training contents as a software should be considered its unique characteristics while testing.

Hoberge [5] distinguishes unique characteristics for game contents testing as following Fig. 1 although there are similarities in testing with general software.

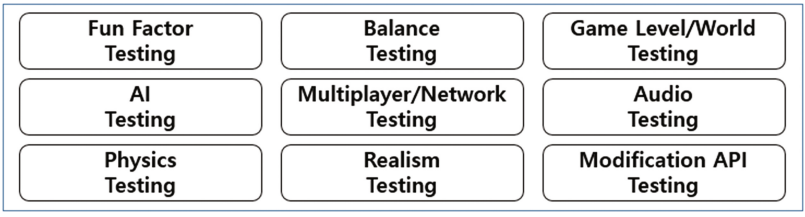


Fig. 1. Unique characteristics of contents testing

Malon emphasized the importance of fun factor on the entertainment game contents in previous research to increase motivation [6, 7]. Other researches related to test criteria and methods about entertainment game contents, J.Y. Park et al. [8] suggested the quality model on ISO/IEC 9126 'Software engineering - Product quality'. Rido et al. [9] presented a test method for quality measurement of entertainment game contents by using ISO/IEC 25010:2001 SQuaRE.

On the other hand, VR based training contents has different testing characteristics with entertainment game contents. Realism for computer generated visualization environment is most important characteristic on training contents therefore it requires quantitative test metrics because objective testing of emotional factor is difficult.

3 Training Contents Test Model and Design Document

In order to develop the VR based training contents considering unique characteristics testing, this research suggests test models referring to quality in use model and product quality model on ISO/IEC 20010:2011 SQuaRE [10] as Tables 1 and 2.

Table 1. Training contents test model referring to quality in use model characteristics

Quality characteristics	Test model
Effectiveness	Does VR environment provide similar training effect with real environment?
Efficiency	Is it operated without unnecessary control?
Satisfaction	Does VR look like real? And does it provide immersion without inconvenience?
Freedom from risk	Is it safe to use?
Context coverage	Is it possible to provide a comprehensive performance about quality factor for usage and response flexibly?

Table 2. Training contents test model referring to product quality model characteristics

Quality characteristics	Test model
Functional suitability	Are all requested functions for operating training implemented accurately?
Performance efficiency	Does it satisfy requested performance including the speed of response and the ratio of video playing on regulated hardware condition?
Compatibility	Does it run with other software without any collision on regulated system and interact correctly with other required system?
Usability	Does it easy and convenient to use?
Reliability	When an error occurs, is it possible to maintain the defined performance or recover back to the defined state?
Security	Does it provide requested usage restriction and any arbitrary change restriction?
Maintainability	Is it implemented as a convenient structure for correcting the error and upgrading?
Portability	Is it installed on regulated hardware and possible to adapt to the system upgrade?

Applying above suggested test models, functional and performance requirements analysis result for VR based training contents are described to SRS (Software Requirements Specification). The SRS describes detailed implementation requirements of software capabilities and related virtual visualization environment defined in SSS (System and Subsystem Specification). After that, the design results to be implemented as coding are described to SDD (Software Design Document) from functional requirements specified in SRS. But the SDD cannot describe design results to meets whole requirements of SRS because it has requirements for visualization environment similar to real. So, the visualization environment design results shall be described including below items to CDD (Contents Design Document) which is suggested uniquely on this research for representing suggested test model above.

Modeling Levels: The computer generated virtual environment shall reflect most detail about real world to increase training fidelity but there is limitation because of computing resource. The modeling levels describe modeling scope considering limited computing resources. Define the modeling area range as first level and the detail object as final level, which is top-down method.

Modeling Features: The feature of VR based training contents to be generated by computer graphic is virtual visualization of real world. Modeling features show source configuration what is referenced for computer graphic modeling. It will be used as a comparison of similarity with real world. Modeling features for each modeling level is traceable with the SRS to cover remained requirements which are not traceable by the SDD.

Modeling Properties: Modeling properties show quantitative properties of static or dynamic modeling object which is presented in modeling features. It describes the number of polygon for modeling, size of texture, and the material for shader as quantitative design metrics. It also describes existence or none of interaction by trainee's action like collision mesh, graphic effect, animation and sound effect.

4 Experiment and Analysis

What this research suggested is experimented by developing training contents of VR based plant safety training system like Fig. 2.

Fig. 2. VR based plant safety training system and contents

The SRS is written through requirements analysis about virtual plant environment and interaction functions of trainee's action like valve control, and the propriety of software requirement analysis is validated through traceability between the SSS.

Then software architecture and function design are written to the SDD, and propriety of it is validated through traceability between the SRS.

According to suggestion of this research, CDD was written as quantitative design results for 3D modeling considering limited computing resource of VR based plant safety training system. Table 3 shows an example description of modeling levels, which could be used to describe the quantitative limits for the modeling scope. A photo of the plant equipment is presented as modeling features for each modeling level with its requirement traceability like example of Table 4. Modeling features were described as design references for computer graphic modeling and could be used as a comparison of realism in training contents. Properties of modeling object is defined quantitatively as modeling properties like example of Table 5 and reference of each property is described like example of Table 6. Modeling properties were described as design results for modeling considering limited computing resource and could be used as quantitative test metrics for training contents testing.

Table 3. Modelling levels description on CDD (example)

Level 1	Level 2	Level 3	Level 4
1. 100 m radius area around gas plant building	1. Gas plant	1. Building	1. Gas plant building
			2. Electric & communication room
		2. Facility	1. High pressure gas pipe
			2. High pressure gas backup pipe
			3. Analog pressure gauge
			4. Digital pressure gauge
	2. Route	1. Main road	1. Vehicle access
			2. Parking lot

Table 4. Modelling features description on CDD (example)

Level	Features	Requirement traceability
1.1.2.3 Analog pressure gauge		[SFR-101], [SFR-102], [SFR-111]

Table 5. Modelling properties description on CDD (example)

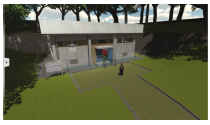
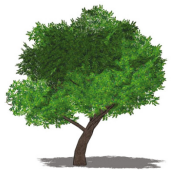
Object	Reference	POL	TEX	MAT	COL	EFF	ANI	SND
1.1.2.3 Gauge	Photo	L	M	2, 4	Y	3	1, 2	N
1.1.2.5 Lever	Photo	L	L	2	Y	3	3, 4	Y
1.1.1.5 Tree 1	Arbitrary	M	H	3	Y	–	–	N

Table 6. Modelling property reference table on CDD (example)

Item	Criteria	Description	Reference Index
POL	H/M/L	No. of Polygon	H(>3K), M(1K–3K), L(<1K)
TEX	H/M/L	Texture Image Size	H(>1,024K), M(>256K), L(>64K)
MAT	1/2/3/4	Material for Shader	1. Stone, 2. Steel, 3. Wood, 4. Glass
COL	Y/N	Collision Mesh	Collision Mesh Existence and Nonexistence (Y/N)
EFF	1/2/3	Graphic Effect	1. Oil Leak, 2. Gas Leak, 3. Explosion
ANI	1/2/3/4	Animation	1. Right Rotate, 2. Left Rotate, 3. Up, 4. Down
SND	Y/N	Sound Effect	Sound Effect Existence and Nonexistence (Y/N)

There were difficulties to test VR based training contents due to lack of test metrics that could be described quantitatively. The similarity between real plant facility and virtual plant, in other word 'realism' was easily tested exampled below Table 7 from using the CDD which described quantitative test metrics as contests test results. After then, the validation of VR based training contents was completed by testing that it meets the software functional requirements described in SRS.

Table 7. CDD descriptions and its tested VR based training contents (example)

Modeling level	Modeling feature	Modeling property		Quantitative test results & contents scene	
LV1: 100 m radius area around gas plant building LV2: gas plant LV3: building LV4: gas plant building		POL TEX MAT COL EFF ANI SND	M H 1 Y None None N	1,592POL 3,792 KB 1 Y None None N	
LV1: 100 m radius area around gas plant building LV2: gas plant LV3: facility LV4: analog pressure gage		POL TEX MAT COL EFF ANI SND	L M 2,4 Y 3 1,2 N	438POL 684 K 2,4 Y 3 1,2 N	
LV1: 100 m radius area around gas plant building LV2: gas plant LV3: building LV4: tree1		POL TEX MAT COL EFF ANI SND	M H 3 Y None None N	2,364POL 3,946 KB 3 Y None None N	

This experiment described quantitative design about VR environment to CDD and it was useful by using it as test metrics to validate realism that is characteristic for effectiveness, satisfaction and functional suitability in suggested test model.

5 Conclusion

VR based training system is similar to a simulator as its feature but it should be classified as advanced serious game because major characteristic of it is similar to entertainment game contents. This research showed test model for training contents with using ISO/IEC 25010:2001 SQuaRE. And also it showed what content shall be included in SRS, SDD and CDD with using suggested test model. After that, suggested methods were experimented through developing a VR based plant safety training system. As a result, especially the CDD that is described quantitative design metrics for computer graphic modeling was verified of its usefulness that makes possible to test realism of VR based training contents.

The market related to VR is dramatically growing, and the various development and diffusion of contents providing visual experience similar to real environment are expected. There is a plan to continue research the general way of quantifying realism which is emotional factor of training contents and the standards that describe them.

Acknowledgments. This research was supported by a grant (10048341) from industry core technology development project funded by Ministry of Trade Industry and Energy (MOTIE) of Korea government.

References

1. Neiborg, D.B.: America's Army: more than a game? In: Conference Proceedings of International Simulation and Gaming Association (ISAGA 2004) (2004)
2. Hsiao, K.-F., Rashvand, H.F.: Integrating body language movements in augmented reality learning environment. Hum.-Centric Comput. Inf. Sci. **1**(1), 1–10 (2011). Springer, Heidelberg
3. Havakli, M.: A people-centric framework for mobile augmented reality systems (MARS) design: ArcHIVE 4Any. Hum.-Centric Comput. Inf. Sci. **5**(37), 1–21 (2015). Springer, Heidelberg
4. Narayanasamy, V., Wong, K.W., Fung, C.C., Rai, S.: Distinguishing games and simulation games from simulators. ACM Comput. Entertainment **4**(2) (2006). ART 6A. ACM, New York
5. Hoberg, J.: Software testing and game testing (2014). http://www.slideshare.het/JohanHoberg/software-testing-and-game-testing
6. Malone, T.W.: Toward a theory of intrinsically motivating instruction. Cogn. Sci. **5**(4), 333–369 (1981). Elsevier, Amsterdam
7. Christou, G.: A comparison between experienced and inexperienced video game players' perception. Hum.-Centric Comput. Inf. Sci. **3**(15), 1-15 (2013). Springer, Heidelberg
8. Park, J.-Y., Shin, S.-W., Lee,N.-Y.: A study about quality models of game software based on ISO/IEC 9126. In: Korea Society of IT Services Fall Conference, pp. 197–202 (2002)
9. Ramadan, R., Hendradjaya, B.: Development of game testing method for measuring game quality. In: 2014 International Conference on Data and Software Engineering (ICODSE). IEEE (2014)
10. ISO/IEC 25010:2011 System and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models (2011)

Progressive Motion Artifact Removal in PPG Signal for Accurate Heart Rate Estimation

Ji Hun An, Heemang Song, and Hyun-Chool Shin(✉)

Department of Electronic Engineering, Soongsil University, 156-742 Seoul, South Korea
{jh.an, shinhc}@ssu.ac.kr, oojs8@naver.com

Abstract. This paper proposes a motion artifact (MA) removal method in the photoplethysmographic (PPG) signal for accurate heart rate estimation. PPG signal is easy to acquire, but it is easily distorted by body movement. In this study, MA is analyzed using acceleration signals and removed in the PPG spectrum for accurate heart rate estimation. The proposed method progressively removes three-axis acceleration spectra in order of spectral power. The performance was confirmed by comparing heart rate estimation errors one case that MA was removed with another case that MA was not removed. After removing MA and applying two peak tracking methods in 12 data sets, the mean absolute error (MAE) of the beat per minute (BPM) is lower than conventional methods.

Keywords: Photoplethysmographic · Motion artifacts · Heart rate · 3-axis acceleration signals

1 Introduction

The photoplethysmographic (PPG) signal is a physical quantity representing the change of the blood vessel volume according to the absorbance of the light projected on the skin. Since the change in the blood vessel volume occurs by the contraction and relaxation of the heart, the heart rate can be estimated using the PPG signal [1]. If the Heart rate during exercise can be checked, immoderate exercise can be prevented. However, the PPG signal acquired from wearable devices is likely to be distorted by body movements. This motion artifact (MA) causes interference to the heart rate information in the PPG signal, which makes it difficult to estimate an accurate heart rate [1, 2].

To solve this problem, researches on the MA removal in the PPG signal have been conducted [2–10]. Typical methods for removing MA from the time domain are the independent component analysis (ICA)-utilized method [3] and the Kalman smoother method [4]. The ICA based method assumed that PPG signal and MA are independent, but it is difficult to support the assumption mentioned above because PPG signal is interfered by MA. The Kalman smoother method models MA by using sinusoidal function, but this is difficult to reflect the characteristics of irregular MA. Because of these problems, studies for removing MA not from time domain, but from frequency domain have been conducted [2, 5–7]. The MA removal method supposed by R. Krishnan [5] and Lopez [6] is effective in a situation with a weak MA, but it is difficult to estimate accurate heart rate in a situation with a strong MA. The removal

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_92

methods that can be applied in a situation with a strong MA are TROIKA [2] and SPECTRAP [7]. These methods use acceleration data to reduce MA from PPG signal. TROIKA consists of three stages such as singular spectral analysis (SSA) to remove MA, sparse signal reconstruction (SSR) to estimate spectrum, and peak tracking to estimate heart rate. However, SSR takes lots of time due to high-resolution spectral estimation, so it has a problem of estimating heart rate in real-time. MA is removed through asymmetric least squares method after obtaining spectrum by using fast fourier transform (FFT) in SPECTRAP and heart rate is estimated by using Bayesian theory.

In this paper, a new MA removal method that has smaller BPM errors is proposed. MA is a noise generated by the motion and its pattern can be identified through the acceleration signal. After analyzing the distortion caused by MA through the spectrum of the acceleration signal, MA is removed from the PPG spectrum for the accurate heart rate estimation. Based on the power of the three-axis acceleration spectra, the method for progressively removing the three-axis acceleration spectra is proposed. The MA removal efficiency was confirmed by the comparison of errors between the actual heart rate and the estimated heart rate before and after MA removal. Also, the performance was evaluated by the comparison with the results in the conventional methods [2, 7].

2 PPG Signal and Motion Artifacts

Data set used in this study consists of one channel ECG of 12 subjects, 2-channel PPG, and 3-axis acceleration signal and was obtained from zhilinzhang.com/spcup2015/data.html [11]. 12 data sets were obtained from the wearable device according to the exercise order set for about 5 min.

Figure 1 shows the relationship between PPG and acceleration signal in the time and frequency domain. It is difficult to intuitively identify how three-axis acceleration signals affect PPG in the time domain as MA becomes larger. However, the PPG spectrum includes MA information in the frequency domain. It can be confirmed that this is significantly correlated with the spectra of three-axis acceleration signals. Thus, it is appropriate to remove MA not in the time domain, but in the frequency domain.

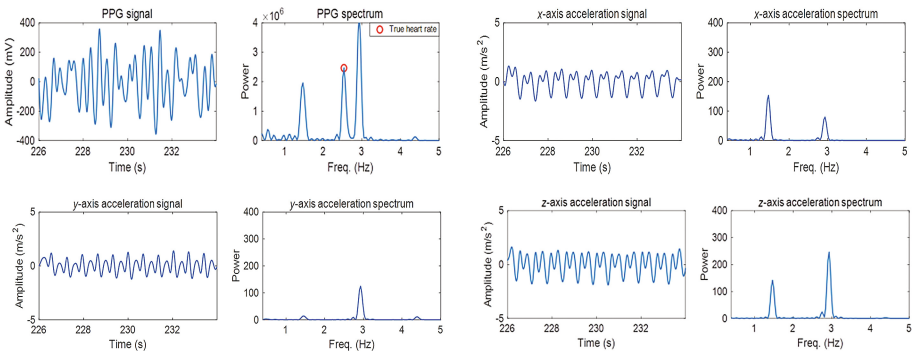


Fig. 1. PPG and 3-axis acceleration signals in the time and frequency domain. (a) PPG signal and spectrum. (b), (c) and (d) 3-axis acceleration signals and spectra.

The power of PPG spectrum in Fig. 1 is larger in the part influenced by MA than actual heart rate. If the maximum value of the PPG spectrum is estimated as heart rate, heart rate estimation error occurs. Thus, we need to remove MA components from PPG spectrum for the accurate heart rate estimation.

3 Proposed Method

Figure 2 shows the overall block diagram of the proposed MA removal method. Bandpass filter with 0.4–5 Hz passband was applied to the PPG signal and three-axis acceleration signal of i th block as a preprocessing. The reason why passband was set as 0.4–5 Hz is because it is within adult heart rate generally [2]. The PPG signal of i th block that bandpass filter was applied and three-axis acceleration signals were expressed as $p_i(n)$ and $a_i^j(n)$ $(j \in \{x, y, z\})$, respectively. We obtained the power spectra of $p_i(n)$ and $a_i^j(n)$ by using fast fourier transform (FFT). The power spectra of PPG and three-axis acceleration were expressed as $P_i^1(f)$ and $A_i^j(f)$, respectively.

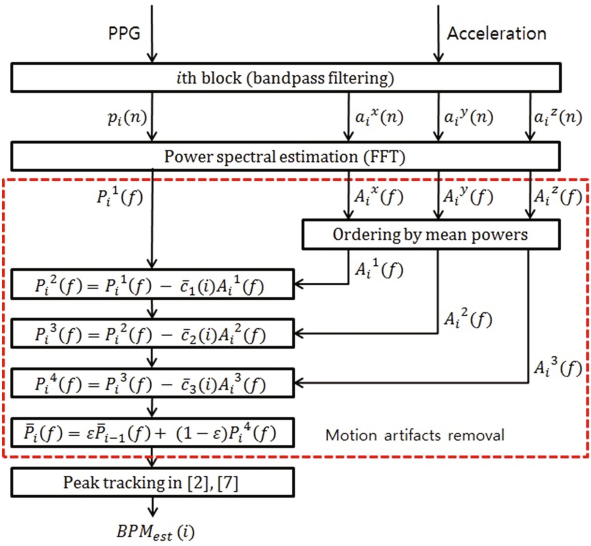


Fig. 2. Proposed MA removal method.

In order to remove MA components, a method for progressively removing acceleration spectra from low average power of MA to high average power of MA from $P_i^1(f)$ is proposed. To do this, $A_i^j(f)$ is rearranged to $A_i^k(f)$ $(k = 1, 2, 3)$ where k represents the order of average power. The greater k in $A_i^k(f)$ is, the greater average power is.

Also, if actual heart rate and principal frequency of MA are overlapped, actual heart rate information is removed. Thus, we used a constraint condition that sets the guard band of the current window by using heart rate estimated in the previous window similar to the conventional method [2], in order to keep the information of the heart rate. $P_i^k(\Delta f)$ and $A_i^k(\Delta f)$ used in the constrain condition are the spectra that set guard band of Δf based on heart rate estimated in the PPG spectrum of the previous window. The proportion of MA included in the PPG signal about each three-axis is calculated through a constrained least square as follows.

$$c_k(i) = \underset{c}{\operatorname{argmin}} \left\| P_i^k(f) - cA_i^k(\Delta f) \right\| \text{ constrained to } \left(\frac{1}{2}\max\left(P_i^k(\Delta f)\right) - cA_i^k(\Delta f) \right) \geq 0 \tag{1}$$

Then, $\bar{c}_k(i)$ is obtained by using autoregressive model about the coefficient $c_k(i)$ through (1) and $\bar{c}_k(i-1)$.

$$\bar{c}_k(i) = \mu \bar{c}_k(i-1) + (1-\mu)c_k(i) \tag{2}$$

where μ is a parameter used in an autoregressive model. The MA removal from the PPG signal is made by removing the results of multiplying $\bar{c}_k(i)$ by $A_i^k(f)$ from $P_i^k(f)$. This was expressed in (3).

$$P_i^{k+1}(f) = P_i^k(f) + \bar{c}_k(i)A_i^k(f) \tag{3}$$

where $P_i^4(f)$ is the spectrum after MA components about three-axis are removed. If $P_i^{k+1}(f)$ is less than zero, its value sets to zero.

Another problem to estimate heart rate by using $P_i^4(f)$ is the case that the power of $P_i^4(f)$ in actual heart rate is very small. In this case, it is difficult to estimate accurate heart rate even though MA is removed. As shown in (4), smoothing process is performed by assuming that there was no significant heart rate change between two consecutive windows.

$$\bar{P}_i(f) = \epsilon \bar{P}_{i-1}(f) + (1-\epsilon)P_i^4(f). \tag{4}$$

$\bar{P}_i(f)$ is the final PPG spectrum with MA removal and ϵ is a parameter controlling the degree of spectral smoothing.

Figure 3 shows examples of progressively removing MA through the proposed method. Figure 3(a) is the case where there is a significant difference between the frequency components of MA and frequency components corresponding to actual heart rate. Figure 3(b) refers to the case where distortion is given in the vicinity of the frequency information corresponding to actual heart rate. In both cases, a large error in the heart rate estimation occurs in the $P_i^1(f)$ before the MA is removed. However, it can be seen that accurate heart rate estimation is possible through the proposed gradual MA removal method.

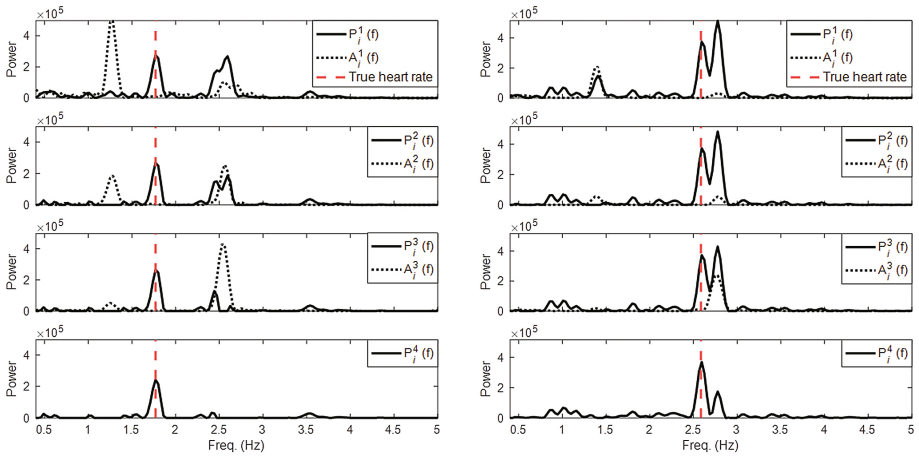


Fig. 3. Examples of MA removal in PPG spectra. (a) The case with a large difference of the frequency component corresponding to the frequency components of MA and actual heart rate. (b) The case that MA gives distortion in the vicinity of the frequency information corresponding to actual heart rate.

Also, peak tracking process is required for the heart rate estimation. Peak tracking [2, 7] plays a role of finding the peak corresponding to heart rate in the $\bar{P}_i(f)$. Peak tracking in [2] performed peak tracking by setting the harmonic relation between peaks and particular rules. Peak tracking in [7] used maximum likelihood estimation by modeling the characteristics of actual heart rate. Here, we adopted the two peak tracking methods for performance evaluation.

4 Experimental Results

Accuracy of heart rate estimation was evaluated by using PPG signal [11] of 12 subjects. The sampling frequency of each data is 125 Hz. Individual heart rate of each subject was estimated by moving the window of 8 s by 2 s. The number of FFT points is set to 4096. Δf is a parameter used in the constraint condition means guard band that is $[prevHR - 12, \ldots, prevHR + 12]$ where $prevHR$ means the heart rate estimated in previous window similar to the conventional method [2]. The parameters used in the autoregressive models are set as $\mu = 0.88$ and $\epsilon = 0.18$.

In order to evaluate the MA removal efficiency of the proposed algorithm, heart rate estimation results before and after MA removal about 12 subjects were compared with actual heart rate. Figure 4 shows heart rate estimation results and actual heart rate before and after the MA removal in 4 subjects among 12 subjects. 4 subjects were selected as subjects that have a significant difference in heart rate estimation results. In Fig. 4, it can be confirmed that there is a significant difference in actual heart rate in the case that MA is not removed.

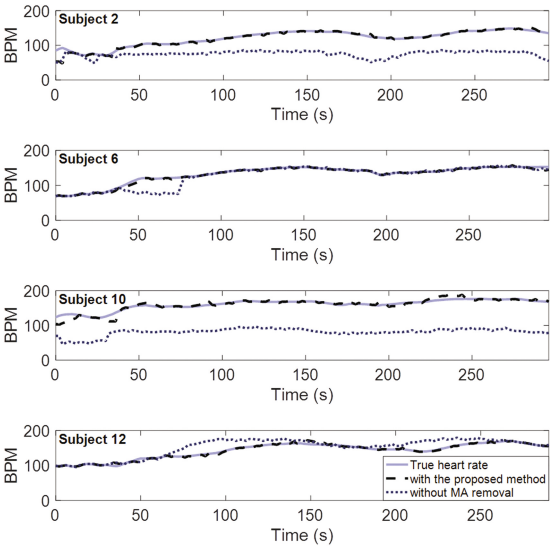


Fig. 4. BPM estimation results using peak tracking [2] with and without the proposed MA removal.

In order to quantitatively evaluate the proposed method, the mean absolute error (MAE) between actual and estimated heart rate was used. It was defined in (5).

$$MAE = \frac{1}{W}\sum_{i=1}^{W} |BPM_{est}(i) - BPM_{true}(i)| \tag{5}$$

where $BPM_{est}(i)$ represents the estimated heart rate and $BPM_{true}(i)$ means actual heart rate obtained by electrocardiogram (ECG). W refers to the width of window.

The conventional methods for removing MA in the frequency domain were compared with heart rate estimation results. Typical methods of the conventional methods are TROIKA [2] and SPECTRAP [7]. Table 1 shows the average and standard deviation of MAE for whole 12 subjects. After applying the peak tracking in [2, 7] to the proposed algorithm, each result was compared with the MAE of the conventional methods. The MAE results of the proposed algorithm are smaller than the MAE of the two other conventional algorithms and have an advantage that the results of the proposed algorithm do not vary greatly according to subjects through a small standard deviation.

Table 1. Performance comparison in terms of MAE on 12 subjects.

Subject #	1	2	3	4	5	6	7	8	9	10	11	12	Mean ± std
Proposed + peak tracking [2]	1.38	1.95	0.90	1.20	1.04	1.36	0.75	0.66	0.68	3.87	0.89	1.48	1.40 ± 0.89
TROIKA [2]	2.29	2.19	2.00	2.15	2.01	2.76	1.67	1.93	1.86	4.70	1.72	2.84	2.34 ± 0.82
Proposed + peak tracking [7]	1.32	1.07	0.98	1.09	1.16	1.12	0.67	0.96	0.77	2.21	0.77	1.60	1.14 ± 0.42
SPECTRAP [7]	1.18	2.42	0.86	1.38	0.92	1.37	1.53	0.64	0.60	3.65	0.92	1.25	1.50 ± 1.95

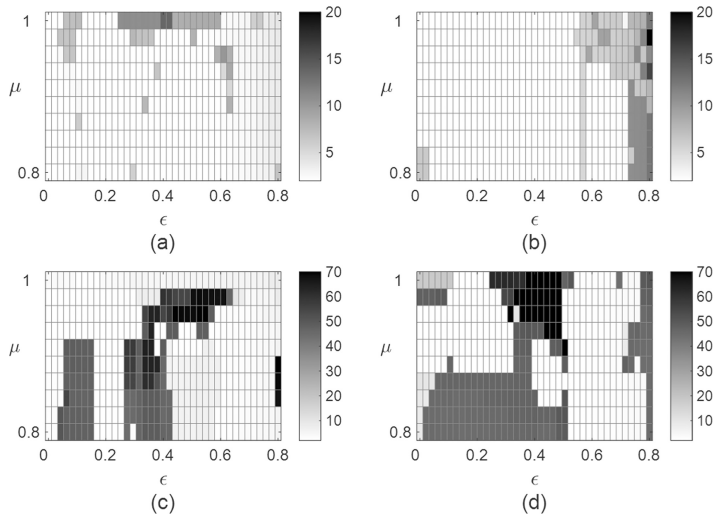


Fig. 5. MAE of BPM depending on the change of parameters μ, ϵ. (a) and (b) Average MAE in 11 subjects who are stable in the parameter change. (c) and (d) MAE of subject 1 and subject 2 who are sensitive to parameter change. (a) and (c) refers to the case that peak tracking in [2] is used and (b) and (d) refers to the case that peak tracking in [7] is used.

In order to investigate the performance variation on the parameters in the proposed algorithm, MAE was, as shown in Fig. 5, calculated by changing μ from 0.8 to 1 and ϵ from 0 to 0.8. The results that peak tracking in [2] was applied to the proposed algorithm were (a) and (c) and the results that peak tracking in [7] was applied were (b) and (d). (a) and (b) represent the average MAE of 11 subjects that performance changes were not large according to parameters. (c) and (d) represent the results of subject 1 and 2 who had a large MAE in each parameter change. However, as shown in (a)–(b), it was verified that the proposed method also showed stable performance even in the parameter change, except for two subjects showing a large MAE.

5 Conclusion

In this paper, a method for progressively removing MA in PPG signal for the estimation of the accurate heart rate was proposed. The MA removal method is performed in the frequency domain and investigates the correlation between the PPG signal and MA. The better performance of the proposed algorithm was demonstrated by comparing the proposed algorithm with the existing methods in terms of MAE.

References

1. Allen, J.: Photoplethysmography and its application in clinical physiological measurement. Physiol. Meas. **28**(3), R1–R39 (2007)
2. Zhang, Z., Pi, Z., Liu, B.: TROIKA: a general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Trans. Biomed. Eng. **62**(2), 522–531 (2015)
3. Kim, B.S., Yoo, S.K.: Motion artifact reduction in photoplethysmography using independent component analysis. IEEE Trans. Biomed. Eng. **53**(3), 566–568 (2006)
4. Lee, B., Han, J., Baek, H.J., Shin, J.H., Park, K.S., Yi, W.J.: Improved elimination of motion artifacts from a photoplehysmographic signal using a Kalman smoother with simultaneous accelerometry. Physiol. Meas. **31**(12), 1585–1603 (2010)
5. Krishnan, R., Natarajan, B., Warren, S.: Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data. IEEE Trans. Biomed. Eng. **57**(8), 1867–1876 (2010)
6. López, S.M., Giannetti, R., Dotor, M.L., Silveria, J.P., Golmayo, D., Miguel-Tobal, F., Bilbao, A., Galindo Canales, M., Martn Escudero, P., et al.: Heuristic algorithm for photoplethysmographic heart rate tracking during maximal exercise test. J. Med. Biol. Eng. **32**(3), 181–188 (2012)
7. Sun, B., Zhang, Z.: Photoplethysmography-based heart rate monitoring using asymmetric least squares spectrum subtraction and Bayesian decision theory. IEEE Sens. J. **15**(12), 7161–7168 (2015)
8. Ram, M., Madhav, K.V., Krishna, E.H., Komalla, N.R., Reddy, K.A.: A novel approach for motion artifact reduction in PPG signals based on AS-LMS adaptive Iter. IEEE Trans. Instrum. Meas. **61**(5), 1445–1457 (2012)
9. Yousefi, R., Nourani, M., Ostadabbas, S., Panahi, I.: A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors. IEEE J. Biomed. Health Inform. **18**(2), 670–681 (2014)
10. Reddy, K.A., George, B., Kumar, V.J.: Use of fourier series analysis for motion artifact reduction and data compression of photoplethysmographic signals. IEEE Trans. Instrum. Meas. **58**(5), 1706–1711 (2009)
11. Zhang, Z.: Undergraduate students compete in the IEEE signal processing cup: part 3. IEEE Signal Process. Mag. **32**(6), 13–116 (2015)

A Case Study on How to Predict Café Profit: A Dimension Reduction via Factor Analysis

Jeong-Hyeon Moon[1], Chae-Young Yun[2], Seon-Joo Park[2], and Kyung-Ah Sohn[2(✉)]

[1] Department of Mathematics, Ajou University, Suwon, South Korea
[2] Department of Software and Computer Engineering, Ajou University, Suwon, South Korea
kasohu@ajou.ac-kr

Abstract. The purpose of this paper is to confirm the improvement of accuracy in predicting the profit of a café by using dimensionality reduction features through Factor Analysis. Profit forecasts for retailers have always been of great interest. We limit the discussion to the prediction of a café profit. We show that dimensional reduction through Factor Analysis is useful for various types of data. After that, we compare the SVM with the linear regression and show that using a good kernel trick of the SVM improves accuracy.

Keywords: Profit prediction · Principal component analysis · Factor analysis · Support vector machine · Dimension reduction

1 Introduction

As demand for coffee increases, many cafes have been set up. But some fail while others last for a long time becoming popular. Therefore, we consider the problem of finding influential features affecting café's sales among a number of external variables and predicting whether a certain café will have good results or not when the café is set up.

To achieve the goal of this work, predicting sales, an essential part is to collect features which are actually affecting sales. So we need to decide which feature we have to collect at first. In a previous work of Shin and Moon [1], it categorized features into shop, approach, and locational characteristic and analyzed their effects on sales. We collect feature data based on the suggestion of this previous work.

We will choose certain features which have high impact on café's success among all features we collected to solve overfitting-problem. In this process, we are going to reduce dimensions by Factor Analysis (FA) and perform Support Vector Machine (SVM) on the features obtained by factor analysis.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_93

2 Methods

2.1 Factor Analysis

First, we talk about Principle Component Analysis (PCA), which is the basis of FA. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [2]. Factor analysis is similar to principal component analysis in that factor analysis also involves linear combinations of variables. Different from PCA, factor analysis is a correlation-focused approach seeking to reproduce the inter-correlations among variables, in which the factors "represent the common variance of variables, excluding unique variance". In terms of the correlation matrix, this corresponds with focusing on explaining the off-diagonal terms, while PCA focuses on explaining the terms that sit on the diagonal. We used FA to identify the relationship of each features (Fig. 1).

```
Loadings:
                     Factor1 Factor2 Factor3
TypeA                 0.101           0.123
TypeB
TypeC                 0.401
TypeD
TypeE                                 0.158
Area                          0.269
ResidentNum                   0.248
WorkerNum             0.514   0.343
MainFacility          0.132   0.512   0.197
VisitorFacility               0.581
Selected201406        0.265   0.936
Selected201610        0.132   0.954
SDS
District201406        0.353           0.931
District201610                0.104   0.933
ADS                  -0.742          -0.222
UseNum                0.728
MaleFlowPopulation    0.681           0.170
FemaleFlowPopulation  0.570           0.145
Brand                 0.147          -0.131
Rent                  0.683           0.422
SimilarSales          0.851           0.239
y                     0.699
```

Fig. 1. Example of FA results: it can be seen that features that are expressed by three factors after applying FA, and factor 1 have the greatest effect on the sales of similar-area merchandise.

2.2 Support Vector Machine

To predict the sales using the extracted features, we used linear regression as a baseline. The linear regression is a regression model that models the linear relationship between

the dependent variable y and one or more independent variable x. We use a multiple linear regression since many variables are used in this study. We also employ a widely-used prediction model of SVM [4]. The SVM algorithm creates a model for predicting new data based on a given set of data. The generated model is represented as a boundary in the space where data is mapped. The SVM algorithm is the algorithm for finding the boundary having the largest margin [3]. SVMs are typically made up of a collection of hyperplanes.

The Linear kernel $k(x, x') =< x, x' >$
The Gaussian RBF kernel $k(x, x') = \exp(-\sigma\|x - x'\|^2)$
The ANOVA RBF kernel $k(x, x') = \sum_{1 \le i_1 \ldots < i_D \le N} \prod_{d=1}^{D} k(x_{id}, x'_{id})$ where k(x,x) is a Gaussian RBF kernel.

Fig. 2. The kernel functions used in the SVM algorithm

There are several kinds of kernel functions that can be used in SVM. We use the three main functions shown in Fig. 2. The first one is a linear kernel. As a basic kernel function, it can provide good predictions on linearly separable data. The second and third are kernel functions that assume a Gaussian distribution. It is a function that uses the kernel trick using the variance and average value of the data.

2.3 Data Description

We collected data provided by 'portal of small business leaders' (http://www.sbiz.or.kr/). We analyzed the type of business (coffee shops and café) in a certain area throughout the county and collected data among the analysis results. The followings are the data we collected: Market type, Area, Resident population, Worker population, Main facility, Gathering facility, Change in coffee shops in selected area, Changes in coffee shops in administrative area, Number of purchase, Sales of similar business, Floating population, Brand index, Rent. So far is the data used for the feature, and for the y value, we use the average sales in that area. For prediction, the number of purchase was removed according to the basic conditions. After that, through FA, we removed 5 features such as Brand index, Change in coffee shops in selected area, and three market types that have very little influence on each factor formed by three.

3 Results

The overall result is summarized in Fig. 3. Predictions without dimension reduction showed the lowest result. After reducing the dimension using FA, the correlation is more than 0.8. Also, we can obtain correlation up to 0.87 using the optimized SVM algorithm which used linear kernel function.

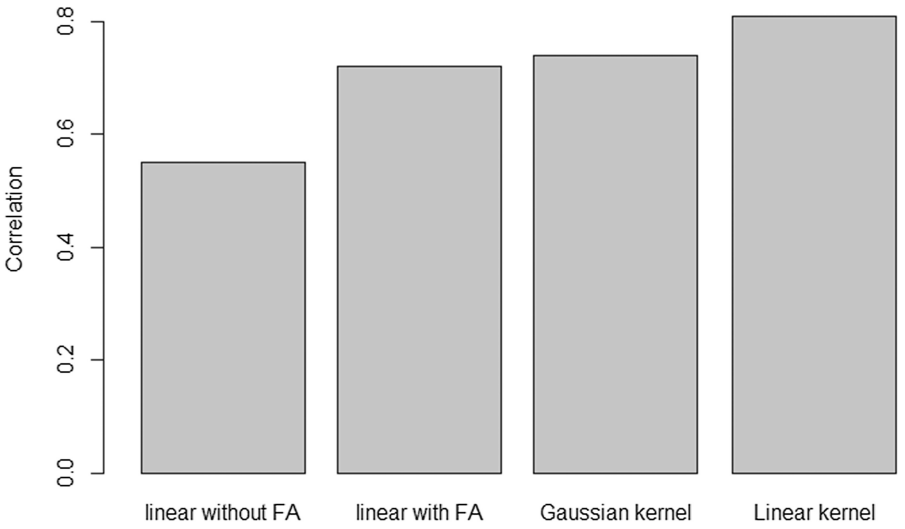


Fig. 3. Performance comparison between the baseline algorithm and the algorithm improved through this study.

To be more specific, first, the number of use of cafes was removed. The degree of correlation between the number of use cases and the sales amount is very high, so a very good value can be predicted. However, it should be assumed that the number of cafes in the area where sales of cafes are not surveyed and sales should be investigated is not investigated.

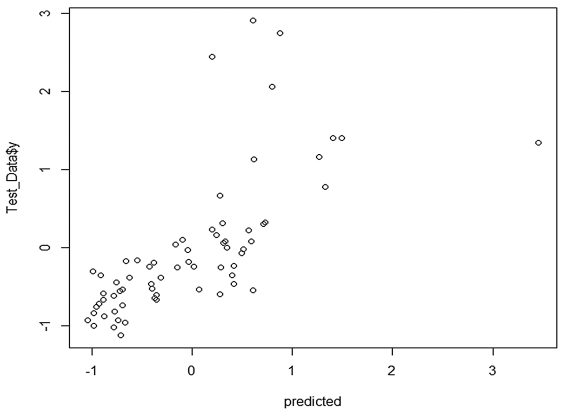


Fig. 4. Scatter plot of predicted and actual values using linear regression without FA.

The scatter plot in Fig. 4 shows the predicted and actual values of the most basic linear regression without feature validation. The correlation was 0.55.

Figure 5 is a model that applies linear regression except for some categorical variables and continuous variables using FA. The correlation was 0.72. In other words, using FA to remove five features, it can be seen that rose by about 0.2.

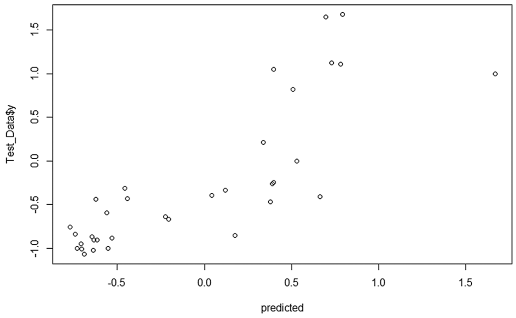


Fig. 5. Scatter plot of predicted and actual values using linear regression with FA.

Figure 6 is a correlation graph of SVM applying three kernel functions that sequentially change soft margin after excluding some categorical and continuous variables using FA. In the first case, we can see that applying the linear function gives the best result.

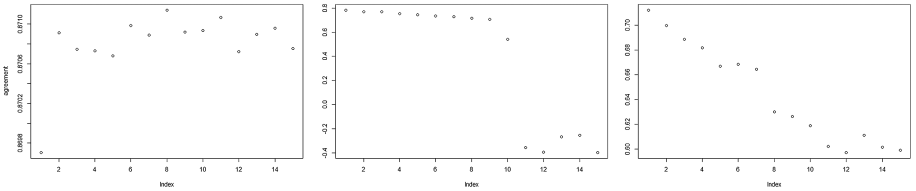


Fig. 6. SVM prediction rate which shows the correlation value according to the soft margin applying FA.

Finally, Fig. 7 is a scatter plot when applying a linear function, which is the best correlation function, and applying a soft margin of 9. The correlation was 0.815.

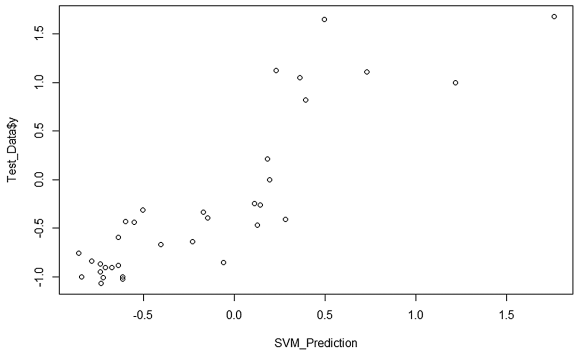


Fig. 7. Scatter plot of SVM applying linear kernel function with optimal soft margin.

4 Conclusion

In fact, there are areas where it is not easy to get a lot of data. In the case of data with geographical and economic factors, such as this study, there may be many difficulties in terms of security problems and accuracy. In this study, to recognize these situations and to overcome them, FA is proposed as a method to remove unnecessary features among various features. Using this FA, we could perform actual predictions better. In addition, we applied various kernel functions on SVM using these selected features and compared it with the linear regression method, which is a baseline function. As a result, it was found that the SVM applying the linear function as the kernel function obtained the best result. The reason why these results are shown seems to be because the relationship of the actual variables is linear. In addition, we can confirm the usefulness of SVM by showing the better result than the linear regression method which can be applied similarly, even though we have this linear relation.

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) (R2215161002000 1002).

References

1. Woojin, S., Soyoun, M.: A study on the effects of locational characteristics on the sales of a coffee shop franchise. J. Korea Real Estate Anal. Assoc. **17**, 111–123 (2011)
2. Jolliffe, I.T.: Principal component analysis and factor analysis. In: Principal Component Analysis, pp. 150–166 (2002)
3. Schölkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge (2002)
4. Lantz, B.: Machine Learning with R. Packt Publishing Ltd., Birmingham (2013)

Motion Blurred Shadows Using a Hybrid Approach

MinhPhuoc Hong(✉) and Kyoungsu Oh

50, Sadang-ro, Dongjak-gu, Seoul 07027, Republic of Korea
hmphuocl985@gmail.com, oks@ssu.ac.kr

Abstract. In this paper, we propose a new algorithm that renders motion blur and motion blurred shadows at the same time using a hybrid approach. Our algorithm generates a shadow map which stores a list of visible time ranges along with depth values at each pixel. In the subsequent pass, we use this shadow map to perform shadow tests at a receiver sample's position and at its time. Our results show that our algorithm addresses some problems that a previous work does not. In addition, our algorithm runs completely on the current GPUs.

Keywords: Real-time rendering · Motion blur · Motion blurred shadows · Visibility

1 Introduction

According to recent research described by ThanhBinh [1], Agarwal and Bedi [2], image processing is an important part of modern graphics and motion blurred shadows is an essential effect in that field. Motion blur is an important effect in computer graphics and it enhances the sense of realism experienced by users. When a geometry is blurred, its shadows should be blurred as well. However, there are a few proposed algorithms for rendering motion blurred shadows.

A brute force method renders a scene with shadow many times then averages the results to produce a correct motion blur and motion blurred shadows. However, this approach is very slow so it does not suit for the real-time rendering. Stochastic sampling based approaches use multi samples per pixel with each sample has a unique random time to render motion blur and motion blurred shadows. Thus, these approaches improve performance but produce noise images with self-shadow artifacts.

In this paper, we introduce a new algorithm that renders motion blur and motion blurred shadows simultaneously using a hybrid approach. Our algorithm combines a per-pixel linked list approach and the stochastic sampling approach. During one frame rendering, at each pixel each moving triangle is visible in a range of time. First, we render a scene from the light and store a list of such visible time ranges along with depth values at each pixel of the shadow map. Then we render the scene from camera using the stochastic sampling approach and do the shadow lookup. Thereafter, we load a visible time range and depth values of a geometry to perform the shadow tests. Finally, all visible samples are averaged to produce the pixel color.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_94

2 Related Works

Many algorithms are proposed for motion blur rendering and motion blurred shadows. Therefore, we refer readers to a survey done by Navarro et al. [3] and a book written by Eisemann et al. [4] for an overview of motion blur and shadow mapping, respectively.

Haeberli and Akeley [5] render a scene with shadow many times and average the results to produce blurred images with motion blurred shadows. However, it has ghosting artifacts at low sampling rates but increasing the sampling rates impact performance substantially.

For each pixel in a shadow map, deep shadow map [6] stores a list of semi-transparent surfaces. The visibility of a surface at a given depth is computed as $\prod_{p_z < z_i} (1 - \alpha_i)$, where z_i and α_i are the depth and opacity of a surface. To render motion blurred shadows, authors assign a random time for each sample and all samples at the same depth are averaged together to an opacity of a surface. Thus such surfaces are regarded as transparent blockers. This approach only works for static receivers. As receiver moves, time dimension is collapsed and motion blurred shadows are rendered incorrectly.

Akenine-Möller et al. [7] use the stochastic rasterization to render motion blurred shadows using time-dependent shadow maps. This algorithm uses many samples per pixel with each sample has a random time. As generating the shadow map and rendering from the camera, each sample has a random time t_s and t_r, respectively. This algorithm uses the stratified sampling to ensure that t_s and t_i belong to the same segment of the exposure interval. Thus, this algorithm can render the correct motion blurred shadows when there are many samples per pixel. However, this approach suffers might suffer self-shadow artifacts. Later, McGuire et al. [8] implement this idea on the current GPUs.

Inspiring the idea of Akenine-Möller et al. [7], Andersson, Hasselgren, Munkberg and Akenine-Möller [9] render motion blurred shadows using depth layers. First, this approach to generate time-dependent shadow maps (TSM), with each sample has motion vectors, and then cluster all samples in each tile of shadow maps into (up to) four depth layers using a method described by Andersson, Hasselgren and Akenine-Möller [10]. Next, for each layer this approach calculates an average motion vector d and re-projects all samples along this vector (to $t = 0.5$). Depth values of these samples are also computed at $t = 0.5$. At each texel of the shadow map, this method accumulates a weight, sample contributions and store together with two depth moments of the variance shadow map [11]. In the subsequent pass, they again re-project a receiver sample along the layer's motion vector to $t = 0.5$ and apply the camera filter as sampling the shadow map. To approximate the visibility of the receiver sample, they accumulate as follow: $V_{final} = \prod_l [1 - \alpha(1 - V)]$, where V is the variance shadow map visibility and α is an opacity of the layer l. This approach renders motion blurred shadows in less noise images but has potential problems when samples move in different direction in a tile. Authors alleviate but do not address completely this problem using a tile-variance approach in [12].

3 Algorithm

First, we present our main idea and then describe the details of the shadow pass in Sect. 3.1 and the lighting pass in Sect. 3.2. Throughout the presentation, we use the term triangle but it can naturally extend to a general geometry which might have animation data defined by Myeong-Won et al. [13].

3.1 Shadow Pass

We assume a triangle linearly moves from the beginning ($t = 0$) to the end frame ($t = 1$). The position of this triangle at $t = 0$ and $t = 1$ is ABC and $A'B'C'$, respectively. To generate motion blur and motion blurred shadows for this triangle, a brute force method renders this triangle many times then average the results. The goal is to find a visible time range of this triangle at each pixel and compute average color along this time range. At the pixel P, this triangle is visible through five intersection points at five times t_1, t_2, t_3, t_4 and t_5, in Fig. 1(a). From this observation, our main idea is to render this triangle only once and get a visible time range of this triangle by finding the first and the last intersection points (F_1 and F_2) at the first time (t_1) and the last time (t_5), respectively. So at the pixel P we can compute the visible time range of this triangle and know a depth range from F_1 and F_2.

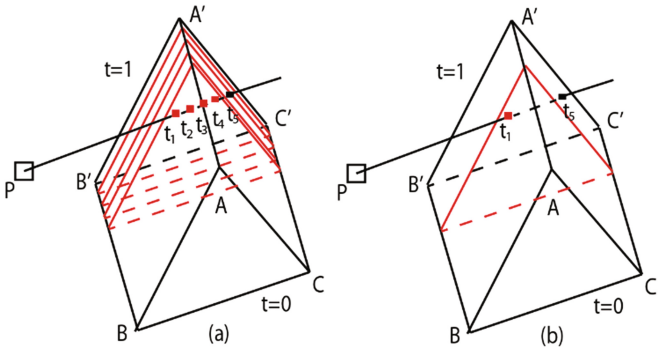


Fig. 1. A triangle moves from the beginning ($t = 0$) to the end ($t = 1$) of a frame. At $t = 0$ and $t = 1$, this triangle is ABC and $A'B'C'$, respectively. (a): A brute force method renders this triangle many times. At the pixel P, this triangle is visible at times t_1, t_2, t_3, t_4 and t_5. (b): Our algorithm renders this triangle only once then finds the first (t_1) and the last times (t_5) this triangle is visible at the pixel P.

To implement our main idea, first we assign a time and texture coordinates to each vertex of triangles ABC ($t = 0$) and $A'B'C'$ ($t = 1$). Next, we use ABC and $A'B'C'$ to form a prism, in Fig. 1(b), then triangulate this prism and send to the rasterization stage in GPU. For each pixel GPU generates two points (F_1 and F_2), each point has an interpolated time, interpolated texture coordinates and a depth value. Finally, the visible time is computed as $|t_1 - t_5|$.

With this main idea, we render a scene from the light to generate a shadow map. In our algorithm, each pixel in the shadow map stores a list of tuples with six values on the form: ($t_1\_t_2$, $z_1\_z_2$), where (t_1, z_1) and (t_2, z_2) are an interpolated time and a depth value of a generated points such as F_1 and F_2.

3.2 Lighting Pass

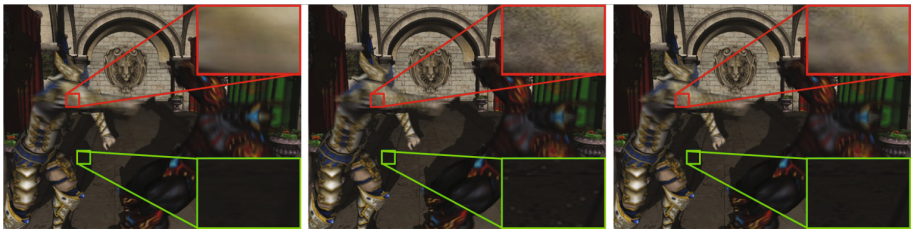
In this pass, we use a stochastic rasterization [8] to render a scene from the camera. A triangle covers a set of pixels when moving from the start ($t = 0$) to the end of frame ($t = 1$). We use two positions of this triangle at $t = 0$ and $t = 1$ to make a convex hull to cover all such pixels. There are multi samples per pixel with each sample has a random time. To check whether the current sample is visible or not, we perform a ray-triangle intersection with this ray is shot from the camera through the current sample. If there is an intersection, the current sample is visible.

To perform the shadow lookup at a visible sample we do as follows. First, we project this sample to the shadow map and load each tuple ($t_1\_t_2$, $z_1\_z_2$). Next, if the visible sample's time (t_s) is inside the visible time range [t_1, t_2], we find a depth value at t_s using linear interpolation along the depth range (z_1, z_2) and compare the interpolated depth with the sample's depth. Finally, we perform shading and average all samples' color in a pixel. To address the self-shadow artifacts in time-dependent shadow mapping method (TSM), we check if the current sample does not belong to the current triangle prior the shadow test.

4 Results

Our algorithm is implemented using DirectX 11, HLSL 5.0 with a GTX 980 Ti 6 GB graphics card. In the shadow pass, we generate and store a shadow map in the memory using a per-pixel linked list described by Barta et al. [14], Burns [15] and Salvi et al. [16]. In the lighting pass, we use the stochastic rasterization [8] with a fast ray-triangle intersection [17] and multi-sampling. For comparisons, we implement a brute force method [5] using 3000 samples to generate reference images and the time-dependent shadow mapping (TSM) using the stochastic rasterization and multi sampling. All result images are rendered at 1024×768 resolution and the shadow map used in the TSM have the same resolution.

Figures 2 and 3 show image quality comparisons between our algorithm and TSM using multi-sampling with the same number of samples per pixel. Due to the insufficient number samples per pixel, both images have noise but images rendered by TSM



Reference TSM Ours

Fig. 2. Image quality comparison between our algorithm and TSM using multi-sampling 40 samples per pixel. Images rendered by TSM has self-shadow artifact (red highlighted area) and visual artifacts (green highlighted area) while ours does not.

Reference TSM Ours

Fig. 3. Image quality comparison between our algorithm and TSM using multi-sampling 80 samples per pixel. Our result image has the similar quality with the reference image, while TSM has visual artifacts.

have visual artifacts (green highlighted area in Fig. 2 and red highlighted area in Fig. 3) in the shadow areas while ours does not. The reason for this is that TSM uses two random times, t_s and t_r, for the same sample. t_s and t_r are used when rendering from the light and from the camera, respectively. Time mismatch results in incorrect shadow tests. Additionally, red highlighted area in Fig. 2 shows that TSM have self-shadow artifacts.

Figure 4 shows the performance comparison between our algorithm by varying the number of samples per pixel. As increasing the number samples per pixel, performance of both algorithms is reduced but in the shadow pass the overhead of draw calls and state changes is higher than ours. The reason for this is that in TSM the number of draw calls is proportional to the number of samples per pixel. Meanwhile, our algorithm uses only one draw call.

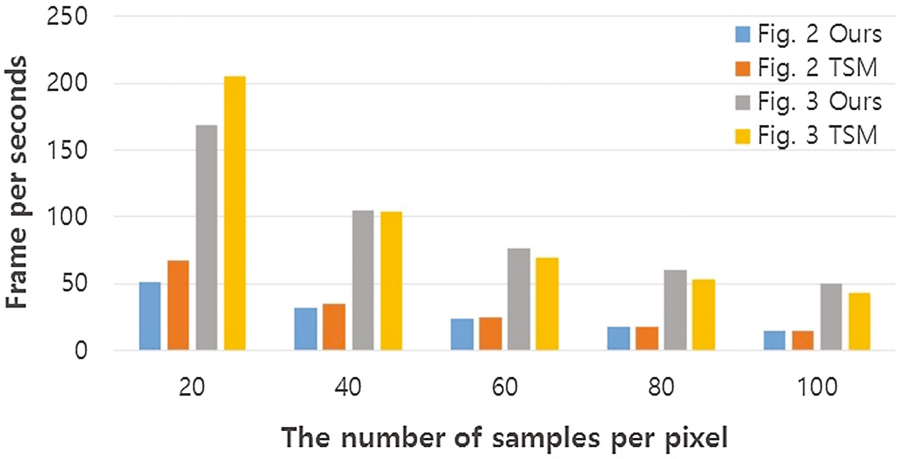


Fig. 4. Performance comparisons between our algorithm and TSM by varying the number samples per pixel in Figs. 2 and 3. The number of triangles in Figs. 2 and 3 are 268 k and 70 k, respectively.

5 Conclusion

We have presented a hybrid algorithm that renders motion blur and motion blurred shadows at the same time. First, we generate a shadow map which stores many time ranges at each pixel. Each time range represents a period of time that a geometry is visible for a given pixel during one frame rendering. In the second pass, we use multi sampling with each sample has a random time to render motion blur and motion blurred shadows. For each visible sample, we project to the light space then load each visible time range along with depth values to perform the shadow tests. All test results are averaged to produce the final pixel color.

References

1. ThanhBinh, N.: Image contour based on context aware in complex wavelet domain. Hum.-Centric Comput. Inf. Sci. **5**(1) (2015). Article 14
2. Agarwal, J., Bedi, S.: Implementation of hybrid image fusion technique for feature enhancement in medical diagnosis. Hum.-Centric Comput. Inf. Sci. **5**(1) (2015). Article 3
3. Navarro, F., Serón, F.J., Gutierrez, D.: Motion blur rendering: state of the art. Comput. Graph. Forum **30**(1), 3–26 (2011)
4. Eisemann, E., Schwarz, M., Assarsson, U., Wimmer, M.: Real-Time Shadows. AK Peters Ltd./CRC Press, Boca (2011)
5. Haeberli, P., Akeley, K.: The accumulation buffer: hardware support for high-quality rendering. In: ACM SIGGRAPH Computer Graphics 1990, vol. 24, New York, pp. 309–318 (1990)

6. Lokovic, T., Veach, E.: Deep shadow maps. In: Proceedings of SIGGRAPH 2000, pp. 385–392, ACM Press/ Addison-Wesley Publishing Co (2000)
7. Akenine-Möller, T., Munkberg, J., Hasselgren, J.: Stochastic rasterization using time-continuous triangles. In: Graphics Hardware, San Diego, California, pp. 7–16 (2007)
8. McGuire, M., Enderton, E., Shirley, P., Luebke, D.: Real-time stochastic rasterization on conventional GPU architectures. In: High Performance Graphics, Saarbrucken, Germany, pp. 173–182 (2010)
9. Andersson, M., Hasselgren, J., Munkberg, J., Akenine-Möller T.: Filtered stochastic shadow mapping using a layered approach. In: Computer Graphics Forum, vol. 34(8), pp. 119–129, Wiley-Blackwell Publisher (2015)
10. Andersson, M., Hasselgren, J., Akenine-Möller, J.: Depth Buffer compression for stochastic motion blur rasterization. In: High Performance Graphics, Vancouver, British Columbia, Canada, pp. 127–134 (2011)
11. Donnelly, W., Lauritzen A.: Variance shadow maps. In: Symposium on Interactive 3D Graphics and Games, Redwood City, California, pp. 161–165 (2006)
12. Guertin, J.P., McGuire, M., Nowrouzezahrai, D.: A fast and stable feature-aware motion blur filter. In: High Performance Graphics, Lyon, France, pp. 51–60 (2014)
13. Myeong-Won, L., Chul-Hee, J., Min-Geun, L. Brutzman, B.: Data definition of 3D character modeling and animation using H-Anim. J. Convergence **6**(3), 19–29 (2015)
14. Barta, P., Kovacs, B., Szecsi, S.L., Szirmay-kalos, L.: Order Independent transparency with per-pixel linked lists. In: Proceedings of CESCG, Vinič n, Slovakia, pp. 51–57 (2011)
15. Burns, C.A.: The visibility buffer: a cache-friendly approach to deferred shading. J. Comput. Graph. Tech. (JCGT) **2**(2), 55–69 (2013)
16. Salvi, M., Montgomery, J., Lefohn, A.: Adaptive transparency. In: High Performance Graphics, Vancouver, Canada, pp. 119–126 (2011)
17. Laine, S., Karras, T.: Efficient Triangle Coverage Tests for Stochastic Rasterization, Technical report, NVIDIA (2011)

Design of ECG Data Compression Algorithm for Efficient M2M-Based Mass Biometric Data Transmission

Jae-Sung Shim[1], Seung-Su Yang[1], Young-Hwan Jang[1], Yong-Wan Ju[2], and Seok-Cheon Park[3(✉)]

[1] Department of IT Convergence Engineering, Gachon University, Seongnam, South Korea
{llsjs28, davichi0803, jang0h}@naver.com
[2] Correspondence Center of Korea Internet and Security Agency, Seoul, South Korea
ywju@naver.com
[3] Department of Computer Engineering, Gachon University, Seongnam, South Korea
scpark@gachon.ac.kr

Abstract. Thanks to the design of different portable subminiature sensors and wired and wireless communication technology, the U-Healthcare service is getting vitalized. A mass amount of raw data is processed in real time when this U-Healthcare service is provided, and efficient processing and storage technologies are required accordingly. Therefore, this paper proposed an ECG data compression algorithm that is improved to efficiently transmit M2M-based mass biometric data.

Keywords: ECG · M2M · Biometric data · Data compression · Algorithm

1 Introduction

Due to the design of different portable subminiature sensors and wired and wireless communication technology, the U-Healthcare service is becoming more active [1].

As a mass amount of raw data is processed in real time when this U-Healthcare service is provided, the necessity of efficient processing and storage technologies is increasing [2].

Especially, the ECG (Electrocardiogram) signals are used to decipher any abnormality in each waveform among biometric data and considered as the most significant factor to diagnose heart diseases [3].

The ECG data compression algorithms to compress this are divided into direct and converted compression methods; they have some weaknesses that the direct compression method can distort the main properties of ECG if its compression rate is increased, and the converted compression requires a relatively large amount of computation [4].

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_95

Thus, in order to transfer efficient M2M-based medical information, this paper proposed an ECG data compression algorithm, which is the combination of the overlapping-pattern compression algorithm that is a lossless compression technique and the compression algorithm to decide existence of data.

2 Related Studies

2.1 M2M

M2M is a technology that mutually transfers information after collecting, manipulating, and processing through communication between objects without human intervention by attaching communication modules to different devices and equipment.

It implies an environment that can provide various fields with spontaneous/intelligent custom services through Internet-connected intelligent objects, away from the existing on-demand paradigm [5].

2.2 U-Healthcare

U-Healthcare is an abbreviation of ubiquitous healthcare, and it is a product/service system that connects information/communication technology with health/medical industry to enhance the quality of life of users and reduce medical costs by obtaining health information for humans anytime and anywhere and utilizing it for prevention, diagnosis, treatment, and follow-up management [1].

2.3 Overlapping-Pattern Compression Algorithm

Lempel-Ziv is a lossless data compression algorithm suggested by ZIV and Lempel in 1978, and it is a method that replaces an overlapping pattern at the current location, if such pattern exists, by finding a relative location to it and the length of the pattern [6].

2.4 Compression Algorithm Deciding the Existence of Data

There are a lot of efficient cases if data is expressed bitwise. The bit-vector technique is used in this type of bitwise computations; it is a method expressing sets containing non-overlapping integers in bits [7].

3 Designing M2M-Based ECG Data Compression Algorithm

3.1 Outline of ECG Data Compression Algorithm

This study suggested a data compression algorithm using the overlapping-pattern compression algorithm, which is a lossless compression technique, and the algorithm deciding the existence of data.

First, the 4-stage preprocessing is performed regarding the ECG signals obtained in advance.

In the first stage, the difference between the currently obtained signals and previously obtained ones.

In the second stage, electrocardiogram signals are classified in sections of one cycle each, starting from the R-wave that has the biggest amplitude among the meaningful sections in ECG, such as P, Q, R, S, and T. Then, the classified ECG signals are stored in buffers for DCT conversion.

In the third stage, in order to improve the compression rate, DCT conversion is performed and the stored data is put into buffers based on different sizes of window filters.

The fourth stage conducts window-filtering and assembles the 100 data stored in buffers to perform the proposed compression algorithm in real time.

The algorithm suggested in this paper is run through this 4-stage preprocessing. Figure 1 shows the outline of the overall algorithm.

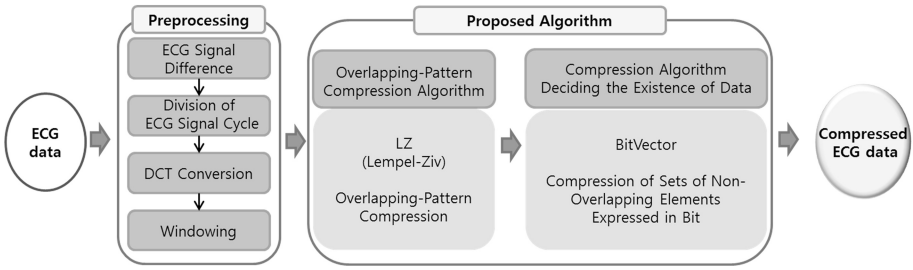


Fig. 1. Algorithm overview

3.2 Requirements for ECG Data Compression Algorithm

This study proposed an ECG data compression algorithm that combined the overlapping-pattern compression algorithm and the compression algorithm deciding the existence of data for the efficient M2M-based transmission of medical information. The requirements for designing this algorithm are as follows.

- Biometric data (ECG) should be preprocessed to use the suggested ECG Data Compression Algorithm.
- Compress the overlapping data using the overlapping-pattern compression algorithm, which is a lossless compression technique.
- There should be no overlapping data to use the compression algorithm deciding the existence of data.
- Indicate the existence of data and the compress the data by 1bit based on the indications using the compression algorithm deciding the existence of data.
- CR and PRD should be enhanced compared with the existing studies.

3.3 Efficient Design of ECG Data Compression Algorithm

In this section, we explain the M2M-based ECG data compression algorithm to efficiently transfer medical information and its specific procedures. The overall content of the ECG data compression algorithm is equivalent to Fig. 2.

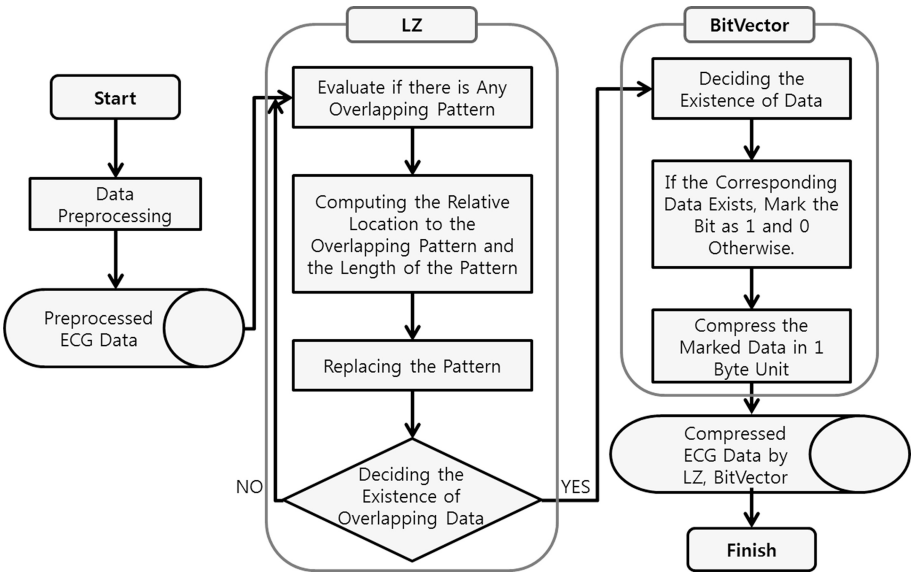


Fig. 2. ECG data compression algorithm

3.3.1 Preprocessing of ECG Signals

This paper named the following procedures "preprocessing of ECG signals."

Regarding the acquired ECG signals, find the difference between the currently obtained signals and previously obtained ones. Then, classify electrocardiogram signals in sections of one cycle each, starting from the R-wave that has the biggest amplitude among the meaningful sections in ECG, such as P, Q, R, S, and T.

Store the classified ECG signals in buffers for DCT conversion, and in order to improve the compression rate, perform the DCT conversion and put the stored data into buffers based on different sizes of window filters.

Finish the preprocessing procedure while conducting window-filtering of stored data and assembling the 100 data stored in buffers.

3.3.2 Overlapping-Pattern Compression Algorithm

The overlapping-pattern compression algorithm proposed in this study can be divided into three stages: evaluating overlapping patterns, computing the relative location and length, and replacing the pattern. The proposed algorithm is designed as shown in Fig. 3.

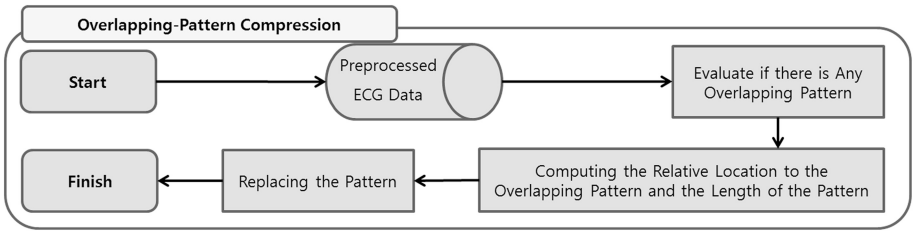


Fig. 3. Overlapping-pattern compression algorithm

In order to compress the overlapping data of ECG data with completed preprocessing, we have to evaluate whether there is any overlapping pattern at the current location.

If such pattern exists, we compute the relative location to the overlapping pattern and the length of the pattern and replace the pattern with <relative location, length>.

As the compression method for overlapping patterns dynamically constitutes a dictionary by using cues from prototype when reading data, this technique can find the overlapping information for each object and compress it.

3.3.3 Compression Algorithm Deciding the Existence of Data

The compression algorithm to decide the existence of data proposed in this study can be divided into three stages: evaluating whether data exists or not, indicating based on the existence of data, and the bit-compression based on the indications. The proposed compression algorithm to evaluate the existence of data is designed as shown in Fig. 4.

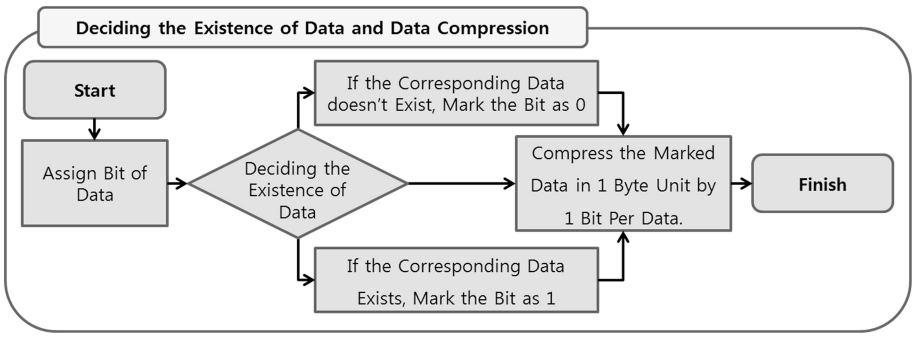


Fig. 4. Compression algorithm deciding the existence of data

Decide whether exists the data with the overlapping data compressed by the overlapping-pattern compression algorithm. Assign one bit for each dataset after the judgment, and express the sets as a series of bits. If the corresponding data exists in a set, mark the bit as 1 (TRUE) and 0 (FALSE) otherwise. The marked data in 1 byte (8 bit) unit are compressed by 1 bit per data.

4 Conclusion

Recently, following the rapid growth of ITC fusion technology, the design of different portable subminiature sensors and wired-wireless communication technology are vitalizing U-Healthcare, and the paradigm of medical services are changing from diagnosing and treating diseases to preventing and managing them.

This design of U-Healthcare technology enables people to monitor not only the whole medical field but also their own health promotion contents individually anytime and anywhere.

Thus, a huge amount of important medical information gets processed in real time wirelessly, and as a result, a mass biometric data is generated, requiring efficient processing and storage techniques.

In addition, if a biometric database is built to manage patients requiring long-term management or those with chronic conditions, there is an issue of the storage capacity. Also, the larger the new data when compared to the existing data and evaluated in order to distinguish whether there is an anomaly or not, the longer the detection time. Therefore, based on this, the necessity of efficient compression of biometric data is on the rise.

Therefore, we conducted a study to efficiently compress ECG signals, which are used as a very significant factor in diagnosing heart diseases, among all the biometric data.

Through this, in order to transmit efficient M2M-based medical information, this study suggested an ECG data compression algorithm, which is the combination of the overlapping-pattern compression algorithm, also known as a lossless compression technique, and the compression algorithm to decide existence of data.

References

1. Jang, S.H., Kim, R.H., Lee, C.W.: Effect of u-healthcare service quality on usage intention in a healthcare service. Technol. Forecast. Soc. Change **13**, 396–403 (2016)
2. Yin, L., Liu, C., Lu, X., Chen, J., Liu, C.: Efficient compression algorithm with limited resource for continuous surveillance. KSII Trans. Internet Inf. Syst. (TIIS) **10**(11), 5476–5496 (2016)
3. El-Saadawy, H., Tantawi, M., Shedeed, H.A., Tolba, M.F.: Electrocardiogram (ECG) classification based on dynamic beats segmentation. In: Proceedings of the 10th International Conference (2016)
4. Kumar, V., Saxena, S.C., Giri, V.K., Singh, D.: Improved modified AZTEC technique for ECG data compression: effect of length of parabolic filter on reconstructed signal. Comput. Electr. Eng. **31**(4), 334–344 (2005)
5. Song, J.S., Kunz, A., Prasad, R.R.V., Sheng, Z., Yu, R.: Research to standards: next generation IoT/M2M applications, networks and architectures. IEEE Commun. Mag. **54**(12), 14–15 (2016)
6. Kärkkäinen, J., Kempa, D., Puglisi, S.J.: Lazy Lempel-Ziv factorization algorithms. J. Exp. Algorithmics (JEA) **21**, 1–19 (2016)
7. Handley, J.C.: Bit vector architecture for computational mathematical morphology. IEEE Trans. Image Process. **12**(2), 153–158 (2003)

Design of Clustering Algorithm for Efficient Energy Management in Wireless Sensor Network Environments

Seung-Su Yang[1], Jae-Sung Shim[1], Young-Hwan Jang[1], Yong-Wan Ju[2], and Seok-Cheon Park[3(✉)]

[1] Department of IT Convergence Engineering, Gachon University, Seongnam, South Korea
{davichi0803, llsjs28, jang0h}@naver.com
[2] Correspondence Center of Korea Internet and Security Agency, Seoul, South Korea
ywju@naver.com
[3] Department of Computer Engineering, Gachon University, Seongnam, South Korea
scpark@gachon.ac.kr

Abstract. Recently, there has been an active research effort on Wireless Sensor Network (WSN) where the sensor nodes consume energy efficiently by communicating between the nodes directly without a network infrastructure. However, previously proposed protocols require regular re-establishment of clusters, which leads to unnecessary energy consumption. Moreover, there is a large energy consumption because a cluster head that is placed far apart from a sink node directly transmits data to the sink nodes. Therefore, in this paper, we analysis the problems of the previous clustering techniques and protocols, and designed a clustering algorithm for efficient energy consumption through the use of an energy threshold during cluster re-establishment and data transmission route selection.

Keywords: Energy management · Clustering algorithm · LEACH · WSN · Multi-hop

1 Introduction

Following the recent advances in wireless communication and sensor technology, it has become possible to use WSN in various applications [1].

Because a WSN consists of communication between sensor nodes without an infrastructure such as a base station, energy of a node can represent the energy of a network [2].

Therefore, there is a research effort on efficient use of energy of nodes in order to increase the lifetime of a network.

The most notable clustering technique is to divide the sensors into clusters, and allow only the cluster head to communicate with sink nodes in order to reduce the amount of energy consumed by each node during data transmission and reception.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_96

The Low-Energy Adaptive Clustering Hierarchy (LEACH) protocol is used as a clustering method, and distributes the energy consumption to all nodes through periodic cluster head re-establishment. This resolves the imbalance in energy consumption and increases the network lifetime.

However, the LEACH protocol requires periodic re-establishment for selecting the cluster head. As such, unnecessary energy is consumed for the re-establishment.

Moreover, a single-hop transmission method is used where the cluster head transmits data directly to the sink node. Hence, there is a significant energy consumption if the cluster head is placed far from the sink nodes.

Therefore, in this paper, we designed an algorithm that improves on the problems of the LEACH protocol, by re-establishing the clusters based on an energy threshold value, and using a multi-hop method for data transmission that considers energy consumption.

2 Related Works

2.1 LEACH Protocol

The LEACH protocol is one of the most notable methods among clustering algorithms. In the LEACH protocol, multiple sensor nodes constitute a cluster, and the sensor nodes within a cluster become the cluster head through periodic re-establishment [3].

The member nodes that belong to a cluster head transmit their data to the cluster head, which is received and merged for transmission to the base station [4].

In the LEACH protocol, cluster heads are replaced through re-establishment during each round for the purpose of balanced energy consumption. Equation 1 is used for re-establishment of cluster heads.

$$\mathrm{Y(n)} = \begin{cases} \frac{R}{1-R\times(v \bmod \frac{1}{r})} : if\ n \in K \\ 0 : otherwise \end{cases} \tag{1}$$

Here, $\mathrm{Y(n)}$ is the threshold for the selection of a cluster head, R. the ratio of cluster heads, v. the current round, and K. the set of all sensor nodes other than cluster heads [4].

3 Design of Clustering Algorithm

3.1 Overview of the Clustering Algorithm

In this paper, we consider the threshold for cluster heads, and re-establish the clusters only when the remaining energy is below the threshold, while omitting the re-establishment process when it is above the threshold. This not only reduces energy consumption, but also increases the lifetime of the entire network. An overview of the algorithm is shown in Fig. 1.

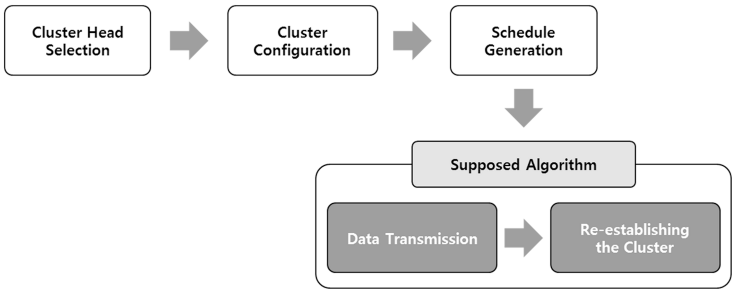


Fig. 1. Clustering algorithm overview

3.2 Functional Block Diagram of the Clustering Algorithm

The proposed clustering algorithm for efficient energy management adopts a multi-hop technique to balance the energy management of cluster heads.

By setting a threshold for cluster heads, the optimal route and the route with the maximum power capacity are selected, and the cluster is re-established only when the remaining energy of the cluster head is below the threshold. An FBD of the proposed clustering algorithm is shown in Fig. 2.

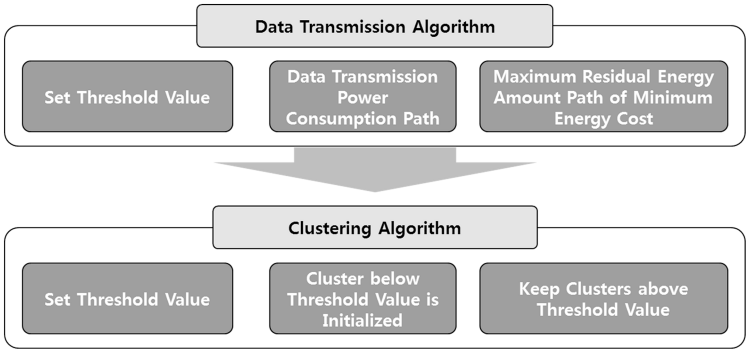


Fig. 2. Functional block diagram

3.3 Data Flow Diagram

The clustering algorithm proposed in this paper defines the re-establishment period through data flow chart, and performs cluster head selection and configuration following node activation.

Moreover, clusters are re-established based on the threshold value setting and remaining energy, in order to achieve regular energy management.

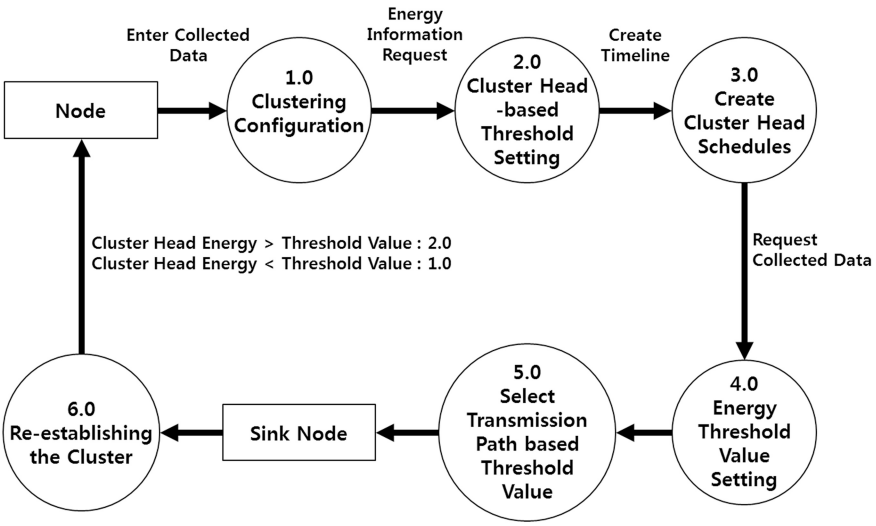


Fig. 3. Proposed algorithm DFD

Subsequently, the cluster head creates a schedule and notifies the member nodes. When data collection is finished, the data is transmitted by selecting transmission routes to the sink nodes. A DFD of the proposed algorithm is shown in Fig. 3.

3.4 Design of the Overall Algorithm

In this paper, we designed a clustering algorithm for efficient energy management in WSN environments. The overall algorithm as designed is illustrated in Fig. 4.

Following the establishment of a cluster, the cluster is re-established if the remaining energy of the cluster head is lower than the threshold given by the following equation. Equation 2 is used for re-establishment.

$$\mathrm{Ecv} = \mathrm{CHaE} - \left(\frac{CHaE}{CHN} \cdot 2\right) \tag{2}$$

In Eq. 2, Ecv represents the energy threshold, CHaE is the average energy of a cluster head, and CHN is the number of cluster heads.

After first establishing a cluster, the time of re-establishment and replacement of the cluster head is determined by comparing Ecv to the remaining energy.

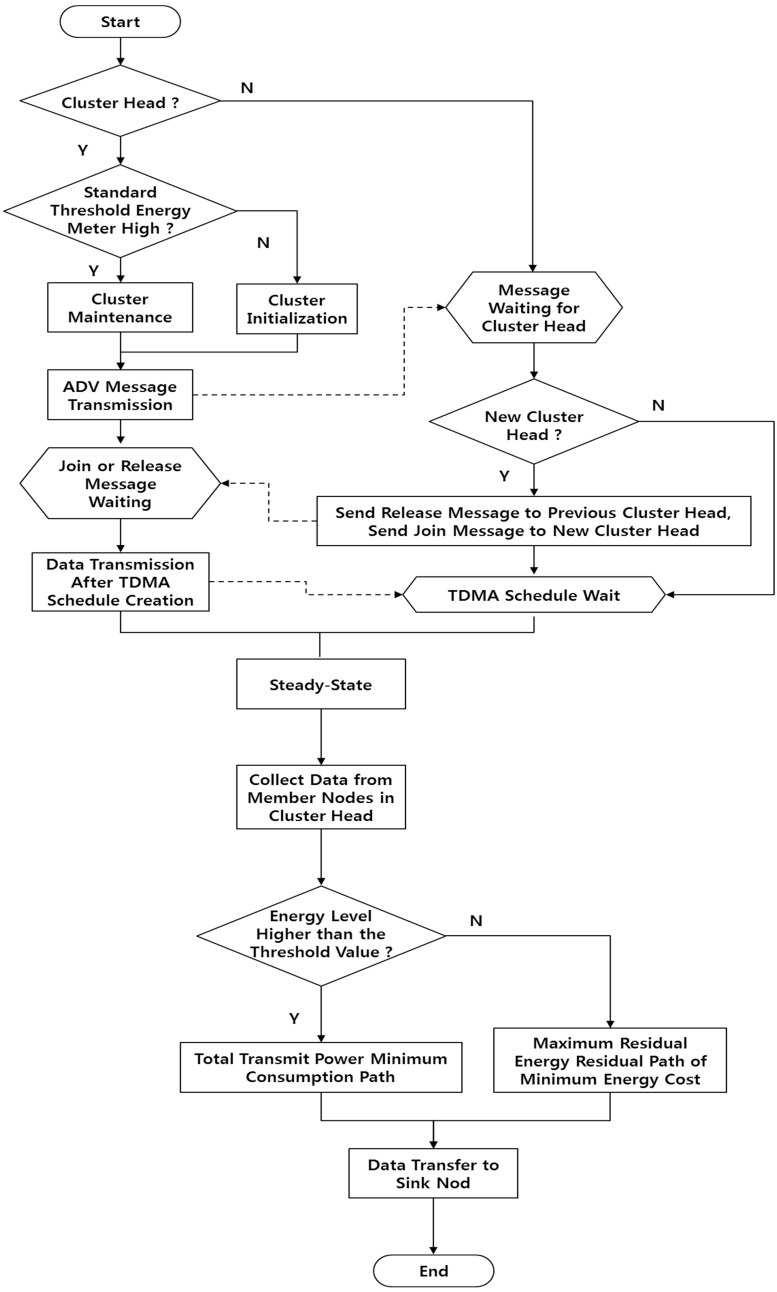


Fig. 4. Full algorithm

4 Conclusion

The LEACH protocol, which has been used frequently thus far, resolves energy imbalance between nodes and increases the lifetime of the entire network through effective cluster establishment and efficient merging of data.

However, because the LEACH protocol selects cluster heads probabilistically and re-establishes clusters regularly, unnecessary energy is consumed.

Therefore, in order to resolve this issue, this paper designed a clustering algorithm for efficient energy management in WSN environments.

In order to design the proposed algorithm, we set an energy threshold value to resolve the problem of unnecessary energy consumption in the previous LEACH protocol.

Unnecessary energy consumption was reduced by re-establishing a cluster only if the remaining energy is below the set threshold, while omitting re-establishment if it is above the threshold.

Moreover, in order to resolve the problems of a single-hop transmission method, a multi-hop method was used, which allows for efficient energy management. Based on the multi-hop method, we designed a clustering algorithm that can maximize the power efficiency of data transmission by selecting the minimal and optimal route based on the threshold.

References

1. Kim, J.-H., Lee, C.-H., Shim, K.-S.: Time series prediction using clustering algorithm. J. Korea Inf. Sci. Soc, **20**, 191–195 (2014)
2. Mamurjon, D., Ha, I.-K., Ahn, B.-C.: A mobile data gathering method for clustered wireless sensor networks. J. Korea Inf. Sci. Soc. **41**, 138–144 (2014)
3. Chung, K.-S., Lee, W.-S., Song, C.-Y.: An energy efficient clustering scheme for WSNs. J. Inst. Electron. Inf. Eng. **50**, 252–258 (2013)
4. Kim, S.-C., Choi, S.-K., Cho, Y.-H.: Clustering algorithm for extending lifetime of wireless sensor networks. J. Korea Soc. Comput. Inf. **20**, 77–85 (2015)

Design of a Framework for Security Enhancement in Telematics Control Units (TCUs)

Kiyoung Jang, Suhong Shin(✉), and Byoungsoo Koh(✉)

DMC R&D Center, 8th FL, 37, Maebongsan-ro, Mapo-gu, Seoul, Korea
{kyjang, shshin, bskoh}@digicaps.com

Abstract. Cars are increasingly being equipped with a variety of driver convenience and safety features, with many vehicles evolving into what are now called "smart" cars. The convenience and safety systems built into these vehicles, which include infotainment, connected car, and autonomous navigation systems, are being actively developed by combining them with mobile communication technology. However, utilizing such technology can potentially leak personal, or vehicle information. A mobile communication module for example, can be used to attack the electronic control device of a vehicle; the attacker then has the ability to endanger the driver by gaining control of the vehicle brakes and steering devices. To protect the driver and the vehicle from such risks, we have designed a technology that creates a secure zone for the storage of important information, restricting external access to the telematics control unit.

Keywords: Secure area · File system · Certificate · File encryption/descript · Sandbox

1 Introduction

The automobile industry has been influenced by recent advancements in mobile communication, which has been incorporated into various technologies including infotainment, connected car, and autonomous driving systems. For example, Intelligent Transport Systems (ITS) are being introduced around the world, and Cooperative-Intelligent Transport Systems (C-ITS) - the next generation ITS - are currently under active research in Korea. The C-ITS technology is a system that provides the driver with traffic conditions and accident risk information, such as sudden stops or fallen objects, in real time. With the advancement of mobile communication and sensor technology, automobiles are no longer simple means of transportation, instead becoming "smarter" with increasingly advanced safety and convenience features. As these technologies evolve, Telematics Control Units (TCUs) for automobile communication are also continuously being developed. Existing TCUs manage radio, GPS, Bluetooth, Wi-Fi and USB, however with the advancement of mobile communication it has become possible to use faster networks, including LTE, 5G and WAVE, than in previous generations. Because the TCU uses bi-directional mobile communication, both the information in the vehicle and the information received from outside the vehicle is

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_97

expected to increase. For example, to update information in a vehicle, one could either download data by connecting to an external update server using a TCU, or check the current vehicle status by sending information externally. As technology advances, ensuring the security of the vehicle or personal information has become critical. There is a need to determine whether the information coming from the update server is reliable and safe to use, and whether the vehicle information and personal information is securely stored.

In this paper, we design a secure zone to securely store data in the TCU responsible for communication. The data in the secure zone is encrypted by default. The key used for encryption is extracted through an internal algorithm, and the decryption function is provided to other devices requiring use of the information in the TCU.

2 Related Studies

This section describes: (1) The file system security used to securely store data in the TCU; and (2) Seed generation to encrypt the data.

2.1 File System

A file system is defined as a system used to arrange and manage data within a partition of a physical disk. In general, a hierarchical file system in a directory structure is used. File system types include FAT, NTFS, EXT, and HFS, depending on the operating system. A file system may perform various functions on a given file, such as inputting, outputting, saving, deleting, and searching. In this paper, these functions are wrapped to restrict access from the outside, other than to a specified user. By adding encryption to a commonly used file system and providing a new wrapped API, the system will provide important information to specific users only.

2.2 Seed Generation

In this paper, the key used for encryption is generated by a random number generator, and is used to store the encrypted data. The random number generator requires the seed value, which is generated by an internal algorithm and special information from the TCU. The internal algorithm generates a new seed value by combining information that is only known to the developer who provided the special information to the TCU, and the module that extracts seed values. Even if the special value of the TCU is known to an outsider, the seed value will still only be known to the developer. Therefore, the seed value used for the random number generator should be configured such that it will not be recognized by an outsider, and the value must remain unchanged once generated. If the seed value is exposed, the previously encrypted data should be discarded.

3 Design

In this section, we design the module that securely stores data in the TCU. This module stores and encrypts at the same time, rather than protecting files in the existing operating system. Some existing operating systems use Trust Zone or Secure Area technologies to secure important data. However, for an embedded system that is expensive and does not have hardware support available, a different approach is necessary. Thus, we propose a method that is a software technology, and is designed to be convenient for the developer. We have designed a module that emphasizes the security of encrypted data by applying a unique key generation method and encryption level, thereby enhancing security strength and efficiency. This section describes the detailed design of the proposed module, which we have named the Extended Encrypting File System (hereinafter referred to as "EEFS").

3.1 EEFS Module Architecture

Figure 1 shows the proposed EEFS module as compared to: (1) The method of storing the existing plain data without encrypting; and (2) The method of separately encrypting and storing using two separate steps. The proposed method does not read, encrypt, and then re-save the existing file. Rather, it immediately and simultaneously performs encryption and storage of the memory data in the file system, providing convenience and secure storage of the data to the developer.

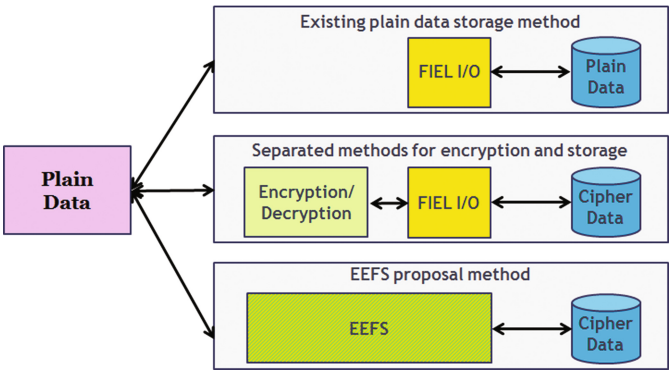


Fig. 1. EEFS module configuration diagram

3.2 EEFS Detailed Design

The detailed design of the EEFS module allows one to select the security strength and processing time for each system, based on the use of a unique seed, the encryption module and file I/O unification method, the encryption key & IV generation method, selection of the secure encryption algorithm, and the encryption level.

3.2.1 Special Encryption Seed

As shown in Fig. 2, two encrypted seeds—seed1 and seed2—are used in EEFS to prevent the risk of exposing the seed. Seed1 inputs a unique value from the corresponding device at the initialization of the EEFS module, so that even if the encrypted cipher data is exposed by hacking, it cannot be decrypted by other devices. Seed2 is used as the final encrypted seed in case seed1 is also hacked, by mixing with seed1 in the EEFS module.

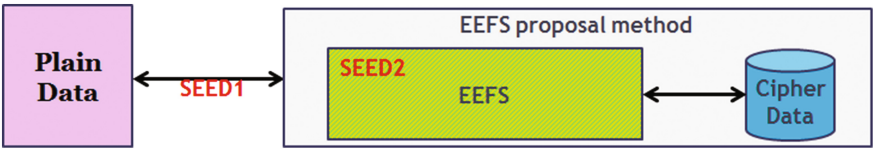


Fig. 2. Unique encryption seed

3.2.2 Encryption and File I/O Unification

The File I/O-related functions use standard C language functions (fopen, fwrite, fread, fflush, fseek, fclose), and provide the EEFS function with module initialization, resource release, and encryption/decryption techniques. Table 1 shows the API provided by the EEFS module. The developer can perform the file I/O at the same level as the existing standard function; however, it contains encryption/decryption functions and provides secure management of data.

Table 1. EEFS provided API

API	Description
xxx_initialize()	EEFS module initialize
xxx_terminate()	EEFS module quit
xxx_fopen()	File create/open
xxx_fwrite()	File write
xxx_fread()	File read
xxx_fflush()	Buffered data is reflected in the actual file
xxx_fseek()	Change the value of the file location indicator
xxx_ftell()	Return the value of the current file location indicator
xxx_fclose()	File close

3.2.3 Key and IV Generation Method

The symmetric key and initialization vector used for encryption/decryption are generated using seed1 and seed2. The generated Key & IV are unique, and can be used only in the corresponding device. Depending on the Generate option, a 16-byte key, 32-byte key, or a 16-byte IV can be generated. Moreover, the key and IV generation complexity can be controlled by adjusting the number of message digest iterations, based on the Count item in the Generate option.

3.2.4 Selection of Symmetric Encryption Algorithm

EEFS uses the AES algorithm based on the crypto library in openssl, which is an open source encryption/decryption algorithm. AES is a 128-bit block cipher adopted in 2000 as a new US standard block cipher, as the result of a proposed security problem in the previous US standard block cipher DES. Most known attacks on DES have already been applied to AES.

* Brute-Force Attack: AES is safer than DES, because AES uses larger keys.
* Statistical Attacks: Numerous tests have failed to statistically analyze the ciphertext.
* Differential and Linear Attacks: Differential and linear attacks against AES are still unknown.

3.2.5 Correlation Based on Encryption Level

The EFFS module is designed to use four levels of encryption, allowing one to select and use the security strength and processing time for each system. As the encryption/decryption process becomes more complex, higher security strength can be achieved, but processing time will increase. Since processing in the system may have life-or-death consequences, the response should be real-time; in low-cost embedded systems or automobiles, processing time can be a significant issue.

mode / key size	ECB	CBC
128 bit	128bit & ECB	128bit & CBC
256 bit	256bit & ECB	256bit & CBC

EEFS_LEVEL	Key size / Mode
Level1	128bit & ECB
Level2	128bit & CBC
Level3	256bit & ECB
Level4	256bit & CBC

Fig. 3. Encryption level

As shown in Fig. 3, the four proposed encryption levels are determined per the combination of the block cipher mode and key size. ECB is the simplest block cipher mode structure and is encrypted by dividing the plaintext into several blocks. Since the same encryption key is used, if multiple blocks have the same value, the encryption value will also be the same. Thus, the ECB encryption mode can be analyzed easily and the key can be exposed; its advantage, however, is fast performance due to parallel processing, in that it is possible to process each block simultaneously by dividing data into blocks. With Cipher Block Chaining (CBC) mode, before each block is encrypted, it is XORed with the encryption value of the previous block. The first block uses the

initialization vector (IV) since there is no previous block to XOR. This mode can be seen as a slightly upgraded version of ECB, however one disadvantage of CBC is that it can prevent known plaintext attacks. Because the blocks are connected to each other, there is a high probability that the ciphertext and plaintext will not match one-on-one. Another disadvantage is that processing speed is slower than ECB; the CBC cannot be performed in a parallel process, because it requires the encryption data of the previous block to encrypt the current block. As the encryption key size increases, processing speed slows, but security increases. Three versions of AES exist, using 128, 192, and 256-bit keys, and with 10, 12, and 14 rounds, depending on the key size. Even though the size of the master key is different, the round keys are all 128 bits. The encryption level can be set when the EEFS module is initialized, and the processing time and security strength can be adjusted per each feature. The processing time for the level can vary for each system, and the user can select the appropriate level.

4 Experiment and Evaluation

Experiments were conducted using real systems. The system environment used for the experiment is as follows:

* OS: Mac OS X EI Capitan (Version 10.11.2)
* Processor: 2.7 GHz Intel Core i5
* Memory: 8 GB 1867 MHz DDR3
* Compiler: Xcode (Version 7.1.1)

In this paper, we compare the encryption/decryption time based on the encryption level, and confirm that encryption and decryption are performed normally using the proposed module.

4.1 Comparison of Computation Time Based on Encryption Level

Table 2 shows the average time taken to encrypt/decrypt 4 MB of data per the encryption level. The results may vary slightly with the performance of a system, but differences in the level will be similar throughout usage of the verification algorithm. Although the processing time increases as the level increases, the security strength also increases.

Table 2. Comparison of computation time based on encryption level (msec)

	EEFS_LEVEL1	EEFS_LEVEL2	EEFS_LEVEL3	EEFS_LEVEL4
Encryption	49.508	53.778	55.977	71.364
Decryption	43.663	52.673	56.038	74.048

4.2 Encryption/Decryption Result of the Proposed Module

Figure 4.

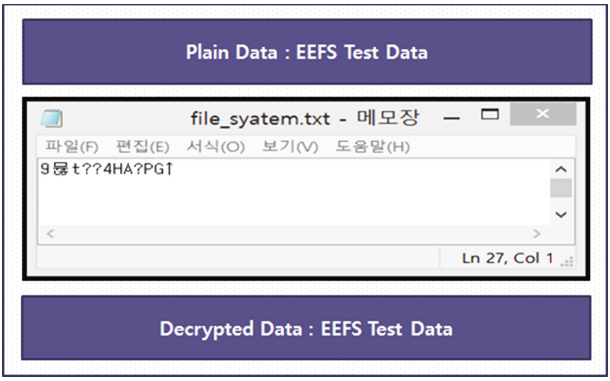


Fig. 4. Encryption/decryption result

5 Conclusion

In this paper, we designed and implemented a module that can securely store information used in TCU equipment - which for the first time vehicles are using to communicate externally - and that can perform as a security function for important data. The implemented security function is designed to be easy to use, by combining standard C Language file I/O related functions with an encryption function. To minimize exposure of the encryption/decryption key and maximize security strength, we use a unique seed method with the optimal algorithm. We have also added a level of encryption that can be controlled by appropriately adjusting the relationship between security strength and computation processing time, for each system. In response to the threat of information leakage of data stored in the user terminal, the proposed module secures stored data using a simple user interface, and verified password protection algorithm.

Acknowledgements. This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program (10070156, Extended service framework for telematics in a connected car environment) funded By the Ministry of Trade, industry & Energy (MOTIE, Korea).

Improved Data Stream Clustering Algorithm for Anomaly Detection

Chunyong Yin[1(✉)], Sun Zhang[1], and Jin Wang[2]

[1] School of Computer and Software,
Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, Nanjing, China
yinchunyong@hotmail.com
[2] College of Information Engineering, Yangzhou University, Yangzhou, China

Abstract. Intrusion detection provides important protection for network security and anomaly detection as a type of intrusion detection, can recognize the pattern of normal behaviors and label the behaviors which departure from normal pattern as abnormal behaviors. We think that the traditional methods based on dataset do not satisfy the needs of dynamic network environment. The network data stream is temporal and cannot be treated as static dataset. The concept and distribution of data objects is variety in different time stamps and the changing is unpredictable. Therefore, we propose an improved data stream clustering algorithm and design the frame of anomaly detection according to the improved algorithm. It can modify the established model with the changing of data stream and detect abnormal behaviors in time.

Keywords: Intrusion detection · Anomaly detection · Data stream · Clustering

1 Introduction

The popularity of internet application brings great challenge to network security. In recent years, hacker intrusion, network paralysis and user information leakage have caused extensive damage to society and economy. Wenke Lee [1] proposed the concept of intrusion detection in 1998 and it provides important protection for network. Intrusion detection can be divided into two types: misuse detection and anomaly detection. Misuse detection analyzes the characteristics of known attack behaviors and builds rule base which is used to match with behaviors. The behavior with higher similarity will be labeled as abnormal behavior. Anomaly detection recognizes the pattern of normal behaviors and label the behaviors which departure from normal pattern as abnormal behaviors.

Early anomaly detection has high false alarm rate and the introduction of data mining makes it a great development. The application of data mining in anomaly detection can be divided into two types according to the processing objects. One is the method based on dataset and most research results is based on dataset. The other one is the method based on data stream. The dataset is static and the data model based on dataset is permanent. Conversely, data stream is temporal and it is changing by time [2]. The data model based on data stream is variable in different time stamps. Because

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_98

the concept and distribution of data objects is variety in different time stamps and the changing is unpredictable [3]. We think that data object is transmitted in network as the form of stream, so the method based on data stream is more appropriate.

The main methods of data mining are classification analysis, clustering analysis and regression analysis. Classification analysis trains the model by labeled datasets and recognizing unlabeled data records in testing phase. Clustering analysis belongs to unsupervised method and it can divide data points into different clusters according to their similarity. The data points in different clusters will have farther distance and the cluster has high similarity inside. The data model based on data stream need to be modified in time. The adjustment of classification model is harder than clustering model, because classification model is supervised and it needs extra label resources. Thus, we improve clustering algorithm based on data stream as the core of anomaly detection and design the anomaly detection frame.

2 Anomaly Detection Model

Recent research works of anomaly detection and clustering algorithm mainly focus on datasets and it can obtain perfect performance in simulate experiments. We summarize and compare the difference between the methods of dataset and data stream as shown in Table 1.

Table 1. Comparison of methods based on dataset and data stream

Types	Methods based on dataset	Methods based on data stream
Data model	Permanent	Temporary
Mathematical relation	Data objects set	Data objects sequence
Saved all	Yes	No
Processing	One-time	Continuous
Results	Accurate	Approximate
Time	Longer	Shorter
Memory	More	Less

We compare them from seven points: (1) The dataset is static and the data model based on dataset is permanent. The data stream is changing with time and the data model based on data stream is temporary; (2) Dataset is the set of data objects and data stream is the sequence of data objects. The definition of data stream is a sequence that constructed by continuous and ordered data points; (3) The method based on dataset usually read all data records into the memory. But for data stream, it is continuous and infinite. The memory consumption will increase with time. Therefore, the method based on data stream utilize the fixed memory to save summary statistics information; (4) The process of dataset is one-time and that of data stream is continuous. However, the process for each data object in data stream is one-time; (5) The results on dataset is always accurate, because it is calculated by accurate value of data objects. In the process of data stream, it only saves summary statics information of data objects that

will cause the approximate result. But the approximate result does not affect the detection of anomaly behaviors; (6) the time consumption on dataset is longer than that on data stream. In some related works, the data mining methods can get pretty better results, but it consumes too long time. To process the data object in time, the algorithm on data stream should have high efficiency; (7) Because of the reason in third point, the method based on dataset takes much more memory than that based-on data stream.

From the comparisons above, we conclude that the method based on data stream need to establish an efficient data stream analysis model, that is, the algorithm has a smaller time and space complexity. The limited computer storage capacity cannot save infinite data objects in data stream, which requires that the memory consumption of the algorithm dose not increasing with time and it can be some fixed value [4].

We design the improved anomaly detection model according to the characteristics of data stream clustering algorithm and it is shown in Fig. 1.

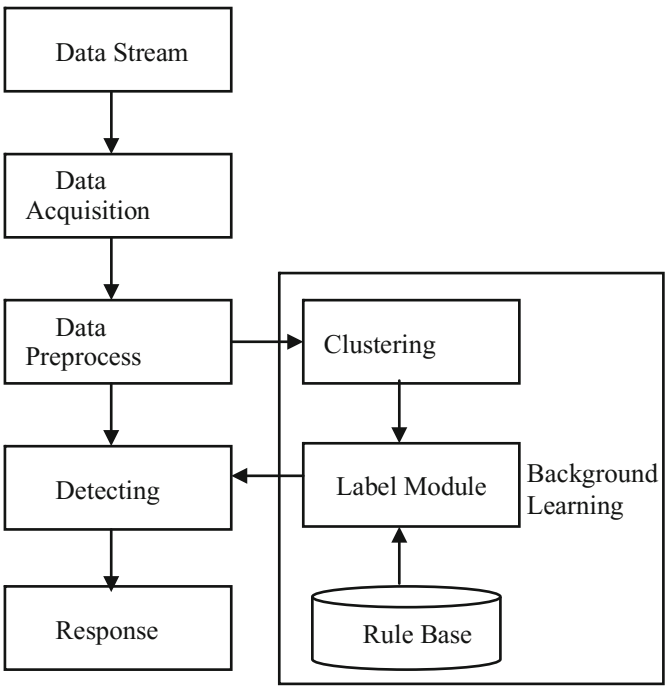


Fig. 1. Anomaly detection model. This shows the anomaly detection model which consists of several modules. Clustering module is the core of anomaly detection model.

The improved anomaly detection model consists of four modules. Data acquisition module collects data objects from data stream and data preprocess module preprocess data objects, including data cleaning and feature selection. The detecting module is composed of three parts. The clustering module is the core of model and it assign new data object into certain cluster. Suspected cluster will be sent to label module and it is

matched with the rules from rule base. Clustering algorithm is unsupervised and it cannot label these data objects. So, label module will label data objects by rule base.

3 Improved Data Stream Clustering Algorithm

In this paper, we propose the improved data stream clustering algorithm for anomaly detection. Figure 2 shows the diagram of clustering algorithm application.

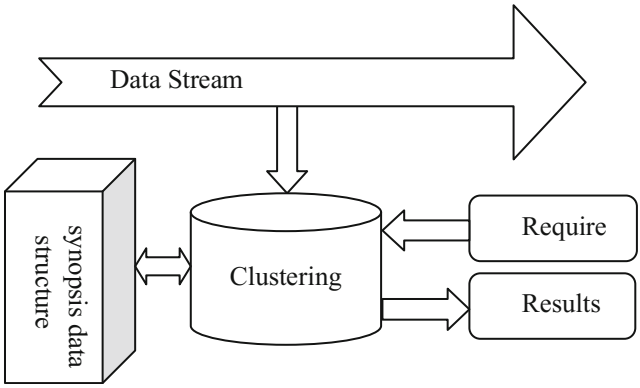


Fig. 2. Diagram of clustering algorithm application. This shows the basic application diagram of data stream clustering algorithm. *Synopsis data structure* is applied to store summary statics information.

Data stream S is a sequence of data objects o, which can be denoted as $S = \{o_1, o_2, o_3, \cdots, o_h\}$, and each data object has n features. The synopsis data structure is the important part of data stream clustering algorithm which is utilized to save the summary statistics information of clusters. In our improved data stream clustering algorithm, we use two types of synopsis data structures to store the summary statistics information of normal clusters and suspected clusters [5].

Normal cluster is denoted as $n - cluster$ which can save the necessary information. As Formula 1 shows, it has four attributes. δ is the number of data objects in the cluster; μ is the center of cluster and it is the average value of data objects; SS is the quadratic sum of data objects in the cluster. We add the attribute *flag* to identify the type of cluster.

$$n - cluster : (\delta, \mu, SS, flag). \tag{1}$$

Suspected cluster is denoted as $s - cluster$ and it is suspected to be abnormal behaviors. Formula 2 shows the five attributes of suspected cluster.

$$s-cluster:(\delta,\mu,SS,flag,list). \tag{2}$$

The attribute *list* stores the identifiers of data objects in suspected cluster and the detailed information of data objects are saved in the disk instead of memory. This attribute will increase the consumption of memory, but $s-cluster$ can be converted to $n-cluster$ and this attribute will be deleted in time. Besides, the amount of abnormal behaviors is far less than that of normal behaviors. This attribute will not occupy the memory for a long time. The main idea of improved data stream clustering algorithm is as follows and the flow diagram is shown in Fig. 3.

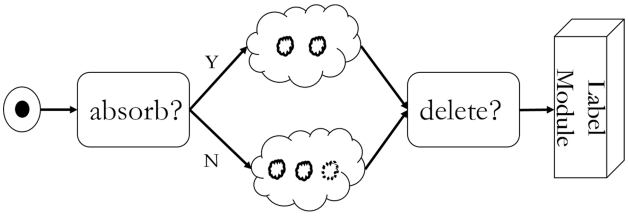


Fig. 3. Flow diagram of data stream clustering algorithm algorithm. This shows the processing of new arrived data object. The clustering algorithm decides it is absorbed by existed clusters or become new cluster. The redundant cluster should be deleted in time and sent to label module.

Step 1: At the beginning of algorithm, it waits to receive a certain number of data objects and the clustering algorithm based on dataset will generated some clusters. These clusters will be labeled as $n-cluster$ or $s-cluster$ according to the number of data objects in the cluster;

Step 2: When the new data object o arrives, it calculates the distance between data object and existed clusters. According to the distance, existed clusters decide whether to absorb this data object;

Step 3: If the distance is less than threshold, data object chooses the nearest cluster to be integrated with. The information of selected cluster should be updated as Formula 3;

$$(\delta,\mu,SS)\rightarrow(\delta+1,\frac{\mu\times\delta+o}{\delta+1},SS+o^2). \tag{3}$$

Step 4: If data object cannot be absorbed by existed clusters, it will be added into the memory as the center of new cluster and it is labeled as $s-cluster$. When the number of data objects in is more than the threshold, $s-cluster$ is transformed into $n-cluster$ and deletes the attribute from the memory;

Step 5: In order to limit the number of clusters in the memory, it should delete redundant clusters. Secure deletion method is to select $s-cluster$ which does not update for a long time;

Step 6: The center of clusters to be deleted is sent into label module. Because of the high similarity in clusters, data objects will be labeled according to the center.

4 Conclusion

In this paper, we propose the improved data stream clustering algorithm and design corresponding anomaly detection model. The clustering algorithm based on data stream is more appropriate than that based-on dataset. The clusters can be updated in time according to new arrived data object. The improved clustering algorithm requires less memory and time cost. The next works of us are to consider the situation of multiple concurrent and the improvement of label module. An appropriate method of feature selection will further improve efficiency.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (61373134). It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). We declare that we do not have any conflicts of interest to this work.

References

1. Lee, W., Stolfo, S.J., Mok, K.W.: Mining audit data to build intrusion detection models. In: International Conferenc e on Knowledge Discovery and Data Mining, pp. 66–72 (1998)
2. Silva, J.A., Faria, E.R., Barros, R.C.: Data stream clustering: a survey. ACM Comput. Surv. **46**, 125–134 (2013)
3. Aggarwal, C.C, Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: International Conference on Very Large Data Bases, VLDB Endowment, pp. 81–92 (2003)
4. Guha, S., Meyerson, A., Mishra, N.: Clustering data streams: theory and practice. IEEE Trans. Knowl. Data Eng. **15**, 515–528 (2003)
5. Chen, Y., Tu, L.: Density-based clustering for real-time stream data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, DBLP, San Jose, California, USA, pp. 133–142, August 2007

Application of an Improved Data Stream Clustering Algorithm in Intrusion Detection System

Chunyong Yin[1(✉)], Lian Xia[1], and Jin Wang[2]

[1] School of Computer and Software,
Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, Nanjing, China
yinchunyong@hotmail.com
[2] College of Information Engineering, Yangzhou University, Yangzhou, China

Abstract. With the continuous development of computer network technology, traditional intrusion detection system is short of good adaptability. Aiming at the traditional intrusion detection system is difficult to adapt to the increasing amount of data demand for real-time processing capability, this paper proposes a clustering algorithm based on sliding window data streams, based on which we build the IDS network security defense model. The experiment results show that the model is able to adapt to the high-speed network intrusion detection requirements.

Keywords: Data streams · Data mining · Clustering algorithm · Intrusion detection

1 Introduction

Intrusion detection is a kind of security mechanism which is used to detect attacks and intrusion behaviors [1]. Intrusion detection system (IDS) has become a needful component in terms of computer and network security IDS [2]. As a network security assistance mechanism, IDS can monitor the network and detect the unauthorized use or abnormal condition to achieve the purpose of network security defense by countering abnormal behaviors without affecting the performance of network.

With the development of network technology and the continuous improvement of performance, the function of network operating system is becoming increasingly complex, as well as the data source generated from the Internet becomes much huger than before [3]. Traditional IDS is difficult to adapt to the demand for real-time processing of the increasing amount of data. It's the key to ensure the effectiveness and real-time of IDS that how to extract a pattern with a certain characteristic in the massive data to describe the user behavior more accurately. Therefore, the theoretical research and practical application of intrusion detection technology in high-speed network environment are particularly important, which have become the frontier issues waiting to be solved in the field of network security. The breakthrough of related theories and technologies is of great significance for network intrusion detection, behavior analysis, content inspection and network management and control.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_99

Data mining is a technology committed to the analysis and comprehension of massive data, and revealing the inherent knowledge of data. Applying it to IDS aims to extract the information of detected user's behavior from massive data, analyze its behavior or characteristics, construct the detection engine of intrusion detection system, and finally realize the recognition and alarm of intrusion.

Although the detection research based on data mining has made many theoretical achievements, some systems have also been applied to a certain extent, but there are still many problems remaining to be solved in this field. First, the traditional data mining intrusion detection system can only be achieved through training data and assumptions. For example, we suppose that the training data with label is easy to obtain, the quality of the training data is very simple, the training data is representative and well-distributed, etc. However, it's very difficult to obtain the training data with label in actual network. On the one hand, manual data classification makes error prone. On the other hand, the detection model and the detection effect generated by training are often not ideal enough even though the virtual training data with label can be obtained. Second, the traditional intrusion detection system using data mining is often based on static data mining, which will result low speed of the response to new technology of network attacks and the disadvantage of poor real-time and so on. Detection systems, and the resulting defensive actions taken, are only effective if they can detect intrusions accurately and in a timely manner, minimizing the impact of the attacks [4]. How to improve the protocol identification and processing ability of intrusion detection system has become a hot spot in network security research, because of the characteristics like greater transmission and faster speed of data, more diverse means of attack and constantly updated network protocols in network and so on.

In the paper [5], it proposes the improved model which integrate K-anonymity with L-diversity and can solve the problem of imbalanced sensitive attribute distribution. It uses K-member clustering algorithm to realize the improved anonymity model which can reduce the algorithm execution time and information loss.

2 Data Flow Mining

2.1 Traditional Algorithm

Traditional data mining algorithms are based on static data with the feature of a small amount of real-time updates, which will be inquired repeatedly. General static data mining process can be divided into two parts. Data is firstly collected in database or other fixed place, then the information contained is excavated through a variety of analysis and mining technology. Data flow is defined as a sequence of real-time, continuous and sequential records whose arrival order is uncontrollable and they cannot be stored in permanent media due to limited resources such as memory and hard disks. As the amount of data grows indefinitely, data flow is limited to single pass scanning, and each data can be processed only once. The fast liquidity of data flow requires that the rate of analysis and processing of the algorithm cannot be lower than the flow velocity of data flow. Data flow mining algorithm framework is shown in Fig. 1.

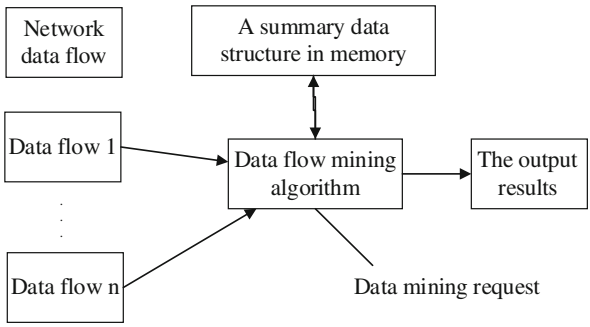


Fig. 1. Data flow mining algorithm architecture diagram

Data flow mining algorithm can be divided into clustering and classification algorithms. Many anomaly detection methods based on the clustering are applied to detect intrusion for reducing the false positive rate and improving the detection efficiency [6]. The clustering analysis divides data objects into several groups to make the similarity of them in the group as high as possible while the similarity outside is as low as possible. Typical representative algorithms include K-means, DBSCAN, and CluStream. Both K-means and DBSCAN algorithm have problems of high computational complexity, high temporal complexity and high spatial complexity because data need to be once processed when the amount of data is large [7]. In addition, the two algorithms cannot deal with the clustering problem in the evolutionary data stream at different time intervals. CluStream algorithm can cluster the data flow in the landmark window while it cannot meet the clustering requirements under the sliding window. The micro-clusters formed by CluStream do not accurately reflect the data distribution in the current data stream. The micro-clusters and outliers in the data flow may be interchanged for the data flow changes continuously. In addition, the CluStream algorithm maintains a fixed number of micro-clusters. When there is noise in the data flow, the algorithm will set up micro-clusters for noise which results in a reduction in the number of true clusters, so that the algorithm becomes unstable.

2.2 Data Stream Clustering Algorithm Based on Attenuating Sliding Window

Cluster analysis of data flow is performed at a specific time interval (called a window). This time interval (window) can be divided into landmark windows, sliding windows, and attenuating sliding windows by type. The landmark window is the set of tuples with the data range of the mining data set ranging from the data stream of the starting landmark to the ending landmark (current), whose size changes as the data flows in and out. While the mining data set range starting from the current landmark and pushing w tuples forward, we call it sliding window, which is also defined as the set of data objects of the most recently arrived w tuples. W is the sliding window, the location of the window changes continuously with the flow of data objects. The attenuation sliding

window is the set of mining tuples for all data sets ranging from the starting point of data stream to the currently arriving data point. It set a different weight for each tuple, which decays with time according to a certain attenuation function.

The data stream clustering algorithm is used to collect all the data in each sub-window in the sliding window in order to cluster them when processing the data flow. Results are divided into multiple candidate micro-clusters and critical candidate micro-clusters according to different features, which are stored in different caches. After all the data in the sliding window has been processed, then the micro-clusters stored in the respective buffers are waiting for processing. The micro-clusters stored in each buffer area are combined to obtain new micro-clusters while some non-merged micro-clusters (such as: critical candidate micro-clusters) are preserved. The algorithm then merges these micro-clusters into the pyramidal structures that have been established in an off-line way. The current clustering is preserved in the form of snapshots and the historical results of clustering are preserved using an attenuation model. These results have a certain influence decreasing with time on the current clustering process. The process of micro-cluster formation algorithm is as follows:

```
SWCStream (DS, λ1, β, λ2, μ, δ)
Get the next point p at current time t from data stream DS;
Merging (p);//Merge the received data;
if rp (the new radius of Cp) < λ1
Merge p into Cp
else
if ro (the new radius of Co) < λ2
    Merge p into Co;
if w (the new weight of Co) > β*μ
    Delete Co;
end if
else
    end if
end if
if (t mod Tp ) = 0 then //Query Cp every Tp
for each p-micro-cluster Cp do
    if wp (the weight of Cp) < β*μ
        Delete Cp;
    end if
end for
for each o-micro-cluster Co do
    if wo (the weight of Co) < ξ
        Delete Co;
    end if
end for
end if
if a clustering request arrives then
    Generating clusters;
end if
```

In this algorithm, C_p represents candidate micro-clusters, C_o represents critical candidate micro-clusters, T_p represents the time required for candidate micro-clusters to

degenerate to the critical candidate micro-cluster points and β is the lower bound of weight of outliers. The attenuation factor is α, λ_1 and λ_2 respectively represents the thresholds of the candidate micro-cluster C_p and the critical candidate micro-cluster C_o.

$$\xi = 2^{-\delta\left(t-t_0+T_p\right)-1} / \left(2^{-\delta T_p}-1\right) \tag{1}$$

$$T_p = \left[\frac{1}{\alpha}\lg\left(\frac{bu}{bu-1}\right)\right] \tag{2}$$

2.3 IDS Model of Data Stream Clustering Algorithm Based on Attenuating Sliding Window

Aiming at the nature and characteristic of data flow, this paper designs IDS model of data stream clustering algorithm based on attenuation sliding window. The model is

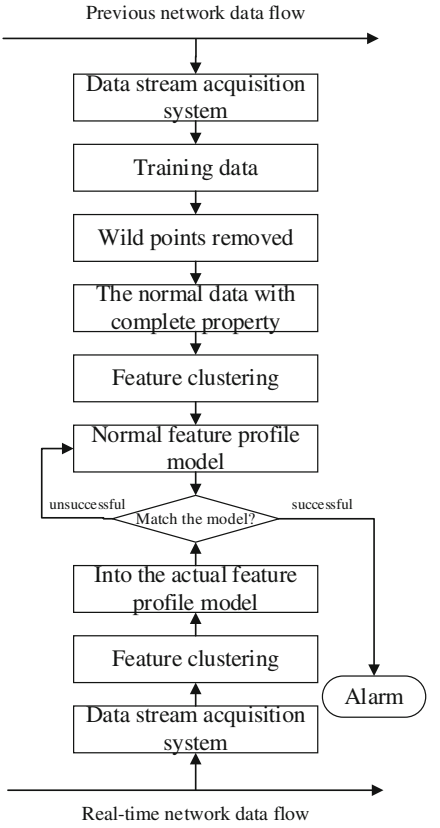


Fig. 2. Intrusion detection system model diagram based on data flow mining

mainly composed of data acquisition, feature clustering, pattern matching and alarm. There are two stages of intrusion detection. In the first stage, the data of the network data flow is sampled as the training data, and clustered by the data stream clustering algorithm using in the sliding window in the environment that the network is not under attack, to form the normal feature pattern, which is the basis of the detection. In the second stage, the incoming real network information is collected by the acquisition device. The clustering algorithm is also used to cluster the information characteristics of incoming data. Matching the feature with the normal training characteristics of the existing system, if succeeded, the information flowing into the system does not contain intrusion information and on the contrary, it contains intrusion behavior. Then the system alerts, prompting the administrator to deal with the intrusion behavior and obtaining the latest data samples to rebuild the cluster model regularly from the sliding window. Figure 2 depicts an intrusion detection model based on data flow mining.

3 Verification

In data stream mining, it is synchronized from establishment to application in the model, which will be dynamically updated with the data flow in real time. In actual operation, the data in KDD CUP 1999 are still used as experimental data. The data consists of the following two parts:

(1) 10% of kddcup.data.gz (training data with class labels) data, a total of 494021 records;
(2) All data in kddcup.testdata.unlabeled.gz (test data without class labels), a total of 197608 records.

The attack detection results of the four major attack types are shown in Table 1.

Table 1. Test results of four major attack types (%)

Attack type	The proportion	Average false positive rate	Average missed alarm rate
DOS	7.92	2.37	1.55
U2R	0.54	0	1.02
R2L	0.72	0.45	0.67
Probing	0.82	1.12	1.13

As shown in Table 1, it is the average false alarm rate and the average missed alarm rate for the four types of attacks. We can see that the detection rate of DOS is lower than that of other three types, and the detection rate of R2L attack is higher.

Table 2. Statistical results of IDS model (%)

Accuracy	False positive rate	False negative rate
98.18	0.81	1.01

As shown in Table 2, the IDS network security defense model of data stream clustering algorithm based on attenuating sliding window can well meet the requirements of intrusion detection in high-speed networks. System based on the model showed better detection performance, higher detection rate and lower false alarm.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (61373134). It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). We declare that we do not have any conflicts of interest to this work.

References

1. Yin, C.Y., Ma, L.Y., Feng, L.: A feature selection method for improved clonal algorithm towards intrusion detection. Int. J. Pattern Recogn. Artif. Intell. **30**(5), 1–14 (2016)
2. Rajput, R., Mishra, A., Kumar, S.: Optimize intrusion prevention and minimization of threats for stream data classification. In: Fourth International Conference on Communication Systems and Network Technologies, pp. 408–413 (2014)
3. Yin, C.Y., Feng, L., Ma, L.Y., Kim, J.Y., Wang, J.: An effective feature selection and data-stream classification model HDP. J. Internet Technol. **17**(4), 695–702 (2016)
4. Chen, Z.J., Zhang, H.L., Hatcher, W.G., Nguyen, J., Yu, W.: A streaming-based network monitoring and threat detection system. In: IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 31–37 (2016)
5. Yin, C.Y., Zhang, S., Xi, J.W.: An improved anonymity model for big data security based on clustering algorithm. Concurr. Comput. Pract. Exp. **29**(7), 1–13 (2016)
6. Gao, H.W., Zhu, D.J., Wang, X.M.: A parallel clustering ensemble algorithm for intrusion detection system. In: Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, pp. 450–453 (2011)
7. Zhu, C.S., Dun, X., Zhu, L.: A study on the application of data stream clustering mining through a sliding and damped window to intrusion detection. In: Fourth International Conference on Information and Computing, pp. 22–26 (2011)

Short Text Classification Technology Based on KNN+Hierarchy SVM

Chunyong Yin[1(✉)], Lingfeng Shi[1], and Jin Wang[2]

[1] School of Computer and Software,
Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, Nanjing, China
yinchunyong@hotmail.com
[2] College of Information Engineering, Yangzhou University, Yangzhou, China

Abstract. A short text classification method based on combination of KNN and hierarchical SVM is proposed. First, the KNN algorithm is improved to get the K nearest neighbor class labels quickly, so as to effectively filter the candidate classes of documents. And then classify them from top to bottom using a multi-class sparse hierarchical SVM classifier. By this way, the document can be classified efficiently.

Keywords: KNN · Hierarchical SVM · Candidate classes · Short text

1 Introduction

The popularity of the use of Internet demands the technology of short text classification to deal with the ubiquitous data, such as Internet news, blog and mail, etc. The technology, known as text mining, is that automatically extracts valuable information and knowledge from those data mentioned which has been mentioned above. According to the length of text, text mining can be divided into long text mining and short text mining, while this two text mining methods did not clearly distinguish in the early stages of this technology research [1]. With the rise of social media, mobile text messages [2], Tweet and microblogging and other short text are emerging uncontrollably. The growing number of users of these applications makes the size of short texts larger and larger. In addition, the short text in the search engine, automatic questioning and topic tracking and other fields play a critical role. By and large, short text mining is increasingly concerned by researchers [3].

The popular short text classification algorithms include K Nearest Neighbor (KNN) algorithm and Support Vector Machines (SVM) algorithm. Specifically, KNN and SVM methods have a huge advantage on the recall rate and accuracy. Although KNN algorithm is simple in principle and its classification efficiency is high enough, it is an instance-based statistical learning method which is not very accurate for classifying samples at class boundaries. The SVM classification algorithm aims to maximize the distance between the classification boundaries, so the classification accuracy is relatively high. However, it also reduces to the process of training classifier

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_100

relatively slow. In short, the use of either of these two methods alone is difficult to achieve the desired classification efficiency and effectiveness.

Therefore, by combining KNN and SVM algorithms, the researchers can not only improve the accuracy of classification, but also improve the efficiency of classification, which can automatically classify the mass documents to achieve better results. When the classification structure of the document forms a hierarchical directory, the hierarchical classification algorithm not only can significantly improve the classification efficiency, but also improve the classification accuracy. However, in most cases, it is difficult to use a flat class directory to contain it because the rich document set contain too much types of semantics. In this paper, we propose a short text classification technique based on KNN and hierarchical SVM classification [4]. In the training phase, we use the multi-class SVM algorithm to uniformity study the hierarchical directory of the samples, rather than independently learning multiple binary classifiers, so that we can compare the output of the classification surfaces more effectively. In the classification stage, the KNN is used to filter the labels of classified samples, and then the studied SVM classifier is used to classify the label of classification candidate sets which are to be classified from top to bottom. In this way, the number of candidate SVM classification face is reduced effectively and accelerate classification process. At the same time, the interference of some unrelated classification face is eliminated, and the accuracy of automatic classification is improved.

2 Related Content

2.1 KNN Classification Algorithm

K Nearest Neighbor (KNN) algorithm is an instance-based statistical learning method, and its main idea is to calculate the K neighbors nearest to the samples which are to be tested in the input feature space and to "vote" the final class label of the new sample through these nearest neighbor class labels [5].

Let the total number of training samples be n, and the posterior probability of the sample x to be sampled from the labeled sample class w_i is $P_n(w_i|x)$, that is the probability that the tested sample x belongs to class w_i. Since the text is represented by the vector space model, it can be assumed that the space which volume is V and is around the sample point x can contain exactly k samples, usually $k = \sqrt{n}$. Where k_i samples belong to class w_i, then $\sum_{i=1}^{c} k_i = k$. Thus, the joint probability density of samples and class w_i is estimated as:

$$P_n(x, w_i) = \frac{k_i/n}{V} \tag{1}$$

The posterior probability is estimated as:

$$P_n(w_i|x) = \frac{P_n(x, w_i) \times V}{P_n(x) \times V} = \frac{P_n(x, w_i)}{\sum_{j=1}^{c} {}_n(x, w_j)} = \frac{k_i}{k} \tag{2}$$

That is to say, the posterior probability of the point x belonging to the class w_i is the ratio of the number of the sample points marked w_i in the volume V to the number of all the sample points in the volume. In this way, in order to achieve the minimum error rate, we choose the class make the ratio largest as the final classification of the results of discrimination.

However, the KNN algorithm usually needs to compute the distance of all the training samples to be sampled and sort them, so as to select the nearest K neighbors. Assuming that the feature dimension of each sample is d, so the time complexity of the above steps is $n \times d + n \log n$. In the classification of mass text, the value of n is often large, the feature dimension is relatively high. Therefore, in order to speed up the efficiency of KNN algorithm, the classification efficiency of the algorithm is usually improved from these two aspects: (1) Reduce the dimension of the sample and select the smallest features to represent the text vector. This method is often more intuitive, but when the dimension is too small, the classification effect will be apparently reduced; (2) Merge similar texts in the training set and then treat it as a document appropriately. In this way, the number of documents that need to compare is significantly reduced. Here we use the following method. We classify the internal documents of each category in each natural category, and then cluster them into j subclasses. And then calculate the center vector of each subclass. Finally, we calculate the distance between the samples to be classified and the center vectors of these subclasses, so as to find out the nearest neighbor centers. The efficiency of KNN classification algorithm is obviously improved because the number of text contained in each category decreases sharply after clustering.

2.2 Hierarchical SVM Classification Algorithm Formulas

SVM method has a more complete theoretical basis and SVM also shows the superior performance of classification in a variety of practical applications [6]. What's more, SVM also has high computational efficiency and can handle large-scale data efficiently. The SVM uses the training data to model the maximum interval hyper plane, and then classifies the unclassified data using the hyper plane as the decision boundary. The maximum interval is the largest value of the minimum geometric interval between the training set sample points and the hyper plane, and the larger the interval, the smaller the generalization error, and then the stronger the ability of classifying the new data. The final classification of the hyper plane model actually only needs to use a few training samples that are closest to the hyper plane, these samples is the "support vector", the other which is not support vector training sample points have no effect on the classification of hyper planes, so support vector machine method has a high stability. Figure 1 shows the support vector machine.

In practice, the category structure of the document usually has a clear hierarchical distribution rather than a single flat structure. When multiple classes form a hierarchical tree, the researchers found that the hierarchical classification model is faster and sometimes even more accurate than its corresponding single-level classification model. Therefore, our classification model is based on the hierarchical classification framework. In this paper, we will study a multi-class hierarchical SVM classifier. The objective

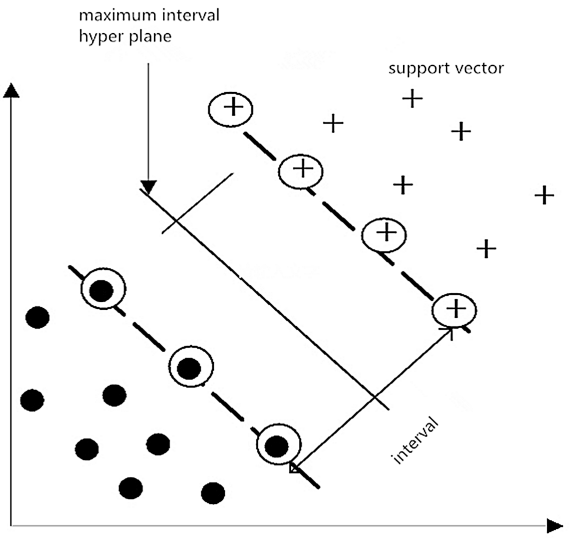


Fig. 1. Sketch map of the maximum interval hyper plane and support vector

function of each classifier is implemented in the same optimization function, instead of training multiple binary classifiers separately.

Let $A(i), C(i), D(i), S(i)$ denote the ancestor, child, descendant and sibling of node i in the hierarchical category, and $A^+(i) = A(i) \cup i$. $x \subset R^d$ denotes a set containing d-dimension training text feature vector. $Y = \{1, 2, \ldots, m\}$ is the number corresponding to the node category in the hierarchical category directory excluding the root node. The training process of hierarchical SVM classification is as follows: Given a set of training text $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}, x_k \in X, y_k \in Y, k \in \{1, \ldots, N\}$. Learning m SVM classification surface $w = \{w_i\} \subset R^d, i = 1, \ldots, m$. Each classification surface correspond to level of a directory the node i in hierarchical directory. We need to solve the following optimization goals:

$$\text{minimize } C_1 \sum_{i=1}^{m} \|w_i\|_1 + \frac{1}{2} C_2 \sum_{i=1}^{m} w_i^T w_i + \frac{C_3}{N} \sum_{k=1}^{N} \zeta_k \tag{3}$$

subject to

$$\begin{array}{cc} w_i^T x_k - w_j^T x_k \geq 1 - \xi_k & \forall j \in S(i) \\ \forall i \in A^+(y_k) & \forall k \in \{1, \ldots, N\} \\ \xi_k \geq 0 & \forall k \in \{1, \ldots, N\} \end{array}$$

In which, the first two terms are mixed $L1$ sparse regularization and $L2$ regularization terms, and the third term is the loss function. C_1, C_2 and C_3 are parameters that control the balance of regularization terms and loss functions. The loss function will punish this case in which difference between the classification output of correct label at the current level and it of other easily confused sibling node is less than 1, when a

training text k corresponds to all nodes on the hierarchy tree from the correct leaf node y_k to the root node path. The smaller the difference is, the larger the corresponding loss item is, so as to effectively increase the discrimination ability of similar categories at each level. In multi-level hierarchical classification, support vector in support vector machines tend to become more intense. From the point of view of reducing the storage cost and the time of classification, we choose to study a frugal model in which each classification surface is composed of only the weight of several sparse features. Therefore, in the hierarchical learning framework, we introduce the $L1$ sparse regularization term to punish the parameters of the classification surface. Although the $L1$ regularization term contains absolute value operations, it is not difficult to prove that the objective function of multi-class SVM classification is still a convex optimization problem for the parameter w. Therefore, we can easily use some effective optimization methods to solve the above optimization problems. We will describe the training process for this model in the next section.

3 KNN+Hierarchical SVM Classification Framework

3.1 Training Hierarchical SVM Classification Model

In order to train the hierarchical SVM model, we rewrite the objective function:

$$\text{minimize}\, J(w) = \Omega(w) + H(w) + r(w) \tag{4}$$

In which:

$$\Omega(w) = \frac{1}{2} C_2 \sum_{i=1}^{m} w_i^T w_i$$

$$H(w) = \frac{C_3}{N} \sum_{k=1}^{N} \max\{0, \max_{i \in A+(y_k)}^{j \in S(i)} \{1 - w_i^T x_k + w_j^T x_k\}\}$$

$$r(w) = C_1 \sum_{i=1}^{m} ||w_i||_1$$

Since the first two equations can be derivative on the right, we can solve it by calculating the sub-gain of $\Omega(w)$ and $H(w)$. We use a two-stage algorithm to solve the non-derivable $r(w)$, that is the sparse problem in the regularization term. That is to say, in each iteration t, we firstly ignore $r(w)$ and update the parameter w^t in $\Omega(w)$ and $H(w)$ using the regularized dual averaging method RDA and get the temporary intermediate variable $w^{t+1/2}$. Then the FOBOS update mode is used to solve the $L1$ regularization term in $r(w)$, and the new parameters in the $(t+1)th$ iteration are obtained as follows:

$$w_{ij}^{t+1} = sign(w^{t+\frac{1}{2}}) \left[w^{t+\frac{1}{2}} - \lambda \right]_+ \tag{5}$$

3.2 KNN+Hierarchical SVM Algorithm Flow

In the training phase of the algorithm, the two parts are carried out separately. KNN training process is to cluster the subclasses of each class and find the optimal K value [7]. The training of SVM classifier has been shown in A, and we main obtain the parameters of each classification surface from the hierarchical tree. In the actual classification stage, the algorithm first computes its nearest neighbor centers by KNN classification algorithm, then counts all the classes in its K nearest neighbors, and calls the corresponding hierarchical SVM classifiers to classify each class. The flow of the KNN+Hierarchy SVM algorithm is as follows:

Steps:

(1) The nearest training sample center (subclass center vector) is selected by the distance function, where k is the optimal parameter trained by KNN.
(2) For each class w_i corresponding to the k sample centers, we retain its corresponding hierarchical path as candidate set, input the feature vectors of the sample x to the SVM classifier of each level corresponding to the candidate set, calculate the similarity between sample x and path on all types.
(3) If the similarity is the largest, the hierarchical path category label corresponding to the category w_i is taken as the classification result of the sample x, and the algorithm ends.

"KNN+Hierarchy SVM" classification algorithm combines the timeliness of KNN algorithm and the accuracy of SVM algorithm. The SVM classifier is used to classify neighbor labels obtained by the KNN classifier as the candidate label set, and the accuracy is high. This method is more effective especially when class labels are large. KNNs can be used to filter out SVM classifiers corresponding to classes that do not need to be called explicitly.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (61373134). It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). We declare that we do not have any conflicts of interest to this work.

References

1. Zhang, H., Zhong, G.: Improving short text classification by learning vector representations of both words and hidden topics. Knowl. Based Syst. **102**, 76–86 (2016)
2. Yin, C., Xi, J.: Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm. Multimedia Tools Appl., 1–17 (2016)
3. Altınel, B., Ganiz, M.C.: A new hybrid semi-supervised algorithm for text classification with class-based semantics. Knowl. Based Syst. **108**, 50–64 (2016)
4. Yin, C., Xiang, J., et al.: Short text classification algorithm based on semi-supervised learning and SVM. Int. J. Multimedia Ubiquit. Eng. **10**(12), 195–206 (2015)

5. Tang, X., Xu, A.: Multi-class classification using kernel density estimation on K-nearest neighbours. Electron. Lett. **52**, 600–602 (2016)
6. Dong, C., Zhou, B., Hu, J.: A hierarchical SVM based multiclass classification by using similarity clustering. In: International Joint Conference on Neural Networks, pp. 1–6 (2015)
7. Tanveer, M., et al.: An efficient regularized K-nearest neighbor based weighted twin support vector regression. Knowl. Based Syst. **94**, 70–87 (2016)

Spectral Response Based Regularization Parameter Selection for Total Variation Image Restoration

Yuhui Zheng[1(✉)], Min Li[1], Kai Ma[1], Shunfeng Wang[2], and Jin Wang[3]

[1] College of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
zheng_yuhui@nuist.edu.cn, limin_513@126.com, mkl9940208@163.com

[2] College of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
wsfnuist@126.com

[3] College of Information Engineering, Yangzhou University, Yangzhou 215127, China
jinwang@yzu.edu.cn

Abstract. This letter introduces a total variation (TV) image restoration method with adaptive Regularization parameter. Our main contributions are two folds: (1) a novel selection scheme that determines the Regularization parameter of TV model in a global way through exploiting the concept of TV spectral response; (2) an efficient algorithm integrating an estimation-and–renewal strategy and the alternating minimization numerical technique to fast calculate the model solution. Experimental results on degraded images indicate the improved performance of our method, both in visual effects and in quantitative evaluations.

Keywords: Total variation · Regularization parameter · Spectral response index

1 Introduction

Image is unavoidably degraded during its acquirement process The classical TV model [1] is written as

$$\min_I \left\{ \sum_{i=1}^{N} \|D_i I\|_2 + \frac{\lambda}{2} \|HI - g\|_2^2 \right\} \tag{1}$$

where $I \in \mathbb{R}^N$ and $g \in \mathbb{R}^N$ are the restored image and the degraded image, respectively, $D_i I$ represents the gradient of image I at point i, N is the size of image I, $H \in \mathbb{R}^{N \times N}$ denotes a linear blurring operator, and $\lambda > 0$ is the Regularization parameter. To date, various schemes on estimating Regularization parameter have been presented, including the discrepancy principle based [2], the L-curve based [3], the structure tensor based [4] and the residual image statistics (RIS) based [5]. However, most of

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_101

image restoration methods with these schemes are relatively high computational cost. Although fast algorithms [6, 7] for TV model using the variable-splitting-and-penalty (VDP) strategy have drawn much attention recently, it seems to be intractable to straightforward combine them with the above-mentioned selection schemes. Determination of the Regularization parameter for fast TV image restoration is till an open problem. We attempt to address the problem in this work.

2 Motivation

Lately, based on the scale space produced by the TV flow (TVF), Gilboa [8] suggested a TV transform defined by $\phi(T) = TI_{tt}(T)$, where T $\in (0, \infty)$ is the evolution time and $I_{tt}(T)$ is the second time derivative of TVF at scale (time) T. Owing to the fact that $I_{tt}(T)$ generates impulse response to the elementary TV components of image, the TV transform can be interpreted as a spectral domain, therefore allowing one to explore certain spectral information measure for image analysis, such as the TV spectral response (SR) [8]

$$S(T) = \sum_{i=1}^{N} |\phi_i(T)| \tag{2}$$

$S(T)$ (hereafter called SR-index for short) describes the amplitude of $\phi(T)$. From the view of the scale space, the SR-index plays a role of "sensor" to show the smoothed-out acceleration of image details in restoration. Naturally, in order to enhance the restored result of The TV model, it is plausible to limit the penalty extent of the Regularization term according to the SR-index. In other words, during image restoration, the higher the SR-index, the more important the data-fidelity term.

TV Image Restoration Using Proposed Selection Scheme: With the aforementioned consideration, a new TV model is presented as follows:

$$\begin{cases} \min\limits_{I} \left\{ \sum\limits_{i=1}^{N} \|D_i I\| + \frac{\psi(\bar{S})}{2} \|HI - g\|_2^2 \right\} \\ \psi(\bar{S}) = 1 - exp\left(-\frac{\bar{S}}{\gamma}\right), \gamma = 0.02\delta_g^2 \end{cases} \tag{3}$$

where $\psi(\bar{S})$ a scalar function of SR-index, $\bar{S} = S/N$ is called mean SR index here, $\gamma > 0$ is an empirical parameter tuning the decay of the exponential expression, and δ_g^2 denotes the variance of image g. Note that $\psi(\cdot)$ is a monotonically increasing function. Therefore, high SR index means that large weight $\psi(\bar{S})$ will be assigned to the data-fidelity term for perserving detail structures and features of image.

Here we employ the alternating minimization (AM) algorithm to solve (3). At first, (3) is equivalently transformed into the following function by introducing an auxiliary variable $w_i = (w_1, w_2)_i$ as follows:

$$\min_{I,w}\left\{\sum_{i=1}^{N}\|w_i\|_2+\frac{\beta}{2}\|w_i-D_iI\|_2^2+\frac{\psi(\bar{S})}{2}\|HI-g\|_2^2\right\} \tag{4}$$

where $\beta>0$ is a penalty parameter, set to be large value to ensure that the solution of (4) approximates to that of (3). Then, the minimization of (4) is accomplished by alternatively updating I and w as:

$$w_i^{n+1}=max\left\{\left(|D_iI^n|-\frac{1}{\beta}\right),0\right\}\frac{D_iI^n}{\|D_iI^n\|} \tag{5}$$

$$\begin{cases} I^{n+1}=\mathcal{F}^{-1}\left(\frac{A^w+B^w+\left(\frac{\psi(\bar{S}^n)}{\beta}\right)G}{P+Q+\left(\frac{\psi(\bar{S}^n)}{\beta}\right)M}\right) \\ A^w=\mathcal{F}(D^1)^*\circ\mathcal{F}\left(w_1^{n+1}\right),P=\mathcal{F}(D^1)^*\circ\mathcal{F}(D^1) \\ B^w=\mathcal{F}(D^2)^*\circ\mathcal{F}\left(w_2^{n+1}\right),Q=\mathcal{F}(D^2)^*\circ\mathcal{F}(D^2) \\ G=\mathcal{F}(H)^*\circ\mathcal{F}(g),M=\mathcal{F}(H)^*\circ\mathcal{F}(H) \end{cases} \tag{6}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the Fourier transform and its inverse transform, respectively, * stands for the conjugate transpose operator,D^1 and D^2 respectively compose of the first and second elements of D_i for all pixel i, $\circ$ denotes the component-wise multiplication, and the division is also component-wise. In this letter $\bar{S}^n$ is computed by

$$\bar{S}^n=\frac{n\Delta t}{N}\sum_{i=1}^{N}\left|\left(I_{tt}^n\right)_i\right|,I_{tt}^n=\frac{I^{n+1}-2I^n+I^{n-1}}{(\Delta t)^2} \tag{7}$$

where Δt is the time-step. From (7) we can see that image I^{n+1} is actually unknown beforehand in iteration. To deal with this problem, we carry out a strategy of estimation-and-renewal, which firstly utilizes the AM algorithm with constant Regularization parameter to pre-estimate next-scale image I^{n+1} for $\bar{S}^n$ computation, and then update image I^{n+1} using the AM algorithm with adaptive Regularization parameter $\psi(\bar{S}^n)$.

In summary, our introduced algorithm is implemented as follows:

Step1. Input captured image g, blurring matrix H, parameters Δt, β and λ, and iteration stopping tolerance ε;
Step2. Calculate w_i^1 with (5);
Step3. Compute I^1 using (6) with constant Regularization parameter λ, that is replacing $\psi(\cdot)$ of (6) with λ;
Step4. Compute w_i^{n+1} with (5);
Step5. Pre-estimate image I^{n+1} according to (6) using constant λ;
Step6. Compute $\psi(\bar{S}^n)$ according to (3) and (7);
Step7. Renew I^{n+1} using (6) with adaptive parameter $\psi(\bar{S}^n)$;
Step8. Repeat Steps 4–7 until satisfying stopping criterion.

3 Experimental Results

In experiments a closed related work called TV_RIS [5] is choosed to be compared with the herein proposed method, for the reason that its scheme on estimating the Regularization parameter exploits the scale space technique as well. In addition, we also evaluate two TV image restoration approaches with VDP strategy based fast algorithms, that is TV_AM [6] and its modified version TV_ADM [7], which utilize contant Regularization parameter through whole recovering process. The main parameters of our method are set as: parameters β and λ are determined according to the typical AM algorithm in [6], the time-step is $\Delta t = 1$, and the stopping tolerance is $\varepsilon = 10^{-10}$. All experiments are carried out with Matlab R2012b, on a computing platform with Intel Pentium Dual-Core 3.3 GHZ CPU.

Figure 1 illustrates the restored results on a corrupted satellite image, Fig. 1b, generated by firstly blurring the original image, Fig. 1a, with a Gaussian kernel of size 11 and then adding Gaussian noise with zero mean and variance $\sigma^2 = 1$. The quantitative comparison in terms of peak signal to noise ratio (PSNR) and time-consumption are shown in Fig. 1a–f Original image, degraded image and restored images of TV_AM, TV_ADM, TV_RIS and our method, respectively.

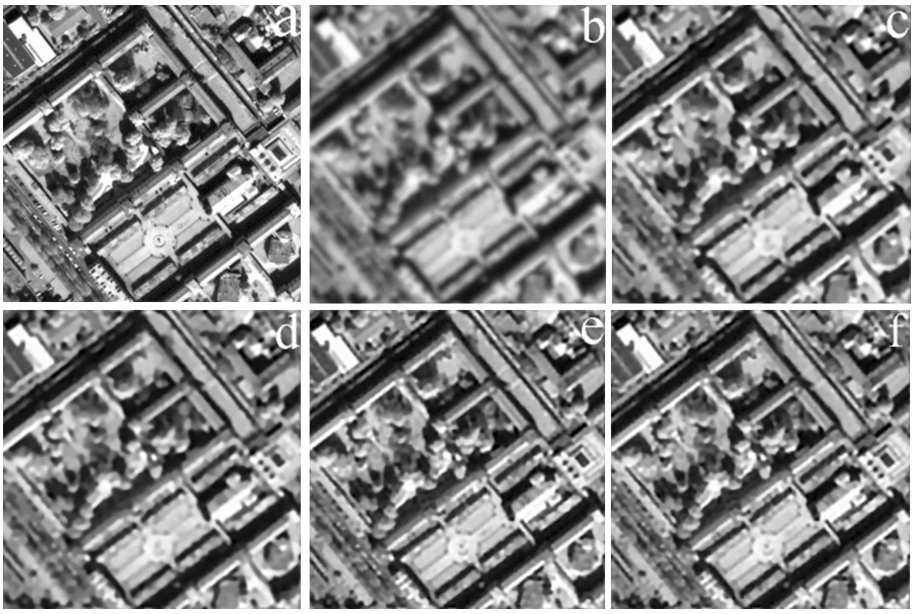


Fig. 1. Restoration results on Test 1 image (satellite image)

Figure 2a–f Original image, degraded image and restored images of TV_AM, TV_ADM, TV_RIS and our method, respectively.

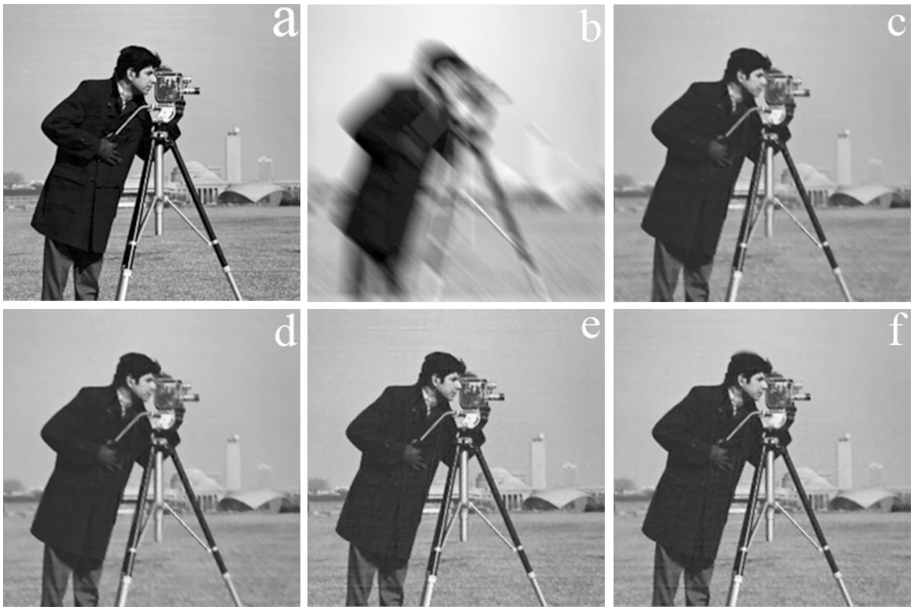


Fig. 2. Restoration results on Test 2 image (standard natural image)

Table 1. Comparison of the four methods in PSNR (dB)

Image	Degraded image	TV_AM	TV_ADM	TV_RIS	Our method
Test 1	12.18	20.23	20.45	25.69	25.26
Test 2	8.261	17.50	17.64	21.54	20.72

Table 2. Comparison of the four methods in time-consumption (s)

Image	TV_AM	TV_ADM	TV_RIS	Our method
Test 1 (400 × 400)	1.802	1.134	15.90	8.84
Test 2 (256 × 256)	1.323	0.749	9.29	4.180

Tables 1 and 2, respectively. We can observe from Fig. 1c to f that TV_RIS and our method outperform TV_AM and TV_ADM. The edges of buildings and streets in Fig. 1e and f appear more clear. This is mainly contributed to the adaptive Regularization parameter schemes using by TV_RIS and our method, which can balance the two term of the TV model in image restoration. As displayed in Table 1, the PSNR value of Fig. 1e is slightly higher than that of Fig. 1f, by the reason that TV_RIS locally determines the regularization parameters, which takes into account the local image contents. However, from Table 2 we can see that TV_RIS comes at a cost of high time-consumption. In contrast, our method with global selection scheme based on SR index not only produces good results but also costs relatively less computing time.

In Fig. 2 we also compares the four methods on a degraded natural image, Fig. 2(b), which is yielded by motion blur with moving direction $\frac{3}{4}\pi$. In addtion, the noise level of Fig. 2(b) is $\sigma^2 = 1$. Once more, this experiment verifies that our method can recover image in a comparatively high cost-effective way.

4 Conclusion

We have developed an selection scheme of Regularization parameter and a fast algorithm using the AM technique for TV image restoration. The scheme employs the SR-index to construct an adaptive function, which can adjust the relative weights of the two terms of the TV model to restore more image details. Since the SR-index calculation requires the image of next iteration, the introduction of an estimation-and–renewal strategy enable the AM algorithm to be applied to fast solve our restoration model. The numerical results show that our method can yield visually satisfying images while with relatively low time consuming. It should be mentioned that to further enhance our method performance, our next work will focus on exploring localized SR-index to estimate Regularization parameter in a local manner.

Acknowledgments. This work was supported by the NSFC under Grants 61402235, the National Research Foundation of Korea under Grant 2011-001-7578, the Natural Science Foundation of Jiangsu Province under Grand BK20150923 and the PAPD.

References

1. Hao, W., Li, J.: Alternating total variation and non-local total variation for fast compressed sensing magnetic resonance imaging. Electron. Lett. **51**, 1740–1742 (2015)
2. Wen, Y.W., Chan, R.H.: Parameter selection for total-variation-based image restoration using discrepancy principle. IEEE Trans. Image Process. **21**, 1770–1781 (2012)
3. Rezghi, M., Hosseini, S.M.: A new variant of L-curve for Tikhonov regularization. J. Comput. Appl. Math. **231**, 914–924 (2009)
4. Zheng, Y., Byeungwoo, J., Zhang, J., Chen, Y.: Adaptive determining regularization parameters in non-local total variation regularization for image denoising. Electron. Lett. **51**, 144–145 (2015)
5. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Variational denoising of partly textured images by spatially varying constraints. IEEE Trans. Image Process. **15**, 2281–2289 (2006)
6. Wang, Y., Yang, J., Yin, W., et al.: A new alternating minimization algorithm for total variation image reconstruction. Siam J. Imaging Sci. **1**, 248–272 (2008)
7. Yang, J., Zhang, Y., Yin, W.: A fast alternating direction method for TVL1-L2 signal reconstruction from partial Fourier data. IEEE J. Sel. Topics Signal Process. **4**, 288–297 (2010)
8. Gilboa, G.: A total variation spectral framework for scale and texture analysis. Siam J. Imaging Sci. **7**, 1937–1961 (2014)

Face Recognition for Mobile Self-authentication with Online Model Update

Seon Ho Oh(✉) and Geon-Woo Kim

Electronics and Telecommunications Research Institute,
Daejon, Republic of Korea
{seonho,kimgw}@etri.re.kr

Abstract. Face recognition system encounters complex change that varies over time, due to a limited control over the environment. So, the facial model of an individual tends to diverse from underlying distribution that collected during initial enrollment. However, new samples that are obtained each time people try to recognize or authenticate can be used to update and refine the models. In this paper, an efficient semi-supervised learning strategy is proposed to update the face recognition model. To maintain a high performance, we exploit a probability based update approach. Performance is assessed in terms of accuracy and equal error rate (EER). Experimental results illustrate that the proposed method effectively update the classifiers.

Keywords: Face recognition · Self-authentication · Semi-supervised learning · Adaptive biometric system

1 Introduction

Rapid development of mobile technology enables many complex tasks such as identification and authentication on the mobile device [1, 2]. Especially, biometric recognition using mobile devices is a convenient and import means for self-authentication [3]. As biometric traits can change over time due to aging and change of lifestyle, however, the performance of a biometric system may degrade substantially [4]. Additionally, it has to deal with uncontrolled environments, e.g., various postures and varying illumination condition due to device mobility. Moreover, lack of computing power on mobile environment makes the problem more difficult. In this context, we aim to develop an efficient semi-supervised learning strategy to update the face recognition models from new samples.

2 Methodology

Face recognition models are initially learned during enrollment using labeled training data, and then updated with unlabeled samples obtained during authentication. To minimize the probability of including imposter data into updated models, two criteria are considered. One is the confidence threshold that determines whether or not to

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_102

include the samples for updates, and the other is the interval that indicates how often the model will be updated. We use 0.7 for threshold and 100 samples for interval.

In order to evaluate the strategy, mobile face dataset for self-authentication is collected. The faces are captured from 10 identities with a number of distinct sessions. Feature is extracted by compact deep neural network. Linear classifier with stochastic gradient descent (SGD) training is used as our baseline recognition model. Modified Huber and log loss functions are tested, and L_2 norm is used as regularization term. To determine the confidence of unlabeled samples, class membership probability estimates [5] is used.

To learn initial model, randomly selected 20 faces per identity from the first sessions is used. And faces in the remaining sessions are randomly interleaved with maintaining the order and total 6,698 images are tested. As show in Table 1, total accuracy and EER are significantly improved in both loss functions when update the face recognition models for each interval.

Table 1. Accuracy and EER improvement on different loss functions.

Loss function	Accuracy	EER	Accuracy (update)	EER (update)
Modified Huber	0.9036 ± 0.0547	0.2643 ± 0.0628	**0.9787 ± 0.0107**	**0.1649 ± 0.0614**
Log	0.9564 ± 0.0164	0.1447 ± 0.0360	**0.9968 ± 0.0023**	**0.0613 ± 0.0395**

3 Conclusion

This paper introduces an efficient semi-supervised learning strategy to tackle the problem of online classifier update. The proposed strategy adopted class membership probability estimates as the confidence value of unlabeled samples and two criteria regulate the update. On a mobile face dataset for self-authentication, the proposed strategy significantly improves the performance both in terms of accuracy and EER.

Acknowledgments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0126-16-1007, Development of Universal Authentication Platform Technology with Context-Aware Multi-Factor Authentication and Digital Signature and No. B0717-16-0107, Development of Video Crowd Sourcing Technology for Citizen Participating-Social Safety Services).

References

1. Rischan, M., de Dharma, N.I.G., Deokjai, C.: Modeling and discovering human behavior from smartphone sensing life-log data for identification purpose. Human-centric Computing and Information Sciences **5**, 31 (2015)
2. Yusuf, A., Mohammad, M.H.K., Athanasios, B., Sotirios, K., Nhan, N., Ruhua, J.: Designing challenge questions for location-based authentication systems: a real-life study. Hum. Centric Comput. Inf. Sci. **5**, 17 (2015)

3. Ramon, B.-G., Norman, P., Rita, W., Raul, S.-R.: Time evolution of face recognition in accessible scenarios. Hum. Centric Comput. Inf. Sci. **5**, 24 (2015)
4. Roli, F., Didaci, L., Marcialis, G.L.: Adaptive biometric systems that can improve with use. In: Ratha, N.K., Govindaraju, V. (eds.) Advances in Biometrics: Sensors, Algorithms and Systems, pp. 447–471. Springer, London (2008)
5. Zadrozny, B., Elkan, C.: Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In: The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 694–699. ACM, New York (2002)

Prototype System Design for Large-Scale Person Re-identification

Seon Ho Oh(✉), Seung-Wan Han, Beom-Seok Choi, and Geon-Woo Kim

Electronics and Telecommunications Research Institute, Daejon, Republic of Korea
{seonho, hansw, bshoi, kimgw}@etri.re.kr

Abstract. Identifying a person across cameras in disjoint views at different time and location has important applications in visual surveillance. However, it is difficult to apply existing methods to the development of large-scale person identification systems in practice due to underlying limitations such as high model complexity and batch learning with the labeled training data. In this paper, we propose a prototype system design for large-scale person re-identification that consists of two phases. In order to provide scalability and response within an acceptable time, and handle unlabeled data, we employ an agglomerative hierarchical clustering with simple matching and compact deep neural network for feature extraction.

Keywords: Person re-identification · Prototype system · Large-scale · Visual surveillance

1 Introduction

Person re-identification (re-id) is a problem of finding a person across cameras in disjoint views at different times and locations. It has many applications such as video surveillance for security, public safety, human-computer interaction, robotics, content-based video or image retrieval [1], and etc. Despite the best efforts, re-id still remains an unsolved problem due to dramatic variations in visual appearance and ambient environment caused by different view point from different camera, significant body pose across time and space, illumination changes, background clutter, occlusions and so on.

The advances in mobile technology enabled a new paradigm for accomplishing large-scale sensing, known in literature as participatory sensing. The key idea of participatory sensing is allowing the ordinary citizen to use their mobile phones to collect and share the data from their surrounding environments. From this point of view, crowdsourced participatory sensing [2] combined with automated person re-id has great potential for public security and safety area. Meanwhile, it is interesting to note that unpaid crowdsourcing yields results of similar or higher quality compared to its paid counterpart [3].

Tremendous studies have been conducted on person re-id, however, existing approaches are unsuitable for large-scale re-id system in practice. Specifically, the re-id system for large-scale environments requires: low model complexity with reasonable

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_103

computational cost and memory usage for operation to provide a response within an acceptable time, and flexibility to adapt any new and massive unlabeled data from a number of cameras [4]. To that end, we propose a prototype system design for a large-scale re-id that consists of two phases.

2 System Design

Person re-id process typically takes the following steps: (1) detect or track individual person from images of different locations over time captured by distinct cameras, (2) extract visual features that are capable to describe and discriminate individuals, and (3) matching specified probe images or tracks against a gallery database of persons by measuring the similarity between features.

In this work, we follow typical re-id pipeline. Figure 1 illustrates our prototype system design for person re-id. Our system consists of two phases: learning and searching.

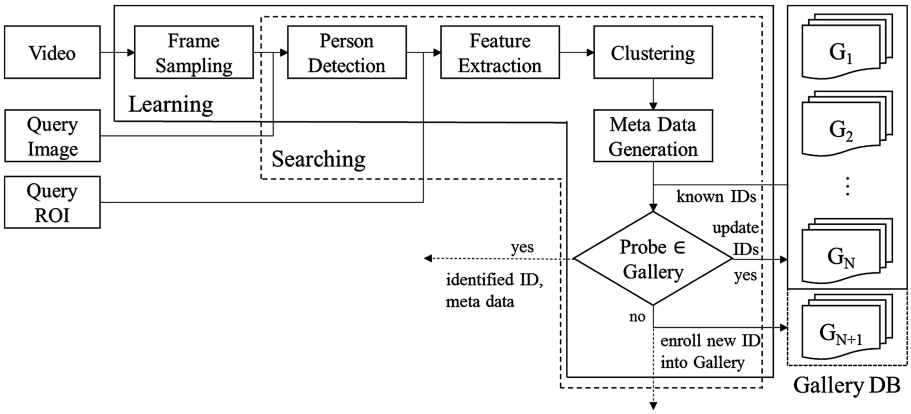


Fig. 1. Diagram of a large-scale person re-identification prototype system.

Learning phase is comprised of six steps. First, frame sampling is conducted since it is difficult to process entire massive image frames captured from a number of cameras. Then person detection is performed to obtain the bounding box of individual persons. We use one of the current state of the art detector YOLO [5] to detect multiple person in real-time since it is currently the fastest approach. For feature extraction, we develop compact deep feature learning framework for re-id, which is simplified variants of GoogleNet [6]. Our deep neural network for feature extraction for re-id is tens of times smaller than exiting AlexNet [7] or VGGNet [8] for re-id. The network extracts a 256-dimensional embedding on a unit hypersphere and represent similarity between images effectively. In order to handle massive unlabeled data that are obtained from a number of cameras on the fly, we employ an agglomerative hierarchical clustering approach. Clustering is made up of two sub parts: local and global. The former is local clustering for features from the same camera or an adjacent camera, and the latter is global clustering for the results of local clustering. Once the clusters are created,

metadata generation, including visual feature and additional information such as clothing, height, and gender is performed. Finally, meta data is matched with gallery database. If the metadata is similar to the existing one, the database is updated, and otherwise, enroll new meta data into gallery database.

Searching is consisting of five or four steps. If query image is given, person detection is performed to extract bounding box of individual. If query ROI is given, person detection is omitted. The steps from feature extraction to meta data generation are the same as in the learning phase. Clustering can be omitted in searching phase. If the given query is similar to the existing one from the gallery, identified id and meta data are returned, otherwise the system fails to identify a query.

3 Conclusion

In this paper, we propose a prototype system design for large-scale re-id that consists of two phases: learning and searching. For preprocessing, state of the art person detector is used and deep neural network for feature extraction is designed and used. In addition, local and global agglomerative hierarchical clustering are employed to provide scalability and response within an acceptable time, and handle unlabeled data.

Acknowledgments. This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0126-16-1007, Development of Universal Authentication Platform Technology with Context-Aware Multi-Factor Authentication and Digital Signature and No. B0717-16-0107, Development of Video Crowd Sourcing Technology for Citizen Participating-Social Safety Services).

References

1. Rao, L.K., Rao, D.L.: Local quantized extrema patterns for content-based natural and texture image retrieval. Hum. Centric Comput. Inf. Sci. **5**, 26 (2015)
2. Ogie, R.I.: Adopting incentive mechanisms for large-scale participation in mobile crowdsensing: from literature review to a conceptual framework. Hum. Centric Comput. Inf. Sci. **6**, 1 (2016)
3. Borromeo, R.M., Toyama, M.: An investigation of unpaid crowdsourcing. Hum. Centric Comput. Inf. Sci. **6**, 1 (2016)
4. Wang, H., Gong, S., Xiang, T.: Highly efficient regression for scalable person re-identification. In: British Machine Vision Conference (2016)
5. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection, arXiv preprint, arXiv:1506.02640 (2015)
6. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–523 (2015)
7. Schroff, F., Kalenichenko, D., and Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
8. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision, vol. 1, no. 3 (2015)

Topic Modeling for Learner Question and Answer Analytics

Kyungrog Kim[1(✉)], Hye Jin Song[2], and Nammee Moon[2]

[1] Department of Electronic Engineering, Hoseo University, Asan, Korea
it4all@hoseo.edu
[2] Department of Computer Software, Hoseo University, Asan, Korea
songl20408@naver.com, mnm@hoseo.edu

Abstract. There is increasing interest in text analysis based on unstructured data such as articles and comments, questions and answers. This is because they can be used to identify, evaluate, predict, and recommend features from unstructured text data, which is the opinion of people. The same holds true for TEL, where the MOOC service has evolved to automate debating, questioning and answering services based on the teaching-learning support system in order to generate question topics and to automatically classify the topics relevant to new questions based on question and answer data accumulated in the system. To that end, the present study proposes an LDA-based topic modeling. The proposed method enables the generation of a dictionary of question topics and the automatic classification of topics relevant to new questions.

Keywords: Data mining · Text mining · Similarity · Topic modeling · Question and answer

1 Introduction

As is often the case, people solve problems they face in everyday life by asking and answering questions. With the advancement of ICT environment, they turn to online communities to find the answers to their questions. Likewise, the field of TEL (Technology Enhanced Learning) witnesses various questions asked and answered regarding educational environment and system use as well as instructional content using the teaching-learning support system [1, 2]. Such activities are conducted on a continuous basis. Questions generated in the process are the results of collective intelligence. In order to make efficient use of such questions and answers, researchers have explored the methods of using machine learning to automatically classify and answer the questions on the grounds that the automatic classification of questions and the implementation of automatic answering system will accelerate the problem solving [3]. In the same vein, the present study proposes a topic modeling method for automatic classification based on the question data generated in the teaching-learning support system.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_104

2 Literature Review

2.1 Topic Modeling

Topic modeling is an algorithm for finding topics in a large unstructured document. This is a way of inferring a subject in a way that clusters words with similar meanings using context-related clues [4]. That is, topic modeling is an algorithm designed to analyze unstructured data. Topic modeling is largely sub-classified into vector-based techniques and probabilistic techniques. Vector based techniques include Latent Semantic Analysis (LSA). Probabilistic techniques include Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA). In particular, LDA is an unsupervised generative topic model and an algorithm widely used for topic classification and computation of similarities between documents [5–7].

Documents are sets of topics. Topics are sets of words. That is, words converge on topics, which in turn form documents. Each word can represent a topic. A set of topics makes a document. Here, as topics are latent, the LDA algorithm is used to uncover the topics [6, 8]. The algorithm determines the frequency of each word constituting a topic in a document. From the perspective of questions and answers, the questions-topics-words relationship is shown on the left in Fig. 1 below. Moreover, the relationship between questions is represented on the right side in Fig. 1. Using these relationships, the present study is to model the topics relevant to questions.

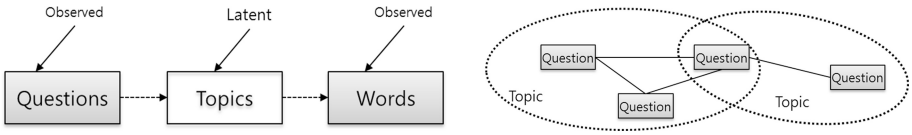


Fig. 1. (Left) Questions-topics-words relationship, (Right) Topic-topic relationship

2.2 Similarity Measure

Similarity measurement is a method of calculating how similar each document is by analyzing frequency of words contained in each document. This is widely used in information retrieval and text mining [9]. The similarity measurement method is classified into correlation-based method and vector-based method. Correlation-based techniques include Pearson Correlation and Spearman Rank Correlation. Based on the vector, there are cosine similarity, Euclidean distance, and Inner product. In particular, cosine similarity is the most widely used method on a vector basis because it is simple and computationally fast. The calculation of cosine similarity is shown in the following Eq. (1) [10]. This study is used to calculate the similarity between constituent words in a topic.

$$\text{Similarity} = \cos(\mathrm{A}, \mathrm{B}) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{1}$$

3 Question Data Topic Modeling Based on LDA

Up until now, a large number of questioning and answering services in the teaching-learning support system have had persons in charge read questions, classify topics and present answers. In order to automate the procedure, questions need to be automatically classified. To that end, the present study draws on the Latent Dirichlet Allocation (LDA) algorithm suggested by Blei et al. (2003) [6]. LDA is straightforward, useful for reducing the dimensions of data and good at extracting semantically coherent topics. Thus, it is widely used for text minings [6]. The present study applies the LDA to the topic classification modeling based on the question data in teaching and learning. Also, we use the cosine similarity method to measure the similarity of constituent words by topic.

First, there are a number of questions in the teaching-learning support system. Each question consists of a set of words. The set of words involves particular topics. Questions may share each topic. That is, question items can be classified into similar questions by topic. Figure 2 shows a procedure of generating a dictionary of questions by analyzing the question topics based on the question data accumulated in the system.

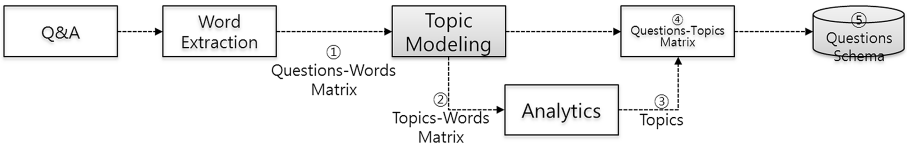


Fig. 2. Topic-based question recognition and classification procedure

- Words are extracted based on existing questions to construct a questions-words matrix.
- N initial topics are set and repeatedly output to select M component words. Based on the set topics and selected words, a topics-words matrix is built. Here, the homogeneity of words and the heterogeneity of topics should be considered.
- Based on the similarities of component words of each topic, the final topic is chosen. Here, the similarities are computed using machine learning.
- A topics-questions matrix is generated.
- Based on the matrix, a schema for the dictionary of questions can be generated.

Based on the schema, topics of new question inputs are surmised and classified.

4 Conclusion

A range of questions and answers are generated in teaching-learning activities in TEL. The generated data automatically accumulates in the system. To repurpose and use such accumulated data, an automatic classification model is required. The present study proposes a topic-driven method for the automatic classification of questions. The proposed model is noteworthy in that it is applicable without the reference data for

initial classification and that it becomes sophisticated over time. The proposed method is intended to use the question data accumulated in the system so as to analyze question topics, generate an initial dictionary of questions, and automatically and inferentially classify topics relevant to new questions. Further studies on the relevance of topics and the empirical verification of data are needed.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government MSIP) (No. 2016015499).

References

1. Cerulo, L., Distante, D.: Topic-driven semi-automatic reorganization of online discussion forums: a case study in an e-learning context. In: 2013 IEEE Global Engineering Education Conference (EDUCON). IEEE (2013)
2. Ezen-Can, A., et al.: Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. ACM (2015)
3. Wang, G., et al.: Wisdom in the social crowd: an analysis of quora. In: Proceedings of the 22nd International Conference on World Wide Web. ACM (2013)
4. Hu, Y., et al.: Interactive topic modeling. Mach. Learn. **95**(3), 423–469 (2014)
5. Yeon, L.S., Lee, K.M.: A reply graph-based social mining method with topic modeling. J. Korean Inst. Intell. Syst. **24**(6), 640–645 (2014)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
7. Park, J.D.: A study on mapping users' topic interest for question routing for community-based Q&A service. J. Korean Soc. Inf. Manag. **32**(3), 397–412 (2015)
8. Anoop, V.S., Asharaf, S., Deepak, P.: Learning concept hierarchies through probabilistic topic modeling. arXiv preprint, arXiv:1611.09573 (2016)
9. Lin, Y.-S., Jiang, J.-Y., Lee, S.-J.: A similarity measure for text classification and clustering. IEEE Trans. Knowl. Data Eng. **26**(7), 1575–1590 (2014)
10. Sidorov, G., et al.: Soft similarity and soft cosine measure: similarity of features in vector space model. Computacióny Sistemas **18**(3), 491–504 (2014)

A Model of Energy-Awareness Predictor to Improve the Energy Efficiency

Svetlana Kim and Yong-Ik Yoon(✉)

Department of IT Engineering, Sookmyung Women's University,
Seoul, South Korea
{xatyna, yiyoon}@sm.ac.kr

Abstract. The data centers contribute to high operational costs and electrical energy will be consumed in enormous amounts. One of the most complex challenges of energy consumption is power management. Many different methods have been applied in order to reduce energy consumption. In this paper, we propose the architecture framework focuses on analyzing the EAP (Energy-Awareness Predictor) to improve the energy efficiency. Through analysis and various integrated sensor devices, the EAP architecture framework can understanding of the consumption patterns and can better controlling of the major energy consuming. Based on inputs independent variables (value of external and internal environmental) is prediction and implement refrigeration and process control, optimization and energy management.

Keywords: Context-aware · Tensor factorization · Energy-Awareness Predictor · Energy efficiency

1 Introduction

Data centers are an increasingly important part of most business operations in the 21st century. Although IT energy consumption covers a considerable range, despite the work of Europe and other administrative organs, there is little information on the total size, power consumption, or efficiency of the data center market as a whole [1–3]. Without this information, it is difficult to protect the data center industry, predict growth, and set effective metrics and objectives. In order to deal with these problems, it is necessary to understand the scale of the problem appropriately and to provide improvements. It is essential that the measurement and the initial set of metrics are agreed and data collection be started on a large scale. However, these metrics often are not applied consistently at a global level.

The energy uses of the data center and its impact on the environment have recently become an important issue for operators and policy makers [4]. The data center is a complex environment designed to IT equipment. Most of the power used in the facility will be converted to heat, the capacity of the cooling system will be significantly required, and additional load will be placed on the conventional recirculation air data center [5]. In addition, the data center has various auxiliaries' support systems such as generator pre-heaters, lighting, fire suppression system, and occupied areas of people that require electrical power. In order to understand, measure or model the data center's

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_105

overall energy efficiency it is necessary to understand the relationship between an electrical load of the housed equipment and the utility power that the facility draws to supply power to the housed equipment for cooling. The power used in the IT equipment in the data center is supplied with a series of power conditioning devices and power distribution devices, each of which is inefficient and thus a certain amount of power is lost. Since this equipment is usually housed in a data center, these losses converted into heat will be added to the thermal output of IT equipment. The industry should improve IT measurement capabilities to ultimately enable taking the measurement directly at the IT load (such a servers, storage, networking, etc.). The seemingly simple PUE (Power Usage Effectiveness) concept requires additional details to guarantee consistent application for the data center. The PUE metric is the de facto standard for measuring data center energy efficiency. PUE compares the amount of power used to power IT equipment (server, storage, and network) with the total power sent to the data center. Also, consistent calculation of PUE promotes performance transmission of data center owners and operators [6–8].

In this paper, we propose the architecture framework focuses on analyzing the EAP (Energy-Awareness Predictor) to improve the energy efficiency. Using analysis and various integrated sensor devices, the EAP architecture framework can better understand the consumption patterns and better control the main energy consumption. On the basis of inputs, independent variables (the value of the external environment and the internal environment) are predicted, and refrigeration and process control, optimization and energy management are implemented. The EAP framework is centered on the concepts and is oriented for the energy efficiency, in spite of being oriented for technologies.

2 Energy-Awareness-Predictor Framework

The Energy-Awareness-Predictor (EAP) Framework is a profiling-based energy saving research that can efficiently utilize energy by converging and analyzing data onto various energy metering devices. Research on energy efficiency management through the management of sensors is underway to form sensor networks in servers, storage, networks, internal, external and peripherals for the purpose of reducing energy in the data center. In such an environment, the necessity to study energy prediction technology applying the machine learning method is very high. The research on efficient energy management method applying machine learning method based on information on the data center related situation is required. It is a strategy method that can apply energy efficient PUE of the data center by applying machine learning method.

The importance of recent artificial intelligence and machine learning has been analysis and prediction is advanced by introducing machine learning methods of processes such as analysis, reasoning, and prediction in the process of big data. Therefore, the existing statistical analysis method PUE also needs the changes to extract indicators and predict the future depending on whether situation by introducing machine learning technology. Traditional analysis is being analyzed by defined algorithms without self-modification of data that is continuously collected. In other words, the self-adjust function, which is the autonomous definition of the data for analysis, is excluded. When using the machine learning technique, the algorithm that is used in the

analysis is proposed to change dynamically in consideration of changes in the surrounding environment and user requirements. In other words, it has a self-adjust function that can change the data before the analysis when new data according to changes in environment and requirements are generated. In addition, it is possible to support the self-optimizing algorithm that optimizes itself considering the changes of the environment and the requirements of the user.

The EAP framework is designed to extend the life of data centers using 3 Tensor technologies. EAP is designed with energy-awareness to improve energy efficiency and can reduce and manage energy use efficiently. With energy-awareness, the EAP is improving energy efficiency and can reduce and manage energy use efficiently. First, it extracts the total energy usage per server and improves energy efficiency by capturing changes in energy usage over time (consumption patterns). The data center energy usage per server can be calculated in terms of the number of servers and can be roughly calculated not only by the energy consumed but also networking, storage, power equipment, cooling power (see the Fig. 1).

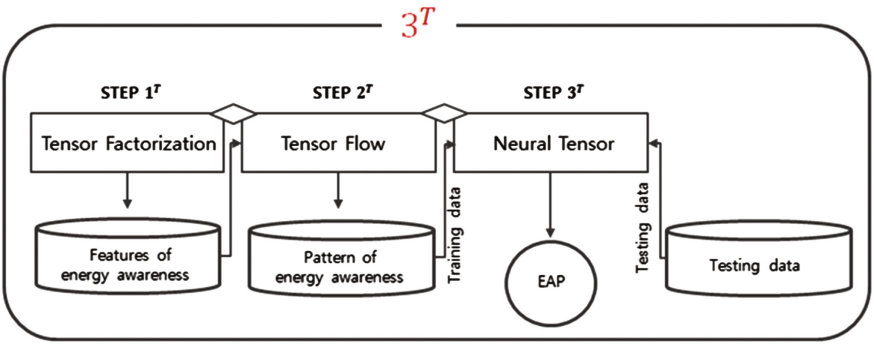


Fig. 1. Architecture of EAP framework

As shown in the Fig. 1, the EAP framework based on cognitive profiling consists basically of three stages: (1) Tensor Factorization to extract energy usage; (2) Tensor Flow for generated features of energy usage; (3) Neural Tensor for Energy Optimization stage. The EAP framework can intelligently and automatically enable configuration management of the data center.

3 Proposed Method

3.1 Tensor Factorization

Tensor Factorization extracts energy usage characteristics by factorizing several servers (chillers, IT load, pump speed, wet bulb) to extract energy usage patterns. Tensor factorization is a generic model for multidimensional systems, such as fast prediction computations and simple optimization techniques. The multidimensional rating

provides more information that a single rating system. There are different types of tensor decomposition models, such as the PARAFAC, also called CANDECOMP (Canonical Decomposition), Tucker, and HOSVD (Higher-Order Singular Value Decomposition). In our approach, we follow the HOSVD formulation. The HOSVD is one of the most powerful tensor decomposition methods, which based on a successive application of the matrix SVD decomposition to the flattened matrices of a given tensor [9]. HOSVD can be used to build orthogonal spaces which can be then used for reduction analysis in a way similar to the subspace projection method server (formulation shown in Fig. 2).

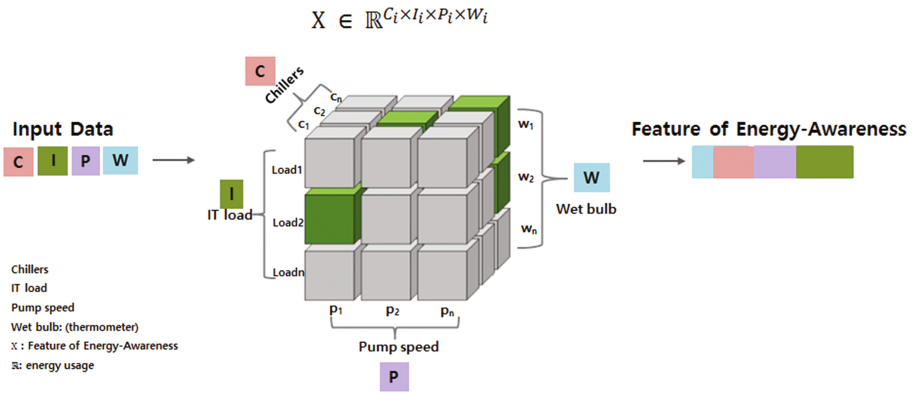


Fig. 2. Tensor factorization for extracting energy usage

Figure 2 is an illustration of tensor factorization into four matrices. Composed values of the measures pre-computed by applying dimension attributes as analysis perspectives are well as various aggregate operations. A tensor factorization performs aggregate operations to extract all possible combinations of the dimension attributes, given factors extracted for the servers, dimensional and context's features to provide rapid responses to any analysis queries.

3.2 Tensor Flow

The Tensor flow [10] is a typical deep learning framework that can represent mathematical calculations and data flow using directional graphs using nodes and edges. Nodes can perform operations such as mathematical calculations, data I/O and reading/storing of data. The edge represents the input and output relationship of data between nodes, and all data can be represented internally by multidimensional array tensor. Tensor flows are designed to be used not only for deep learning but also for other machine learning algorithms, allowing multiple machines learning algorithms to be executed in various languages. The tensor flow has the following characteristics:

- Rich expressiveness through data flow graph;
- Operates in CPU/GPU mode without code modification;
- Define only the calculation structure and the target function, it will automatically process the differential calculation;
- Supports Python/C++ and supports various languages through SWIG (Simplified Wrapper and Interface Generator);

Tensor flow can reduce unnecessary operation of the data center, minimize power waste, and maximize energy saving efficiency, by understanding the same pattern automatically generated by tensor factorization. The modeling of energy usage pattern is generally to determine the relationship between energy consumption data and specific variables that indicate the main cause of energy consumption. In addition, time information is added to extract changes in energy usage over time. It is characterized by the ability to identify energy usage over time for energy efficiency. Based on this, we can suggest ways increase energy efficiency over time. Figure 3 below shows an example. Because energy consumption is different for each hour of the same operation, energy consumption per hour can be extracted. Every time it repeatedly learns data by time zone, that can gradually change the weight and obtain training data for each time period.

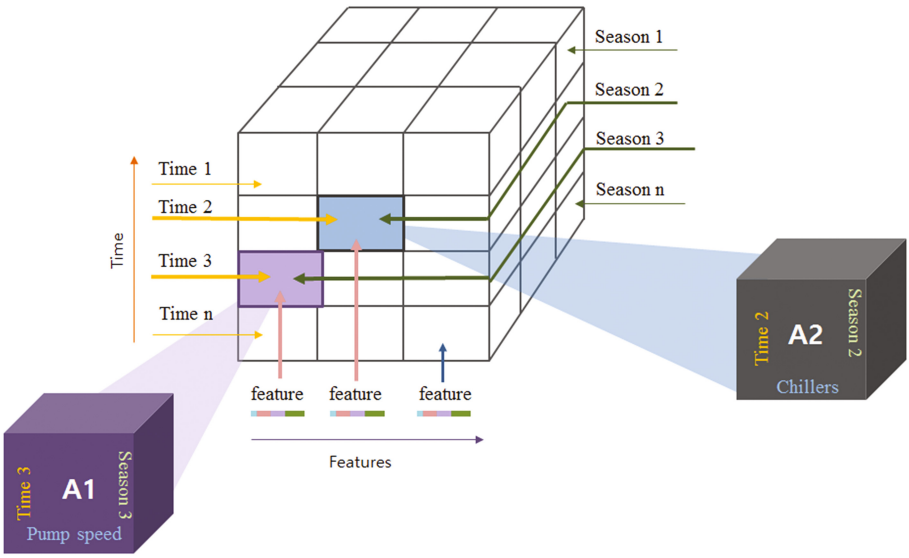


Fig. 3. Tensor flow for energy pattern extraction

3.3 Neuron Tensor

The neural tensor [11] utilizes an integrated similarity function that reflects energy usage at various times. Energy consumption can be predicted through key perceptions that primarily affect energy consumption. A typical structure of a neural tensor consists of neurons with multiple layers such as sublayers that receive inputs and upper layers

that produce result values. Neural tensor hidden layer can have multiple intermediate matching score layers. In such a neural tensor, the neurons of the layer communicate with neurons of the previous layer to receive information and communicate with neurons of the next layer to export results. In this paper, we define the computational structure of the neural sensor prediction model, connect it to the target function, and input only the data. The neural tensor uses a patch filter to compare while scanning the entire layer to find out what kind of features the input data has, and generates a feature map. As shown in Fig. 4 below, the feature map is cut to a certain size like a tile, and the energy consumption is predicted with the maximum value in each piece as a representative value.

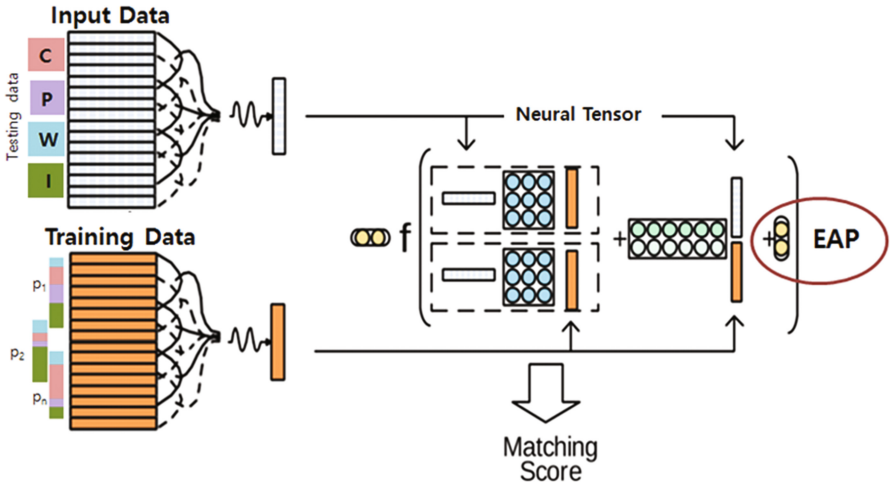


Fig. 4. Neural tensor for energy optimization (prediction)

4 Conclusion

This paper proposed the concept of EAP (Energy-Awareness Predictor) for improve the energy efficiency. We explored the use of machine learning methods to increase the energy efficiency of the data center, PUE. In recent years, the use of machine learning has increased and the PUE has been lowered by applying deep-learning techniques to the data centers operated by Google itself. It is necessary to present an optimum cognitive learning model by establishing the applied characteristics based on the collected data and continuously executing the instruction process accompanied by this. Through analysis and various integrated sensor devices the EAP framework can understanding of the consumption patterns and can better controlling of the major energy consuming. Based on the input variables (the values of the external and internal environment) the independent variables were predicted and showed refrigeration and process control, optimization and energy management. To the future work, we will be adapting implement to the proposed architecture and to develop EAP framework.

Acknowledgement. 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-R2718-16-0004-0001002) supervised by the IITP (National IT Industry Promotion Agency).

References

1. Belady, C., Pflueger, J: Green Grid: Enabling the Energy-Efficient Data Center. http://www.dell.com/downloads/global/power/ps1q08-20080199-GreenGrid.pdf
2. The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE. http://www.thegreengrid.org/gg_content/TGG_Data_Center_Power_Efficiency_Metrics_PUE_and_DCiE.pdf
3. Gartner Press Release (2015). http://www.gartner.com/newsroom/id/3055225
4. Technavio Global Data Center Market 2014–2018, November 2014. http://www.technavio.com/report/global-data-center-market-2014-2018
5. ISO/IEC JTC 1/SC 39 (Sustainability for and by Information Technology). http://www.iso.org/iso/standards_development/technical_committees/other_bodies/iso_technical_committee.htm?commid=654019
6. Jeong, S., Kim, Y.-W.: A holistic investigation method for data center resource efficiency. In: ICTC 2014, pp. 548–549 (2014)
7. Blackburn, M., Azevedo, D., Ortiz, Z., Tipley, R., Van Den Berghe, S.: The Green Grid Data Center Compute Efficiency Metric: DCcE (2010)
8. Beitelmal, P.: Model-Based Approach for Optimizing a Data Center Centralized Cooling System (2006). http://www.hpl.hp.com/techreports/2006/HPL-2006-67.pdf
9. Lathauwer, L., Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
10. TensorFlow™. https://www.tensorflow.org
11. Qiu, X., Huang, X.: Convolution neural tensor network architecture for community-based question answering. In: Proceedings of IJCAI 2015, pp. 1305–1311 (2015)

A Novel Structure Tensor Using Nonlocal Total Variation Operator

Yin Yu[1(✉)], Kai Ma[2], Yuhui Zheng[2], and Jin Wang[3]

[1] The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China
yuyin2001@126.com
[2] CICAEET, Nanjing University of Information Science and Technology, Nanjing 210044, China
mkl9940208@163.com, zheng_yuhui@nuist.edu.cn
[3] College of Information Engineering, Yangzhou University, Yangzhou 215127, China
jinwang@yzu.edu.cn

Abstract. The structure tensor known as a second moment matrix which integrates the local data information of the image. It has been a well-established tool in the image processing field. To date a variety of nonlinear structure tensors have been emerged. Among them, the non-local structure tensors (NLSTs) is the focus of researching for the reason that it explores the spatial interactions in images. However, the performance of the existing NLST in image analysis is limited. In this paper, we propose a new structure tensor calculation method by using the nonlocal means filter to smooth the matrix-valued data. The resulting nonlocal structure tensor is effective in the orientation estimation and structural analysis of the image. Meanwhile, the nonlocal TV model based this structure tensor has been successfully applied in noise removal. Experimental results show that our model has better performance in preserving the structures, details and textures.

Keywords: Structure tensor · Image analysis · Nonlocal mean filter · Nonlocal TV model

1 Introduction

The classical structure tensor [1–6] applies Gaussian convolution which is a linear technique for averaging information within a neighbourhood. The benefit of Gaussian smoothing is simple and robust, but it lacks for debluring and locating structures. This is result from the fact that the local neighbourhood for the integration is fixed in both specific size and Specific position. Consequently, it fails to adapt to the data and tends to blur the discontinuities information embedded in tensors.

To address the limitations of Gaussian convolution, different adaptive local structure tensors which using nonlinear diffusion techniques have been proposed in a variety of backgrounds. Nagel and Gehrke [7] utilized an adaptively local Gaussian kernel to smooth the tensor for optic flow calculation. Based on this work, Middendorf [8].

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_106

Köthe [2] proposed a structure tensor by using an hour-glass shaped filtering. Zhang [3] put forward a bilateral filtering based nonlinear structure tensor. The adaptive local structure tensor proposed by Van den Boomgaard and Van de Weijer can extract image orientation adopt robust statistics. Brox [9] respectively employed nonlinear diffusion equations to obtain structure tensor which applied to image feature detection, opticflow estimation and image denoising. Though the existing adaptive local structure tensors have shown their effectiveness in many image processing application fields, they often fail to extracting the local information of image. It should be noticed that these construction methods of tensors are local which only consider the neighboring pixels in the calculations. In fact, pixels sharing similar characteristic often replicate in a non-local way. The local construction methods ignore the connection between the centered pixel and the others who outside the scope of consideration. As a result, the performances in image analysis of the local structure tensors are restricted.

Naturally, the efforts of the researchers have concentrate on the non-local structure tensors (NLSTs) which take the above mentioned characteristic into consideration. Doré [10] constructed an NLST by utilizing the non-local means filter (NLMF) [11] to smooth the tensor field. S. Lefkimmiatis and S. Osher [12] join the NLST in the designing of regularization operator. Chierchia [13] and Zheng [14] work out the NLST by applying the non-local total variation regularizations. Euclidean Distance is used to measure the similarity between the matrix-valued data. Unfortunately, the above NLSTs are isotropic and these TV-based construction method do not preserve the fine details [17–20].

In this paper, we utilized nonlocal operators to construct an anisotropic non-local structure tensor (ANLST). The nonlocal operators take the similarity of image patch which is measured by a weight function of the patches into consideration. Thereby, the useful spatial information in image can be exploited. We then extend the TV-based model to a nonlocal setting based on the proposed ANLST applications to edge detection.

2 Preliminaries

2.1 Structure Tensor

Let I denote the reference image and $\nabla I_{(v,u)} = \left(\left(I_{v,u}\right)_x, \left(I_{v,u}\right)_y \right)^{\mathrm{T}}$ denote the gradient at (i,j), T is the transpose. The structure tensor S of I can be defined as:

$$S = \begin{bmatrix} \left(I_{i,j}\right)_x^2 & \left(I_{v,u}\right)_x \left(I_{i,j}\right)_y \\ \left(I_{i,j}\right)_x \left(I_{i,j}\right)_y & \left(I_{i,j}\right)_y^2 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \tag{1}$$

In order to suppress noise, filtering technique is used to smooth matrix data. The uniform expression:

$$g(S) = g\begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \tag{2}$$

When $g(\bullet)$ is linear filtering, $g(S)$ is linear structure tensor. When $g(\bullet)$ is nonlinear filtering, $g(S)$ is nonlinear structure tensor.

The eigenvalues of structure tensor can be computed by:

$$\lambda_{1,2} = \frac{1}{2}\left[u_{11} + u_{22} \pm \sqrt{(u_{11} - u_{22})^2 + 4u_{12}^2}\right] \tag{3}$$

And the corresponding orthogonal eigenvectors can be computed by:

$$\gamma_1 = \left[\frac{2u_{12}}{u_{22} - u_{11} + \sqrt{(u_{11} - u_{12})^2 + 4u_{12}^2}}\right], \gamma_1 \perp \gamma_2 \tag{4}$$

2.2 Nonlocal Operator

The idea of the nonlocal means filter introduced by Buades et al. [15] denoises a current point by replacing the noisy pixel with a weighted average of nonlocal neighboring pixels with similar structures. It is similar to the Yaroslavsky filter [16] and patch-based methods which was used for texture synthesis at first. They all consider the similarities of the patch around the pixel instead of the only pixel itself. The nonlocal means method can be written as follow:

$$NL_p(x) := \frac{1}{\int_\Omega w_p(x,y)dy} \int_\Omega w_p(x,y)p(y)dy \tag{5}$$

$$w_p(x,y) = \exp(-\left\|G_\sigma * |p(x+\cdot) - p(y+\cdot)|^2\right\|/(a * a)) \tag{6}$$

Where $p(x+\cdot)$ denotes an image patch centered at current point x with specific size in the reference image p, the weight function $w_p(x,y)$ measures the similarity between points (x,y). $\int_\Omega w_p(x,y)dy$ is the normalizing factor, G_σ is the Gaussian kernel with standard deviation σ and h is a filtering parameter which set according to the standard deviation of noise.

Let $\Omega \subset R^2, x, y \in \Omega, w(x,y)$ is a weight function. The nonlocal gradient is:

$$(\nabla_w U)(x,y) = (U(y) - U(x))\sqrt{w(x,y)} \tag{7}$$

The nonlocal divergence can be written as:

$$(div_w \overrightarrow{v})(x) = \int_\Omega (v(x,y) - v(y,x))\sqrt{w(x,y)}dy \tag{8}$$

The nonlocal type of TV regularization function based on the nonlocal operator is as follows:

$$J_{NLTV}(U) = \int \|\nabla_w U\| = \int_\Omega \sqrt{\int_\Omega (U(x) - U(y))^2 w(x,y) dy dx} \tag{9}$$

3 Proposed Structure Tensor

The linear structure tensors suppress noise effectively while it blur the structural information and inaccurate in edge and corner location. In Eq. (2), when $g(\bullet)$ is linear filtering, it can be written as linear PDE of matrix data:

$$\frac{\partial S_{i,j}}{\partial t} = div(\nabla S_{i,j}) \ \ (i,j = 1,2) \tag{10}$$

In Eq. (3), The each channel of the structure tensor was smoothed by the linear diffusion filter. As a result, the implicit structural information in the tensor data tend to blurred such as edge and corner.

To solve the problem, we obtain a new nonlocal structure tensor as follow:

$$\frac{\partial S_{i,j}}{\partial t} = div\left(\frac{1}{\sqrt{\varepsilon^2 + \sum_{k,l=1}^{2} |\nabla_w U(x,y) S_{kl}|^2}} \nabla_w U(x,y) S_{i,j}\right) \tag{11}$$

$$w_u(x,y) = \exp\left(-\left\|G_\sigma * |U(x+\cdot) - U(y+\cdot)|^2\right\| / (a * a)\right) \tag{12}$$

The Eq. (11) smooths the structure tensor with nonlinear diffusion filter. The diffusion coefficient decided by the channels information. $(\nabla_w U)(x,y)$ denotes the nonlocal gradient operators measured by the weight function $w_u(x,y)$.

4 Experimental Results

In the experiments, the performances of our proposed nonlocal structure tensor is evaluated on the standard test images, considering two popular applications of structure tensor: corner detection and orientation estimation.

4.1 Corner Detection

Figure 1 illustrates a corner detecting experiment on a noisy synthesis image shown in Fig. 1(b). The noisy image is generated by adding the white Gaussian noise with standard variance $\sigma = 10$ into the original image depicted in Fig. 1(a). The ground-truth

corners are manually labeled and displayed in Fig. 1(a). Figure 1(c)–(e) respectively show the corner detection results of the three structure tensors. In this experiment, corner is detected at local maxima of the smallest eigenvalue of structure tensor. It can be found in Fig. 1(c) that the TV-ST loses some corners in the low contrast or seriously degraded regions. The NLST and our proposed structure tensor, by contrast, detect all the corners from the noisy image. Table 1 shows the quantitative comparision between the detection results. Obviously, we can see from the Mean Localization Error values in Table 1 that the herein proposed structure tensor performs the best in term of corner localization accuracy.

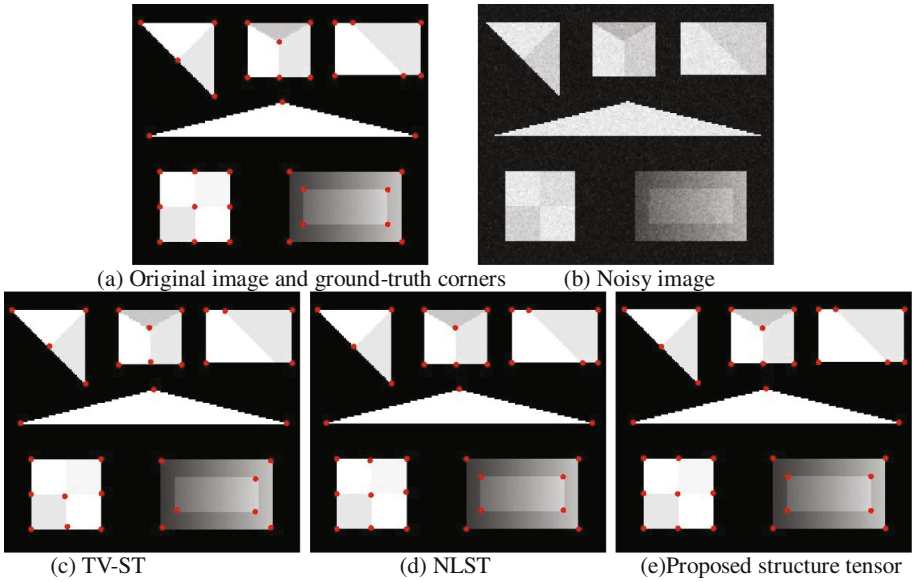


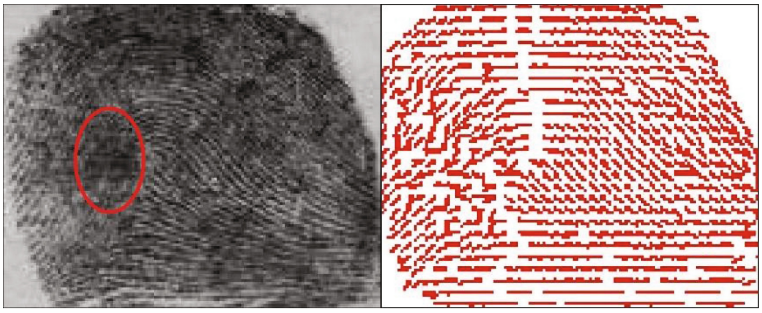
Fig. 1. Corner detection results by using the three nonlinear structure tensors on a noisy image

Table 1. Comparisons between the corner detection results

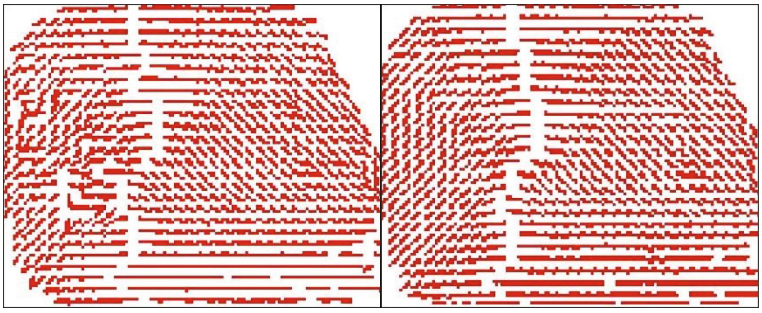
Algorithm	Detected corners	Lost corners	Mean localization error
LAST	33	3	1.0882
NLST	36	0	0.5278
ADNLST	36	0	0.3056

4.2 Orientation Estimation

Figure 2 shows the orientation estimation results of the three structure tensors on a fingerprint image shown in Fig. 2(a). The orientation is estimated by using the eigenvector corresponding to the smallest eigenvalue of structure tensor. Obviously, compared with the other two nonlinear structure tensors, our proposed method can obtain smooth and precise orientation estimation from the low quality fingerprint image.



(a) Fingerprint image (b)TV-ST



(C)NLST (D) Proposed structure tensor

Fig. 2. Orientation estimation results using the three nonlinear structure tensors on a real fingerprint image

5 Conclusions

As a useful tool for image analysis, the structure tensor has been successfully applied in various fields of computer vision and image processing, including feature detection, optic flow computation and orientation estimation. To further improve the performance of structure tensor, we introduce the nonlocal TV operator into structure tensor computation, giving rise to a novel nonlocal structure tensor. Experimental results on the test images show that our proposed structure is superior to the TV-ST and NLST in corner detection and orientation estimation.

References

1. Förstner, W., Gülch, E.: A fast operator for detection and precise location of distinct points, corners and circular features. In: ISPRS Intercommission Workshop, Interlaken, pp. 281–305 (1987)

2. Köthe, U.: Edge and junction detection with an improved structure tensor. In: Krell Pattern Recognition Proceedings of Dagm Symposium. LNCS, vol. 2781, pp. 25–32. Springer (2003)
3. Zhang, L., Zhang, L., Zhang, D.: A multi-scale bilateral structure tensor based corner detector. In: Computer Vision - ACCV 2009, Asian Conference on Computer Vision, Xi'an, China, 23–27 September 2009, Revised Selected Papers. DBLP, pp. 618–627 (2009)
4. Wu, X., Liu, S., Wu, M., et al.: Nonlocal denoising using anisotropic structure tensor for 3D MRI. Med. Phys. **40**, 101904 (2013)
5. Estellers, V., Soatto, S., Bresson, X.: Adaptive regularization with the structure tensor. IEEE Trans. Image Process. **24**, 1777–1790 (2015)
6. Brox, T., Weickert, J., Burgeth, B., et al.: Nonlinear structure tensors. Image Vis. Comput. **24**, 41–55 (2006)
7. Nagel, H.H., Gehrke, A.: Spatiotemporally adaptive estimation and segmentation of OF-fields. In: Computer Vision—ECCV 1998, pp. 1145–1149. Springer, Heidelberg (1998)
8. Middendorf, M., Nagel, H.H.: Estimation and interpretation of discontinuities in optical flow fields. In: 2001 Proceedings, Eighth IEEE International Conference on Computer Vision (ICCV), IEEE Xplore, pp. 178–183 (2001)
9. Brox, T., Weickert, J., Burgeth, B., et al.: Nonlinear structure tensors. Image Vis. Comput. **24**, 41–55 (2006)
10. Chierchia, G., Pustelnik, N., Pesquetpopescu, B., et al.: A non-local structure tensor based approach for multicomponent image recovery problems. IEEE Trans. Image Process. **23**, 5531–5544 (2014)
11. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 60–65. IEEE Computer Society (2005)
12. Lefkimmiatis, S., Osher, S.: Non-local structure tensor functionals for image regularization. IEEE Trans. Comput. Imaging **1**, 1 (2015)
13. Chierchia, G., Pustelnik, N., Pesquetpopescu, B., et al.: A non-local structure tensor based approach for multicomponent image recovery problems. IEEE Trans. Image Process. **23**, 5531–5544 (2014)
14. Zheng, Y., Jeon, B., Zhang, J., et al.: Adaptively determining regularisation parameters in non-local total variation regularisation for image denoising. Electron. Lett. **51**, 144–145 (2015)
15. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Siam J. Multiscale Model. Simul. **4**, 490–530 (2005)
16. Yaroslavsky, L.P.: Digital picture processing: an introduction. In: Digital Picture Processing, pp. 113–116. Academic Press, New York (1976)
17. Lou, Y., Zhang, X., Osher, S., et al.: Image recovery via nonlocal operators. J. Sci. Comput. **42**, 185–197 (2010)
18. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D Nonlinear Phenom. **60**, 259–268 (1992)
19. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Siam J. Multiscale Model. Simul. **7**, 1005–1028 (2008)
20. Tschumperlé, D., Deriche, R.: Vector-valued image regularization with PDE's: a common framework for different applications. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 506–517 (2005)

Gaussian Mixture Model Based Image Denoising Method with Local Constraints

Min Li[1], Yuhui Zheng[1(✉)], Shunfeng Wang[2], and Jin Wang[3]

[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
limin_513@126.com, zheng_yuhui@nuist.edu.cn
[2] College of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China
wsfnuist@126.com
[3] College of Information Engineering, Yangzhou University, Yangzhou 215127, China
jinwang@yzu.edu.cn

Abstract. Recently, the image denoising methods based on patch priors have received extensive attention. Among these methods, expected patch log likelihood (EPLL) has achieved great success, using Gaussian mixture priors by the Gaussian mixture model (GMM). In the paper, we observe that GMM model requires the estimation of a global parameter λ, rather than locally adaptive parameters. Based on this, we propose a modification of the GMM model which is imposed the local constraints on partition of the image. The experimental results illustrate that our proposed method performs comparatively well.

Keywords: Image denoising · Gaussian mixture model · Expected patch log likelihood · Image patch · Local constraint

1 Introduction

Among many image restoration methods, patch-based methods [1–3] have offered effective ways, for example, log likelihood (log L) probability [4] and maximum a posteriori (MAP) [5] method. We usually use the Markov random field (MRF) [6] for the whole image processing directly. At this point, log L probability and MAP are difficult to calculate accurately. Thus, a general optimization framework based on patch prior has been widely put forward, the most representative, Field of experts (FOE) [7] framework.

Expected log patch likelihood (EPLL) [8] is also an optimization framework using Gaussian mixture priors [9] learned by the Gaussian mixture model (GMM) [10, 11]. We find that it is a global parameter λ in the GMM model, instead of local adaptive parameters. The influence is that the denoising performance of different image regions is inconsistent. For avoiding this effect, we propose a novel GMM model with local constraints. We use a set of constraints λ_i and each λ which permits to satisfy the constraint for one region does not serve for other ones. Certainly, a different selection of the regularization parameters [12] may give better results in some region, but no

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_107

single parameter can give an improvement for all regions at the same time. In the Lagrange multiplier formulations, we need several the Lagrange multipliers to be able to impose locally that noise variance is given by δ^2. Therefore, parameter is now spatial adaptive.

2 Proposed Method

2.1 Background of Expected Patch Log Likelihood

Expected patch log likelihood (EPLL) is a general optimization framework based on patch priors for image restoration. Given a natural image u and known priori p, EPLL can be defined as:

$$EPLL\,(u) = \log\, p(u) = \sum_i \log\, p(P_i u) \tag{1}$$

Where P_i denotes an operator for extracting image patch u_i from image u. The patch prior used in the joint conditional density with the EPLL is a Gaussian Mixture Model (GMM) given by:

$$p(u) = \prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j N(u_i \mid \mu_j, \Sigma_j) \tag{2}$$

Where π_j are the mixing coefficients, μ_j and $\sum_j$ are the corresponding mean and covariance matrix.

Given a noise image u_0, the degradation model can be expressed by $\|u - u_0\|^2$. In order to achieve good results for restoration, we should maximize log likelihood (log L) probability of the image patch, while keeping the u and u_0 as consistent as possible. Therefore, EPLL model based on priori p is represented as follows:

$$\min_u \left\{ \frac{\lambda}{2} \|u - u_0\|^2 - \sum_i \log\, p(P_i u) \right\} \tag{3}$$

Where λ is a regularization parameter. The equation can be solved by Half Quadratic Splitting [8] which introduces a set of auxiliary variables z^i and changes the cost function into the following form:

$$\min_{u,\{z_i\}} \left\{ \frac{\lambda}{2} \|u - u_0\|^2 + \sum_i \{ \frac{\beta}{2} (\|P_i u - z_i\|^2) - \log\, p(z_i) \} \right\} \tag{4}$$

Where β is the penalty parameter which often is set to be large enough to ensure that the solution of (4) is close to that of (3). Then formula (4) can be minimized by alternatively updating z^i and u_i.

2.2 Proposed Method with Local Constraints

Suppose that $\{O_1, \ldots, O_r\}$ is a partition of image. Given $\lambda_1, \ldots, \lambda_r$, we consider the following problem:

Whether there are values $(\lambda_1, \ldots, \lambda_r)$ satisfying local constraints [13, 14] as follows

$$\frac{1}{|O_i|}\|u - u_0\|^2 = \delta^2, \forall i = 1, \ldots, r. \tag{5}$$

We proposed to solve the following constrained problem:

$$\begin{aligned} &\max_u \sum_i \log\, p(P_i u) \\ &s.t. \|u - u_0\|^2 = \delta^2 |o_i| \end{aligned} \tag{6}$$

In case that we answer the above question in the affirmative sense, the solution of (6) would give a solution of the problem:

$$\min_u \left\{ \frac{\lambda(x)}{2}\|u - u_0\|^2 - \sum_i \log\, p(P_i u) \right\} \tag{7}$$

Where $\lambda(x) = \sum_{i=1}^{r} \lambda_i \chi_{o_i}$. For simplicity we shall write $\vec{\lambda} = \{\lambda_1, \ldots \lambda_r\}$ and $\vec{\lambda} \geq 0$ if $\lambda_i \geq 0$ for all $i = 1, \ldots, r$. The novel model that we propose is an extension of GMM model, where the parameter λ takes different values for different regions. To solve (7) we use the same numerical approach we used to solve (3). The Eq. (7) is equivalently transformed into the following function:

$$\min_{u,\{z_i\}} \{ \frac{\lambda(x)}{2}\|u - u_0\|^2 + \sum_i \{ \frac{\beta}{2}(R\|Ru - z_i\|^2) - \log\, p(z_i) \} \} \tag{8}$$

For solving (8), at first, we choose the most likely Gaussian mixing weight j_{max} for each patch $R_i u$. Then Eq. (8) is minimized by alternatively updating z_i and u:

$$z_i^{n+1} = (\Sigma_{j_{\max}} + \frac{1}{\beta} I)^{-1} \cdot (R_i u^n \Sigma_{j_{\max}} + \frac{1}{\beta}\mu_{j_{\max}} I) \tag{9}$$

$$u^{n+1} = u^n + \Delta t[\lambda(x)(u_0 - u^n) - \sum_i \beta R_i^T (R_i u^n - z_i^n) \tag{10}$$

Where I is the identity matrix, Δt is the time step. In practice, for updating the parameters λ_i, we use Uzawa's method [15]. In summary, the algorithm can be implemented as follows:

Step1. Input corrupted image u_0, model parameters $\beta, \Delta t$ and iteration stopping tolerance ε;
Step2. Choose the most likely Gaussian mixing weights j_{max} for each patch $R_i u$;
Step3. Initially, we take the values of $\lambda_i > 0$ small enough so that

$$Q_{Oi}\left(u^{\lambda}\right) = \frac{1}{|O_i|}\left\|u^{\lambda} - u_0\right\|^2 > \delta^2, \forall\, i = 1, \ldots, r$$

Step4. For each set of values $\lambda_i > 0$, we alternatively update (9) and (10), until we reach the asymptotic state u^{λ}.
Step5. For each $i \in \{1, \ldots, r\}$ recompute $\lambda_i = \max(\lambda_i + \rho\left(Q_{Oi}\left(u^{\lambda} - \delta^2\right), 0\right)$ (with $\rho > 0$ small enough)
Step6. Iterate steps 4–5 until the λ_i' satisfying stopping criterion.

3 Implementation and Experiment Results

In experiments, we compare our proposed method with the original EPLL in image denoising. The GMM with 200 mixture components is learned from 2×10^6 images patches which are sampled from the Berkeley Database. The experimental pictures are added Gaussian noise with zero mean and standard variance $\delta = 25$.

Figures 1 and 2 show the performance of the EPLL with Gaussian mixture priors and our method respectively on Test1 image (i.e. No. 37073) and Test2 image (i.e. No. 103070) in denoising. We can find that our proposed method outperforms the original EPLL in the denoised result. Because that the local constraints are equivalent

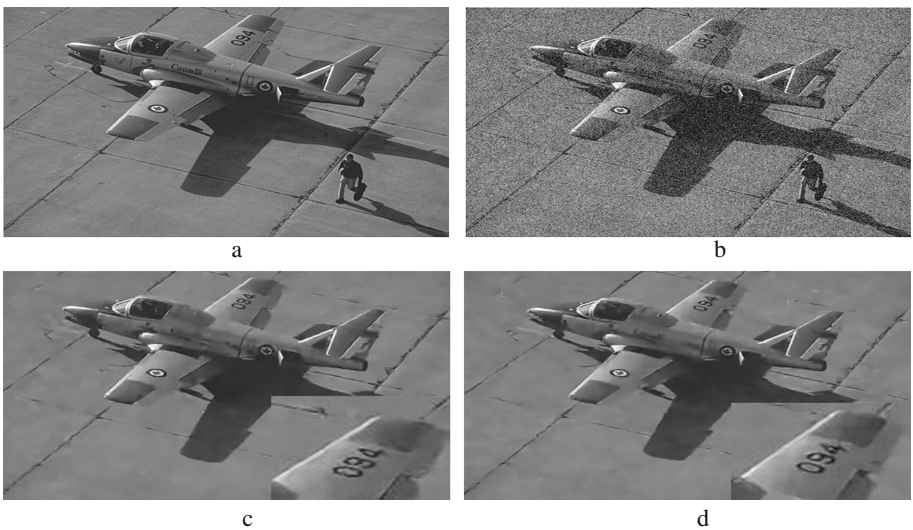


Fig. 1. Denoising results on Test1 image

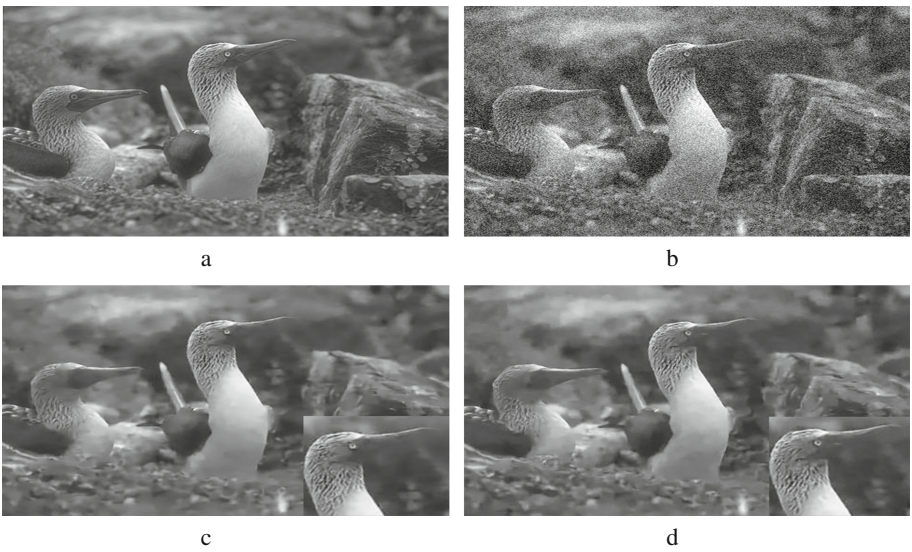


Fig. 2. Denoising results on Test2 image

Table 1. The PSNR results of different denoising models

Image	EPLL	Our method
Test1	30.37	30.54
Test2	29.89	30.22

to the fidelity term with spatial adaptive parameters by Lagrange multiplier method. For the related quantitative comparison, as demonstrated in Table 1, the peak signal to noise ratio (PSNR) value of our method is also higher than the original EPLL.

4 Conclusions

Image priors play a vital role in image restoration tasks. In this paper, we devote to researching on Gaussian mixture model based on local constraints. We construct an adaptive regularization parameter coupling the local entropy of the image, which varies with different regions of the image and each λ corresponds to a region. The numerical results show our proposed method achieves a satisfying denoised result, compared with the original EPLL algorithm with fixed regularization parameters.

References

1. Vaksman, G., Zibulevsky, M., Elad, M.: Patch ordering as a regularization for inverse problems in image processing. Siam J. Image Sci. **9**(1), 287–319 (2016)
2. Dameravenkata, N., et al.: Image quality assessment based on a degradation model. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **9**(4), 636–650 (2000)
3. Boulanger Jr., C.K., Bouthemy, P.: Space-time adaptation for patch-based image sequence restoration. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1096–1102 (2007)
4. Yavuz, M., Fessler, J.A.: Maximum likelihood emission image reconstruction for randoms-precorrected PET scans. Nucl. Sci. Symp. Conf. Rec. IEEE **15**(2), 229–233 (2000)
5. Chantas, G., Galatsanos, N., Likas, A.: Maximum a posteriori image restoration based on a new directional continuous edge image prior. In: IEEE International Conference on Image Processing IEEE Xplore I-941-4 (2005)
6. Sigelle, M.: A cumulant expansion technique for simultaneous Markov random field image restoration and hyperparameter estimation. Int. J. Comput. Vis. **37**(3), 275–293 (2000)
7. Roth, S., Black, M.J.: Fields of experts. Int. J. Comput. Vis. **82**(2), 205–229 (2009)
8. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 479–486. IEEE (2011)
9. Kurisu, K., et al.: Image segmentation using a spatially correlated mixture model with Gaussian process priors. In: Pattern Recognition, pp. 59–63. IEEE (2013)
10. Nguyen, T.M., Wu, Q.M.J.: Gaussian-mixture-model-based spatial neighborhood relationships for pixel labeling problem. IEEE Trans. Syst. Man Cybern. Part B Cybern. Publ. IEEE Syst. Man Cybern. Soc. **42**(1), 193–202 (2012)
11. Yu, G., Sapiro, G., Mallat, S.: Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity. IEEE Trans. Image Process. **21**(5), 2481–2499 (2012)
12. Zheng, Y., et al.: Adaptively determining regularization parameters in non-local total variation regularisation for image denoising. Electron. Lett. **51**(2), 144–145 (2015)
13. Bertalmio, M., et al.: TV based image restoration with local constraints. J. Sci. Comput. **19**(1), 95–122 (2003)
14. Li, F., Zeng, T.: Image restoration via tight frame regularization and local constraints. J. Sci. Comput. **57**(2), 349–371 (2013)
15. Ciarlet, P.G.: Introduction to numerical linear algebra and optimization. Math. Comput. **55** (191), 159 (1990)

A Study on Secure Protocol Techniques Supporting TCUs in a Telematics Environment

Byoungsoo Koh and Suhong Shin(✉)

DMC R&D Center, 8th FL. Maebongsan(St) 37, Mapo-gu, Seoul 03909, Korea
{bskoh, shshin}@digicaps.com

Abstract. As mobile communications develop, automobiles are becoming increasingly computerized. Connected cars can communicate in various ways, with the range of communication also expanding. Onboard communication enables the exchange of information between various control devices, including instruments in the car, engine control units, and steering and braking devices. External communication can inform the driver of traffic-related, or accident-prevention information. By connecting externally using telematics, new services including software updates, remote diagnosis, emergency calls (eCall), payment, internet, infotainment, and automobile related apps have become available. Whilst connected cars are evolving into communication devices with many capabilities, in the future connected cars may include many interfaces and gateways, in which internal and external operations can be manipulated. Therefore, a comprehensive security architecture is required within the vehicle at points of contact with the outside world, and security enhancements are required for secure storage of keys used for authentication and communication security. An approach that takes into account both functional safety and information security has thus far not been found; there is therefore a need for specialized hardware and software to protect the relevant data, along with the solutions listed, to meet the functional safety required by ISO 26262.

Keywords: Secure protocol · Telematics control unit · Secure framework

1 Introduction

Connected technology allows communication with not only the network inside a vehicle, but also concurrently with traffic control centers and other nearby vehicles. As vehicles gain more connectivity with their surrounding environment, it has become possible to provide various types of improvements for drivers. Connectivity provides additional information on the environment to a vehicle, which can improve the driving operation and provide services based on the vehicle using a multitude of information.

An example of this is a remote software update, which eliminates the need for costly recalls when a problem occurs in a vehicle network or when a sudden software update is required. Moreover, it allows for additional features, such as online services or toll collection, to be applied. However, such external vehicle access also increases the risk of unauthorized manipulation or cyber-attacks. Therefore, a cutting-edge

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_108

semiconductor solution must also provide a security system for the protection of not only human lives and property along the road, but also private information.

In the future, car-to-car (C2C) and car-to-infrastructure (C2I) communication technologies will enhance the road safety and efficiency. For example, the system may alert the driver regarding approaching anomalies or accidents in advance, or allow a remote diagnosis through a wireless interface such that components requiring preventive measures can be serviced. For this purpose, sensitive data such as location and speed must be exchanged, and such data must be protected from illegal attacks. However, one problem is that the security of such data has yet to be perfected. Moreover, although such systems have been implemented on actual cars through numerous studies, future smart cars will require even more network resources, and continual research is therefore required regarding the corresponding security mechanisms. In this paper, we analyze the existing technologies of secure communication protocols required in such telematics environments, and propose a security framework for safe storage and management of keys used for security protocols applied both inside and outside a vehicle.

This paper is organized as follows. In Sect. 2, we describe the communication and security technologies related to smart cars. In Sect. 3, we study the security frameworks used for security protocols of in-vehicle networks (IVN). Finally, Sect. 4 provides some conducing remarks regarding the present study.

2 Technological Trend

This chapter describes the communication and security technologies required for telematics. Telematics technology can be classified into previous IVNs and external vehicle communication, and its adopted protocols comply with various standards.

2.1 WAVE Communication Technology

For security in WAVE communication technology, IEEE 1609.2 defines the format for secure messages and the procedure for secure communication. IEEE 1609.2 also provides an authentication mechanism for WAVE messages and a user authentication mechanism. Moreover, an anonymity mechanism is included in the standard for user protection. The greatest technical issue to be resolved in WAVE security regarding the protection of messages between vehicles moving at high speeds is to minimize the processing time so that packet losses do not occur.

The WAVE security layer encrypts the messages through the following procedure.

- The hash value of the plaintext message is calculated.
- The digital signature value of this hash value is calculated.
- A ciphertext is created through a combination of the plaintext and digital signature.

2.2 Authentication

Each element comprising a V2X network (vehicle-to-vehicle, vehicle-to-infrastructure, and vehicle-to-mobile device) must be able to provide mutual authentication of the exchanged information, as well as a trust relationship. For suitability with the particular environment of high-speed vehicles, ongoing effort has been made regarding the standardization of a wireless communication protocol between WAVE vehicles, where the channel bandwidth is reduced to 10 MHz and the RF output power is increased to 44.8 dBM. Although this standardization has been currently omitted to reduce the wireless link connection time, this problem must be resolved through research into high-speed message encryption technologies. The crash avoidance metrics partnership (CAMP) standard of the US has an architecture for generating IDs using the linkage authority (LA) in a traditional public-key infrastructure (PKI) and generating certificates based on such IDs. The types of certificates include certificate signing request (CSR) certificates used for vehicle identification, and pseudonym certificates used for establishing trust in the communication between vehicles.

In the car-to-car communication consortium (C2C-CC) of Europe, CSR certificates are generated from the long-term CA (LTCA), and pseudonym certificates are generated from the pseudonym CA (PCA).

The telecommunication technology association (TTA) of Korea defines an authentication service architecture as similar to that used for V2V communications in C2C-CC.

2.3 Security Protocol

It is clear that strong security is required in all V2V systems to prevent communication theft or modifications because a communication failure may lead to a criticality. The United States National Highway Traffic Safety Administration (NHTSA) has already devised a rough security framework for supporting real-time exchanges and authentication. NHTSA has examined a few alternatives, and has decided to adopt an asymmetric PKI encryption system. The PKI was designed to support secure communication through public networks, and checks the identity using a digital signature.

2.4 Authentication and Security Protocols in IVN Environments

Controller area network (CAN) communication, which is the most frequently used protocol in a vehicle network, does not provide data encryption or authentication features despite being a broadcast communication protocol. Ford separates the network used for braking devices from the network used for Internet access, thereby preventing cyber-attacks from impeding the operation of the vehicle. Toyota allows the use of only their own proprietary app store to prevent virus infections occurring through downloaded apps. In Korea, a government research project illustrating the vulnerabilities of CAN communication was conducted, and encryption and authentication techniques providing protection from such threats were designed. In the field of vehicle-ICT convergence security, there is currently ongoing research into the technology for secure communication between the vehicle engine control units (ECUs).

3 Security Framework for Vehicle Communication

In this chapter, we describe the architecture design and a framework for secure communication with support for telematics and key updates. In previous vehicles, vulnerabilities against network exploits and reverse-engineering attacks occurred because IVNs were used without the support of external networks. The system proposed in the present paper was designed such that the internal protocol performs secure communication through authentication and hashing, and the engines or keys required for security are regularly downloaded from a cloud. This technique resolves the security vulnerabilities in existing smart cars by enhancing the security of the in-vehicle security protocol and supporting security key updates through an external network, which also ensures the system availability.

The following provides explanations of each security component.

3.1 Secure IVN Communication

Figure 2 illustrates the functionality of secure communication in an IVN environment. Currently, IVN communication for telematics can be divided into the communication between the telematics control unit (TCU) and human-machine interface (HMI),

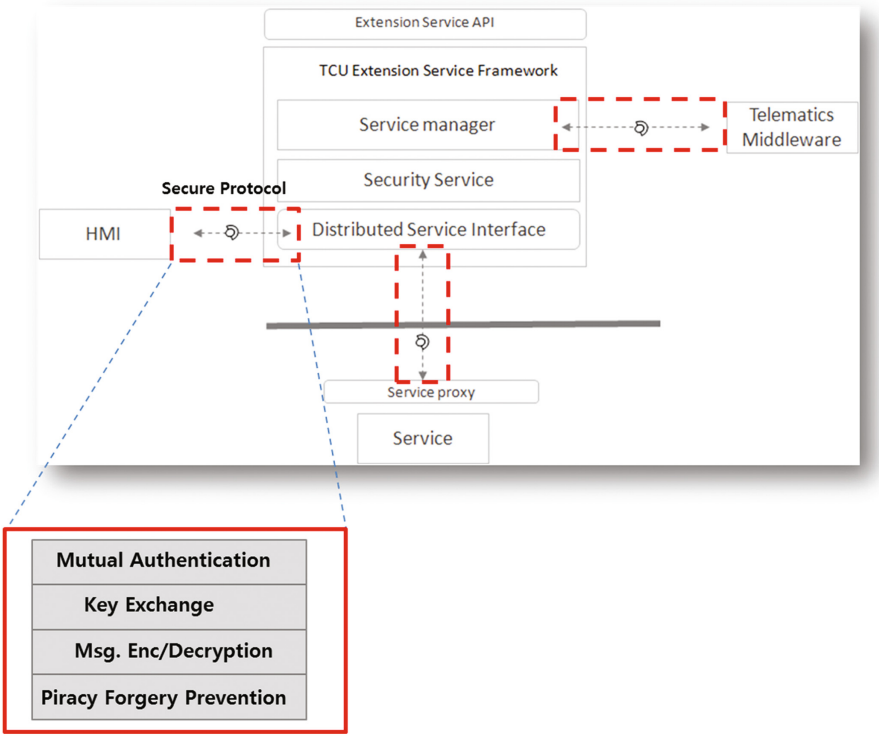


Fig. 1. End-to-end secure communication block

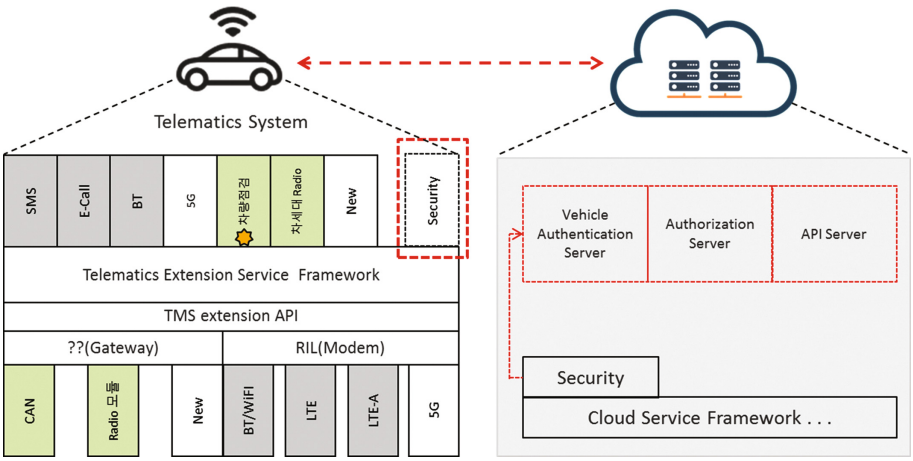


Fig. 2. TCU-cloud data download security block

between the TCU and middleware, and between the TCU and service modules. Safety must be guaranteed for message transmission and reception based on each individual communication protocol.

The red block in (Fig. 1) therefore supports security protocols, which include the following features.

- Mutual Authentication
- Key Exchange
- Message Encryption/Decryption
- Piracy Forgery Prevention

3.2 Secure Communication Between TCU and Cloud

Telematics is connected to external systems through an LTE network from the interior of the vehicle, and various types of information are collected or stored based on such connectivity. During this process, the vehicle and external system frequently attempt to network, which necessitates the safe downloading of relevant data. In this study, we designed a security feature for safe updates between the telematics module and cloud, allowing security modules and key information for security reinforcement to be safely downloaded and maintained from within the vehicle.

3.2.1 Security Module

The security module block is located in the TCU, and is responsible for the safe delivery of data during communication with the cloud system. The security module block supports an encryption algorithm and authentication features. This module requests and processes the API and key information required by the TCU through the cloud system.

3.2.2 Vehicle Authentication Server

The vehicle authentication server is a server that authenticates the TCU modules. It supports verification of the trustworthiness of a TCU based on the unique information of the vehicle passed from the security module of the TCU, and on the previously issued TCU digital certificate (Fig. 3).

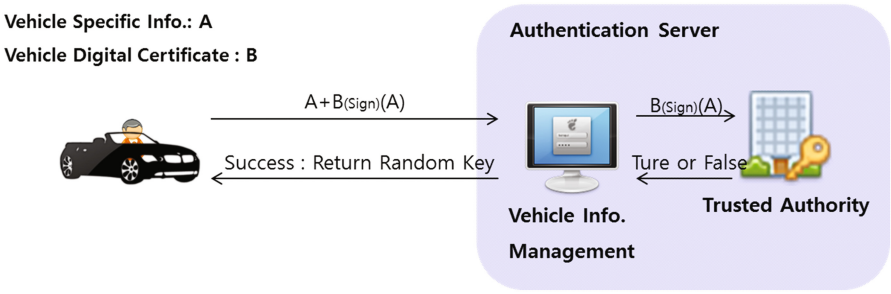
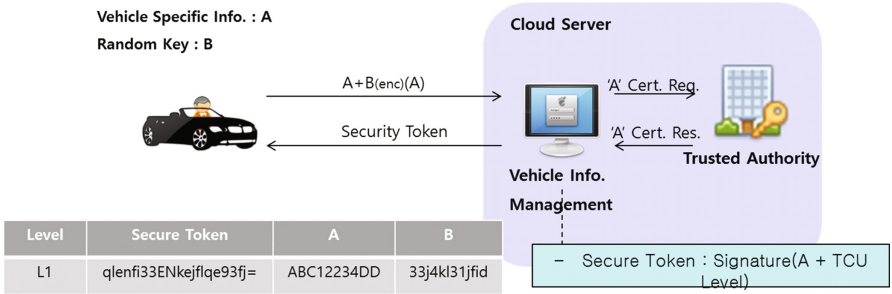


Fig. 3. Authentication server

3.2.3 Authorization Server

The authorization server generates and manages secure tokens for controlling the access information to the APIs based on the TCU information. The authorization server verifies the API access authority based on a combination of the unique vehicle information and a random key, and passes the secure token to the TCU (Fig. 4).



Level	Secure Token	A	B
L1	qlenfi33ENkejflqe93fj=	ABC12234DD	33j4kl31jfid
L2	qelfjelj91393EEKJkdf=	ABE12333DJ	13lkjklfkIda
L3	asdlj1kl3j4QQWEdkfj=	IEJ353134EK	1lk3jflkdak

Fig. 4. Authorization server

3.2.4 API Server

The security module sends API requests to the API server on the cloud along with the API profile required by the TCU and the secure token assigned by the authorization server. The API server on the cloud verifies the received secure token, and passes the resulting value of the information requested by the client (APIs or a key) to the security module.

4 Conclusion

This paper presented a study on a security framework supporting secure communication technologies for the safe transmission and reception of data during communication inside and outside a vehicle in a telematics environment. Through this study, we proposed a cloud-based update system for supporting section-wise key generation and updates, as well as enhanced security robustness. This will allow for an active defense and countermeasures against the intrusion of various internal and external protocols in a smart vehicle, thereby supporting renewability in an SW platform environment. In the future, safe communication with other control units will need to be improved by combining security features based on hardware security modules, and by using digital signatures for messaging, or the overall encryption through the application of such modules. Discrete hardware solutions, such as a security controller or a controller combining TPM functionality, enable the application of various devices with verified functionality, from IC cards to automobile security, and are hence expected to provide optimized security frameworks in connectivity environments.

Acknowledgements. This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program (10070156, Extended service framework for telematics in a connected car environment) funded By the Ministry of Trade, industry and Energy (MOTIE, Korea).

Development of Test Agents for Automated Dynamic Testing of UNIWAY

Seoung-Hyeon Lee(✉) and Jung-Chan Na

Electronics and Telecommunications Research Institute, Daejeon, South Korea
{duribun2,njc}@etri.re.kr

Abstract. UNIdirectional security gateWAY (UNIWAY) is a technology that makes it possible to communicate only in one direction physically, unlike a general network that communicates bi-directionally. It is a technology to transmit data from the safe area to the non-safety area. Since it is an industrial system that operates continuously for a long time, it is necessary to grasp the normal functional operation of the system and satisfy the performance requirement in advance. In order to test that UNIWAY software operates stably for various network situations, it is necessary to elicit test items such as communication node, transmitted data size, transmission data type, and prepare test data in various combinations. In addition, since it takes a lot of time and effort to check the stability, it is necessary to automate the test and confirm the test results. In this paper, we design an automated agent that can test automatically, and repetitively using predefined test scripts and data.

Keywords: UNIWAY · Test agent · Automated testing

1 Introduction

In recent software development, processes and development tools are similar because the environment is generalized, and many automation tools necessary for software development are being developed.

An automated tool that replaces the developer's own tests reduces time and space constraints. It also provides more and more accurate results than developers do.

In recent years, projects that require the use of automation tools are increasing. The system configuration and interface are also determined by the developer's propensity and cannot be fully developed and tested consistently. Automated tools can consistently apply a large number of cases and provide faster and more accurate results than human tests [1, 2].

The automation tools started in the field of testing are expanding to component-based development and UI/UX automation tools to improve development efficiency.

A software test verifies whether a program is working properly by comparing the expected result with the actual result using a sample data set. However, because

This work was supported by the ICT R&D program of MSIP/IITP-[R0126-15-1095, Unidirectional Security Gateways developments in cyber-physical systems].

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_109

developers need to test themselves, they cannot target all data sets and spend a lot of time building a systematic test plan for consistent testing. Repetitive and simple test work is a very important task indispensable for a project, which gives developers difficulties of simple labor. Test automation can test all cases consistently and quickly, reducing the burden on developers and increasing software reliability.

UNIWAY collected a lot of industrial sensor data, and the collected data is transmitted reliably to a central server [3, 4]. If the central server fails to provide reliable data, the user may be confused about the provided information [5].

Therefore, in this paper, we conducted research on the automatic dynamic test agent for the development efficiency and consistent testing, and implemented the automated dynamic test agent.

2 Conditions for Automated Test Tools

After designing for the test automation, create test cases and create a script that fits the case. Test scripts should be written in accordance with the business and development language. The results of the test automation should be the same even if the viewpoint and the environment are changes.

Here are some things to keep in mind when writing test scripts:

The user does not intervene during the test. It is difficult to automate when user intervention during testing.

If the user enters a script during the test, the results will vary depending on the input. It should be written to the minimum extent only when user intervention is absolutely necessary.

Do not take a long time to perform the test. If the test takes a long time to run, you might suspect that it is working properly and you will not be able to do anything else until the test is complete.

Sometimes it takes a long time, but it's a good idea to avoid testing as long as possible. Also, when access to the actual database is affected and the test is dependent on the outside, the test may fail at the same time when the service is down.

It is advisable to avoid external access as much as possible because it can damage the outside or affect the test result when connecting from outside. However, writing a script with too many constraints does not make sense for the tests, so you need to set a minimum limit. It does not assume that resources are necessarily present.

Assuming that the data is in the database, testing can be successful or unsuccessful. This is because you have to put the data in the database. In this case, it is recommended to write the data in a form to directly input data while testing. You can reduce the error by not assuming possible resources when you are testing (Fig. 1).

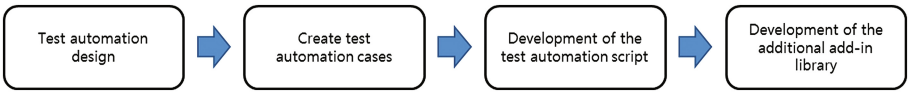


Fig. 1. Procedures of test script development

3 Design for Automated Dynamic Test Agent

3.1 System Under Test

As shown in Fig. 2, the UNIWAY system is constructed in the network environment to prevent external threats, so that data can be physically transmitted in only one direction. The system consists of the following three items.

① Unidirectional communication server

- Transmission from the sending communication server to the receiving communication server
- Physical unidirectional processing only

② Two-tier process within a communications server
③ Proxy process: Communication with external AP server
④ Middleware process: Communications server communication

Figure 2 shows the environment configuration for dynamic test automation scripts.

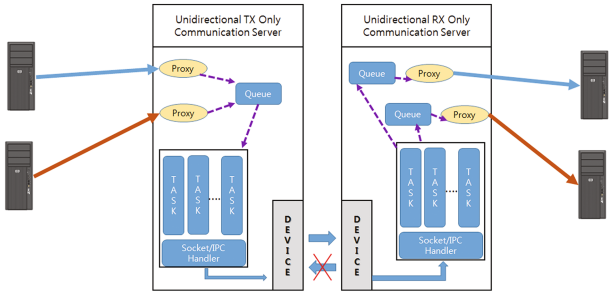


Fig. 2. Overview of the UNIWAY

3.2 The Environment of Dynamic Test Automation Scripts and Data Development

The test environment is configured as shown in Fig. 3.

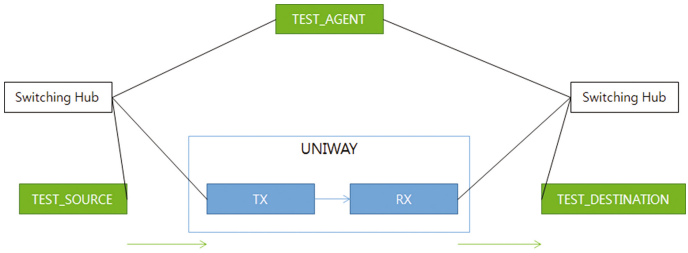


Fig. 3. Overview of the UNIWAY

- TEST_AGENT: Execute automatic test execution command with TEST_SOURCE, TEST_DESTINATION using test script and test data
- TEST_SOURCE: Perform data transmission. Data transfer origin
- TEST_DESTINATION: Perform data reception, data transfer destination
- HUB: Switching hub for transmitting test related information
- UNIWAY: System under test. The system consists of TX (transmit) and RX (receive) communication servers and transmits data in one direction.

3.3 Define Test Items and Create Test Scenarios

3.3.1 Functional Test Items

The items to be input in the functional test are SOURCE/DESTINATION count, SOURCE/DESTINATION connection, packet count, packet size, time interval of packet transmission and data type. The input items are as follows.

- SOURCE/DESTINATION count: 1, 2 and 10
- SOURCE/DESTINATION connection: 1:1, 2:1 and 10:1
- Packet count: 1,000 and 10,000
- Packet size: 128 byte and 1024 byte
- Time interval of packet transmission: none, 0.1, 0.2 and 0.5 s
- Data type: Plain text, Special Character, video, Random Binary, Photo.

3.3.2 Performance Test Items

The performance test measures bandwidth and data loss rate by changing the number of SOURCE and DESTINATION, and the connection between SOURCE and DESTINATION.

3.4 Creating a Test Script

The test script is defined by the test item definition and the test scenario. The defined test script is shown in Fig. 4.

```
TEST_BEGIN : TC01_TCPF_0002
        RX_UNIWAY_CONFIG_PROCESS_FILE_NAME :
TEST_RX_UNIWAY_CONFIG_PROCESS_0001
        TX_UNIWAY_CONFIG_PROCESS_FILE_NAME :
TEST_TX_UNIWAY_CONFIG_PROCESS_0001
        RX_UNIWAY_CONFIG_IPC_FILE_NAME :
TEST_RX_UNIWAY_CONFIG_IPC_0001
        TX_UNIWAY_CONFIG_IPC_FILE_NAME :
TEST_TX_UNIWAY_CONFIG_IPC_0001
        TEST_DATA_FILENAME : GENE_000000128000
        INTERVALS : 0
        DATA_SIZES : 128
        RX_TCP_BEGIN : TC01_TCPF_0002_0001
                RX_TCP_DESTINATION_IP_ADDRESS : 192.168.1.6
                RX_TCP_DESTINATION_IP_PORT : 24010
        RX_TCP_END
        TX_TCP_BEGIN : TC01_TCPF_0002_0001
                TX_TCP_SOURCE_IP_ADDRESS : 192.168.0.3
                TX_TCP_DESTINATION_IP_ADDRESS : 192.168.0.4
                TX_TCP_DESTINATION_IP_PORT : 24010
        TX_TCP_END
TEST_END
```

Fig. 4. The example of test script

4 Experimental Results

The UNIWAY TCP/UDP functional test performs tests using test scripts and test data in the test automation tool. Figure 5 shows the test environment using test automation tools.

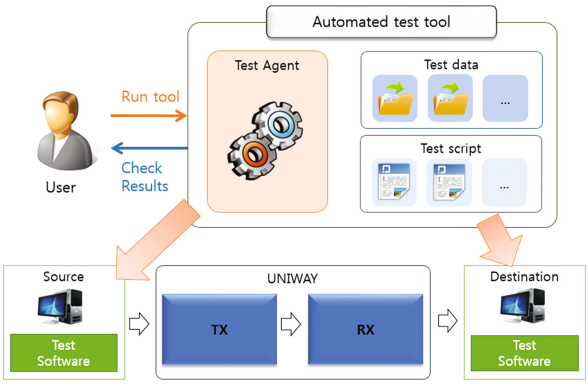


Fig. 5. The test environment using test automated tools

The test tool is implemented so that the user can selectively execute the test through the Command Line Interface, and displays the test case information in progress on the screen in real time. The source and destination PCs include test software to receive commands from the test automation tool and perform data transfer. When the user passes the test execution command to the test automation tool, the test automation tool performs the following actions.

- Start the test: Load the test script to load the test run information.
- Test data distribution: Copy the data to be used for the test to the source PC.
- Run test: First run the test software on the destination PC and prepare to receive the file, then issue a data transfer command from the source PC test software to the destination PC.
- Result collection and judgment: Compare and determine the original test data and the data transmitted via UNIWAY (Figs. 6 and 7).

TC_ID_NAME	SUB_TC_NAME	SOCKET_TYPE	TX-START_TIME	TX-END_TIME	RX-START_TIME	RX-END_TIME	PASS/FAIL	DATA_RATE
TC01_TCPF_0001	TC01_TCPF_0001_0001	TCP	20022141.141373	20022141.198221	20022143.764145	20022144.930925	PASS	3.177738
TC01_TCPF_0002	TC01_TCPF_0002_0001	TCP	20022321.300733	20022321.374014	20022323.956743	20022325.136578	PASS	3.209287
TC01_TCPF_0003	TC01_TCPF_0003_0001	TCP	20022501.383276	20022501.454453	20022504.053698	20022505.216974	PASS	3.216472
TC01_TCPF_0004	TC01_TCPF_0004_0001	TCP	20022641.415024	20022641.483681	20022644.076722	20022645.261577	PASS	3.219797
TC01_TCPF_0005	TC01_TCPF_0005_0001	TCP	20022821.515853	20022821.572894	20022824.157270	20022825.339234	PASS	3.203878
TC01_TCPF_0006	TC01_TCPF_0006_0001	TCP	20023001.685709	20023001.804288	20023004.347343	20023005.529172	PASS	3.193259
TC01_TCPF_0007	TC01_TCPF_0007_0001	TCP	20023141.886933	20023141.976076	20023144.513915	20023145.686526	PASS	3.168716
TC01_TCPF_0008	TC01_TCPF_0008_0001	TCP	20023321.963340	20023322.068774	20023324.618153	20023325.730175	PASS	3.158107
TC01_TCPF_0009	TC01_TCPF_0009_0001	TCP	20023502.227861	20023502.313283	20023504.876086	20023506.020782	PASS	3.177862
TC01_TCPF_0010	TC01_TCPF_0010_0001	TCP	20023642.328942	20023642.407524	20023644.961109	20023646.137007	PASS	3.180825
TC01_TCPF_0011	TC01_TCPF_0011_0001	TCP	20023822.442011	20023823.445623	20023825.099215	20023831.709331	FAIL	5.460456
TC01_TCPF_0012	TC01_TCPF_0012_0001	TCP	20024003.061120	20024004.094074	20024005.699965	20024012.314247	FAIL	5.429509
TC01_TCPF_0013	TC01_TCPF_0013_0001	TCP	20024143.719268	20024144.667287	20024146.379003	20024153.010893	FAIL	5.501670
TC01_TCPF_0014	TC01_TCPF_0014_0001	TCP	20024323.976822	20024324.869469	20024326.598168	20024333.232811	FAIL	5.492344
TC01_TCPF_0015	TC01_TCPF_0015_0001	TCP	20024504.940147	20024505.945916	20024507.604576	20024514.154858	FAIL	5.436685
TC01_TCPF_0016	TC01_TCPF_0016_0001	TCP	20024645.292929	20024645.860666	20024647.895588	20024659.496529	PASS	8.119261

Fig. 6. Result of functional test

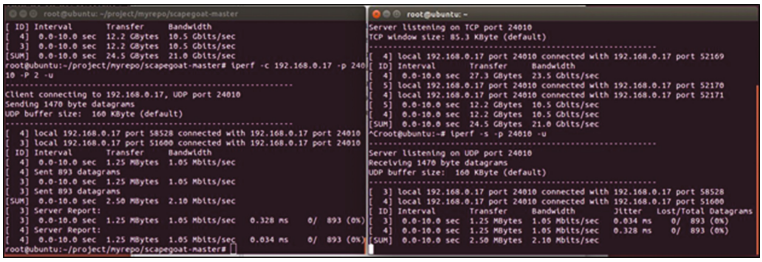


Fig. 7. Result of performance test

5 Conclusion

Repetitive and simple test work is a very important task indispensable, which gives developers difficulties of simple labor and test automation can test all cases consistently and quickly, reducing the burden on developers and increasing software reliability.

In this paper, we have conducted a study on the implementation of an automatic dynamic test agent for the development efficiency and consistent testing. The developed test agent showed satisfaction with UNIWAY's functional and performance test.

References

1. Jovanovi, M.: Software testing methods and techniques. IPSI BgD J. **5**, 30–41 (2009)
2. Software Testing Tools List. http://www.softwaretestingclass.com/software-testing-tools-list/. Accessed June 2012
3. Maity, S., Park, J.-H.: Powering IoT devices: a novel design and analysis technique. J. Converg. **7**, 1–17 (2016)
4. Gaur, M.S., Pant, B.: Trusted and secure clustering in mobile pervasive environment. Hum. Centric Comput. Inf. Sci. **5**, 32 (2015)
5. Im, H., Kang, J., Park, J.H.: Certificateless based public key infrastructure using a DNSSEC. J. Converg. **6**(3), 26–33 (2015)

Multi-step Prediction for Time Series with Factor Mining and Neural Network

Mingji Zhou[1], Jin Liu[1(✉)], Fei Li[1], and Jin Wang[2]

[1] College of Information Engineering, Shanghai Maritime University, Shanghai, China
{zhoumingji,liujin,feili}@shmtu.edu.cn
[2] College of Information Engineering, Yangzhou University, Yangzhou, China
jinwang@yzu.edu.cn

Abstract. Multi-step prediction for time series is a challenging research area with broad applications which can provide important information for relevant decision-makers. Many works extended different architecture of artificial neural networks to perform time series prediction, but they mostly only consider the time series itself, does not weigh the impact of time series of relevant factors. In this paper, a new method of time series prediction based on factor mining is proposed. By analyzing target time series, the means of discovering factors influencing time series and pinned down the most relevant factors was proposed. In the end, a method to do multi-step prediction with artificial neural networks, MTPF is proposed to conduct the time series prediction, create time series model and forecast time series. The proposed method is applied for a shipping price index time series prediction. Results show that this method can improve accuracy of prediction when compared with traditional methods.

Keywords: Time series · Multi-step prediction · Factor mining · Neural network

1 Introduction

Time series [1] refers to a series of observation values obtained in chronological order, and many areas are related to the time series. In the research field of natural science and social science, a large number of decision-making problems are inseparable from the forecast. Time series prediction uses several historical observations values of time series to predict the future value. It appears in many real-word problems, such as weather forecasting, stock index prediction, shipping index prediction.

Time series can be understood as a sequence of vectors, $x(t)$, $t = 0,1,\ldots$ where t represents elapsed time. Theoretically, x may be a value that varies continuously with t, such as temperature. In practice, it will be sampled to give a series of discrete data points which is equally spaced in time. Neural networks have been widely used as time series forecast. Neural network has concentrated on forecasting future developments of the time series from values of x up to the current time. Formally this can be stated as: find a function f: $R^N \rightarrow R$ such as to obtain an estimate of x at time $t + d$ $t+d$오류! 참조 원본을 찾을 수 없습니다., from the N time steps back from time t t 오류! 참조 원본을 찾을 수 없습니다., so that:

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_110

$$x(t+d) = f(x(t), x(t-1), \ldots, x(t-N+1)) \tag{1}$$

$$x(t+d) = f(\mathrm{y}(t)) \tag{2}$$

Where $y(t)$ is the N-dimensional vector of lagged x values. Normally d will be one, so that f will be forecasting the next value of x.

Although some nonlinear time series prediction method can simulate complex relationships, most methods only consider the time series itself, does not weigh the impact of relevant time series. In order to solve this problem, a method of multi-step prediction for time series with factor mining and neural network is proposed which can significantly improve the accuracy of both single time series and multiple time series.

The remaining of this paper is organized as follows. Section 2 presents the related work in the literature. Section 3 explains our proposed method. Section 4 presents the experiment results. The paper ends up with some conclusions in Sect. 5.

2 Related Work

Auto regressive (AR) model is one of linear prediction model. In 1927, the British mathematician Yule [2] proposed this model to predict the variation of the market, marking the time series analysis method production. Another linear prediction model is auto regressive moving average (ARMA) model. In 1938 the Swedish statistician Wold [3] completed a systematic study of the discrete stationary time series, made famous wold decomposition and gave ARMA model. The last one is auto regressive integrated moving average (ARIMA) model, by the American statistician and Box and Jenkins [4] proposed a time series prediction method in the 1970s. They made a systematic and thorough discussion for time series analysis method theory and applications, which is considered to be a leap in the history of the development of the timing analysis. Linear models whose mathematical model is not linear have no ability to cope with certain non-stationary signals.

Radial basis function (RBF) neural network is a kind of nonlinear prediction model. In 1988, Broomhead and Lowe firstly applied RBF to the neural network design constituted a radial basis function neural network [5]. Radial basis function network is a kind of local approximation network, which only a small number of neurons in the input space determine the output of the network.

Support vector machine (SVM), which was first proposed by Cortes and Vapnik [6]. It not only can be applied in other machine learning problems such as fitting function, but also be applied in the prediction of the time series.

Hrasko et al. [7] used Restricted Boltzmann Machine and the Back propagation algorithm for time series prediction. Pisoni et al. [8] used nonlinear auto regressive models (NARX) and artificial neural networks (ANNs) for environmental prediction. Yu et al. [9] proposed a new hyper-parameters selection approach for support vector machines to predict time series. Faruk [10] proposed a hybrid neural network and ARIMA model for water quality time series prediction.

Li et al. proposed a method [11] based on the similarity score set to predict missing ratings and they also presented a method based on the support vector regression to

adjust the deviation. The results show that the algorithms can increase the accuracy of recommendation versus the traditional CF (Collaborative filtering). Barzaiq et al. proposed a trajectory prediction algorithm [12] called personal destination pattern analysis to analyse the different destinations in various trajectories of an individual, and use it for mobile ad serving. Loke proposed the algorithms [13] for finding regions of interest using mobile crowdsourcing, which reduce the number of questions required and increasing the efficiency.

3 Multi-step Time-Series Prediction Based on Factor Mining

The detail of our proposed method for Multi-step time-series forecasting MTPF is presented in this section.

3.1 Factor Analysis

Factor analysis is a method for figuring out a small number of variables which can be used to describe the correlation between multiple variables. According to the correlation size, the variables are classified to find the main factors affecting the time series. With n samples, p indicators, $X = (X_1, X_2, X_3, \ldots X_p)^T$ is used to observe the random variables, find the common factor for $F = (F_1, F_2, F_3, \ldots F_m)$ (m < p), the model is shown as follows.

$$X = \mu + AF + \varepsilon \tag{3}$$

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \tag{4}$$

F is a common factor of p and A is the factor load matrix. During factor analysis, factor loading matrix, orthogonal transform and factors are given a reasonable explanation. For any one factor, there are a number of variables on which have high load, and the remaining variables in the load are very low or zero.

3.2 Factor Mining

In the multi-step prediction of multivariate time series, the correlation factors and the target time series are analyzed firstly. Then using correlation analysis method to find correlation between other sequence and the target time series. The flow chart for the algorithm is shown in Fig. 1.

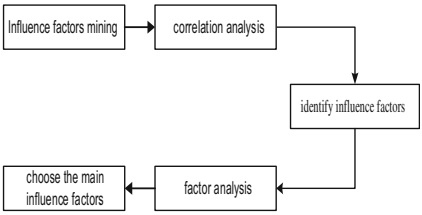


Fig. 1. Algorithm flow chart

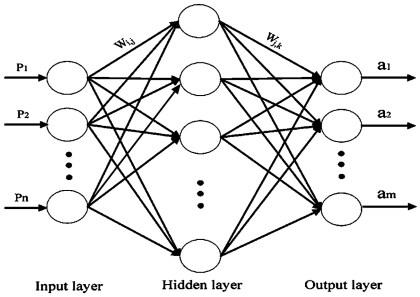


Fig. 2. BP neural network

3.3 Multi-step Forecasting with Neural Network

We used BP neural network for the prediction. For training purposes, a dynamic back-propagation algorithm is required to compute the gradients, which is more computationally intensive than static back-propagation and spends more time. In addition, the error surfaces for dynamic networks can be more complex than those for static networks. The connection weights and bias values of ANN are initially chosen as random numbers and then fixed by the results of training process. Since many alternative training are available, back- propagation is used to train our models. Assume there are P input-output pairs of vectors for ANN. The goal of any training algorithm is to minimize the global (mean sum squared) error E between the real network output $o^{(p)}$ and the desired output $d^{(p)}$. The formula is defined as follows:

$$E = \frac{1}{2P}\sum_{p}\sum_{k}\left(d_k^{(p)} - o_k^{(p)}\right)^2 \tag{5}$$

Where p p오류! 참조 원본을 찾을 수 없습니다. is the index of the P training pair of vectors, k is the index of elements in output vector, $d(p)$ k $d_k^{(p)}$ 오류! 참조 원본을 찾을 수 없습니다. and $o(p)$ k 오류! 참조 원본을 찾을 수 없습니다 are the k_{th} element of p_{th} desired output vector and real output vector respectively. In addition, mean square error (MSE) is used to evaluate the learning effects of BPNN. The architecture of the BP network is shown in Fig. 2. The training is continued until the MSE drops below a certain threshold or tolerance level.

The multi-step prediction task can be achieved by two methods. One is directly prediction, which is explicitly training a model to predict several steps. Another one is the iterative method, which is doing repeated one-step predictions up to the desired horizon [14]. The multi-step forecasting (predicting more than one day in advance) is iterative process where the output of system is fed again as input. And it is commonly used for short forecast horizons [15].

Time series is dynamic, there is a relationship between the sequence of the day and the previous day's changes, and there is also a relationship between the sequence of the day and the changes in the previous period of time. In MTPF, the method of sliding

window simulation is used to construct the input of the neural network. In prediction, the actual observation value x_n, x_{n+1}, x_{n+2}, ..., x_{n+m} is used as the input of the network, the output value for the next moment predicted is x_{n+m+1} and the actual observation value x_{n+1}, x_{n+1}, ..., x_{n+m}, x_{n+m+1} is used as input, x_{n+m+2} is the output, and so on. According to the specific circumstances of the time series to carry out the K step prediction and K is the moving window for multi-step time series prediction.

4 Experiment

In the experiments, we used time series of a container freight index (CCFI sequence) as the prediction target. For the impact factors of shipping price index, we choose six factors time series, which are shipping volume sequence, international oil price index (WIF sequence), consumer price index(CPI sequence), producer price index(PPI sequence), NASDAQ sequence, the exchange rate sequence, which are well known relevant to predict shipping index sequence. We gathered historical data from January 1, 2014 to February 20, 2015, which from WIF sequence, NASDAQ sequence daily data. We gathered historical data from January 1, 2012 to September 30, 2015, which from shipping volume sequence, PPI sequence, the exchange rate sequence monthly data. In order to eliminate the influence of the unit of data during training and predicting, the input of neural networks need to be normalized, and the output of neural networks need to be anti-normalized. In this paper, we used the following formula to normalize the data processing:

$$\tilde{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6}$$

And we used following formula to anti-normalize the data processing:

$$\tilde{y} = \mathrm{y} * (x_{max} - x_{min}) + x_{min} \tag{7}$$

A correlation analysis of influencing factors are conducted and part of results are shown above. It can be analyzed in Table 1 that Shipping Volume sequence and CCFI sequence at the level of a significant correlation. And in Table 2, we can find USD to RMB sequence and CCFI sequence are at the level of a significant correlation.

Table 1. Shipping volume sequence and CCFI sequence descriptive statistics

	Mean	Std. deviation	N
CCFI	1073.1151	121.13327	45
Shipping volume	6912598.667	628936.4512	45

Table 2. USD to RMB sequence and CCFI sequence descriptive statistics

	Mean	Std. deviation	N
CCFI	1073.1151	121.13327	45
USD to RMB	6.214187	.0863590	45

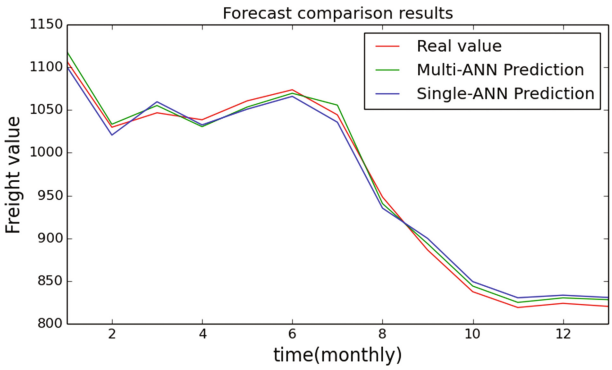


Fig. 3. Multi-step prediction results of USD to RMB sequence

Based on the mined factors, the result of multi-step time series prediction is shown in Fig. 3. The red curve represents the real value, the green one represents predictive value of multiple artificial neural networks and the purple one represents the predictive value of single neural networks. The diagram shows that the MTPF model prediction result is more accurate than the single ANN model prediction result. Predictive value by MTPF and the real value are very close. It shows the accuracy of the model prediction.

5 Conclusion

The completed MTPF model was successfully predicted the time series. Time series neural networks provided predication successfully as well. It can be concluded that, prediction made by MTPF can be used as an effective time series analysis and prediction tools. We will keep working on the improvements of MTPF such as multiple neural network prediction. In the future, the MTPF method can be extended to other areas such as financial market or weather forecasts for uncertainty modeling.

References

1. Weigend, A.S.: Time series prediction: forecasting the future and understanding the past. Santa Fe Institute Studies in the Sciences of Complexity (1994)
2. Gershenfeld, N.A., Weigend, A.S.: The future of time series: learning and understanding. No. 93-08-053 (1993)
3. Wold, H.: A study in the analysis of stationary time series, pp. 113–115 (1939)
4. Box, G.E.P., Jenkins, G.M.: Time series models for forecasting and control. J. Time **3**(2), 199–201 (1970). San Francisco
5. Lowe, D.: Multi-variable functional interpolation and adaptive networks. Complex Syst. **2**, 321–355 (1988)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)

7. Hrasko, R., Pacheco, A.G.C., Krohling, R.A.: Time series prediction using restricted Boltzmann machines and backpropagation. Procedia Comput. Sci. **55**, 990–999 (2015)
8. Pisoni, E., Farina, M., Carnevale, C.: Forecasting peak air pollution levels using NARX models. Eng. Appl. Artif. Intell. **22**(s4–5), 593–602 (2009)
9. Yu, Y., Song, J., Ren, Z.: A new hyper-parameters selection approach for support vector machines to predict time series. In: Zu, Q., Hu, B., Elçi, A. (eds.) Pervasive Computing and the Networked World, pp. 775–787. Springer, Heidelberg (2013)
10. Faruk, D.Ö.: A hybrid neural network and ARIMA model for water quality time series prediction. Eng. Appl. Artif. Intell. **23**(4), 586–594 (2010)
11. Li, W., Li, X., Yao, M., et al.: Personalized fitting recommendation based on support vector regression. Hum. Centric Comput. Inf. Sci. **5**(1), 21 (2015)
12. Barzaiq, O.O., Loke, S.W.: Personal destination pattern analysis with applications to mobile advertising. Hum. Centric Comput. Inf. Sci. **6**(1), 17 (2016)
13. Loke, S.W.: Heuristics for spatial finding using iterative mobile crowdsourcing. Hum. Centric Comput. Inf. Sci. **6**(1), 4 (2016)
14. Aizenberg, I., Sheremetov, L., Villa-Vargas, L.: Multilayer neural network with multi-valued neurons in time series forecasting of oil production. Neurocomputing **8495**, 61–70 (2015)
15. Cadenas, E., Rivera, W.: Short term wind speed forecasting in La Venta, Oaxaca, México, using artificial neural networks. Renew. Energy **34**(1), 274–278 (2009)

A Vision-Based Approach for Deep Web Form Extraction

Jiachen Pu[1], Jin Liu[1(✉)], and Jin Wang[2]

[1] Shanghai Maritime University, Shanghai, China
{pujiachen, jinliu}@shmtu.edu.cn
[2] College of Information Engineering, Yangzhou University, Yangzhou, China
jinwang@yzu.edu.cn

Abstract. The World Wide Web is a large source of information that contains data in either Surface Web or Deep Web. Compared with the data in the Surface Web, the Deep Web contains a greater amount of structured data with higher quality, but it is difficult to use directly. Studies in this field have revealed some methods for Deep Web Form Extraction, they may fall into the following categories which are HTML-based, vision-based, ontology-based, ML-based, NLP-based and so on. This paper try to combine the DOM tree and the convolutional neural network together and then find out the form in the Web page. This paper proposed a vision-based method VBF, which figures out the form from the Web page through the acquisition of the HTML code and screenshots of Web pages, establishment of the DOM tree and the calculation of the neural network and form recognition, matching, and generation.

Keywords: Deep Web · Form extraction · Vision-based

1 Introduction

The World Wide Web is a large source of information that contains data in either Surface Web or Deep Web. As for the information in the Deep Web, those pages are generated through dynamic data query using the back-end database, and it cannot be used by search engines. Compared with the data in the Surface Web, the Deep Web contains a greater amount of structured data with higher quality.

Although the data in the Deep Web is structured, high-quality and has good characteristics, it is difficult to use it directly due to the differences of the developers lead to the great differences in the structure of the pages. So we need to find a way to perform Deep Web data extraction effectively and quickly regardless of the differences in the structure of the query pages. Using convolutional neural networks for identification is an efficient method that has been developed in recent years and aroused extensive attention, it can automatically recognize visual patterns from the original image. As a deep learning framework, it was developed to minimize data pretreatment requirements. Convolutional neural networks have a certain degree of invariability for geometric transformations, deformation and light, and less computational cost to scan the whole image.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_111

The remaining of this paper is organized as follows: Sect. 2 mainly talking about our data extraction method VBF and Sect. 3 involves the details of our experiment.

1.1 Related Work

A lot of domestic and foreign scholars and research institutions have studied in Deep Web data extraction. These methods fall into the following categories which are HTML-based, vision-based, ontology-based, ML-based, NLP-based and so on:

For the first category, Crescenzi [1] analyzed the page's HTML code and used certain rules for data extraction and then proposed RoadRunner, matching HTML tags to extract data for some of the data in simple deep web surface extraction. Chang [2] proposed IEPAD, solved the problem that the training of the data extraction requires a lot of manual annotations. He used the string pattern matching to extract the data. The model offers the expressiveness and it's easier to understand the meaning of the matching string. The latter two methods are deployed around the DOM. If a DOM tree is generated, extraction work will be carried out efficiently. Liu [3] performed data extraction by analyzing the DOM tree according to certain rules and then proposed MDR, grouping a plurality of similar nodes to divide the data area, in which a node corresponds to the data record. Based on MDR, Zhai [4] put forward DEPTA which only requires more than one data item in the page and uses alignment algorithm to improve the accuracy of data extraction. Using the analysis of the DOM tree for data extraction requires a lot of walking the DOM tree node which leads to high time complexity, and cannot achieve good results.

The second category is the vision-based methods. Cai [5] proposed the VIPS method, which is based on visual semantics, the DOM tree and visual features are combined to form a DOM tree with visual characteristics, from which the data is extracted. VIPS uses page layout features to form the corresponding semantics block, and the correlation threshold value is calculated for the connected semantic blocks to determine whether to merge or split them. Based on VIPS, Liu [6] put forward VIDE, using the visual information of the web page, including the position of the feature page layout features, appearance, features and content features for extracting data to solve the problem that VIPS relies too much on the web page. Due to the diversity of web forms, this method is not so adaptable. Furthermore, the dependence of VIDE on VIPS results in poor efficiency.

The ontology-based methods are listed as follows: The BYU [7] system designed by David is directly dependent on the data in the Web page without relying on the structure of the page or the DOM tree, using ontology to locate the page constant, and construct objects for them. Vijayarajan [8] proposed the image ontology model, which improves image retrieval by extracting the O-A-V triple from the user query and then compares with the description of the image stored in the ontology. Ontology-based data extraction method is less affected by the differences brought about by the web pages, but is greatly determined by the correctness and completeness of the ontology database.

Some researchers combine NLP with data extraction. Califf [9] put forward the RAPIER method which considers the whole page as plain text and sets some template to extract the information. Freitag [10] put forward SRV, which applied natural language processing and machine learning to web data extraction. It can be used to extract semi-structured data in the Deep Web and solves the problem that if the data do not follow the stored template format, then it cannot be extracted. Soderland [11] put forward WHISK, by improving training algorithm that allows for free-form text decimated data, which can be used in a wider range. Rafiei [12] proposed a method for extracting data according to semantic similarity, which improves the extraction accuracy but requires corpus and dictionary. The use of natural language processing (NLP) for data extraction ignores the structural features of Deep Web pages and is suitable for handling web pages that contains large amount of text.

Other researchers tried to extract data using machine learning. Zhou [13] and Lafferty [14] respectively used Hidden Markov Model and CRF to extract Web page information. The first one makes use of Markov random fields in the hierarchical tree structure to accomplish the task of data extraction and data unit recording while the second one trains a network to maximize its output conditional probabilities for extracting information. These methods have a good statistical basis. Moreover, there're a lot of robust training algorithms available, but they require a lot of training data.

Borromeo [15] revealed that quality of unpaid crowdsourcing is significantly higher than paid crowd work, while the completion time is remarkably lower. Human can do better than machines in data extraction, but requires much more time.

The aforementioned methods can extract data from web pages. Nevertheless, they either need too much training data or work without high precision or efficiency. To overcome these problems, we proposed a vision-based method VBF. We will present the details in next section.

2 A Vision-Based Approach for Deep Web Form Extraction VBF

The data in the Deep Web is often presented in tabular form, so the form recognition algorithm and the form matching algorithm is especially important. We try to combine the DOM tree and the convolutional neural network together and then find out the form in the Web page. In that way, we proposed a vision-based method VBF.

2.1 Data Extraction Algorithms

The data in the Deep Web is often presented in tabular form, so the form recognition algorithms and form matching algorithm is especially important. We use the form recognition algorithm to extract data in the table in the Web page and the form matching algorithm to match the DOM element of the form for extracting information.

Algorithm 1. Form Recognition Algorithm

1. **Input:** The address of a Web page that contains a table **URI**.
2. **Execute Form Recognition**
   Get the picture of the web page using the **URI**
   Get the HTML source code using the **URI**
   Build the DOM tree using the HTML source
   Get the coordinate of the point related to the form in the web page
   Get the DOM element of the form using Algorithm 2
   Get the HTML source of the DOM element of the form
   Generate the Excel spreadsheet file **table** using the HTML source
3. **Output:** A Microsoft Excel spreadsheet file that contains a Web form **table.**

As shown in Algorithm 1, we get shots and the HTML source code of the entire page and then generate DOM tree of the web pages according to the source code. After that, we do some simple image processing, including normalization and thresholding, and pass the processed image to the convolutional neural network to obtain the table position in the picture. Finally, we use the form matching algorithm to get the DOM element of the form, transform the DOM element of the form into HTML source code and generate the corresponding Microsoft Excel file using the transformed HTML source code;

Algorithm 2. Form Matching Algorithm

1. **Input:** A DOM element associated with the form **DOM-point**.
2. **Execute Form Matching**
   Initialize a new queue **DOM-queue**
   Push **DOM-point** into **DOM-queue**
   While **DOM-queue** is not empty
   Pop up the **DOM-queue** and get the head element **DOM-current**
   Get the children nodes of **DOM-current** and store them into **DOM-list**
   For each **Dom-element** in **DOM-list**
   Push **DOM-element** into **DOM-queue**
   If **DOM-current** is table
   Return **DOM-current**
   Push **DOM-point** into **DOM-queue**
   While **DOM-queue** is not empty
   Pop up the **DOM-queue** and get the head element **DOM-current**
   Get the parent node of **DOM-current** and store it into **DOM-element**
   Push **DOM-element** into **DOM-queue**
   If **DOM-current** is table
   Return **DOM-current**
   Return **null**
3. **Output:** the DOM element of the form **DOM-table**.

As shown in Algorithm 2, we use BFS (Breadth First Search) to search for the DOM element. At first, we list all the children nodes and try to find the DOM element

of the form. If we cannot find the element, we will try to find that in all the parent nodes. Finally, if all the related nodes are searched and the required node still cannot be found, we will confirm that the node of the Web form cannot be found in the web page, otherwise the DOM element of the Web form will be returned.

3 Experiment Result

To verify the effectiveness of form extraction algorithm based on visual semantics, this paper combined with the needs of practical application projects "ocean sailing query". The timetable for the various information on the website has been crawled for experiments, and our method was compared with the MDR, DEPTA and VIDE algorithm. In the experiment, the convolutional neural network used to identify forms is designed as follows. The total number of the images of the table samples is 54000, respectively contain 200 pictures in which contains 280 kinds of forms with different locations and sizes. The size of the images is set to 128 × 128. 40 pages were randomly selected from the experimental details pages of Chinese Sea Timetable Web and Chinese Seaport Web for crawl experiment, time spent is calculated using the automatic page turning techniques. Figure 1 shows the results of the extraction.

The extraction average time spent extracting 40 pages are 4.11 s, 4.51 s and 9.76 s using MDR, DEPTA, VIDE, but our method only takes 1.52 s. As can be seen from the figure, the growth rate of extraction time of our method should be less than that

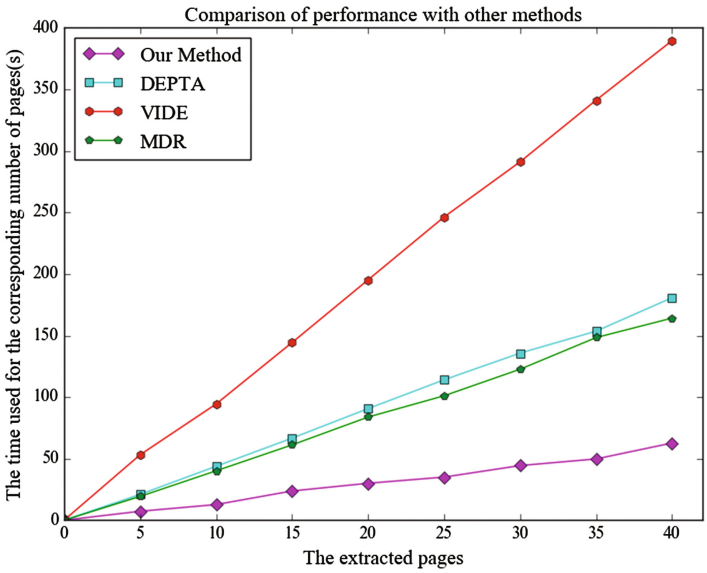


Fig. 1. The performance of data extraction. The horizontal axis represents the extracted pages, while the vertical axis represents the total extraction time used for the corresponding number of pages.

of DEPTA. This is because DEPTA only compares DOM tree in data extraction which requires a lot of computing. If combined with visual information of the Web page, it can reduce the amount of calculation and improve the extraction efficiency greatly.

We also did crawl experiments to test the accuracy and completeness of the proposed method. This article uses the following formula to measure performance: Accuracy, Recall and F-score. The accuracy rate is the proportion of the accurate extracted data record to all extracted record, and the recall rate refers to the proportion of extracted data records to all data records, while F-score is the harmonic mean of precision and recall rates. Table 1 shows the performance comparison of our method with VIDE which is reported to perform well in Web data extraction.

Table 1. Comparison of performance with other methods.

Method	Precision (%)	Recall (%)	F-score (%)
VIDE	99.4	94.8	97.0
Our method	100	97.1	98.5

4 Conclusion

In this paper, we proposed a method to automatically extract data form using vision-based information, which is characterized by the combination of visual information and the DOM node information. Compared with VIDE, it does not need to obtain visual information of all the nodes, so that the extraction efficiency greatly improved.

References

1. Crescenzi, V., Mecca, G., Merialdo, P.: Roadrunner: towards automatic data extraction from large web sites. In: Very Large Data Bases (VLDB), vol. 1, pp. 109–118 (2001)
2. Chang, C.H., Hsu, C.N., Lui, S.C.: Automatic information extraction from semi-structured web pages by pattern discovery. J. Decis. Support Syst. **35**(1), 129–147 (2003)
3. Liu, B., Grossman, R., Zhai, Y.: Mining data records in web pages. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 601–606. ACM (2003)
4. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: Proceedings of the 14th International Conference on World Wide Web, pp. 76–85. ACM (2005)
5. Cai, D., Yu, S., Wen, J.R., et al.: VIPS: a vision-based page segmentation algorithm. Microsoft Technical report, MSR-TR-2003-79 (2003)
6. Liu, W., Meng, X., Meng, W.: ViDE: a vision-based approach for deep web data extraction. IEEE Trans. Knowl. Data Eng. **22**(3), 447–460 (2009)
7. Embley, D.W., Campbell, D.M., Smith, R.D.: Ontology-based extraction and structuring of information from data-rich unstructured documents. In: Proceedings of the Seventh International Conference on Information and Knowledge Management, 2–7 November 1998

8. Vijayarajan, V., Dinakaran, M., Tejaswin, P., et al.: A generic framework for ontology-based information retrieval and image retrieval in web data. J. Hum. Centric Comput. Inf. Sci. **6**(1), 18 (2016)
9. Califf, M., Mooney, R.: Relational learning of pattern-match rules for information extraction. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, Florida, Orlando, pp. 328–334 (1999)
10. Freitag, D.: Machine learning for information extraction in informal domains. J. Mach. Learn. **39**(2–3), 169–202 (2000)
11. Soderland, S.: Learning information extraction rules for semi-structured and free text. J. Mach. Learn. **34**(1–3), 233–272 (1999)
12. Rafiei, M., Kardan, A.A.: A novel method for expert finding in online communities based on concept map and PageRank. J. Hum. Centric Comput. Inf. Sci. **5**(1), 1–18 (2015)
13. Zhou, S.X., Lin, Y.P., Wang, Y.N.: Text information extraction based on active hidden Markov model. J. Hunan Univ. (Nat. Sci.), 601–606 (2007)
14. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289 (2001)
15. Borromeo, R.M., Toyama, M.: An investigation of unpaid crowdsourcing. J. Hum. Centric Comput. Inf. Sci. **6**(1), 11 (2016)

Questions Classification with Attention Machine

Yunlu Liaozheng[1], Jin Liu[1(✉)], and Jin Wang[2]

[1] Shanghai Maritime University, Shanghai, China
{ylliaozheng, jinliu}@shmtu.edu.cn
[2] College of Information Engineering, Yangzhou University, Yangzhou, China
jinwang@yzu.edu.cn

Abstract. Due to the development of deep learning, word embedding has been introduced into nature language process. So, we tried using word embedding to simplify the information extraction of questions classification and take the advantage of big data. Additionally, with the advantages of attention machine, RNN in machine translation could consider the middle states to avoid the problem of bias on inputs. We introduce it into questions classification, and the experiment shows we get a little better performance than the best performance before.

Keywords: Questions-classification · Word-embedding · Attention

1 Introduction

Due to very fast growth of information in the last few decades, getting precise information in real time is becoming increasingly difficult. Search engines such as Google and Yahoo are helping in finding the information but the information provided by them are in the form of documents which consumes a lot of time of the user. Question Answering Systems have emerged as a good alternative to search engines where they produce the desired information in a very precise way in the real time. This saves a lot of time for the user. Question Categorization, is a useful technique in Web-based Question Answering system. Based on the questions, it will be associated to the corresponding category. Earlier approaches for the creation of automatic document classifiers consisted of manually building, by means of knowledge engineering techniques, an expert system capable of taking Document Categorization decisions. The major disadvantage of which was that it required rules manually defined by a knowledge engineer with the aid of a domain expert. To overcome the pitfalls associated with rule-based classification, Machine Learning techniques are currently applied for these purposes. In this approach set of pre-classified questions are fed to the classifier. Another approach that can be taken is context based interpretation. It takes advantage of tracking the contextual meaning of words and phrases during (and after) the development of ontology for that context, and subsequently uses this information as knowledge base for interpretation of free text sentences. The classifiers will be trained by set of training examples for each category, which are predefined. Hence forth the classifiers will be used to classify set of questions. In the recent year, an approach based

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_112

on word embedding was used in this task. It solves a big that the proportion of questions in a language is very limited. Although this problem could be partially solved by SVM, it's still an impediment on the road of performance. In this paper, we will discuss how does the information inside the word embedding solve this problem and what performance can word embedding lead to.

2 Related Works

A variety of researches in QA system has been done in the decade years. And in all the key facts to QA system, we choose to focus on classifying the categories of questions in QA. We think it can contribute to the improvement of the QA systems because it effectively limits the area of the answers.

Question classification in TREC QA has been studied during the past decade. It is to assign one or more categories to a given question written in natural language and the set of predefined categories are usually called question taxonomy or answer type taxonomy, such as "Location," "Human," and so on. Since typical Question Answering systems need to automatically generate the answers, they use the question classification to limit the area of the answers and use the corresponding methods to generate answers [1]. On the other hand, the focus of question classification in QA systems is to classify and recommend the category of the question.

Traditionally, various text classification techniques have been studied [2–6]. They present several statistical methods for text classification on two kinds of textual data, such as newspaper articles and e-mails. But unfortunately, question classification of the QA systems has two special aspects from traditional text classification; the input of question classification is short because it is a question and the corpus of questions is distinctly much smaller than common text.

Zhang [7] used question patterns and rule based question classification mechanism for the QA system related to learning Java programming. Xia [8] have built their question class taxonomy in Chinese cuisine domain and implemented rule based classifier as their primary classification approach. The system also achieved good accuracy within specific domain.

Zhang and Lee [9] performed several experiments on question classification using the same taxonomy as Li and Roth, as well as the same training and testing data. In an initial experiment, they compared different machine learning approaches with regards to the question classification problem: Nearest Neighbors (NN), Naive Bayes (NB), Decision Trees (DT), SNoW, and SVM. NN, NB, and DT are by now standard techniques and good descriptions of them can be found in for instance Mitchell (1997). The feature extracted and used as input to the machine learning algorithms in the initial experiment was bag-of-words and bag-of-n-grams (all continuous word sequences in the question). Questions were represented as binary feature vectors since the term frequency of each word or n gram in a question usually is 0 or 110. The results of the experiments show that SVM gets the best performance which is 85.5% accuracy.

Now, with the feature extraction, the SVM could get 95% accuracy. But the problems remain. First, with the development of network, big data become the most

useful tool to train model. But the corpus of questions is still very small, even tiny. Second, we need to find a more efficient way to do the unsupervised feature extraction.

The problems mentioned above are not only in the questions classification but also in the other NLP tasks. So, unsupervised compressing from bag-of-words into word embedding can improve not only the performance of questions classification but also can be introduced into many NLP related system such as web information search and analyze system [10–15], information processing system [16] and information retrieval and image retrieval in web data [17].

3 Bi-directional Recurrent Neural Networks with Attention

Recurrent neural networks have made a big success in NLP. Unlike feed forward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs. This makes them applicable to tasks such as unsegmented connected handwriting recognition or speech recognition.

It has a good performance about sequence problems. Different from other classification problems. One question contains many words orderly. Change the arrangement of words will significantly affect the result. So, it can be treated as a word sequence.

In this case, recurrent neural network is a good choice as it can store the sequence information in its hidden layer so that we can use the sequence information as feature.

For example, if we take the question "Who is Tom?" as an input, the hidden layer's input h(t) is combined with the input of input layer which is represented by word embedding E(x) and the last hidden layer's output h(t-1). The word "Who" is not only the input of x1, but also determine the result of y2 and y3, so the relationship is marked by dotted lines. The generalize structure of RNN is showing below in Fig. 1:

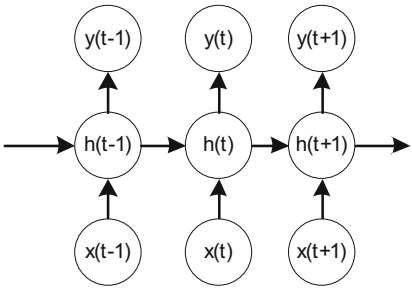


Fig. 1. RNN unfolded form

So, the t-1 input will be stored by parameters in hidden layer, and then still influence the hidden layer next time. This makes the words all stored in the hidden layer and form a context. Training the next word will take all the words before into account. But it still has a disadvantage. If the words before need to consider the next words, in another word, the dependency is reversed, it cannot solve this situation. To overcome this problem, we use bi-directional RNN to process the input sequence. Bi-directional RNN not only scan the input forward but also scan the input backward.

Generally, we expect the final output could contain the information through the whole sentences. That means the final output is the sentence embedding. We are forced to have a large hidden layer to contain this information and still lose some of it. This problem also appears in many other tasks such as machine translation but solved by attention machine recently. RNN with attention machine has a good performance on NMT. We also introduce the attention machine into questions classification to improve the performance on long sentences.

Instead of just use the last output, attention machine allows the classifier to access the intermediate information produced by every step. With an additional hidden layer and a sigmoid function connected to the intermediate state of RNN's hidden layer, the attention machine can identify which word contributes to the result most.

4 Experiment

4.1 Training

We use the training data set used in Li's learning question classification experiments [14]. The building work is supported by [15] (Fig. 2).

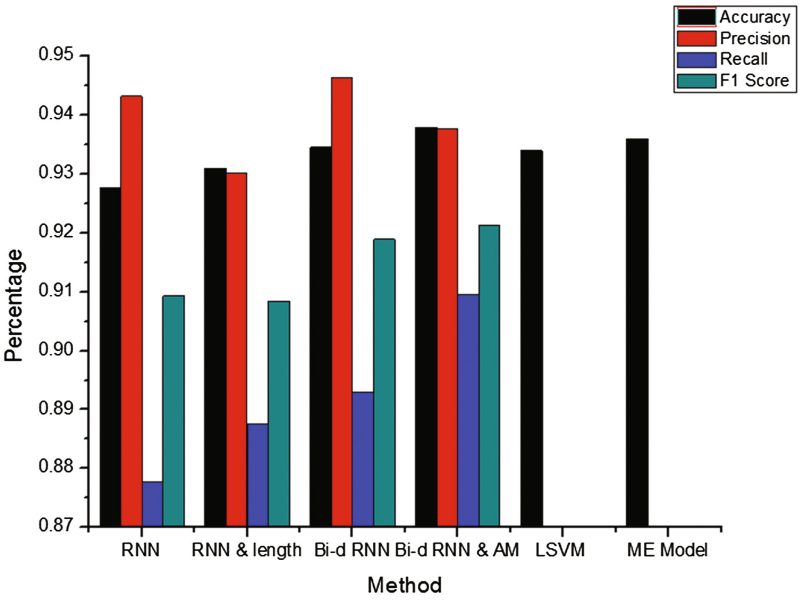


Fig. 2. Comparison

The test data set is the TREC 10 questions.

There are some words which are not contained in the vectors. So, we just exclude those simples (Table 1).

Table 1. Classification accuracy comparison

Description	Accuracy	Precision	Recall	F1 Score
Classification using RNN	0.9276	0.9432	0.8778	0.9093
RNN with length	0.931	0.9302	0.8876	0.9084
Bi-directional RNN	0.9345	**0.9463**	0.893	0.9189
Bi-directional RNN with attention machine	**0.9379**	0.9377	**0.9095**	**0.9214**
Linear SVM	0.934			
Maximum Entropy Model	0.936			

The whole experiment is based on the following configuration:

Weight initializer: Xavier Initializer
Regularization: Dropout 0.5
Optimizer: RMSProp Optimizer
Learning rate: Begin with 0.0005 then multiply 0.98 every 50 steps.

4.2 Result

The experiment shows that with the attention machine, the performance of bi-directional RNN is a little better than Maximum Entropy Model, and the F1-score is also the highest.

Another difference is, if we use clustering approach to classify the questions automatically, the segmented areas contain not only questions but also normal sentences. Then we go back to the purpose of questions classification, we classify the questions to fit them with different ways of answering generation. So, the language model makes the questions close to its similar sentences which are more likely to be related to the answers.

5 Expectation

Beyond the Questions
How does human classify the questions? The questions' classes are related to their corresponding answers. We give a question a "HUMAN" label, means we expect the answer is about a human.

If we can not only be limited by questions, but look for the similar context in the answers. We might could use answers to help classify the questions.

You may not omit references. Instructions as to where to find a fuller version of the references are not permissible.

References

1. Loni, B.: A survey of state-of-the-art methods on question classification. Technical rep., Delft University of Technology, pp. 1–40 (2011)
2. Prasad, R., Natarajan, P., Subramanian, K., Saleem, S., Schwartz, R.: Finding structure in noisy text: topic classification and unsupervised clustering. Int. J. Doc. Anal. Recogn. **10**(3–4), 187–198 (2007)
3. Liu, D., McVeety, S., Prasad, R., Natarajan, P.: Semi-supervised topic classification for low resource languages. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5093–5096 (2008)
4. Bracewell, D.B., Yan, J., Ren, F., Kuroiwa, S.: Category classification and topic discovery of japanese and english news articles. Electron. Notes Theoret. Comput. Sci. **225**(2), 51–65 (2009)
5. Zhang, P., Wu, C., Wang, C., Huang, X.: Personalized question answering system based on ontology and semantic web. In: IEEE International Conference on Industrial Informatics, pp. 1046–1051 (2006)
6. Xia, L., Teng, Z., Ren, F.: Question classification for Chinese cuisine question answering system. IEE J. Trans. Electrical Electron. Eng. **4**(6), 689–695 (2009)
7. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of the 26th Annual International ACM (2003)
8. Xu, W., Rudnicky, A.: Can artificial neural networks learn language models. In: Sixth International Conference on Spoken Language Processing (2000)
9. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. In: Advances in Neural Information Processing Systems, pp. 932–938 (2001)
10. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
11. Niu, Y., Junjie, C,. Liguo, D., Wei, Z.: Study on classification features of Chinese interrogatives. Comput. Appl. Softw. **29**(3) (2012)
12. Li, X., Roth, D.: Learning question classifiers. In: COLING 2002, (2002)
13. Research supported by NSF grants IIS-9801638 and ITR IIS-0085836 and an ONR MURI Award
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Microsoft Research, 770–778 (2016)
15. Shtykh, R.Y., Jin, Q.: A human-centric integrated approach to web information search and sharing. Hum.-Centric Comput. Inf. Sci. **1**(2), 1–37 (2011)
16. Borromeo, R.M., Toyama, M.: An investigation of unpaid crowdsourcing. Hum.-Centric Comput. Inf. Sci. **6**(11), 1-19 (2016)
17. Vijayarajan, V., Dinakaran, M., Tejaswin, P., Lohani, M.: A generic framework for ontology-based information retrieval and image retrieval in web data. Hum.-Centric Comput. Inf. Sci. **6**(18), 1–30 (2016)

Development of Unidirectional Security Gateway Satisfying Security Functional Requirements

Seon-Gyoung Sohn, Jungchan Na, and Kyung-Soo Lim(✉)

Hyper-Connected Communication Research Laboratory,
System Security Research Group, Electronics and Telecommunications Research Institute, Daejeon, Korea
{sgsohn,njc,lukelim}@etri.re.kr

Abstract. A connection between an industrial control network and IT network can expose measurement equipment, control systems and important infrastructure components to various cyber-attacks. Many technologies have been proposed to protect industrial control networks against cyber-attacks and to provide confidentiality, integrity, and availability. Among the technologies, a physical unidirectional security gateway provides protection of critical systems by forcing unidirectional communication between the two networks. The unidirectional security gateway needs to provide safety and reliability, and to guarantee, the common criteria for information technology security evaluation is operated. In this paper, we propose a unidirectional security gateway satisfying security functional requirements derived from CC v3.1.

Keywords: Unidirectional security gateway · One-way data transmission · Common criteria · Security requirement

1 Introduction

A unidirectional security gateway [1] is used for secure communication between networks separated by different security levels. The unidirectional security gateway is a device that transfers data from a high security level area such as a control network to a low security level area while blocking the inflow of data from the outside to maintain security of the secure area. Since the device transmits data only in one direction, it cannot guarantee the transmission reliability provided by a bidirectional protocol such as TCP. Therefore, various mechanisms for ensuring data transmission reliability was considered [2–4].

Because the unidirectional security gateway is used in key infrastructure, it will comply with strict Critical Infrastructure Protection standards or regulations established by NRC, NIST, and NIAP. The Common Criteria for Information Technology Security Evaluation (abbreviated as Common Criteria or CC) is an representative international standard (ISO/IEC 15408) for computer security certification. CC provides assurance that the process of specification, implementation and evaluation of a computer security product has been conducted in a rigorous and standard and repeatable manner at a level

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_113

that is commensurate with the target environment for use. CC is used as the basis for a government driven certification scheme and typically evaluations are conducted for the use of Federal Government agencies and critical infrastructure. Many unidirectional security gateway products such as Fort Fox Hardware Data Diode, Waterfall Unidirectional Security Gateway, OWL DualDiode, and BAE System Data Diode have acquired CC certification [5–8].

In this paper, we choose the security functional requirements (SFRs) provided by CC as a method to guarantee the safety and reliability of unidirectional security gateway and describe the structure and functions of the system developed to satisfy the derived requirements.

2 Related Work

2.1 Waterfall Unidirectional Security Gateway WF-400 Security Target

Waterfall Unidirectional Security Gateway is a product that protects control system and operation network against attacks from external network in industrial network environment. Waterfall Unidirectional Security Gateway WF-400 1 Gbps product has been awarded an EAL 4+ certification under a CC security evaluation. According to the relevant Security Target [9], the SFRs for WF-400 consist of the following components from CC Part 2, summarized in Table 1.

Table 1. SFRs for WF-400

Security function class	Security functional component	
User data protection (FDP)	FDP_IFC.2	Complete Information Flow Control
	FDP_IFF.1	Simple security attributes

2.2 Owl Dual Diode Communication Cards Security Target

OWL DualDiode product is a unidirectional system hardware for unidirectional transmission. OWL DualDiode Communication Cards 10G, 2.5G, and 1.0G v.7 & v.7t models [10] certified of EAL 2. The Table 2 describes the SFRs that are satisfied by DualDiode Communication Card.

Table 2. SFRs for DualDiode

Security function class	Security functional component	
User data protection (FDP)	FDP_IFC.2	Complete Information Flow Control
	FDP_IFF.1	Simple security attributes
	FDP_IFF.5	No illicit information flows
Protection of the TSF (FPT)	FPT_FLS.1	Failure with preservation of secure state
	FPT_PHP.1	Passive detection of physical attack

2.3 SFRs for Unidirectional Security Gateway

Since no Protection Profile (PP) was introduced for developing and evaluating unidirectional security gateway products, a PP based on the CC V3.1 was proposed to help development and evaluation of a unidirectional security gateway. The SFRs in the proposed PP consist of 8 security function classes, as follows;

- Security audit (FAU): FAU_ARP.1, FAU_GEN.1, FAU_GEN.2, FAU_SAA.1, FAU_SAR.1, FAU_SAR.3, FAU_SEL.1, FAU_STG.1, FAU_STG.3, FAU_STG.4
- Cryptographic support (FCS): FCS_CKM.1, FCS_CKM.2, FCS_CKM.4, FCS_COP.1
- User data protection (FDP): FDP_ACC.1, FDP_ACF.1, FDP_IFC.1, FDP_IFF.1, FDP_ITT.1
- Identification and authentication (FIA): FIA_AFL.1, FIA_ATD.1, FIA_SOS.1, FIA_UAU.2, FIA_UAU.4, FIA_UAU.7, FIA_UID.2
- Security management (FMT): FMT_MOF.1, FMT_MSA.1, FMT_MSA.3, FMT_MTD.1, FMT_MTD.2, FMT_SMF.1, FMT_SMR.1
- Protection of the TSF (FPT): FPT_ITT.1, FPT_SMT.1, FPT_TST.1
- TOE Access (FTA): FTA_SSL.3
- Trusted Path/Channel (FTP): FTP_ITC.1

3 Design of Unidirectional Security Gateway

We selected SFRs from [11] to develop a secure and reliable unidirectional security gateway. The selected SFRs are shown in Table 3.

Table 3. Selected SFRs for our system

Security function class	Security functional component	Elements
FAU	FAU_ARP.1	FAU_ARP.1.1
	FAU_GEN.1	FAU_GEN.1.1, FAU_GEN.1.2
	FAU_SAA.1	FAU_SAA.1.1, FAU_SAA.1.2
	FAU_SAR.1	FAU_SAR.1.1, FAU_SAR.1.2
	FAU_SAR.3	FAU_SAR.3.1
	FAU_STG.1	FAU_STG.1.1, FAU_STG.1.2
	FAU_STG.3	FAU_STG.3.1
	FAU_STG.4	FAU_STG.4.1
FCS	FCS_COP.1	FCS_COP.1.1
FDP	FDP_IFC.2	FDP_IFC.2.1, FDP_IFC.2.2
	FDP_IFF.1	FDP_IFF.1.1, FDP_IFF.1.2, FDP_IFF.1.5
FIA	FIA_AFL.1	FIA_AFL.1.1, FIA_AFL.1.2
	FIA_SOS.1	FIA_SOS.1.1

(*continued*)

Table 3. (*continued*)

Security function class	Security functional component	Elements
	FIA_UAU.2	FIA_UAU.2.1
	FIA_UAU.4	FIA_UAU.4.1
	FIA_UAU.7	FIA_UAU.7.1
	FIA_UID.2	FIA_UID.2.1
FMT	FMT_MOF.1	FMT_MOF.1.1
	FMT_MSA.1	FMT_MSA.1.1
	FMT_MSA.3	FMT_MSA.3.1, FMT_MSA.3.2
	FMT_MTD.1	FMT_MTD.1.1
	FMT_PWD.1	FMT_PWD.1.1, FMT_PWD.1.2, FMT_PWD.1.3
	FMT_SMF.1	FMT_SMF.1.1
	FMT_SMR.1	FMT_SMR.1.1, FMT_SMR.1.2
FPT	FPT_PST.1	FPT_PST.1.1
	FPT_TEE.1	FPT_TEE.1.1, FPT_TEE.1.2
	FPT_TST.1	FPT_TST.1.1, FPT_TST.1.2, FPT_TST.1.3
	FPT_ITT.1	FPT_ITT.1.1
FTA	FTA_SSL.5	FTA_SSL.5.1

The selected SFRs can be categorized into requirements for unidirectional data transmission and requirements for maintaining the reliability of the system. The requirements for unidirectional data transfer include the User Data Protection Class, the requirements for system reliability include Security Audit Class, Cryptographic Support Class, Identification and Authentication Class, Security Management Class, TSF Protection Class, and TOE Access Class.

We developed a unidirectional security gateway, UNIWAY, which satisfies the selected SFRs. The functional structure of the system is shown in the Fig. 1.

The Unidirectional Transfer Proxy block implements the FDP_IFC.2, FDP_IFF.1, and FAU_GEN.1 SFRs. The block performs following functions.

- Transfer data from a sending host in the secure area to a receiving host in the non-secure area
- Protocol connection only to the service that the administrator set as “Permit”
- IP based access control
- Audit generation for unidirectional data processing

Unidirectional Transfer Middleware block realizes FDP_IFC.2, FDP_IFF.1, FAU_GEN.1 SFRs and performs common functions related to message transmission between Unidirectional Transfer Proxy block and Unidirectional Network Interface block.

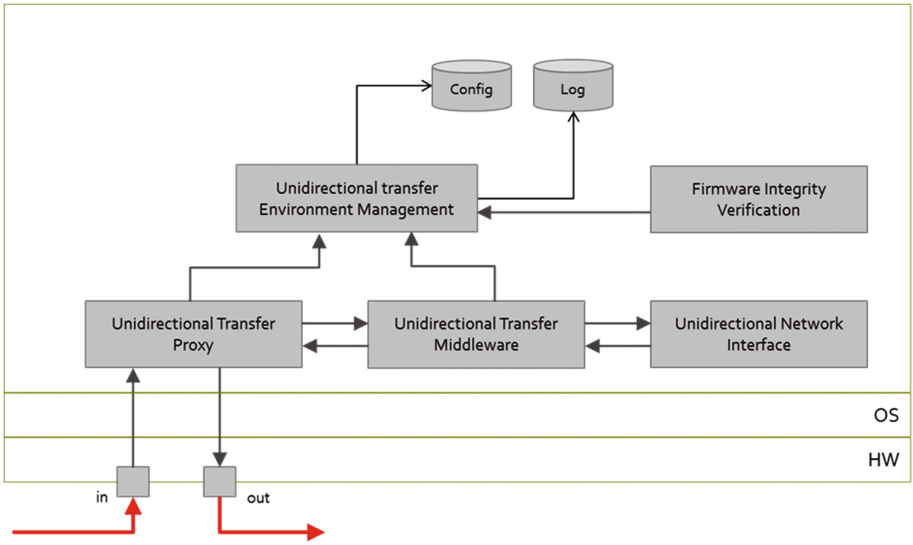


Fig. 1. UNIWAY functional structure

Unidirectional Network Interface implements FDP_IFC.2, FDP_IFF.1 SFRs. It receives data from Unidirectional Transfer Middleware of TX-only UNIWAY and transmits data to Unidirectional Transfer Middleware of RX-only UNIWAY.

Firmware Integrity Verification implements FPT_TEE.1, FPT_TST.1, and FAU_GEN.1 SFRs. The block performs static integrity verification during booting of UNIWAY system and dynamic integrity verification for abnormal access during normal system operation.

Unidirectional transfer Environment Management implements the SFRs of FAU_ARP.1, FAU_GEN.1, FAU_SAA.1, FAU_SAR.1, FAU_SAR.3, FAU_STG.1, FAU_STG.3, FAU_STG.4, FCS_COP.1, FIA_AFL.1, FIA_SOS.1, FIA_UAU.2, FIA_UAU.4, FIA_UAU.7, FMT_MOF.1, FMT_MSA.1, FMT_MSA.3, FMT_MTD.1, FMT_PWD.1, FMT_SMF.1, FMT_SMR.1, FPT_PST.1, FPT_ITT.1 SFRs. This function block performs following functions.

- Audit reception, storage, search and audit storage management
- Generate security threat detection rules (accumulation or combination of audit records) and detect potential security violations according to detection rules
- Administrator management (account registration and identification/authentication)
- Security management (configuration management, security policy management, process setting, etc.).

4 Conclusions

In this paper, we described a unidirectional security gateway, UNIWAY, satisfying SFRs based on CC v3.1. UNIWAY not only provides unidirectional data transmission function but also assures safety and security, we expected that UNIWAY can be utilized as a network separation and a network connection technology between the control network and the IT network of the main infrastructure.

Acknowledgments. This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2016 [R0126-15-1095, Unidirectional Security Gateways developments in cyber-physical systems].

References

1. Barker, R.T., Cheese, C.J.: The application of data diodes for securely connecting nuclear power plant safety systems to the corporate it network. In: 7th IET International Conference on System Safety, incorporating the Cyber Security Conference, IET (2012)
2. Arkhangelskii, V., et al.: Secure one-way data transfer. In: 2016 IEEE NW Russia Young Researchers in Electrical and Electronic Engineering Conference (EIConRusNW). IEEE (2016)
3. Rogowski, D.: Software support for common criteria security development process on the example of a data diode. In: Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, June 30–July 4 2014
4. Common Criteria. https://www.commoncriteriaportal.org
5. Fort Fox Hardware Data Diode. https://www.fox-it.com/en/products/datadiode
6. Waterfall Security Solutions USA. http://www.waterfallsecurity.com/technology
7. Owl Computing Technologies. http://www.owlcti.com/dualdiode_technology.html
8. BAE System Data Diode. https://www.ia.nato.int/niapc/Product/BAE-System-Data-Diode—EAL-7_168
9. Waterfall: Waterfall Unidirectional Security Gateway WF-400 Security Target. V0.72 (2012)
10. OWL: DualDiode Communication Cards 10G, 2.5G, 1.0G v.7 & v.7t Models Security Target. (2014)
11. Lee, H.-J., Won, D.: Protection profile for unidirectional security gateway between networks. Int. J. Secur. Appl. **7**(6), 373–384 (2013)

Context Based Multiple Players Identification in Sports Images

MiYoung Nam(✉), Jun-Young Lee, Sung Youb Kim, and YoungGiu Jung

YM-Naeultech, Inharo, Namgu, Incheon, Republic of Korea
{ym-nam, sykim, youngq.jung}@ym-naeultech.com

Abstract. In this paper we propose to develop technology for recognizing Jersey number of each athlete in real time to do sports video indexing in multi-player game. More than two persons are often detected as one object due to frequent collisions between persons in sports. This system recognizes the jersey number of more than one persons based on learnt context information using texts and uniform color information. The resulting system is able to achieve a jersey number recognition accuracy up to 97% on video and photography.

Keywords: Jersey number recognition · Color context · SVM

1 Introduction

Sports video indexing technology has been researched in various meaning levels. Mainly, there are shot level [1], event level [2], and action level [3]. Recognizing individual player is a very important technology in these various video indexing technologies, and it is used to provide various real time broadcasting services if various broadcasting information are combined [4–8].

This paper is to develop the technology which recognizes Jersey number in real time in the detection area of actually playing athlete and does video tagging by relating this information with existing registered athlete. The athlete is very fast which makes adjacent or overlapping conditions occur frequently. This paper makes it possible to discriminate multiple objects and recognize the Jersey number accurately by extracting uniform and Jersey number in the area of complex objects after learning representative color of uniform and Jersey number.

2 Context Based Jersey Number Recognition Algorithm

The proposing team and Jersey number recognition system is as Fig. 1.

It is composed of algorithm which recognizes team by using color information, module which detects Jersey number area, and algorithm which recognizes Jersey number based on Jersey number detection results.

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_114

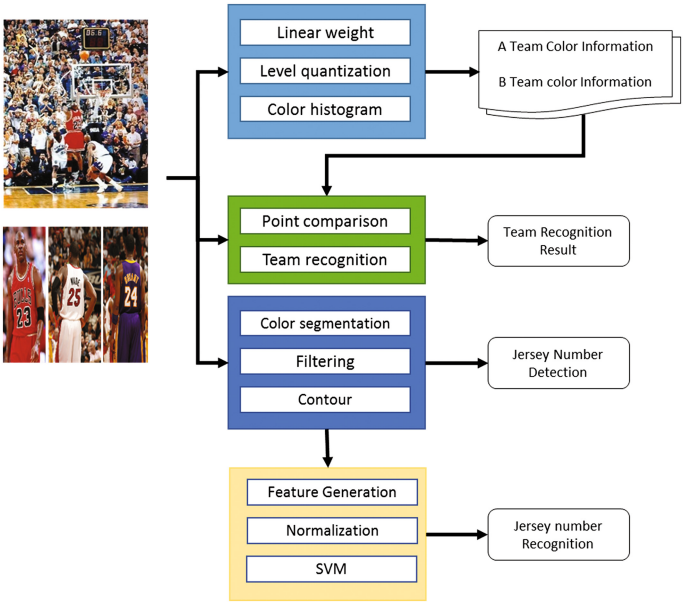


Fig. 1. Team recognition/Jersey team number recognition system architecture

2.1 Color Based Team Recognition Algorithm

In sports indexing system, tagging each athlete's team information automatically is one of the very important functions for various broadcasting services. We develop algorithm for identifying teams automatically using quantized color information. Figure 3 shows the algorithm for team identification.

Team identification is conducted by learning and recognition process. Learning is the process of extracting representative colors for each team. The recognition is based on color value threshold using representative color of each team. The figure below shows the quantized RBG image of input image.

2.2 Context Color Learning Algorithm for Multiple

The learning module for classifying and recognizing Jersey number using context color information in the area of complex objects is as follows: (Quantize and save team Jersey number and Jersey RGB color information) (Fig. 2).

In case of Jersey number color information, extract the color histogram of the number area. Use the histogram value of highest location as representative value and save the uniform color information as histogram range.

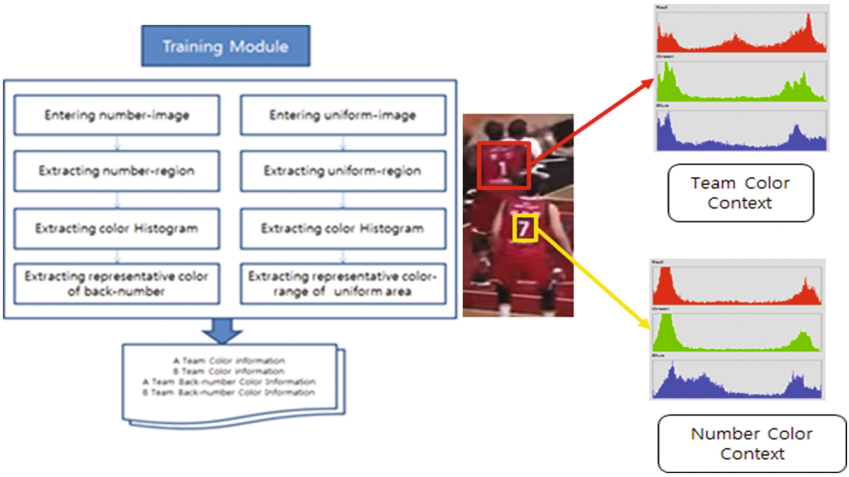


Fig. 2. Extracting and learning color information

2.3 Detection of Jersey Number Using Uniform Color Information

We extract the uniform area in the standard of context color value saved when learning in test videos and then extract the number areas. Jersey number can be detected by using color connection regardless of number of detected objects.

In case of number extraction, we investigate the Euclidean distance of each pixel of representative color information and test video, then extract the area by marking the location of which distance is under particular value.

The methodology of uniform extraction is similar to number extraction method. However, the uniform color can become brighter or darker by the lighting which makes the pixel value include the large range of the uniform colors. Therefore three representative values are used.

The method obtains the standard deviation of pixel values with the pixel of which the value after histogram extraction of uniform area is maximum in the range of ± 64 pixel ranges.

$$\text{Min} = \text{Pixel of which histogram is maximum} - \text{Standard deviation}$$
$$\text{Max} = \text{Pixel of which histogram is maximum} + \text{Standard deviation}$$

The three representative values are as follows:

$$\text{Value1} = (\text{min} + \text{max})/2$$
$$\text{Value2} = (\text{min} + \text{Value1})/2$$
$$\text{Value3} = (\text{max} + \text{Value1})/2$$

The smallest value is selected by calculating the similarity of pixel values and the representative three values. If the selected value is under particular value, the location is marked and extracts the uniform area.

2.4 Statistics Model Based Jersey Number Recognition Algorithm

We developed highly efficient recognition system with the characteristics of contour descriptor using the non-linear classifier, SVM. Jersey recognition module makes the detected number area into feature vector number and learns the number using the SVM, then performs the recognition process.

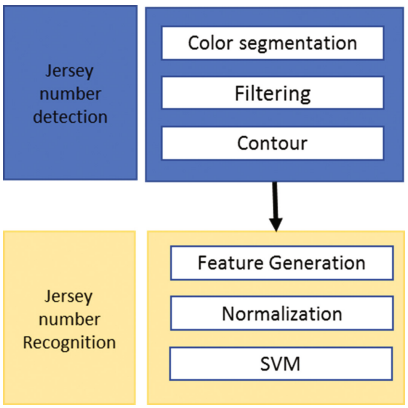


Fig. 3. Jersey number detection and jersey number recognition

3 Results

This system is experimented using the actually filmed broadcasting videos or games. The following shows the result of detecting three Jersey numbers of three persons in the area of one object.

F-measure	0.915
Precision	0.974
Recall	0.863

4 Conclusion

This paper introduces the methodology for providing information in sports highlight videos. We tried to solve the athletes classification problem due to overlapping phenomenon occurred by fast movements of sports video. Jersey number recognition is proved to be possible even when several persons are assembled using color context information as the result of the experiment.

Acknowledgment. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (R0101-16-293, Development of Object-based Knowledge Convergence Service Platform using Image Recognition in Broadcasting Contents).

References

1. Saric, M., Dujmic, H., Papic, V., Rozic, N.: Player number localization and recognition in soccer video using HSV color space and internal contours. In: World Academy of Science, Engineering and Technology, vol. 43 (2008)
2. Ozbay, S., Ereelebi, E.: Automatic vehicle identification by plate recognition. In: World Academy of Science, Engineering and Technology, vol. 9 (2005)
3. Lu, C.-W., Lin, C.-Y., Hsu, C.-Y., Weng, M.-F., Kang, L.-W., Liao, H.-Y.M.: Identification and tracking of players in sport videos. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, ICIMCS 2013, p. 113. ACM, New York (2013)
4. Idan, B.-A., Basha, T., Avidan, S.: Racing bib number recognition. In: British Machine Vision Conference (BMVC) (2012)
5. Ephrati, J., Moghadam, N.: Basketball player identification by Jersey and number recognition (2002)
6. Gerke, S., Muller, K., Schafer, R.: Soccer Jersey number recognition using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2015)
7. Šarić, M., et al.: Player number localization and recognition in soccer video using HSV color space and internal contours. In: The International Conference on Signal and Image Processing (ICSIP 2008) (2008)
8. Chen, R., Ho, A., Hong, J.: Jersey Number Recognition (2014)

A Study of the Factors Influencing Information Security Policy Compliance: Focusing on the Case of 'K' Institute

Gwang-il Ju[1(✉)], Jinhyung Park[1], Wanyoung Heo[1], Joon-Min Gil[2], and Hark-Soo Park[1]

[1] Department of Information Security Service, Korea Institute of Science and Technology Information (KISTI), Daejeon, South Korea
{kiju, ntoskr, wyheo, hspark}@kisti.re.kr
[2] School of Information Technology Engineering, Catholic University of Daegu, Gyeongsan, South Korea
jmgil@cu.ac.kr

Abstract. A variety of information security solutions are being developed to enhance user security from advanced and enhanced security threats. Highlight the aspects of end-user security as the most important aspect of information security. Based on the factors influencing the user's behavior, this study focuses on the impact of information on the impact of information on the time required by users (Higgins 1996). The purpose of this study is to understand the causes of information security compliance and establish effective security systems to address information security policies in the information security section of the Information Security Administration. In particular, it seeks to analyze the security compliance issues and measure the outcome of the survey by focusing on visual factors and opportunities in terms of users' voluntary improvement. To do this, consider influencing factors that affect user decision making based on self-determination theory (SDT).

Keywords: Information security · Compliance · Security framework · Security policy indicator

1 Introduction

The basic departure of information security is based on compliance with employee security policies. The intention of information security policy compliance has been studied in diverse fields. This study has attempted to investigate the case of 'K' Institute, focusing on the influence of self-determination. It proposes the importance of staff training and paradigm shift in that self-directed decision could have influence on the results among the employees' security perceptions.

The conventional studies on information security policies have discussed employees' compliance with information security [1]. In addition, they paid attention to the use of basic sanctions at the violation of security policies [2]. These sanction theories were partially successful, but they were limited in being applied to individuals' violation of information security as latent factors [3]. Furthermore, users usually form a

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_115

favorable attitude toward a product if it has a lot of good attributes or functions, while if it has few good attributes and functions, it forms a relatively less favorable attitude. As a study on user behavior, it is influenced by the information provided when a decision is made [4].

It is confirmed that the self-motivation of information security policy compliance has a positive effect on compliance intention. In addition, the intention of information security policy compliance was high among the employees who were willing to observe the organizational system dutifully and act in a way of benefiting the organization [5]. Self-determination theory (SDT) is a theory which puts emphasis on intrinsic motivation. It is defined as self-determination right on personal information in privacy [6]. This study focuses on investigating how to measure improvement effects through the SDT-based security policy and how to apply it to the security system in a public sector.

2 Information Security Policy Indicators

2.1 Understanding of Risk Management and Information Security

Recently, security policy starts from the concept that important assets are protected from possible threats. Based on the information security management system (ISMS) titled 'BS7799' completed by the BSI in the United Kingdom in 1999 and 'ISO27001' developed by the ISO in 2005, Korea has launched the K-ISMS and applied to public agencies [7, 8]. According to the definition of the risk of 'ISO17799,' risk refers to a possibility that external threats can cause damage on assets by taking advantage of internal vulnerabilities (Fig. 1).

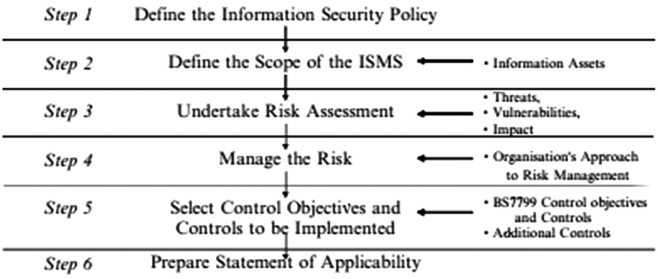


Fig. 1. The BS7799 process model

As stated above, risk starts with the concept of how much the internal assets can be protected from internal weakness and external threats. For this, the development of diverse security systems has originated from the basic principle of information security, which is called 'risk management.' In other words, risk represents a series of processes protecting assets from threats by establishing security policies.

2.2 National Information Security Management System

It's been explained that a national information security management system is structured for the National Intelligence Service (NIS) to plan and coordinate national information security from the perspective of information security governance. [9] Regarding security monitoring-centered cyber incident response duties, in addition, it has a response system against its affiliates and agencies through the cyber response center in each authority. Governmental and public agencies have enacted and operated information security guidelines customized for their circumstances based on the NIS' national information security guidelines. However, the national information security management system managed by the NIS fulfills its duties in a closed and secret manner. In other words, it has not been properly operated (Fig. 2).

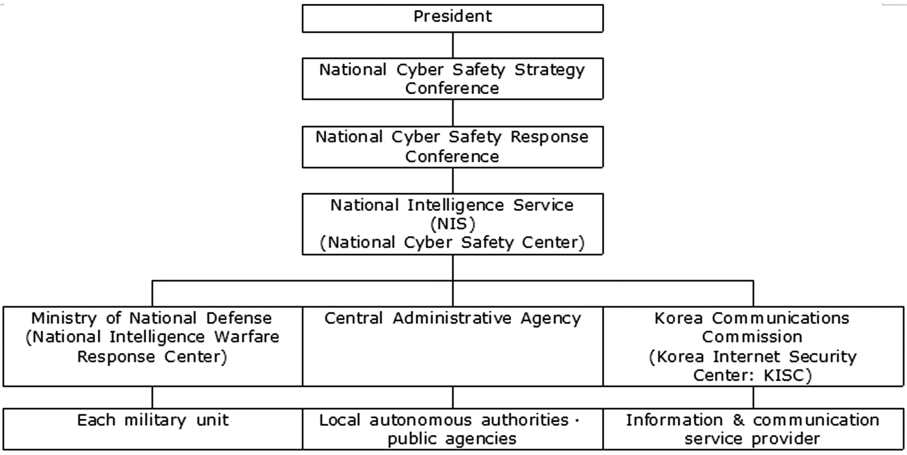


Fig. 2. National information security management

2.3 'K' Institute's Information Security System

'K' Institute has provided more than 100 different types of information services in national science & technology through the Internet. It has established and implemented a security system to be prepared against attacks from diverse websites in cooperation with central bureaus. Based on this system, the agency has also provided information security supports to its research affiliates. In addition, it has provided security monitoring services to 25 agencies in science & technology and R&D services after establishing domestic and international cooperation systems with the NIS and central bureaus.

'K' Institute's security management system can be divided by level as shown in Fig. 3 below. Starting from 'Level IV' which would be available as the records as solid proof which meets security management requirements, it is classified into Levels II and III which include the guidelines & procedures after passing through the manager's approval procedure in the agency which adopts the government's information security

policy and work instructions & checklist. Furthermore, security policy which reflects internal information security policies according to changes in the government's information security and related laws is put as the highest level which requires the CEO's approval.

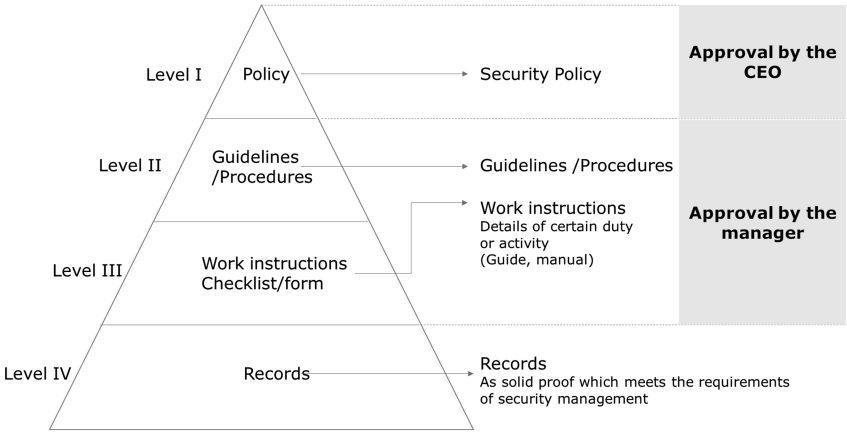


Fig. 3. Information security management frame (process)

2.4 Reinforcement of User Information Security Policies

Recently, 'K' Institute attempted to change two aspects relating to final users for the purpose of strengthening information security systems. First, it targeted to unify user security-related management points with integrated portal services (Fig. 4).

Second, there has been a shift in the system regarding various security contacts which have been applied to user security from the centralized management to the system in which users make a decision on their own and check the results in person. Therefore, there has been the improvement of efficiency in security management, and overall security management has been enabled through the improvement of existing security loopholes (Fig. 5).

It was attempted to analyze diverse security-related performances from security management based on major information security activities by comparing major indicators by the point of time.

2.5 Basic Information Security Policies

'K' Institute has implemented security management in two aspects: administrative security and technical security. In administrative security, security training has been given to all employees twice a year, and separate education has been offered to managers such as service programs for the purpose of improving employees' security awareness. In technical security, on the contrary, the prevention of accidents by checking the weakness of web services and security monitoring including response to

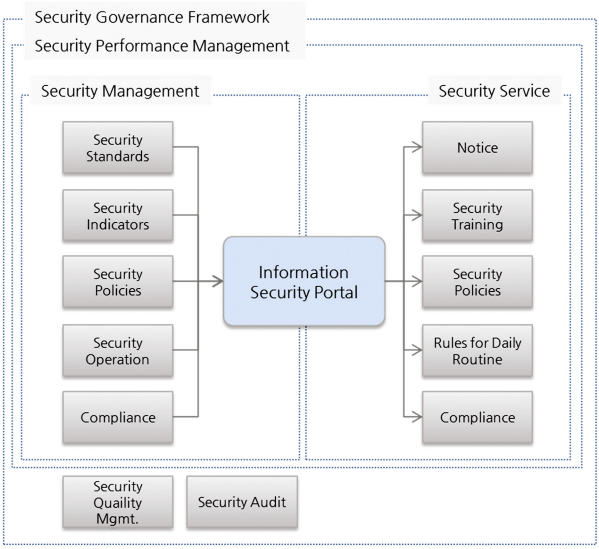


Fig. 4. Information security governance

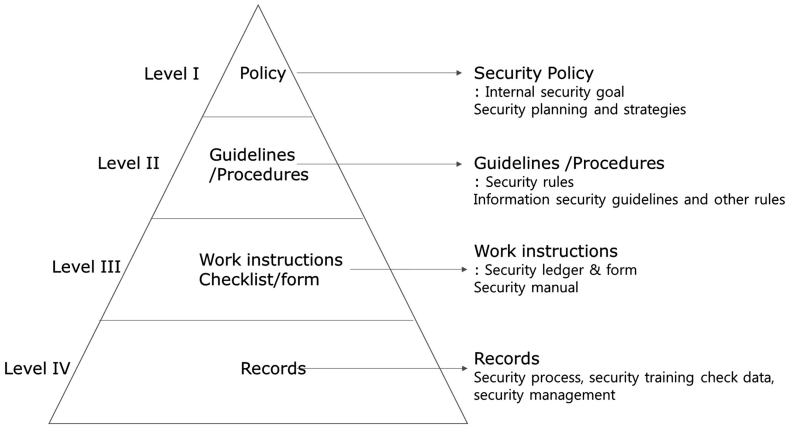


Fig. 5. Information security management frame (task)

security breaches by employees have been supported together with Science & Technology Security Control Center. Furthermore, each department's information security policies have been inspected and analyzed monthly on Cyber Security Assessment Day when employees' PCs are inspected (Table 1).

Table 1. Security inspection solutions

Category	Previous	Now
Manufacturer	'N'	'C'
Year introduced	2009	2015
Feature	Centralized	User check

2.6 'K' Institute's Information Security Inspection Policies

'K' Institute implemented Cyber Security Assessment Day targeted to assess each department's information security level every month and inspected 10 security indicators. For this inspection, the NIS' security indicators were referred to. In terms of setting, the indicators were adjusted according to internal circumstances (Table 2).

Table 2. Information security indicators by category

No.	Description
1	Windows login account password
2	Screensaver password
3	Time of screensaver activation (min)
4	Vaccine installed
5	Firewalls set
6	Shared folder set
7	Shared folder password
8	Windows security update
9	Local system set
10	Conditions of the unused ActiveX

'K' Institute developed and applied a new security inspection solution in 2015 to apply the user security reinforcement policy. The new security inspection solution has three characteristics as follows:

First, there was a shift in the structure from the simple search of user information through the conventional centralized system to the system in which individual users are able to check their scores and detect and correct weaknesses in person (Fig. 6).

Second, an upgraded security policy in which the use of PCs are limited unless certain security scores are observed after acquiring the authority to implement user PCs can be applied (Fig. 7).

Third, it is able to set and apply diverse Help functions for users' better understanding. With easy-to-check and dashboard-like configuration, Help on each indicator and emoticons & colors which explain conditions by scores, the system was designed in a more intuitive and easy-to-understand manner for users.

Fig. 6. Check on PC security vulnerability

Fig. 7. Check on PC security vulnerability by scores

2.7 Changes by the Amended Security Policies

The Cyber Security Assessment Day implemented by 'K' Institute every month is configured to measure and compare score distribution by security vulnerability indicators. In the past, it was able to check the scores through the agent in a user PC after reflecting the government's information security policies on the indicators and having them applied by the central authority. Furthermore, the information security manager had to deliver the feedback on the continued improvement of vulnerability to the target department.

Under the new security system, however, the security indicators in user PCs can be earned in realtime through directly automated assessment by index. In addition, employees' security awareness was enhanced by allowing users to check and reflect the indicators in person.

After all, the new security vulnerability inspection system helped users improve overall security levels from the security management perspective. In statistics collected through the Cyber Security Inspection Day as well, scores were dramatically distributed (Table 3).

Table 3. Acquisitions by security indicator

Category	Previous (A)	New (B)	(B–A)
1	76	77	1
2	45	94	49
3[a]	47	98	51
4	71	99	28
5	88	98	10
6	79	94	15
7[a]	79	98	19
8	77	100	23
9[a]	87	91	4
10	56	65	9
Mean	70.5	91.4	20.9

[a]In some indicators, new indicators may differ from old ones.

In 'K' Institute, there was change in some indicators through the introduction of new systems. However, implementation rates mostly increased. Furthermore, mean value has considerably risen from 70s to 90s. This increase was possible because of the followings: (i) easy implementation of the target indicators through the introduction of a simple method in which users are able to check and apply the value by just clicking a mouse escaping from the one-way approach which does not reveal the conventional indicators to the inspection candidates; (ii) display of users' scores on a computer monitor all the time; (iii) able to use computer if certain scores are earned and kept.

3 Conclusion

These days, the leak of information due to employees' negligence has become more common than a security breach in the outside [10], which means that the internal control of information security should be more reinforced. Therefore, a security manager has promoted diverse activities to reinforce personal security awareness. There should be change in security policies from supplier-oriented to user-centric security policies. If diverse security policies are applied based on the sufficient understanding of users, a security level would spontaneously go up.

The level of information security starts from users' security awareness. In information security policies in public sectors, in particular, there should be effects to increase end-point users' level provided that national information security policies are well observed and implemented.

This study attempted to analyze the factors affecting personal security awareness and behavior based on the better understanding of users. There might be limitations in that there were few studies, and research targets are limited. Even so, this study would provide some suggestions to future studies provided that there might a chance to improve information security environment based on the user's researches.

References

1. Ernst and Young: Ernst & Young 2008 global information security survey (2008). http://faisaldanka.wordpress.com/2008/10/20/ernst-young-2008-global-information-security-survey
2. Kankanhalli, A., Teo, H.H., Tan, B.C.Y., Wei, K.K.: An integrative study of information systems security effectiveness. Int. J. Inf. Manage. **23**(2), 139–154 (2003)
3. Kim, S., Park, S.: Factors affecting information security policy compliance intention. J. Soc. e-Business Stud. **16**(4), 33–51 (2011)
4. Higgins, T.E.: Knowledge activation: accessibility, applicability, and salience. In: Higgins, T.E., Kruglanski, A.W. (eds.) Social Psychology: Handbook of Basic Principles, pp. 133–168. Guilford Press, New York (1996)
5. Myung-gil, C., Hwa-yeong, C.: Neutralization Skills and organization's citizenship against the observation of employees' information security policies. Inf. Syst. Rev. **17**(3), 65–76 (2015)
6. Mayer-Schonberger, V.: Generational development of date protection in Europe. In: Agre, P., Rotenberg, M. (eds.) Technology and Privacy, pp. 219–241. MIT Press, Cambridge (1998)
7. International Standard Organization. http://www.induction.to/bs7799/
8. The British Standards Institution. https://www.bsigroup.com/en-GB/iso-27001-information-security/
9. Kim, J.: National information security agenda and policies. Digit. Policy Stud. **10**(1), 41–50 (2012)
10. Ponemon Institute: Risky Business: How Company Insiders Put High Value Information at Risk (2016)

Safe-Driving Aid System Using Bi-directional Cameras

Byoung Wook Kwon, Kyung Yeob Park, Seo Yeon Moon, and Jong Hyuk Park(✉)

Department of Computer Science and Engineering, Seoul National University of Science and Technology (SeoulTech), Seoul 139-743, Republic of Korea
{bwkwon, kypl, moon, sy0621, jhparkl}@seoultech.ac.kr

Abstract. An accident occurs if you are driving drowsy on a highway or if you are not looking ahead in the school zone. In this paper, we propose 'safe driving aid system. To do this, two cameras are used. One camera recognizes the driver's face and eyes, and the other camera recognizes the Pedestrian in front of the driver. This makes it possible to prevent pedestrian accidents and sleepy driving accidents in advance.

Keywords: Driving system · Sensor network · Bi-directional camera

1 Introduction

The school zone area is the area where pedestrian accidents are the highest. The school zone accident rate accounts for 58% of the total pedestrian accident rate, among them, the accident rate of children aged 7 years old who entered elementary school was the highest at 13.2%. In the case of a pedestrian accident, an accident occurs because the driver does not recognize the pedestrian when he or she is traveling without observing the front. Also, there are a lot of sleepy driving accidents a car is like driving in a state where the driver cannot control that, it can pose a significant threat to the surrounding vehicle or pedestrian [1]. Figure 1 is shown the pedestrian accident statistics. In fact, if a driver suffers about five seconds, a vehicle with a speed of 60 km per hour is about 80 m, and a speed of 100 km per hour is about 140 m without a driver. Therefore, in this paper, we propose an image - based system to assist the driver to avoid such a dangerous situation [2].

2 Safe-Driving Aid System Using Bi-directional Cameras

2.1 System Configuration Diagram

Figure 2 shows the configuration of the 'Safe Driving Aid System' proposed. A two-way camera is installed inside the vehicle so that the driver and the pedestrian are simultaneously recognized. Frames acquired through a two-way camera are delivered to the server via Raspberry Pie, and the server analyzes the image using

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_116

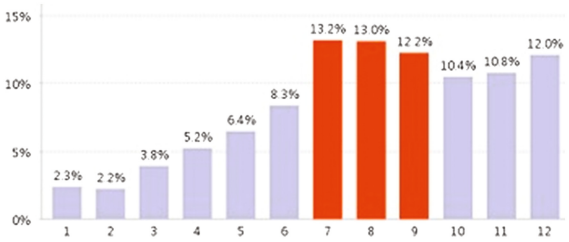


Fig. 1. Graph of pedestrian accident statistics.

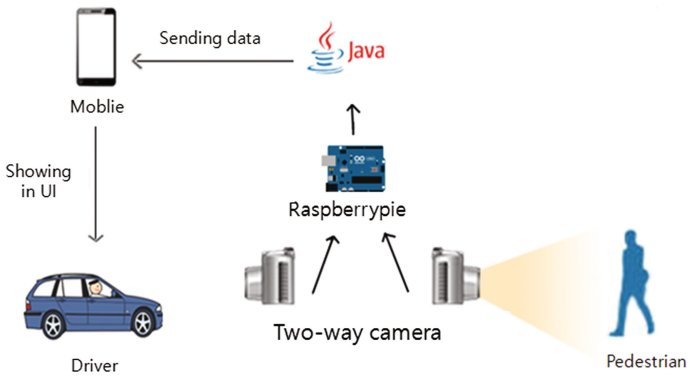


Fig. 2. Safe-driving aid system using bi-directional cameras

image processing algorithms. If it is judged that there is a risk based on the analyzed contents, the information is transmitted to the smartphone and informed to the driver through the smartphone application UI and notification.

The video communication between the camera and the server is done via Raspberry pie based on JAVA language. Raspberry pie is a small computer that can fit inside a limited vehicle and is easy to interface with hardware and software, so it is widely used in video communication.

2.2 Driver Recognition and Event Handling

The internal camera recognizes the driver's face and focuses specifically on the angle and pupil of the head. The pupils are recognized using a cascade classifier using the Haar feature [3]. At this time, /when the eye is closed, /the algorithm does not recognize the eyes because the pupil is not detected by the Haar feature. Therefore, /we write an algorithm to interpret a situation in which the eye is not recognized as a situation in which the eye is blind. In addition, since the facial region is not recognized by the Haar feature even when the angle of the head falls, In addition, since the facial region is not recognized by the Haar feature even when the angle of the head falls, the algorithm is developed so that the situation where the face region is not detected is

interpreted as a situation where the head is distracted. If one or more of the two situations occurs, you can send it to your smartphone via socket communication. The smartphone receiving the value receives the current position of the driver via the Google Map API through event processing, Find the nearest sleepiness rest area or resting-place from your current location and calculate the distance from the driver's current location and output it to the UI. Figure 3 shows the result of face region and eye region recognition by Haar feature according to turning head. The Haar function recognizes the face area according to the arrangement of eyes, nose and mouth. Therefore, when you turn your head, the face area tends to be slightly smaller than in front. Also, when you turn your head, you may not recognize the opposite eye in the direction you turn your head. You can use these two pieces of information to determine whether the driver has turned their head to the right or to the left. Head-turning awareness is used in conjunction with the pedestrian recognition results described in the next section and is used to warn the driver of the pedestrian warning.

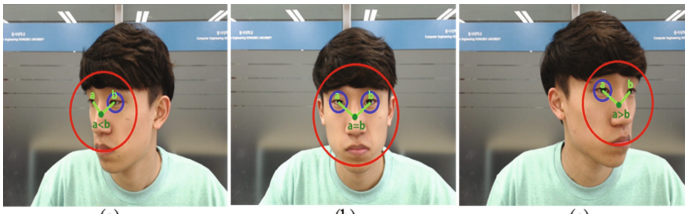


Fig. 3. (a) Turn on left column (b) In front (c) Right column

2.3 Pedestrian Recognition and Event Handling

The driver may not be able to look ahead, paying attention to the driver's left or right field of view. In areas where speed is not high, such as in a school zone or an alleyway, It is easy to relax and look around [4]. However, this school zone and the alleys are the areas where pedestrians are concentrated and large and small accidents occur due to the driver's worries. In order to prevent this, the driver's assistant system algorithm is implemented to warn the driver when the driver is not watching the front using the front camera and the internal camera and the pedestrian appears. A frontal camera recognizing a pedestrian carries out a pedestrian recognition algorithm using a Histogram of oriented gradients (HOG) for a moving object of the photographed image to determine whether the moving object is a pedestrian. Figure 4 shows the result of recognizing the pedestrian by performing HOG on the pedestrian test data. At this time, if it is judged that the driver is not watching the front, for example, the camera on the front side recognizes that a pedestrian appears on the right side. When the driver's head is reversely turned to the left, the smartphone notification gives a warning. Whether the driver's head moves depends on the size of the face area and the perception of the eye, as described in the previous section, or the sum of the distance between the center of the face area and the center of the eye. If the sum of the distance between the center

Fig. 4. Pedestrian recognition using OpenCV

point of the face area and the centers of the eyes is smaller than the previous one, the face can be recognized as a situation turned in one direction.

3 Conclusion

This paper is about driver accident prevention system. Two cameras are installed in the vehicle and the smartphone functions as a head-up display (HUD). By using this system, the presence or absence of a pedestrian ahead can be recognized quickly, and the driver can effectively perceive the drowsy driving. Through accurate recognition, smartphone UI, beeps, and vibrations will be notified to prevent drowsiness and pedestrian accidents, and prevent further accidents.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B4011069).

References

1. Yiannakoulias, N., Scott, D.M.: The effects of local and non-local traffic on child pedestrian safety: a spatial displacement of risk. Soc. Sci. Med. **80**, 96–104 (2013)
2. Koh, P.P., Wong, Y.D., Chandrasekar, P.: Safety evaluation of pedestrian behaviour and violations at signalised pedestrian crossings. Saf. Sci. **7**, 143–152 (2014)
3. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)

Telling Computer and Human Apart: Image-Sound Based CAPTCHA System

Jung Hyun Ryu, Nam Yong Kim, Seo Yeon Moon(✉), and Jong Hyuk Park

Department of Computer Science and Engineering, Seoul National University of Science and Technology (SeoulTech), Seoul 139-743, Republic of Korea
{jh.ryu, kimny, moon, sy0621, jhpark1}@seoultech.ac.kr

Abstract. The Internet is one of important thing in an individual's life and it has made a big changes. As many activities are performed by the Internet, we need specific system to prevent malicious Internet bot programs that take advantage of this convenience. Among them, security technology applying 'Turing Test' to distinguish whether a subject using a specific service is a human or machine is an important thing of security technology of computer science. A representative example is the CAPTCHA. However, the vulnerability was revealed by various studies and cases. For this reason, we propose a new user authentication method using sound and image.

Keywords: CAPTCHA · Computer security · Authentication

1 Introduction

Many websites and applications offer a variety of ways to prevent malicious activity such as advertising postings, account hacking, and online voting manipulation of programs with malicious intent [1]. As a representative example, CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) exists. CAPTCHA identifies human and machines by providing information that only human can judge using texts, images, sounds, and so on. The type of CAPTCHA that is mainly used today is to show an image containing text information and to input the answer. However, this method cannot prove safety because the vulnerability has been exposed several times in various research papers and studies. Since then, Google has announced a new authentication method called reCAPTCHA in 2014, but the same year, the authentication code was cracked, revealing the vulnerability again. For this reason, this paper roughly proposes a new user authentication method that provides a specific sound and replaces the answer with an image [2].

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_117

2 Related Work

We plan to use lightweight encryption algorithm for encryption of sound and images used for authentication and encryption for data transmission. Lightweight encryption algorithm is lightweight and fast in comparison to existing encryption algorithms in high-speed environments such as big data and cloud computing, and lightweight environments such as mobile devices. Among them, we will apply the 'LEA' developed in 2013 by NSRI (National Security Research Institute) of Korea.

The LEA is a symmetric-key encryption algorithm that encrypts/decrypts a 128-bit plaintext/ciphertext block with a 128/192/256-bit master key to generate a 128-bit ciphertext/plaintext. It is based on ARX operation and Feistel-like structure. It does not need decryption module, so it can be lightened and is suitable for mobile devices.

After that, we will research to make it a safer authentication system.

2.1 Text Based Authentication

Many services use Text based CAPTCHA, which is a way to provide randomly distorted string as an image and match it for authentication. Because of the convenience of implementation, many Web services and applications are mainly used for the purpose of automatic subscription prevention. Text based CAPTCHA uses various functions to transform text in the database into images and to distort the image, but it can also be used to restore distorted text [2].

There are many techniques that can read a character string output to an image without entering the inside of the system by applying image processing technology including SVM, and so on. For this reason, the safety of the CAPTCHA is less reliable because it has characteristics that it is easier to attack than other type of CAPTCHAs.

2.2 Audio Based Authentication

Audio based user authentication is mainly based on the differences in the language recognition capabilities of human and computers. It is mainly designed to mix noise with human voice so that only human can recognize it. As a representative example, there is Audio-based CAPTCHA which is also used as supportive way of text-based CAPTCHA. It randomly chooses words, sentences, and random numbers, distorts them into voice files, and then show the user to answer them.

Recently, however, deeprunning-based algorithms such as DNN (Deep Neural Networks), CNN (Convoluted on Neural Network) and LSTM-RNN (Long Short-Term Memory-Recurrent Neural Network) have been able to remove noise and recognize the voice. So the Audio-based user verification method is vulnerable.

2.3 Image Based Authentication

Image-based authentication is a way of asking for answers by showing pictures or some images. Mostly, verification is done by showing an image of a specific object and inputting the name as text, or restoring the rotated image to its original state [3].

Typically, there is an image CAPTCHA and a reCAPTCHA that uses an randomly imported image from Google Maps. However, image-based user authentication system, like the two user verification systems above, has the potential to be attacked.

Artificial intelligence-based image recognition programs, such as the recently announced the Vision algorithm of Facebook and Google's Inception V3, can analyze and even describe the information in the image.

Based on these algorithms, image-based user authentication systems will not be able to distinguish between human and computers. Especially, Google's Inception V3 showed an accuracy of 93.9% on their research [4].

3 New Way of User Authentication

The existing image-based authentication is to show an random image having the same meaning to the user and to match the image. And the audio-based authentication selected a way of inputting a single word or a combination of words corresponding to the correct answer by hearing a voice with noise. The interface we toughly think is shown in the following Fig. 1.

Fig. 1. Proposed authentication system interface

Both types of vulnerabilities have emerged as technology for image analysis and voice recognition evolves. We have considered new user authentication system by merely combining the advantages of both methods. This new user authentication is to prove that a user hears a random sound except for a human voice and selects the corresponding image to be a human. The sounds that are provided to the user are animals, environments, objects, nature, etc., and all sounds except the voice of the human are questions. It mixes the noise and lowers the sound quality, making it difficult to recognize the computer through the voice recognition program.

We use specific images as an answer, which uses simplified images like pictograms with noise to make the computer hard to recognize.

4 Conclusion

In this paper, we propose improvement way of CAPTCHA, which is one of the security technologies for preventing malicious bot programs, to the mobile environment. We roughly studied vulnerabilities of the CAPTCHA systems used now and thought about ways to improve them.

The new CAPTCHA what we proposed is designed attacks caused by image processing technology and voice recognition technology. In order to improve stability and availability in mobile environment, we plan to apply the LEA (Lightweight Encryption Algorithm) registered as a standard to the Korea Telecommunications Technology Association (TTA).

If many studies are done in the future, it is expected that it will provide another security environment for the increasing use of mobile environment.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B4011069).

References

1. Von Ahn, L., Blum, M., Langford, J.: Telling humans and computers apart automatically. Commun. ACM **47**(2), 56–60 (2004)
2. Ved Prakash Singh, P.P.: Survey of different types of CAPTCHA. Int. J. Comput. Sci. Inf. Technol. **5**(2), 2242–2245 (2014)
3. Kambourakis, G., Damopoulos, D., Papamartzivanos, D., Pavlidakis, E.: Introducing touchstroke: keystroke-based authentication system for smartphones. Secur. Commun. Netw. **9**(6), 542–554 (2016)
4. Turkanovi, M., Brumen, B., Hölbl, M.: A novel user authentication and key agreement scheme for heterogeneous ad hoc wireless sensor networks, based on the Internet of Things notion. Ad Hoc Netw. **20**, 96–112 (2014)

Author Index

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1