Jeng-Shyang Pan
Shyi-Ming Chen
Ngoc Thanh Nguyen (Eds.)

LNAI 7197

Intelligent Information and Database Systems

4th Asian Conference, ACIIDS 2012
Kaohsiung, Taiwan, March 2012
Proceedings, Part II

2 Part II

Springer

Lecture Notes in Artificial Intelligence 7197

Subseries of Lecture Notes in Computer Science

Jeng-Shyang Pan Shyi-Ming Chen
Ngoc Thanh Nguyen (Eds.)

Intelligent Information and Database Systems

4th Asian Conference, ACIIDS 2012
Kaohsiung, Taiwan, March 19-21, 2012
Proceedings, Part II

Springer

Volume Editors

Jeng-Shyang Pan
National Kaohsiung University of Applied Sciences
Department of Electronic Engineering
No. 415, Chien Kung Road, Kaohsiung 80778, Taiwan
E-mail: jengshyangpan@gmail.com

Shyi-Ming Chen
National Taichung University of Education
Graduate Institute of Educational Measurement and Statistics
No. 140, Min-Shen Road, Taichung 40306, Taiwan
E-mail: smchen@mail.ntcu.edu.tw

Ngoc Thanh Nguyen
Wrocław University of Technology, Institute of Informatics
Wybrzeże Wyspiańskiego 27, 50370, Wrocław, Poland
E-mail: ngoc-thanh.nguyen@pwr.edu.pl

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-28489-2 e-ISBN 978-3-642-28490-8
DOI 10.1007/978-3-642-28490-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012931775

CR Subject Classification (1998): I.2, H.3, H.2.8, H.4-5, I.4-5, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

ACIIDS 2012 was the fourth event of the series of international scientific conferences for research and applications in the field of intelligent information and database systems. The aim of ACIIDS 2012 was to provide an international forum for scientific research in the technologies and applications of intelligent information, database systems and their applications. ACIIDS 2012 took place March 19–21, 2012, in Kaohsiung, Taiwan. It was co-organized by the National Kaohsiung University of Applied Sciences (Taiwan), National Taichung University of Education (Taiwan), Taiwanese Association for Consumer Electronics (TACE) and Wroclaw University of Technology (Poland), in cooperation with the University of Information Technology (Vietnam), International Society of Applied Intelligence (ISAI), and Gdynia Maritime University (Poland). ACIIDS 2009 and ACIIDS 2010 took place in Dong Hoi and Hue in Vietnam, respectively, and ACIIDS 2011 in Deagu, Korea.

We received more than 472 papers from 15 countries over the world. Each paper was peer reviewed by at least two members of the International Program Committee and International Reviewer Board. Only 161 papers with the highest quality were selected for oral presentation and publication in the three volumes of ACIIDS 2012 proceedings.

The papers included in the proceedings cover the following topics: intelligent database systems, data warehouses and data mining, natural language processing and computational linguistics, Semantic Web, social networks and recommendation systems, collaborative systems and applications, e-business and e-commerce systems, e-learning systems, information modeling and requirements engineering, information retrieval systems, intelligent agents and multi-agent systems, intelligent information systems, intelligent Internet systems, intelligent optimization techniques, object-relational DBMS, ontologies and knowledge sharing, semi-structured and XML database systems, unified modeling language and unified processes, Web services and Semantic Web, computer networks and communication systems.

Accepted and presented papers highlight new trends and challenges of intelligent information and database systems. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to the Honorary Chairs: Cheng-Qi Zhang (University of Technology Sydney, Australia), Szu-Wei Yang (President of National Taichung University of Education, Taiwan) and Tadeusz Wieckowski (Rector of Wroclaw University of Technology, Poland) for their support.

Our special thanks go to the Program Chairs, all Program and Reviewer Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the

selected papers for the conference. We cordially thank the organizers and chairs of special sessions, which essentially contribute to the success of the conference.

We would also like to express our thanks to the keynote speakers Jerzy Swiatek from Poland, Shi-Kuo Chang from the USA, Jun Wang, and Rong-Sheng Xu from China for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, National Kaohsiung University of Applied Sciences (Taiwan), National Taichung University of Education (Taiwan), Taiwanese Association for Consumer Electronics (TACE) and Wroclaw University of Technology (Poland). Our special thanks are due also to Springer for publishing the proceedings, and other sponsors for their kind support.

We wish to thank the members of the Organizing Committee for their very substantial work, especially those who played essential roles: Thou-Ho Chen, Chin-Shuih Shieh, Mong-Fong Horng and the members of the Local Organizing Committee for their excellent work.

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.

Jeng-Shyang Pan
Shyi-Ming Chen
Ngoc Thanh Nguyen

Conference Organization

Honorary Chairs

Cheng-Qi Zhang	University of Technology Sydney, Australia
Szu-Wei Yang	National Taichung University of Education, Taiwan
Tadeusz Wieckowski	Wroclaw University of Technology, Poland

General Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Program Committee Chairs

Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taichung University of Education, Taiwan
Junzo Watada	Waseda University, Japan
Jian-Chao Zeng	Taiyuan University of Science and Technology, China

Publication Chairs

Chin-Shiuh Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Li-Hsing Yen	National University of Kaohsiung, Taiwan

Invited Session Chairs

Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan

Organizing Chair

Thou-Ho Chen	National Kaohsiung University of Applied Sciences, Taiwan

Steering Committee

Ngoc Thanh Nguyen - Chair	Wroclaw University of Technology, Poland
Bin-Yih Liao	National Kaohsiung University of Applied Sciences, Taiwan
Longbing Cao	University of Technology Sydney, Australia
Adam Grzech	Wroclaw University of Technology, Poland
Tu Bao Ho	Japan Advanced Institute of Science and Technology, Japan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, Korea
Jason J. Jung	Yeungnam University, Korea
Hoai An Le-Thi	University Paul Verlaine - Metz, France
Antoni Ligęza	AGH University of Science and Technology, Poland
Toyoaki Nishida	Kyoto University, Japan
Leszek Rutkowski	Technical University of Czestochowa, Poland

Keynote Speakers

– Jerzy Swiatek

President of Accreditation Commission of Polish Technical Universities, Poland

– Shi-Kuo Chang

Center for Parallel, Distributed and Intelligent Systems, University of Pittsburgh, USA

– Jun Wang

Computational Intelligence Laboratory in the Department of Mechanical and Automation Engineering at the Chinese University of Hong Kong, China

– Rong-Sheng Xu

Computing Center at Institute of High Energy Physics, Chinese Academy of Sciences, China

Invited Sessions Organizers

Bogdan Trawiński	Wrocław University of Technology, Poland
Oscar Cordon	Wrocław University of Technology, Poland
Przemyslaw Kazienko	Wrocław University of Technology, Poland

Ondrej Krejcar	Technical University of Ostrava, Czech Republic
Peter Brida	University of Zilina, Slovakia
Kun Chang Lee	Sungkyunkwan University, Korea
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Kuan-Rong Lee	Kun Shan University, Taiwan
Yau-Hwang Kuo	National Cheng Kung University, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taichung University of Education, Taiwan
Chulmo Koo	Chosun University, Korea
I-Hsien Ting	National University of Kaohsiung, Taiwan
Jason J. Jung	Yeungnam University, Korea
Chaudhary Imran Sarwar	University of the Punjab, Pakistan
Tariq Mehmood Chaudhry	Professional Electrical Engineer, Pakistan
Arkadiusz Kawa	Poznan University of Economics, Poland
Paulina Golińska	Poznan University of Technology, Poland
Konrad Fuks	Poznan University of Economics, Poland
Marcin Hajdul	Institute of Logistics and Warehousing, Poland
Shih-Pang Tseng	Tajen University, Taiwan
Yuh-Chung Lin	Tajen University, Taiwan
Phan Cong Vinh	NTT University, Vietnam
Le Thi Hoai An	Paul Verlaine University, France
Pham Dinh Tao	INSA-Rouen, France
Bao Rong Chang	National University of Kaohsiung, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan

International Program Committee

Cesar Andres	Universidad Complutense de Madrid, Spain
S. Hariharan B.E.	J.J. College of Engineering and Technology Ammapettai, India
Costin Badica	University of Craiova, Romania
Youcef Baghdadi	Sultan Qaboos University, Oman
Dariusz Barbucha	Gdynia Maritime University, Poland
Stephane Bressan	NUS, Singapore
Longbing Cao	University of Technology Sydney, Australia
Frantisek Capkovic	Slovak Academy of Sciences, Slovakia
Oscar Castillo	Tijuana Institute of Technology, Mexico
Bao Rong Chang	National University of Kaohsiung, Taiwan

Hsuan-Ting Chang	National Yunlin University of Science and Technology, Taiwan
Lin-Huang Chang	National Taichung University of Education, Taiwan
Chuan-Yu Chang	National Yunlin University of Science and Technology, Taiwan
Jui-Fang Chang	National Kaohsiung University of Applied Sciences, Taiwan
Wooi Ping Cheah	Multimedia University, Malaysia
Shyi-Ming Chen	National Taichung University of Education, Taiwan
Guey-Shya Chen	National Taichung University of Education, Taiwan
Rung Ching Chen	Chaoyang University of Technology, Taiwan
Suphamit Chittayasothorn	King Mongkut's Institute of Technology, Thailand
Tzu-Fu Chiu	Aletheia University, Taiwan
Chou-Kang Chiu	National Taichung University of Education, Taiwan
Shu-Chuan Chu	Flinders University, Australia
Irek Czarnowski	Gdynia Maritime University, Poland
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Jiangbo Dang	Siemens Corporate Research, USA
Tran Khanh Dang	HCMC University of Technology, Vietnam
Paul Davidsson	Malmö University, Sweden
Hui-Fang Deng	South China University of Technology, China
Phuc Do	University of Information Technology, Vietnam
Van Nhon Do	University of Information Technology, Vietnam
Manish Dixit	Madhav Institute of Technology and Science, India
Antonio F.	Murcia University, Spain
Pawel Forczmanski	West Pomeranian University of Technology, Poland
Patrick Gallinar	Université Pierre et Marie Curie, France
Mauro Gaspari	University of Bologna, Italy
Dominic Greenwood	Whitestein Technologies, Switzerland
Slimane Hammoudi	ESEO, France
Hoang Huu Hanh	Hue University, Vietnam
Jin-Kao Hao	University of Angers, France
Le Thi Hoai An	Paul Verlaine University – Metz, France
Kiem Hoang	University of Information Technology, Vietnam
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Ying-Tung Hsiao	National Taipei University of Education, Taiwan

Chao-Hsing Hsu	Chienkuo Technology University, Taiwan
Wen-Lian Hsu	Academia Sinica, Taiwan
Feng-Rung Hu	National Taichung University of Education, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Hsiang-Cheh Huang	National University of Kaohsiung, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Deng-Yuan Huang	Dayeh University, Taiwan
Jingshan Huang	University of South Alabama, USA
Piotr Jedrzejowicz	Gdynia Maritime University, Poland
Albert Jeng	Jinwen University of Science and Technology, Taiwan
Alcala-Fdez Jesus	University of Granada, Spain
Jason J. Jung	Yeungnam University, South Korea
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Radosaw Piotr Katarzyniak	Wroclaw University of Technology, Poland
Muhammad Khurram Khan	King Saud University, Saudi Arabia
Cheonshik Kim	Sejong University, Korea
Joanna Kolodziej	University of Bielsko-Biala, Poland
Ondrej Krejcar	VSB - Technical University of Ostrava, Czech Republic
Dariusz Krol	Wroclaw University of Technology, Poland
Wei-Chi Ku	National Taichung University of Education, Taiwan
Tomasz Kubik	Wroclaw University of Technology, Poland
Bor-Chen Kuo	National Taichung University of Education, Taiwan
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Raymond Y.K. Lau	City University of Hong Kong, Hong Kong
Kun Chang Lee	Sungkyunkwan University, Korea
Chin-Feng Lee	Chaoyang University of Technology, Taiwan
Eun-Ser Lee	Andong National University, Korea
Huey-Ming Lee	Chinese Culture University, Taiwan
Chunshien Li	National Central University, Taiwan
Tsai-Hsiu Lin	National Taichung University of Education, Taiwan
Yuan-Horng Lin	National Taichung University of Education, Taiwan
Chia-Chen Lin	Providence University, Taiwan
Hao-Wen Lin	Harbin Institute of Technology, China
Min-Ray Lin	National Taichung University of Education, Taiwan
Hsiang-Chuan Liu	Asia University, Taiwan

Yu-lung Lo	Chaoyang University of Technology, Taiwan
Ching-Sung Lu	Tajen University, Taiwan
James J. Lu	Emory University, USA
Janusz Marecki	IBM T.J. Watson Research Center, USA
Vuong Ngo Minh	Ho Chi Minh City University of Technology, Vietnam
Tadeusz Morzy	Poznan University of Technology, Poland
Kazumi Nakamatsu	School of Human Science and Environment, University of Hyogo, Japan
Grzegorz J. Nalepa	AGH University of Science and Technology, Poland
Jean-Christophe Nebel	Kingston University, UK
Vinh Nguyen	Monash University, Australia
Manuel Nunez	Universidad Complutense de Madrid, Spain
Marcin Paprzycki	Systems Research Institute of the Polish Academy of Sciences, Poland
Witold Pedrycz	University of Alberta, Canada
Ibrahima Sakho	University of Metz, France
Victor Rung-Lin Shen	National Taipei University, Taiwan
Tian-Wei Sheu	National Taichung University of Education, Taiwan
Chin-Shiuh Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Shu-Chuan Shih	National Taichung University of Education, Taiwan
An-Zen Shih	Jinwen University of Science and Technology, Taiwan
Gomez Skarmeta	Murcia University, Spain
Serge Stinckwich	IRD, France
Pham Dinh Tao	National Institute for Applied Sciences Roue, France
Wojciech Thomas	Wroclaw University of Technology, Poland
Geetam Singh Tomar	Gwalior Malwa Institute of Technology and Management, India
Dinh Khang Tran	School of Information and Communication Technology HUST, Vietnam
Bogdan Trawiski	Wrocaw University of Technology, Poland
Hoang Hon Trinh	Ho Chi Minh City University of Technology, Vietnam
Hong-Linh Truong	Vienna University of Technology, Austria
Chun-Wei Tseng	Cheng Shiu University, Taiwan
Kuo-Kun Tseng	Harbin Institute of Technology, China
Felea Victor	Alexandru Ioan Cuza University of Iai, Romania
Phan Cong	Vinh, NTT University, Vietnam
Yongli Wang	North China Electric Power University, China

Lee-Min Wei	National Taichung University of Education, Taiwan
Michal Wozniak	Wroclaw University of Technology, Poland
Homer C. Wu	National Taichung University of Education, Taiwan
Xin-She Yang	National Physical Laboratory, UK
Horng-Chang Yang	National Taitung University, Taiwan
Shu-Chin Yen	Wenzao Ursuline College of Languages, Taiwan
Ho Ye	Xidian University, China

Table of Contents – Part II

Clustering Technology

Intelligent Digital Watermarking and Image Processing

Intelligent Management of e-Business

Intelligent Media Processing

Modelling and Optimization Techniques in Information Systems, Database Systems and Industrial Systems

User Adaptive Systems(1)

User Adaptive Systems(2)

Advances in Nature-Inspired AutonQomic Computing and Networking

Approach to Image Segmentation Based on Interval Type-2 Fuzzy Subtractive Clustering

Long Thanh Ngo and Binh Huy Pham

Department of Information Systems, Faculty of Information Technology,
Le Quy Don Technical University, No 100, Hoang Quoc Viet, Hanoi, Vietnam
{ngotlong,huybinhhvhc}@gmail.com

Abstract. The paper deals with an approach to image segmentation using interval type-2 fuzzy subtractive clustering (IT2-SC). The IT2-SC algorithm is proposed based on extension of subtractive clustering algorithm (SC) with fuzziness parameter m. And to manage uncertainty of the parameter m, we have expanded the SC algorithm to interval type-2 fuzzy subtractive clustering (IT2-SC) using two fuzziness parameters m_1 and m_2 which creates a footprint of uncertainty (FOU) for the fuzzifier. The input image is extracted RGB values as input space of IT2-SC; number of clusters is automatically identified based on parameters of the algorithm and image properties. The experiments of image segmentation are implemented in variety of images with statistics.

Keywords: Subtractive clustering, type-2 fuzzy subtractive clustering, type-2 fuzzy sets, image segmentation.

1 Introduction

Segmentation is an essential technique in image description or classification. Many different segmentation approaches have been developed using fuzzy logics [20], [21], [22], [23]. Clustering is a common technique in many areas such as data mining, pattern recognition, image processing,. . . , especially image segmentation. In particular, many clustering algorithms have been developed, including k-mean [19], fuzzy c-mean [18] and mountain clustering [13], [14], [15]. However, in most real data exists uncertainty and vagueness which cannot be appropriately managed by type-1 fuzzy sets. Meanwhile, type-2 fuzzy sets allow us to obtain desirable results in designing and managing uncertainty. Therefore, type-2 fuzzy sets have been studied and widely applied in many fields [6], [7], [8], especially pattern recognitions. On that basis, clustering algorithms has been extended and developed into type-2 fuzzy clustering algorithms to identify FOU of fuzzifiers, resulting to managing uncertainty is better.

Subtractive clustering algorithm, proposed in 1994 by Chiu [1,2], is an extension of the mountain clustering methods by improving the mountain function to calculate potential of becoming a cluster center for each data point based on the location of the data point with respect to the other data points. The subtractive clustering algorithm only consider data points, not a grid points, which

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 1–10, 2012.

reduces the computational complexities and gives better distribution of cluster centroids. Kim et al. [5] have improved subtractive clustering algorithm by proposing a kernel-induced distance instead of the conventional distance when calculating the mountain value of data point. J.Y. Chen et al [4] proposed a weighted mean subtractive clustering.

The paper deals with an approach to image segmentation using the proposed interval type-2 fuzzy subtractive clustering. The subtractive clustering algorithm is extended by setting a fuzziness parameter m into mountain function for data points. The fuzziness parameter m can alter the results of the mountain function so it has highly influence to the results of clustering. Through fuzziness parameter m, we can reduce the dependence of clustering results to initial values of parameters of the algorithm. As the adjustment parameter m, we can obtain better clustering results without interesting into setting the initial values of the parameters of the algorithm. Therefore, fuzziness parameter m created uncertainties for the SC. To design and manage uncertainties of fuzzifier m , we extend to interval type-2 fuzzy subtractive clustering (IT2-SC) by using two fuzzifiers m_1 and m_2 which creates a footprint of uncertainty (FOU). Approach to image segmentation is introduced based on the proposed clustering algorithms, in which the algorithm automatically identifies number and centroid of clusters based on image data. Experiments of image segmentation are implemented and summarised in comparison with other clustering.

The remainder of the paper is organized as follows. In Section 2 introduces briefly type-2 fuzzy sets and interval type-2 fuzzy sets, subtractive clustering. In section 3, we discuss how to extend subtractive clustering method with fuzziness parameter m and the proposed interval type-2 fuzzy subtractive clustering. In section 4, we provide several experiments of image segmentation to show the validity of our proposed method. Finally, section 5 gives the summaries and conclusions.

2 Preliminaries

2.1 Type-2 Fuzzy Sets

A type-2 fuzzy set in X is denoted by $\tilde{A}$, and its membership grade of $x \in X$ is $\mu_{\tilde{A}}(x,u), u \in J_x \subseteq [0,1]$, which is a type-1 fuzzy set in [0, 1]. The elements of domain of $\mu_{\tilde{A}}(x,u)$ are called primary memberships of $\tilde{A}$ and memberships of primary memberships in $\mu_{\tilde{A}}(x,u)$ are called secondary memberships of $\tilde{A}$.

Definition 1. *A type* -2 *fuzzy set, denoted by* $\tilde{A}$*, is characterized by a type-2 membership function* $\mu_{\tilde{A}}(x,u)$ *where* $x \in X$ *and* $u \in J_x \subseteq [0,1]$*, i.e.,*

$$\tilde{A} = \{((x,u), \mu_{\tilde{A}}(x,u)) | \forall x \in X, \forall u \in J_x \subseteq [0,1]\} \tag{1}$$

or

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{A}}(x,u))/(x,u), J_x \subseteq [0,1] \tag{2}$$

in which $0 \leq \mu_{\tilde{A}}(x,u) \leq 1$.

At each value of x, say $x = x'$, the 2-D plane whose axes are u and $\mu_{\tilde{A}}(x', u)$ is called a *vertical slice* of $\mu_{\tilde{A}}(x, u)$. A *secondary membership function* is a vertical slice of $\mu_{\tilde{A}}(x, u)$. It is $\mu_{\tilde{A}}(x = x', u)$ for $x \in X$ and $\forall u \in J_{x'} \subseteq [0, 1]$, i.e.

$$\mu_{\tilde{A}}(x = x', u) \equiv \mu_{\tilde{A}}(x') = \int_{u \in J_{x'}} f_{x'}(u)/u, J_{x'} \subseteq [0, 1] \tag{3}$$

in which $0 \leq f_{x'}(u) \leq 1$.

In manner of embedded fuzzy sets, a type-2 fuzzy sets [6] is union of its type-2 embedded sets, i.e

$$\tilde{A} = \sum_{j=1}^{n} \tilde{A}_e^j \tag{4}$$

where $n \equiv \prod_{i=1}^{N} M_i$ and $\tilde{A}_e^j$ denoted the j^{th} type-2 embedded set of $\tilde{A}$, i.e.,

$$\tilde{A}_e^j \equiv \{\left(u_i^j, f_{x_i}(u_i^j)\right), i = 1, 2, ..., N\} \tag{5}$$

where $u_i^j \in \{u_{ik}, k = 1, ..., M_i\}$.

Type-2 fuzzy sets are called an interval type-2 fuzzy sets if the secondary membership function $f_{x'}(u) = 1 \ \forall u \in J_x$ i.e. a type-2 fuzzy set are defined as follows:

Definition 2. *An interval type-2 fuzzy set $\tilde{A}$ is characterized by an interval type-2 membership function $\mu_{\tilde{A}}(x, u) = 1$ where $x \in X$ and $u \in J_x \subseteq [0, 1]$, i.e.,*

$$\tilde{A} = \{((x, u), 1) | \forall x \in X, \forall u \in J_x \subseteq [0, 1]\} \tag{6}$$

Uncertainty of $\tilde{A}$, denoted by FOU, is union of primary functions i.e. $FOU(\tilde{A}) = \bigcup_{x \in X} J_x$. Upper/lower bounds of membership function (UMF/LMF), denoted by $\overline{\mu}_{\tilde{A}}(x)$ and $\underline{\mu}_{\tilde{A}}(x)$, of $\tilde{A}$ are two type-1 membership functions and bounds of FOU.

2.2 Subtractive Clustering Algorithm

Subtractive clustering finds the optimal data point to define a cluster centroid based on the density of surrounding data points, see [1,2] for more details. Consider a collection of n data points: $X = \{x_1, x_2, ..., x_n\}$, where, x_i is a vector in M-dimensional space.

The density function of a data point is described as follows:

$$P_i = \sum_{j=1}^{n} e^{-\frac{4}{r_a^2}\|x_i - x_j\|^2} \tag{7}$$

where:

- P_i: the density of surrounding data points of i^{th} data point.
- r_a: is a positive constant defining a neighbourhood radius.
- $\|.\|$: denotes the Euclidean distance.

After the density of every data point has been computed, the data point with the highest density is selected as the first cluster centroid. Let x_i^* is the location of the first cluster centroid and P_i^* is its density value then $P_1^* = \max_{i=1}^{n} P_i$.

The density of all data points has been revised as follows:

$$P_i = P_i - P_1^* e^{-\frac{4}{r_b^2}\|x_i - x_1^*\|^2}; i = 1, ..., n \tag{8}$$

where r_b is is a positive constant and $r_b = \eta * r_a$ with a good choose is $r_b = 1.5 r_a$. When the density of all data points have been reduced by (2), the data point with the highest remaining density is selected as the second cluster centroid.

Generally, after k^{th} cluster centroid has been obtained, the density of each data point is updated by the following formula:

$$P_i = P_i - P_k^* e^{-\frac{4}{r_b^2}\|x_i - x_k^*\|^2}; i = 1, ..., n \tag{9}$$

The subtractive clustering algorithm includes the following steps:

Step 1: Initialization, r_a, η with $\eta = \frac{r_b}{r_a}$, $\overline{\varepsilon}$ and $\underline{\varepsilon}$

Step 2: Calculating density for all data points by using formula (7). Data point with the highest density is selected as the first cluster center. $P_k^* = \max_{i=1}^{n} P_i$ where $k = 1$ and P_k^* is the density of the first cluster center.

Step 3: The density of all data points is revised by using formula (9).

Step 4: Let x^* is a data point with its density is highest and equal P^*.
- If $P^* > \overline{\varepsilon} P^{ref}$: x^* is a new cluster center and back to Step 3.
- Else if $P^* < \underline{\varepsilon} P^{ref}$: back to Step 5.
- Else:
  + Let d_{min} is shortest of the distances between x^* and all previously found cluster centroids.
  + If $\frac{d_{\min}}{r_a} + \frac{P^*}{P^{\text{ref}}} \geq 1$: x^* is a new cluster center and back to Step 3.
  + Else: $P(x^*) = 0$ and select x^* with the highest density, $P(x^*)$, and back to step 4.

Step 5: Output the results of clustering.

The membership degree of data point in each cluster is as follows:

$$\mu_{ik} = e^{-\frac{4}{r_a^2}\|x_i - x_k\|^2} \tag{10}$$

3 Interval Type-2 Fuzzy Subtractive Clustering

3.1 Extending Subtractive Clustering Algorithm

In the subtractive clustering algorithm, we must set four parameters: accept ratio $\overline{\varepsilon}$, reflect ratio $\underline{\varepsilon}$, cluster radius r_a and squash factor η (or r_b). The choice of parameters have considerable influences on results of clustering. If values of $\overline{\varepsilon}$ and $\underline{\varepsilon}$ are large, the number of cluster centroids will be reduced. Conversely small values of $\overline{\varepsilon}$ and $\underline{\varepsilon}$ will increase the number of cluster centroids. If value of r_a is too small, too many cluster centroids will be generated. Conversely, too few cluster centroids will be generated. The value of η has influences to results of clustering as r_a. Thus, we can not know the best parameters to be used for a given data, even a parameter search is performed to make the best clustering results. The influences of the four parameters to clusters have been described detail in papers of Demirli [20,21]. Therefore, these parameters are uncertainties in the subtractive clustering algorithm.

Subtractive clustering estimated the potential of a data point as a cluster centroid based on the density of surrounding data points, which is actually based on the distance between the data point with the remaining data points. When the degree of membership of i^{th} cluster centroid is defined as the formula(10). Therefore, SC includes various types of uncertainty as distance measure, parameters Initialization... So we consider a fuzziness parameters that control the distribution of data points into clusters by making the parameter m in the density function to calculate the potential of a data point as follows:

$$P_i = \sum_{j=1}^{n} e^{-\frac{4}{r_a^2}(x_j - x_i)^{\frac{2}{m-1}}} \tag{11}$$

If x_k is the k^{th} cluster position, has potential P_k^*, then the potential of each data point is revised by the following formula:

$$P_i = P_i - P_k^* e^{-\frac{4}{r_b^2}(x_i - x_k)^{\frac{2}{m-1}}}; i = 1, ..., n \tag{12}$$

Then the choice of the value of the parameter m has greatly influence to results of clustering. If m is small, the number of centroids will be reduced. Conversely, if m is too large, too many centroids will be generated.

3.2 Interval Type-2 Fuzzy Subtractive Clustering Algorithm

In extended subtractive clustering algorithm, the membership degree of a point in k^{th} cluster centroid is defined as following formula:

$$\mu_{ik} = e^{-\frac{4}{r_a^2}(x_i - x_k)^{\frac{2}{m-1}}} \tag{13}$$

where x_k is k^{th} cluster centroid.

According to the formula (13), membership value of a data point in k^{th} cluster centroid depends on the position of k^{th} cluster and the fuzziness parameter m. On the other hand, the position of the k^{th} cluster also depends on the fuzziness parameter m. Thus, the fuzziness parameter m is the most uncertainty element in the expanded subtractive clustering algorithm. Therefore, to design and manage the uncertainty for fuzziness parameter m , we extend a pattern set to interval type-2 fuzzy sets using two fuzzifiers m_1 and m_2 which creates a footprint of uncertainty (FOU) for the fuzziness parameter m. Then the degree of membership of k^{th} cluster centroid is defined as following formula:

$$\begin{cases} \overline{\mu}_{ik} = e^{-\frac{4}{r_a^2}(x_i - x_k)^{\frac{2}{m_1 - 1}}} \\ \underline{\mu}_{ik} = e^{-\frac{4}{r_a^2}(x_i - x_k)^{\frac{2}{m_2 - 1}}} \end{cases} \tag{14}$$

We have two density functions to calculate potential of each data point as follows:

$$\begin{cases} \overline{P}_i = \sum_{j=1}^{n} e^{-\frac{4}{r_a^2}(x_j - x_i)^{\frac{2}{m_1 - 1}}} \\ \underline{P}_i = \sum_{j=1}^{n} e^{-\frac{4}{r_a^2}(x_j - x_i)^{\frac{2}{m_2 - 1}}} \end{cases} \tag{15}$$

If the centroids are identified by the formula (15), we will have centroids v_L and v_R. Thus, we will do type-reduction for centroids as bellows:

$$P_i = \frac{\overline{P}_i * m_1 + \underline{P}_i * m_2}{m_1 + m_2} \tag{16}$$

And when we identified k^{th} cluster center, the density of all data points is revised by using following formula:

$$\begin{cases} \underline{P}_i^{sub} = P_k^* \sum_{j=1}^{n} e^{-\frac{4}{r_b^2} d_{ij}^{\frac{2}{m_1 - 1}}} \\ \overline{P}_i^{sub} = P_k^* \sum_{j=1}^{n} e^{-\frac{4}{r_b^2} d_{ij}^{\frac{2}{m_2 - 1}}} \\ P_i^{sub} = \frac{\underline{P}_i^{sub} * m_1 + \overline{P}_i^{sub} * m_2}{m_1 + m_2} \\ P_i = P_i - P_i^{sub} \end{cases} \tag{17}$$

The interval type-2 fuzzy subtractive clustering algorithm includes the following steps:

Step 1: Initialization, r_a, η with $\eta = \frac{r_b}{r_a}$, $\overline{\varepsilon}$ and $\underline{\varepsilon}$, m_1 and m_2 $(1 < m_1 < m_2)$

Step 2: Calculating density for all data points with two fuzzifiers m_1 and m_2 by using formulas (15) and (16). Data point with the highest density is selected as the first cluster centroid: $P_k^* = \max_{i=1}^{n} P_i$ where $k = 1$ and P_k^* is the density of the first cluster centroid.

Step 3: The density of all data points is revised by using formula (17).

Step 4: Identification of the next cluster centroids are as similar as SC.

Step 5: Output the results of clustering.

4 Experimental Results

In this section, we carry out image segmentations based on fuzzy clustering algorithms as FCM, SC and our proposed IT2-SC algorithm. In our experiments, the RGB model is used to monochrome image segmentation. The experiments have performed on a variety of images and comparison with the segmented image obtained by the proposed IT2-SC method and some other techniques as SC and FCM.

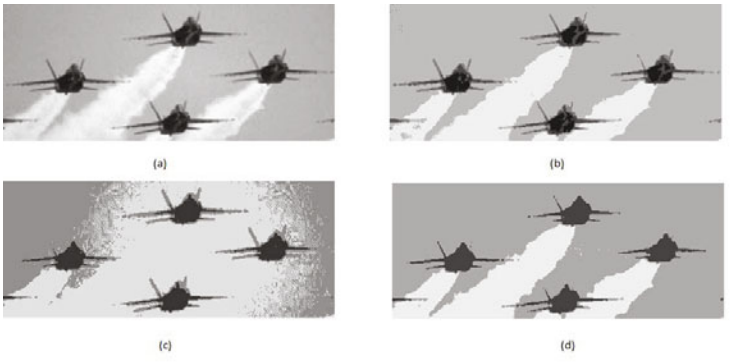


Fig. 1. (a) Original image, (b,c,d) result obtained by applying SC, FCM and IT2-SC

In Fig.1(a) shows an image whose principal entities (sky,smoke, air-planes) were almost homogeneous in the intensity of their gray level. Fig.1 b-d show results for SC, FCM and the IT2-SC. Results of image segmentation were obtained by our proposed IT2-SC and SC with the original initialization parameters, respectively, $\overline{\varepsilon} = 0.5$, $\underline{\varepsilon} = 0.15$, $r_a = 0.5$ and $\eta = 1.5$. For our proposed IT2-SC, we use two fuzzifiers, respectively,$m_1 = 1.7$ and $m_2 = 2.3$. In FCM, we use initialization values with $c = 3$ and $\varepsilon = 0.00005$

We can see that the segmentation result in Fig.1 (d) is the best. Since then, we can easily observe regions of image.

Fig.2(a) shows an image whose principal entities (rocks, sky, ground) were almost homogeneous in the intensity of their gray level. Figs.2 b-d show results for SC, FCM and our proposed IT2-SC, respectively. Results of image segmentation were obtained by our proposed IT2-SC and SC with the original initialization parameters, respectively, $\overline{\varepsilon} = 0.5$, $\underline{\varepsilon} = 0.15$, $r_a = 0.5$ and $\eta = 1.5$. For our proposed IT2-SC, we use two fuzzifiers, respectively,$m_1 = 1.7$ and $m_2 = 2.3$. In FCM, we use initialization values with $c = 3$ and $\varepsilon = 0.00005$

Fig. 2. (a) Original image, (b,c,d) result obtained by applying SC, FCM and IT2-SC

We can see that the segmentation result in Fig.2 (d) is the best. Since then, we can easily observe regions of image.

The summarised results in Table 1 and segmentation results on real images shows that the our proposed IT2-SC algorithm gives good results.

Table 1. Results of image segmentation for SC, FCM and our proposed IT2-SC

Algorithm	Segmentation results (Clusters)	
	Image 1	Image 2
SC	4	5
FCM	3	3
IT2SC	3	3

HSI model is used to segment color images. In our experiments, input color images are converted into HIS color images, which are more closer to human vision. Also, we used our proposed IT2-SC to classify I component which represents intensity of the images. Then, with the membership function of pixel we get by our proposed IT2-SC and H component of the image, we get a new feature for HIS color images. Finally, we converted new HIS color images to RGB color images and used our proposed IT2-SC to RGB color images segmentation.

Fig.3 (a) shows a color image of fruit. Fig.3 (b) shows results of segmentation of SC with the original initialization parameters, respectively, $\overline{\varepsilon} = 0.5$, $\underline{\varepsilon} = 0.15$, $r_a = 0.25$ and $\eta = 1.5$. The experiment only uses RBG model. Fig.3 (c) shows results of image segmentation of our proposed IT2-SC with the original initialization parameters, respectively, $\overline{\varepsilon} = 0.5$, $\underline{\varepsilon} = 0.15$, $r_a = 0.25$, $\eta = 1.25$, $m_1 = 1.85$ and $m_2 = 2.15$ on both HSI model and RGB model. We can see that the proposed algorithm can get more accuracy results.

Fig. 3. Color image segmentation results a) original image using SC c) using FCM d) using our proposed IT2-SC

5 Conclusion

The paper presents a new approach to image segmentation using the proposed IT2-SC. Through fuzziness parameter m, we can reduce the uncertainty of the algorithm in the initialization for parameters. Since then, we have expanded into interval type-2 fuzzy subtractive clustering algorithm by using two different values of fuzziness parameter m. Experiments on image segmentation are implemented in comparisons with other fuzzy clustering such as FCM or subtractive clustering.

For the future works, we will apply the proposed IT2-SC method to model type-2 and interval type-2 TSK fuzzy logic systems by the construction fuzzy rules. The second is to improve the computational performance by speeding up the algorithm using GPU.

Acknowledgment. This paper is sponsored by Vietnam's National Foundation for Science and Technology Development (NAFOSTED), Grant No 102.012010.12.

References

1. Chiu, S.L.: Fuzzy Model Identification Based on Cluster Estimation. Journal on Intelligent Fuzzy Systems 2, 267–278 (1994)
2. Chiu, S.L.: Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification. In: Dubois, D., Prade, H., Yager, R.R. (eds.) Fuzzy Information Engineering: a Guide Tour of Applications, pp. 149–162. Wiley, New York (1997)
3. Demirli, K., Cheng, S.X., Muthukumaran, P.: Subtractive clustering based modeling of job sequencing with parametric search. Fuzzy Sets and Systems 137, 235–270 (2003)
4. Chen, J.Y., Qin, Z., Jia, J.: A Weighted Mean Subtractive Clustering Algorithm. Information Technology Journal 7(2), 356–360 (2008) ISSN 1812-5638
5. Kim, D.W., Lee, K.Y., Lee, K.H.: A Kernel Based Subtractive Clustering Method. Pattern Recog. Lett. 26(7), 879–891 (2005)

6. Mendel, J., John, R.: Type-2 fuzzy set made simple. IEEE Trans. on Fuzzy Systems 10(2), 117–127 (2002)
7. Karnik, N., Mendel, J.M.: Operations on Type-2 Fuzzy Sets. Fuzzy Sets and Systems 122, 327–348 (2001)
8. Mendel, J.M., John, R.I., Liu, F.: Interval Type-2 Fuzzy Logic Systems Made Simple. IEEE Trans. on Fuzzy Systems 14(6), 808–821 (2006)
9. Rhee, F., Hwang, C.: A type-2 fuzzy C-means clustering algorithm. In: Proc. Joint Conf. IFSA/NAFIPS, pp. 1919–1926 (July 2001)
10. Hwang, C., Rhee, F.C.: Uncertain Fuzzy clustering: Intervel Type-2 Fuzzy Approach to C-means. IEEE Trans. on Fuzzy Systems 15, 107–120 (2007)
11. Fazel Zarandi, M.H., Zarinbal1, M., Turksen, I.B.: TypeII Fuzzy Possibilistic C-Mean Clustering. In: IFSA-EUSFLAT (2009)
12. Zhang, W.B., Liu, W.J.: IFCM: Fuzzy Clustering for Rule Extraction of Interval Type-2 Fuzzy Logic System. In: Proc. IEEE Conf. on Decision and Control, pp. 5318–5322 (2007)
13. Shen, H.-y., Peng, X.-q., Wang, J.-n., Hu, Z.-k.: A Mountain Clustering Based on Improved PSO Algorithm. In: Wang, L., Chen, K., S. Ong, Y. (eds.) ICNC 2005. LNCS, vol. 3612, pp. 477–481. Springer, Heidelberg (2005)
14. Shen, H.Y., Peng, X.Q., Wang, J.N.: Quick mountain clustering based on improved PSO algorithm. Journal of Systems Engineering 22(3), 333–336 (2006)
15. Yang, M.S., Wu, K.L.: A modified mountain clustering algorithm. Pattern Analysis and Applications 26(8), 125–138 (2005)
16. Demirli, K., Muthukumaran, P.: Higher Order Fuzzy System identification Using Subtractive Clustering. J. of Intelligent and Fuzzy Systems 9, 129–158 (2000)
17. Demirli, K., Cheng, S.X., Muthukumaran, P.: Subtractive clustering based on modeling of job sequencing with parametric search. Fuzzy Sets and Sys. 137, 235–270 (2003)
18. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
19. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(7), 881–893 (2002)
20. Cinquea, L., Forestib, G., Lombardic, L.: A clustering fuzzy approach for image segmentation. Pattern Recognition 37, 1797–1807 (2004)
21. Ali, M.A., Dooley, L.S., Karmakar, G.C.: Object Based Image Segmentation Using Fuzzy Clustering. In: ICASSP 2006: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 14-19, vol. 2, p. II (2006)
22. Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J.: Fuzzy c-means clustering with spatial information for image segmentation. Computerized Medical Imaging and Graphics 30, 9–15 (2006)
23. Dong, G., Xie, M.: Color Clustering and Learning for Image Segmentation Based on Neural Networks. IEEE Trans. on Neural Networks 16(4), 925–936 (2005)
24. Hong, T.-P., Wu, C.-H.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. Journal of Information Hiding and Multimedia Signal Processing 2(2), 173–184 (2011)
25. Kaganami, H.G., Ali, S.K., Zou, B.: Optimal Approach for Texture Analysis and Classification based on Wavelet Transform and Neural Network. Journal of Information Hiding and Multimedia Signal Processing 2(1), 33–40 (2011)

Improving Nearest Neighbor Classification by Elimination of Noisy Irrelevant Features

M. Javad Zomorodian[1,2], Ali Adeli[2], Mehrnoosh Sinaee[2], and Sattar Hashemi[2]

[1] Institute of Computer Science, Shiraz Bahonar Technical College, Shiraz, Iran
[2] Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran
{javadzomorodian,ali.adeli405,mehrnoosh.sinaee}@gmail.com,
s_hashemi@shirazu.ac.ir

Abstract. This paper introduces the use of GA with a novel fitness function to eliminate noisy and irrelevant features. Fitness function of GA is based on the Area Under the receiver operating characteristics Curve (AUC). The aim of this feature selection is to improve the performance of k-NN algorithm. Experimental results show that the proposed method can substantially improve the classification performance of k-NN algorithm in comparison with the other classifiers (in the realm of feature selection) such as C4.5, SVM, and Relief. Furthermore,this method is able to eliminate the noisy irrelevant features from the synthetic data sets.

Keywords: AUC, Genetic algorithm, Feature selection, Noisy feature elimination, k-NN.

1 Introduction

In non-parametric density estimation algorithms, the distribution of data is calculated without any particular assumption on its parameters. Two popular approaches in these algorithms are: Kernel Density Estimation (KDE) and k-Nearest Neighbor (k-NN). k-NN is a simple classifier that has been used in various real world applications. In some cases k-NN is vulnerable with some problems, such as few instances, noisy data and too many features, which decrease the performance of k-NN. To improve the performance of this classifier, many solutions are introduced. One of the approaches to solve those mentioned problems is searching in feature space to find optimal subset of features that can improve the classification accuracy of k-NN. This goal can be achieved by eliminating the irrelevant features in the noisy data sets.

In this paper, we attack this problem and introduce a novel algorithm to deal with. The proposed approach, firstly performs a feature selection using the GA and the AUC measure as the fitness function. In other words, simple GA (SGA) with statistical measure fitness function has been used to select optimal subset of features to achieve better classification accuracy.

This paper is organized as follows: in section 2, related work in this domain are reviewed. AUC and Receiver Operating Characteristics (ROC) curve are described in section 3. The proposed method is presented in section 4. Sections 5

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 11–21, 2012.

includes data set, experimental results and discussion, and conclusion is mentioned in the last section of paper.

2 Related Work

There is much research considering the problem of feature selection. It has been used when I) the number of features is larger than the number of training data, II) the number of features is too large for feasible computation III) many features includes noisy value. These issues can result in a significant drop in classification performance.

Many methods have been proposed for feature selection, i.e. Sequential Forward Selection (SFS) [17] is a greedy algorithm. The drawback of this method is that it is unable to remove obsolete features. Sequential Floating Forward Selection (SFFS) [19] and Sequential Floating Backward Selection (SFBS) [19] can capable to perform backward steps at each forward step. The neural network pruning is a method that has been introduced in this field [5]. Genetic algorithm is a complex method that has been proposed for feature selection [15]. In this solution, the GA explores the feature space to select optimal subset of features. Another method is innovated which uses correlation criteria such as Fast Correlation-Based filter Selection (FCBS) [16]. A new approach has been proposed by Das is fast filter method that hybrids boosting and incorporates some of the features of wrapper methods [6]. Song et al. [20] introduced a new method for feature selection which uses the Hilbert-Schmidt Independence Criterion (HSIC) as a measure of dependence between the features and the labels. In [20], backward elimination using HSIC (BAHSIC) is a filter method that is used for feature selection.

CFS (Correlation-based Feature Selection) is a method for feature selection that is studied by Hall [13]. This correlation-based method selects features which are highly correlated with the class. The Pearson's correlation was used to compute this correlation. Genetic programming (GP) proposed by Muni et al. to select best features [18]. The algorithm selects best features simultaneously and constructs a classifier based on the subset of features. For c-class problem, it provides a classifier having c-trees. Note that this proposed method introduced two new crossover operations, homogeneous and heterogeneous crossover operations. Bradley and Mangasarian [4] introduced another method, FSV (Feature Selection Concave method), which is a concave minimization approach to minimize a weighted sum of distances of misclassified points to two parallel planes that bound the sets. MFS (Multiple Feature Subsets) [11] that contains an ensemble structure of multiple NN classifiers. The results of all k-NN classifiers are combined in order to achieve strong accuracy. MFS (Multiple Feature Subsets) introduced by Bay [2] that contains an ensemble structure of multiple Nearest Neighbor (NN) classifiers.

3 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a two dimensional illustration of the classifier performance. It is suitable solution to analyze the classification accuracy in the binary class problem. For this purpose, the ROC curve calculates the True Positive rate (Sensitivity) versus False Positive rate (1 - Specificity). The ROC is a strong and statistical tool to compare binary classifiers. Sensitivity and Specificity are described in equations (1). To plot the ROC curve, the sensitivity and specificity need to be calculated as follows [10]: true positive (TP) is the number of predicted positive cases that are actually positive, true negative (TN) is the number of predicted negative cases that are actually negative, false positive (FP) is the number of predicted positive cases that are actually negative and false negative (FN) is the number of predicted negative cases that are actually positive.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}, \quad \text{Specifity} = \frac{\text{TN}}{\text{TN+FP}} \tag{1}$$

The AUC is a part of the area of the unit square. The AUC is a scalar value, in interval [0,1], to show the discriminative power of binary classifiers. If the value of AUC is less than "0.5", it shows undesirable result, but if the AUC value of classifier is close to "1", it shows a remarkable performance for binary classification. Equation (2) shows the AUC formula.

$$AUC = \sum_{k=1}^{n} (FP_k - FP_{k-1})(TP_k - TP_{k-1}) \tag{2}$$

4 Proposed Method

The aim of this study is to improve the classification accuracy of k-NN algorithm. One of the best solutions to this problem is to remove some noisy irrelevant features from the data set which can help the classification to be more accurate. In this respect, a great approach is needed to select the best features easily. For this purpose, this study has focused on the GA-based feature selection. The GA is one of the best and popular search tools. Note that the main difference between our method and the other GA-based feature selection method is that the fitness function of GA. Our new proposed fitness function is AUC. The parameters of the GA are shown in Table 1.

The algorithm works in the following process: First of all, the data set has been split to the unseen data and training sets. Next, the 10-fold cross validation has been used to validate k-NN. Then, a population of n chromosomes has been produced, randomly. After that, in each fold, the fitness function is called that will be described later. After the computation of fitness for all chromosomes, the evaluated population is used for recombination and mutation. The cycle of SGA will be described later. The evolutionary process has continued until the conditions are satisfied, i.e. the number of iteration equals to the predefined

Algorithm 1. Fitness Function

Input(s): Chromosome, Data set
Select features from data set based on chromosome
Compute the score of each instance based on (3)
Sort scores vector (decreasing)
Set $i \leftarrow 1$
for threshold :=1 down to 0 with step length 0.01 **do**
  Set instance $\leftarrow 1$
  while score(instance) $\geq$ threshold **do**
    if label of instance is positive **then**
      TP(i) $\leftarrow$ TP(i)+1
    else
      FP(i) $\leftarrow$ FP(i)+1
    end if
    instance $\leftarrow$ instance+1
  end while
  TP(i) $\leftarrow$ TP(i)/(no. of positive instances)
  FP(i) $\leftarrow$ FP(i)/(no. of negative instances)
  $i \leftarrow i$+1
end for
Call AUC function (TP,FP)
return AUC as the fitness of chromosome
Function AUC (TP,FP)
Set AUC $\leftarrow 0$
for i:=1 to length of TP vector **do**
  Base $\leftarrow$ FP(i+1) - FP(i)
  Height $\leftarrow$ TP(i+1) + TP(i);
  Set AUC $\leftarrow$ AUC + (Base $\times$ Height)/2
end for
return AUC
end function

value. After each fold, the training error should be computed with the validation set. For this propose, the features are selected (based on the genes with values "1" of the best chromosome) from the validation set. After that, the selected features are stored to k-NN algorithm for classification. After 10 folds, the best features are selected from the unseen data and then they have been used in k-NN. Finally, the testing error has been computed.

4.1 Genetic Algorithm (GA)

The evolutionary computing contains four main schemes, i.e. Genetic Algorithm (GA), Evolutionary Strategy (ES), Evolutionary Programming (EP) and Genetic Programming (GP). The GA, developed by Holland, includes several attributed features: it is not too fast, it is a well-laid, heuristic tool for complicated and compound problems [8]. The GA is suitable to optimize the problems with binary or integer representation [8]. Holland's original GA is now known as the simple genetic algorithm (SGA) [8].

In the cycle of SGA, the tournament selection with size k has been used as follows: after the computation of fitness, k chromosomes are selected randomly, and then the best two chromosomes (from this random set) have been selected for recombination and mutation. Note that the name of this method for the parent selection is "tournament 2 of k". The approach used for crossover operation is "uniform". In the "uniform" method, each gene has a probability value (chance) for recombination and if the probability value of the gene is smaller than or equal to the probability of doing the crossover (P_c), then the gene will be selected for crossover operation. In the mutation operation, if the probability value of the gene is smaller than or equal to the probability of doing the mutation (P_m) then the gene mutates. The operator type of mutation applied is "Bitwise bit-flipping". For survival selection, the two new chromosomes (offspring) have been replaced with two worst chromosomes in the population. The parameters of SGA has been outlined in Table 1. Note that in the population, the length of each chromosome is equal to the number of features. The gene with value "1" means that the feature should be selected from the data set while the gene with value "0" shows that the feature should be removed.

4.2 Fitness Function

The fitness function used in the SGA is based on AUC. In the AUC algorithm [1], each instance takes a probability score ($\varphi(i)$) based on its label of neighbors. To compute the probability, first, the distance of each sample to the others has been calculated with (4). Then, the labels of k-NN to this sample are considered. The score of each sample has been computed with (3),

$$Score(i) = \frac{no.\ of\ positive\ NN\ to\ sample\ i}{k} \tag{3}$$

K-NN is so sensitive to distance measure which will be shown in [7]. So, the selected metric measure is an important task. There are four properties that a metric measure must satisfy: (a, b and c are vectors) non-negativity, i.e. $D(a, b) \geq 0$, reflexivity, i.e. $D(a, b) = 0$ if and only if $a = b$, symmetry, i.e. $D(a, b) = D(b, a)$, triangle inequality, i.e. $D(a, b) + D(b, c) \geq D(a, c)$. The metric measure used in this paper is based on Minkowski metric. Equation (4) shows the equation of the noted distance [7].

$$L_k(a, b) = (\sum_{i=1}^{d} |a_i - b_i|^k)^{1/k} \tag{4}$$

It is also referred to as the L_k norm. So, the Euclidean distance is the L_2 norm, and the L_1 norm refers to the Manhattan one.

After the computation of scores, TP and FP rates are measured, and the ROC curve is plotted. At last, to compute the value of AUC, the area of all trapezoids, which is located under the ROC curve, is calculated. The summation of all these areas can be considered as the AUC and fitness of chromosome.

5 Experimental Results

In the testing phase, 10-fold cross validation used to validate empirical results. Table 1 shows the value of parameters in testing step. Meanwhile number of iterations are 200 or 500.

Table 1. Parameter setting of SGA

Parameters	Value (s)
Crossover operation	Uniform
Mutation operation	Bitwise bit-flipping
Probability of crossover, p_c	0.8
Probability of mutation, p_m	0.05 or 0.06
Representation	Binary
Number of population	100
Survival selection	Based on the fitness
Parent selection	Tournament 2 of 5

After each fold, the validation set has been used to compute the training error of k-NN classification. After 10 fold, the testing error of k-NN classification has been calculated using the testing set (unseen data). The results of the testing have been reported in Tables 3, 4, 5 and 6.

5.1 Data Set

The data sets used for the analysis of the model have been indexed in Table 2. Experimental results were achieved in two types. In the first type, our presented method deals with data sets which are mentioned in Tables 4, 5 and 6. All data sets were chosen from UCI repository [3]. The last two data sets of Table 2, "Arcene" and "Madelone", were taken from [12]. Next type of testing is applied in order to analyze the behavior of the proposed method on generated irrelevant features (Table 3). For this purpose, our method tested on the synthetic data sets which are randomly generated with some relevant and irrelevant features [21]. All the synthetic data sets contain 500 samples in the binary class distribution. The values of all features (relevant/irrelevant) are randomly picked from distribution in interval [0,10]. In all synthetic data sets, a data point belongs to positive class if the average value of relevant features for this instance is smaller than the threshold; otherwise it belongs to negative class. The threshold is set as the average values of all features in whole data. So, for each data set, the threshold is deterministic. Table 3 compares the performance of the proposed method and the simple k-NN on the synthetic data sets.

5.2 Discussion

In this section, results of the proposed method are compared with the other feature selection methods. Experimental results show that the proposed method can

Table 2. The detail of each data set used in this paper

Data set	# features	# samples	# class
Glass	10	214	6
WDBC	30	569	2
Ionosphere	34	351	2
Pima	8	760	2
Sonar	60	208	2
WBC	9	699	2
Iris	4	150	3
German	20	1000	2
Vote	16	435	2
Liver	7	345	2
Hepatitis	19	155	2
Credit	14	690	2
Arcene	500	2,600	2
Madelon	10,000	200	2

improve the classification performance of the k-NN. Furthermore, in a number of cases, the k-NN classifier with the proposed method can performs better than other classifiers such as SVM, C4.5, and Relief in the realm of feature selection. The results of the comparison have been demonstrated in Tables 4, 5 and 6.

In Table 3, effect of feature selector on k-NN classification is presented. For this purpose, we generated many data sets with different number of the relevant and irrelevant features. Experimental results show that in all cases of generated data set, the proposed method achieves better performance than the simple k-NN without selection mechanism.

In Table 4, the proposed method outperforms the rest, i.e. BAHSIC, FOHSIC, Pearson's correlation (PC) and mutual information (MI), on some data sets such as Sonar, Ionosphere, Arcence and Madelon. BAHSIC method uses backward elimination to select best features. The BAHSIC approach is unable to reevaluate the features which are removed previously while this GA-based method reevaluates each feature more than one. In BAHSIC, the OVA (One Versus All) has been used for multi class problem which adds extra complexity to the algorithm while in our method, the ROC measure can be used for the both binary and multi class problems without any increasing of complexity. In the each iteration of BAHSIC approach, 10% of current features are removed. This percentage is so critical for feature selection that may be caused inefficient feature selection methodology. In the proposed method, the percentage of the selected features varies according to the fitness function. One of the most important properties of BAHSIC is its ability to deal with the large scale data sets. Table 4 proves that the performance of the proposed method on the large scale data sets, i.e. Madelon and Arcene, is better than the BAHSIC approach. The PC and MI are unable to gain good performance. The reason of their's weak performance is that they process the usefulness of each feature independently. So, the nonlinear

Table 3. Empirical results of classic k-NN and the proposed approach when irrelevant features incorporated into the synthetic data sets

Synthetic data set		k-NN Error rate	
# relevant features	# irrelevant features	Proposed Method	Without Selection
4	6	**10.62**	25.01
5	5	**07.76**	24.54
5	8	**23.74**	24.87
10	10	**22.43**	26.98
10	12	**24.54**	29.73

Table 4. Comparison of classification errors

Data set	Method								
	BAHSIC	FOHSIC	PC	MI	RFE	RELIEF	L_0	R2W2	Proposed method
Ionosphere	12.31	12.85	12.37	13.11	20.27	11.73	35.90	13.71	**10.04**
Sonar	27.95	25.06	25.53	26.99	21.64	24.06	36.53	32.34	**16.71**
WBC	03.82	03.84	04.05	03.57	03.46	**03.17**	03.26	03.73	03.64
WDBC	**05.34**	**05.34**	05.37	06.72	07.76	07.24	16.75	06.83	08.83
Arcene	22.01	19.04	31.05	45.07	34.05	30.09	46.02	32.05	**10.01**
Madelon	37.98	38.07	38.46	51.60	41.58	38.67	51.31	100	**34.46**

interaction between features has been ignored. Furthermore in the both of the Ionosphere and the Sonar data sets, the GA helps k-NN algorithm to reach the lowest classification error.

In Table 5, our algorithm is compared with the CFS approach and FSV approach. In the CFS, the correlation between each feature and the class label is considered independently. In some cases, there are two features which the correlation between each feature and the class label is low, but the correlation between the combination of them as a features subset and the class label is high. However in the proposed method, each chromosome considers all features and their influences together on the classification. In this Table, comparison between the proposed method, the CFS based methods and the FSV algorithm illustrated that our idea outperforms than the others on the Sonar and WBC data sets. It means that consideration of all features together for the classification problem achieve better performance.

In Table 6, the proposed method is compared with five methods such as k nearest neighbor (k-NN), nearest neighbor with forward (FSS) and backward (BSS) sequential selection of features [1] and two types of MFS. The MFS algorithm is divided to two different groups, MFS1 and MFS2. In MFS1, sampling of the features is done with replacement, but in MFS2, sampling is done without replacement. The Table contains 10 data sets (Glass, Ionosphere, Iris, Pima, Sonar, German, Vote, Liver, Hepatitis and Credit). Our algorithm on 7 data sets performs better than others because of the following reasons:

Table 5. Comparison of classification accuracies

Data set	Method				
	CFS-based method			FSV	Proposed method
	NB	k-NN	C4.5		
WBC	72.53	73.26	70.82	92.23	**96.40**
Ionosphere	90.79	89.99	**90.94**	86.5	90.00
Pima	**76.22**	76.06	71.41	75.90	66.24
Sonar	65.46	79.79	70.82	72.25	**83.34**

1. In MFS, each classifier can access only to random subset of features. Random subset of feature is not a good idea to train the k-NN classifiers, because it is so sensitive to selected features. So, random selection approach can select irrelevant features (non-informative features) which reduce the accuracy of k-NN classifier and in this situation, ensemble idea is not efficient to improve the classification rate of k-NN classifier. In our method, feature selection is based on an evolutionary algorithm which chooses the best features finally.
2. Number of individual classifiers is important and effective on the voting decision. In the MFS approach, 100 classifiers is chosen. Experimental results show that the proposed method with a single k-NN classifier which is associated with the GA algorithm is better than the ensemble structure of 100 k-NN classifiers. Selection of classifiers is an external procedure that applies additional complexity to the MFS algorithm.
3. One of the problems of MFS is that the voting technique can be improved if each single classifier selects the correct class more often than any other class. The k-NN classifier requires a strong and intuitive feature selector to classify input instances correctly according to the selected features. In MFS algorithm, each classifier determines a label for each instance without any efficient feature selection algorithm. This can improve the accuracy rate of k-NN classifier. In our approach, the GA with a strong statistical fitness function, AUC, is associated with a single NN classifier to determine the label of instances.
4. As the mentioned in section related work, the main drawback of BSS and FSS methods is that they are unable to remove obsolete features. Furthermore experimental results illustrate that GA with AUC fitness function is able to select optimal feature subsets which improve the classification rate of k-NN in comparison to k-NN algorithms.

Our algorithm shows poor performance on Pima, Vote and Credit data sets. On Vote data set, our idea illustrate reasonable result in comparison with other methods. The reason of indigent performance on the Pima and Credit refers to small number of selected features which are 9 and 2 on the Credit and Pima data sets respectively.

Table 6. Classification accuracy when feature selection methods are used to improve k-NN

Data set	k-NN						
	NN	k-NN	FSS	BSS	MFS1	MFS2	Proposed Method
Glass	67.90	66.80	72.30	72.50	75.80	76.10	**90.70**
Ionosphere	84.50	85.50	85.20	86.90	85.50	86.80	**90.00**
Iris	94.30	95.10	93.70	93.50	92.50	92.70	**96.66**
Pima	69.70	73.60	67.70	68.50	72.50	**72.30**	66.24
Sonar	80.00	81.10	76.00	80.30	82.30	82.50	**83.34**
German	66.50	67.10	67.60	67.80	68.40	68.20	**71.00**
Vote	87.20	88.10	90.80	89.60	**91.90**	91.50	82.61
Liver	60.40	61.30	56.80	60.00	62.40	62.40	**65.87**
Hepatitis	74.20	74.40	76.30	75.20	77.70	76.80	**89.29**
Credit	81.60	83.50	82.70	81.60	**84.30**	83.80	81.74

6 Conclusion

In this paper, a new method has been introduced for feature selection. The proposed method of feature selection is based on SGA. The main property of SGA is the binary representation, which is an intuitive and interpretable way to select features. Length of chromosome is equal to the number of the features. The selection of the features is based on the fitness of the chromosome. The fitness of chromosome has been calculated based on the AUC, a statistical measure to compare classifiers on the binary class problem.

The experimental results show that the proposed method substantially improves the k-NN classification. In some cases, the proposed method helps the k-NN to result in more accurate classification than some other classifiers in the realm of feature selection such as C4.5, SVM, and Belief. Furthermore, experimental results on the synthetic data sets include irrelevant features shows that the proposed method is able to eliminate the irrelevant noisy features.The challenge of this method deals with the multi-class data set. For future work, this method will be developed for multi-class problems.

References

1. Aha, D.W., Bankert, R.L.: Feature Selection for Case-based Classification of Cloud Types: An Empirical Comparison, pp. 106–112. AAAI Press (1994)
2. Bay, S.D.: Combining nearest neighbor classifiers through multiple feature subsets. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998, pp. 37–45. Morgan Kaufmann Publishers Inc., San Francisco (1998)
3. Blake, L., Merz, C.J.: Uci repository of machine learning databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
4. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: ICML 1998, pp. 82–90 (1998)

5. Castellano, G., Fanelli, A.M., Pelillo, M.: An iterative pruning algorithm for feed-forward neural networks. IEEE Transactions on Neural Networks 8(3), 519–531 (1997)
6. Das, S.: Filters, wrappers and a boosting-based hybrid for feature selection. In: ICML 2001, pp. 74–81 (2001)
7. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, vol. 7. Wiley (1973)
8. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing, vol. 12. Springer, Heidelberg (2003)
9. Farzanyar, Z., Kangavari, M.R., Hashemi, S.: Effect of Similar Behaving Attributes in Mining of Fuzzy Association Rules in the Large Databases. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006, Part I. LNCS, vol. 3980, pp. 1100–1109. Springer, Heidelberg (2006)
10. Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. ReCALL 31(HPL-2003-4), 1–38 (2004)
11. Frohlich, H., Chapelle, O., Scholkopf, B.: Feature selection for support vector machines by means of genetic algorithms. In: Proc. 15th IEEE Int. Conf. on Tools with AI, pp. 142–148 (December 2003)
12. Gaudel, R., Sebag, M.: Feature selection as a one-player game. In: ICML 2010, pp. 359–366 (2010)
13. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand (April 1999)
14. Hashemi, S., Kangavari, M.R., Yang, Y.: Class specific fuzzy decision trees for mining high speed data streams. Fundam. Inform. 88(1-2), 135–160 (2008)
15. Lanzi, P.L.: Fast feature selection with genetic algorithms: a filter approach (1997)
16. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering 17(4), 491–502 (2005)
17. Marill, T., Green, D.: On the effectiveness of receptors in recognition systems. IEEE Transactions on Information Theory 9(1), 11–17 (1963)
18. Muni, D.P., Pal, N.R., Das, J.: Genetic programming for simultaneous feature selection and classifier design. IEEE Transactions on Systems Man and Cybernetics Part B Cybernetics a Publication of the IEEE Systems Man and Cybernetics Society 36(1), 106–117 (2006)
19. Pudil, P., Novovicová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters 15(11), 1119–1125 (1994)
20. Song, L., Smola, A.J., Gretton, A., Borgwardt, K.M., Bedo, J.: Supervised feature selection via dependence estimation. CoRR, p. 1 (2007)
21. Vivencio, D., Hruschka, E., Nicoletti, M., dos Santos, E., Galvao, S.: Feature-weighted k-nearest neighbor classifier. In: FOCI 2007, pp. 481–486 (2007)

Lattice Based Associative Classifier

Naveen Kumar, Anamika Gupta, and Vasudha Bhatnagar

Department of Computer Science, University of Delhi, India
{nk,agupta,vbhatnagar}@cs.du.ac.in

Abstract. Associative classification aims to discover a set of constrained association rules, called Class Association Rules (CARs). The consequent of a CAR is a singleton and is restricted to be a class label. Traditionally, the classifier is built by selecting a subset of CARs based on some interestingness measure.

The proposed approach for associative classification, called Associative Classifier based on Closed Itemsets (ACCI), scans the dataset only once and generates a set of CARs based on closed itemsets (ClosedCARs) using a lattice based data structure. Subsequently, rule conflicts are removed and a subset of non-conflicting ClosedCARs which covers the entire training set is chosen as a classifier. The entire process is independent of the interestingness measure. Experimental results on benchmark datasets from UCI machine repository reveal that the achieved classifiers are more accurate than those built using existing approaches.

Keywords: Associative Classification, Non-redundant, Closed Itemsets, Class Association Rules.

1 Introduction

Rule based classification methods aim to discover a set of rules in the database that forms an accurate classifier [13]. Association rule mining (ARM) aims to discover all association rules from a given dataset that satisfy a minimum support and confidence criterion [1]. Association rules for classification, termed as Class Association Rules(CARs), have been viewed as constrained association rules where the consequent must be a 1-itemset that forms the class label [9].

Liu et al. [9] introduced the first model for classification based on associations (CBA). The model generates a set of class association rules (CARs) using apriori based association rule mining algorithm and ranks the CARs according to the confidence-support-antecedent size criterion. Finally, a database coverage pruning mechanism is used to choose a subset of CARs, which forms a classifier, such that each chosen CAR covers at least one training example. In order to predict the class label of an unseen tuple, the tuple is matched against the rules of the classifier and the class label of the rule which matches it most closely is assigned to it. This technique of building classifier is known as associative classification technique and the resulting classifier is called associative classifier [2,4,6,7,8,9,11,12,16]. Understandability and amenability to human reasoning are the major strengths of associative classifiers.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 22–31, 2012.

However the large number of CARs to be examined for inclusion in associative classifiers leads to heavy expense in terms of both storage, and computation time for ranking and pruning [11,12]. The CARs generated using association rule mining algorithms may be redundant in the following sense: a rule is redundant if a more general rule exists with same support and confidence [15,16]. Another issue related to CARs is that there may exist several conflicting rules having same antecedent but different consequents [12]. These two issues restrain adoption of the technique, which otherwise finds strong favor in commercial applications where reasoning for decision is important.

A variety of interestingness measures like support, confidence, correlation coefficient etc. are employed to prune the set of CARs. Heravi and Zaiane [6] describe 53 probability based objective interestingness measures for associative classification and observe that employing different interestingness measures for ranking and pruning the CARs yields associative classifier of different accuracies. The choice of appropriate interestingness measures for the dataset at hand remains an unresolved challenge.

To solve the problem of too many CARs, we propose to generate CARs based on closed itemsets (ClosedCARs) which are inherently non-redundant and smaller in number than the complete set of CARs. This is followed by removal of conflicts from the set of ClosedCARs. Subsequently, a classifier is built by selecting a subset of non-conflicting ClosedCARs using the database coverage pruning mechanism. Generation of non-conflicting ClosedCARs is independent of interestingness measure and hence the resulting classifier is parameterless.

1.1 Proposed Approach

In this paper, we propose a lattice based approach for generating associative classifiers using closed itemsets (ACCI). We introduce the notion of ClosedCARs, which is a set of class association rules based on closed itemsets. Zaki et al. have established that association rules generated using closed itemsets are non-redundant [15]. It follows naturally that ClosedCARs are also non-redundant and hence smaller in number than the whole set of CARs.

The proposed single scan algorithm to generate set of non-conflicting ClosedCARs is used to induce a high quality associative classifier. The proposed approach does not employ any interestingness measure and hence is parameterless. Prediction of an unseen tuple is performed by selecting the rule whose antecedent has maximum number of attributes common with the tuple at hand. Experimental results on several datasets of UCI machine repository reveal that the classifiers generated using proposed ACCI approach are more accurate than the existing associative classifiers and also the classifier generated using traditional C4.5 method.

Rest of the paper is organized as follows: section 2 introduces associative classification and describes the algorithm CLICI for generation of closed itemsets, section 3 presents the related work, section 4 describes the proposed approach for generation of associative classifier using ClosedCARs and section 5 presents the experimental results. Finally section 6 concludes the paper.

2 Background

Associative Classification. First we introduce some terminology. Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n items. Let D denote the set of transactions (the dataset). A transaction $t \in D$ is a set of items. A set of one or more items is termed as an itemset. The support count of an itemset X is defined as the number of transactions in which X occurs. An itemset X is called **closed itemset** if there does not exist a proper superset Y of X such that support of Y is same as that of X.

An association rule is an implication of the form $A \rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \phi$. A and B are termed antecedent and consequent respectively of the rule. A rule has support s if $s\%$ of transactions in D contains $A \cup B$ and has confidence c if $c\%$ of transactions in D that contain A also contain B. Association rule mining aims to discover all rules that have support and confidence greater than the user-specified minimum support and minimum confidence threshold.

A classification rule associates a given set of items to a particular class c_i, where $c_i \in \mathcal{C}$, $\mathcal{C}$ being the set of class labels. Associative classifier is defined as a subset of all association rules where consequent of the rule is restricted to one class label. Thus rules in associative classifier are of the form $A \rightarrow c_i$.

Discovery of Closed Itemsets: CLICI Algorithm. Gupta et al. [5] have proposed an incremental algorithm, CLICI, for discovery of the closed itemsets in a given dataset. The algorithm obviates the need to generate all candidate itemsets. The discovered closed itemsets are stored in a data structure, called CILattice. CILattice has two components: a lattice $\mathcal{L}$ and a table ITable that maps items to their first use in the lattice. $\mathcal{L}$ is a complete lattice, with top node $\top$ and bottom node $\bot$. A node X of $\mathcal{L}$ represents a closed itemset I_X and stores its frequency f_X along with links to its parents and children nodes. ITable aids efficient traversal during search and insert operations.

We reproduce some of the terms used in the algorithm. Let $A(X)$, $D(X)$, $P(X)$ and $C(X)$ denote the set of ancestors, descendants, parents and children respectively of the node X in $\mathcal{L}$.

Definition 1. *A node X is an ancestor of node Y iff $I_X \subset I_Y (I_X \neq I_Y)$.*

Definition 2. *A node X is a descendant of node Y iff $I_X \supset I_Y (I_X \neq I_Y)$.*

Definition 3. *A node X is parent of a node Y if $X \in A(Y)$ and $\nexists$ any $Z \in A(Y) : X \in A(Z)$ and $Z \neq X$.*

Definition 4. *A node X is a child of node Y if $X \in D(Y)$ and $\nexists$ any $Z \in D(Y) : X \in D(Z)$ and $Z \neq X$.*

Observation 1. *If node X has a child node Y in $\mathcal{L}$ then the closed itemset of Y is minimal superset of all descendants of X. Similarly closed itemset of X is maximal subset of all ancestors of Y.*

Definition 5. *First Node F_i of an item i is a node X in the $\mathcal{L}$ where $i \in I_X$ and $\nexists$ a node Y such that $i \in I_Y$ and $Y \in A(X)$.*

It follows from the definition 5 that there is exactly one F_i for each item $i \in I$ and Itable stores the pair (i, F_i). Fig. 1 shows the example dataset and corresponding $< \mathcal{L}, Itable >$. We can observe here that although each node in $\mathcal{L}$ is a closed itemset, it may or may not have a class label. For details of procedures for insert, delete and search in $\mathcal{L}$, please refer to [5].

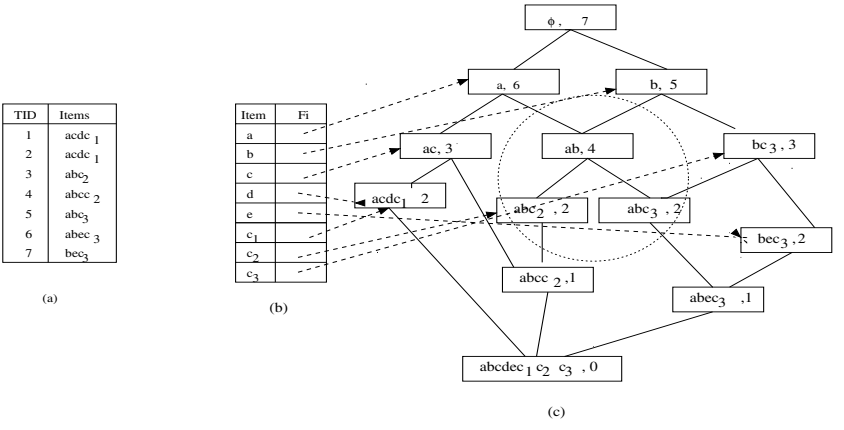


Fig. 1. (a) A toy dataset (b) Header table with items and pointer to nodes having their first use (c) $\mathcal{L}$ with nodes storing closed itemset and its support

3 Related Work

Associative classification is a three step process i) discovery of class association rules (CARs) ii) ranking and sorting the CARs according to a pre determined criterion iii) selecting a subset of rules using a suitable pruning mechanism. The selected rule set forms the associative classifier.

Several associative classification algorithms use apriori based methods for generation of CARs [2,4,7,9,14,16] while CMAR algorithm [8] uses FP-growth method. All these algorithms make multiple passes over the data to discover CARs. CPAR algorithm [14] is based on predictive association rule mining and adopts a greedy approach to generate rules directly from training data.

Different strategies are used for ranking the rules and building classifier. CBA method, developed by Liu et al [9], ranks the rules according to the decreasing order of precedence based on confidence, support and order of generation. CMAR algorithm [8] ranks the rules according to confidence, support, and number of attributes in the antecedent. It selects general and high confidence rules, selecting only positively correlated rules using chi-square testing. For database coverage, given a tuple, upto δ rules covering the tuple are considered unlike CBA [9] which selects only the first rule covering it. CPAR uses expected accuracy measure to select the rules for the classifier. Antonie et al [2] use correlation coefficient as

a measure for pruning the weakly correlated rules. Deng et al. [4] use certainty factor as a measure of pruning while Lan et al. [7] use intensity of implication and dilated chi-square as two interestingness measures. The GEAR algorithm [16] removes the redundant CARs, thus presenting a compact ruleset. Coenen et al. [3] show the effects of changing the support and confidence thresholds on the accuracy of the classifier.

CBA algorithm [9] performs prediction by selecting the first rule of the classifier that matches the unseen tuple. CMAR algorithm [8], CPAR algorithm [14] and the algorithm by Antonie et al [2] select the best k rules using weighted chi-square, expected accuracy and average confidence respectively, as a measure for evaluating the strength of each selected rule. Class label of the strongest rule is assigned as the class label of the unseen tuple.

4 Approach for Generation of Associative Classifier Using Closed Itemsets (ACCI)

The proposed approach, ACCI, introduces the notion of ClosedCARs, which is a set of class association rules based on closed itemsets. A lattice based data structure serves the dual purpose of storing both the closed itemsets and ClosedCARs. Generation of ClosedCARs is accomplished in two stages. Closed itemsets are discovered using the CLICI algorithm described in section 2, followed by generation of non-conflicting ClosedCARs. Subsequently, a subset of ClosedCARs is selected using the pruning mechanism based on database coverage. Following sections describe ACCI approach for generation of associative classifier and prediction method for assigning a class label to an unseen tuple.

4.1 Generate ClosedCAR

A formal definition of Closed Itemsets based Class Association Rules (ClosedCARs) is given below.

Definition 6. *A rule of the form $A \rightarrow c$ is a ClosedCAR if i) it is CAR i.e. $A \subset I$ and c is a class label. ii) either A or $A \cup c$ or both are closed itemsets.*

Table 1 shows the set of ClosedCARs for the toy dataset given in Fig. 1(a). As mentioned earlier, set of ClosedCARs, based on closed itemsets, is a reduced representation of set of CARs. All CARs along with their support and confidence can be derived from the set of ClosedCARs analogous to the work of Pasquier et al. [10]. Further rules generated using the set of ClosedCARs are non-redundant [15]. Before describing genCCAR algorithm, we present some observations.

Observation 2. *For all $X \in P(\perp)$*

1. *$\exists$ only one child of X i.e. $C(X) = \{\perp\}$.*
2. *X has a unique class label $c \in \mathcal{C}$.*
3. *There exist at least two nodes $X_1 \in P(\perp)$ and $X_2 \in P(\perp)$ having different class labels.*

This implies that if I_X does not include a class label i.e. $I_X \cap \mathcal{C} = \phi$ then $X \notin P(\perp)$.

Observation 3. *For a node $X \in L$, if $X \notin P(\perp)$ then $|C(X)| > 1$.*

Observation 4. *For a node $X \in \mathcal{L}$,*

1. *if $I_X \cap \mathcal{C} = c_i$ where $c_i \in \mathcal{C}$ then $I_Y \cap \mathcal{C} = c_i$ for all descendants Y of X.*
2. *if $I_X \cap \mathcal{C} = \phi$ then $|C(X)| > 1$ i.e. $|D(X)| > 1$. Two cases may arise here:*
   (a) all descendants of X have same class label.
   (b) at least two descendant nodes of X have different class label.

Algorithm genCCAR . Algorithm genCCAR traverses $\mathcal{L}$ from the top node $\top$ and visits each node, processing it as per the following conditions:

1. $I_X \cap \mathcal{C} = c_i$, where $c_i \in \mathcal{C}$: Generate a ClosedCAR $R : (I_X - c_i) \rightarrow c_i$ with s(R) = $s(I_X)$, c(R) = $s(I_X)/s(I_X - c_i)$.
2. $I_X \cap \mathcal{C} = \phi$: Search for all descendants Y of X such that Y has a class label. According to observation 4, there exists more than one such descendant.
   (a) Class label of all descendants Y of X is same. A Close CAR is generated $R : I_X \rightarrow c_i$ where c_i is class label.
   (b) There exist more than one descendant Y of X such that their class labels are distinct. For each such descendant Y, a ClosedCAR is generated $R : I_X \rightarrow c_j$ where $c_j \in \mathcal{C}$ is the class label of Y.

For the dataset given in Fig. 1, the generated set of all ClosedCARs by above algorithm is shown in Table 1.

Table 1. Set of ClosedCARs

ClosedCAR (Support (%), Confidence (%))	
1. $a \rightarrow c_1$ (28.6, 33.3)	8. $ab \rightarrow c_2$ (28.6, 50)
2. $a \rightarrow c_2$ (28.6, 33.3)	9. $ab \rightarrow c_3$ (28.6, 50)
3. $a \rightarrow c_3$ (28.6, 33.3)	10. $be \rightarrow c_3$ (28.6, 100)
4. $b \rightarrow c_2$ (28.6, 40)	11. $acd \rightarrow c_1$ (28.6, 100)
5. $b \rightarrow c_3$ (42.85, 60)	12. $abc \rightarrow c_2$ (14.3, 100)
6. $ac \rightarrow c_1$ (28.6, 66.6)	13. $abe \rightarrow c_3$ (14.3, 100)
7. $ac \rightarrow c_2$ (14.3, 33.3)	

Table 2. Set of Non-conflicting ClosedCARs

ClosedCAR (Support (%), Confidence(%))	
1. $a \rightarrow c_1$ (28.6, 33.3)	5. $be \rightarrow c_3$ (28.6, 100)
2. $b \rightarrow c_3$ (42.85, 60)	6. $acd \rightarrow c_1$ (28.6, 100)
3. $ac \rightarrow c_1$ (28.6, 66.6)	7. $abc \rightarrow c_2$ (14.3, 100)
4. $ab \rightarrow c_2$ (28.6, 50)	8. $abe \rightarrow c_3$ (14.3, 100)

4.2 Removing Conflicts

Consider Fig. 1 (c) again, now focusing on the circled nodes. Two conflicting rules $ab \rightarrow c_1$ and $ab \rightarrow c_2$ are generated using the algorithm described in previous section. It is imperative to resolve such conflicts to generate an accurate associative classifier. We define a *Conflict Set* which is a set of nodes that will result into conflicting rules.

Definition 7. *If a node $X \in \mathcal{L}$ has at least two descendant nodes Y_i, $i \in 1,..,|D(X)|$, such that $I_{Y_i} = I_X \cup c_{\lambda_i}$, c_{λ_i} is a class label, then node X and all such Y_is constitute the Conflict Set. X is called the leader of that set.*

The procedure for removing conflicts traverses the lattice $\mathcal{L}$ once and identifies all the *Conflict Sets* using definition 7. It may be noted that a node of $\mathcal{L}$, other than the nodes of *Conflict Sets*, always has a class label and gives rise to exactly one ClosedCAR with class label as the consequent and the remaining items as the antecedent of that rule.

Conflict Set results into conflicting rules. Conflict is resolved by selecting only one rule from each *Conflict Set* . Leader X of the *Conflict Set* constitutes the antecedent of the rule and consequent of the rule is the class label of the most confident node in the *Conflict Set* . Node $X \in \mathcal{L}$ is labeled with that class label and is marked non-conflicting while rest of the nodes of *Conflict Set* are marked conflicting. Nodes marked conflicting do not participate in the classifier building process. The set of non-conflicting ClosedCARs generated using above mentioned procedure is shown in Table 2.

Although conflicts are removed from the set of ClosedCARs, still they may arise during the prediction phase. An unseen tuple may not match fully to any of the rules present in the classifier but may match partly to more than one rule having different class labels. Such conflicts can be resolved only during the prediction phase (Section 4.4).

4.3 Building a Classifier

Database coverage pruning mechanism is used to select a subset of non-conflicting ClosedCARs which acts as a classifier. Database coverage method tests the generated rules against the training dataset and rules that cover at least one training object are kept in the classifier. The selected non-conflicting ClosedCARs, represented as nodes in $\mathcal{L}$, are marked and only the marked nodes of $\mathcal{L}$ are used in prediction of unseen tuple.

4.4 Prediction of Unseen Tuple

In the prediction phase, unseen tuples are assigned class labels. The proposed method assigns the class of the rule that has maximum number of attributes common with the unseen tuple at hand. If there is no rule having attributes common with the tuple at hand, then the majority class in the training data set, called the default class, is assigned as the class of unseen tuple. However, in our experiments with sixteen datasets, such a situation was never encountered. The process of prediction of a class label for an unseen tuple using the lattice $\mathcal{L}$ of ClosedCARs is explained below:

Let t denote an unseen tuple. We need to find, whether a node X exists in $\mathcal{L}$ such that t equals I_X or t is a subset of I_X. If such a node exists then the class label of that node is assigned to t. Otherwise for all items $i \in t$, F_i and all descendants Y of F_i are searched. Class label of the descendant having maximum

number of common attributes is assigned to t. If more than one descendant exists having same number of common attributes then the node denoting rule with maximum confidence is chosen and if confidence is also same then the most frequent rule is chosen. If no such descendant exists i.e. none of the item of t is present in $\mathcal{L}$, then the label of most frequently occurring class in the training set, known as default class, is assigned to t.

5 Experimental Analysis

In this section, we report the experimental evaluation of the proposed approach. The algorithm was implemented in C++ and run on Linux platform. We experimented on 16 well known benchmark datasets in the UCI machine learning repository[1]. The proposed approach, ACCI, is applied after discretizing the numeric datasets using entropy based method by Weka[2]. Missing values are replaced using means for numerical and modes for nominal attributes using Weka.

5.1 Comparison of Number of ClosedCARs and CARs

We first experimented to validate our hypothesis that the number of ClosedCARs is significantly less than the number of CARs. We tried generating CARs using Weka software using the support threshold of 0%. We found that it was not possible to generate the complete set of CARs in some datasets like hepatitis, lymph, zoo etc. Thus we set the support threshold to 1% for generation of CARs. Although ACCI approach is independent of interestingness measures, but for fair comparison, we used the same support threshold for generation of ClosedCARs. Fig. 2 shows the ratio of number of CARs and number of ClosedCARs generated on different datasets. We observe that the number of ClosedCARs is much less than the number of CARs for ten of the sixteen datasets. For rest of the datasets the ratio is higher, maximum being 4.8.

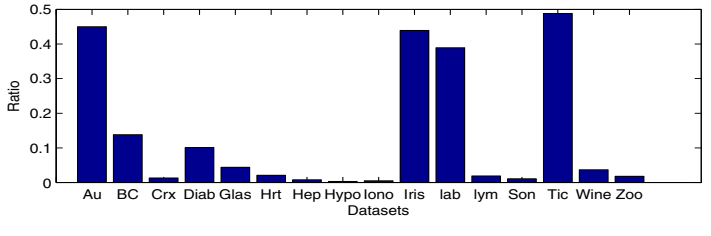


Fig. 2. Ratio of no. of CARs and no. of ClosedCARs

[1] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
[2] http://www.cs.waikato.ac.nz/~ml/weka

5.2 Comparison of Accuracy of ACCI with Existing Approaches

Next we compare the performance of ACCI with other associative classification approaches and one of the most popular tree classification approach i.e. C4.5 (Java version J48). J48 (with optimal parameters using CVParameterSelection option of Weka software) was executed on all the datasets and the accuracies were noted (col 6 of Table 3). CBA algorithm is run using the software available at National University of Singapore[3] with support threshold of 1% and confidence threshold of 50% as experimented by authors in [9] (col 7). Results of CMAR and CPAR algorithm on different datasets are reported from [8] and [14] respectively (col 8 and 9). Col. 10 shows results of proposed ACCI approach.

Table 3 shows that the accuracy of the ACCI classifier is higher than that of other classifiers for thirteen out of sixteen datasets. It is marginally lower for the remaining three datasets. Since ACCI does not employ an interestingness measure for pruning, its performance is independent of the changes in thresholds.

Table 3. Comparison of Accuracy of Proposed Classifier with Existing Classifiers

	Data Set	No. of Attr	Class	No. of Inst	J48	CBA	CMAR	CPAR	ACCI
1	Australian (Au)	14	2	690	86.5	84.9	86.1	86.2	**88.98**
2	BreastC-W (BC)	9	2	699	95.9	96.3	96.4	96.0	**97.91**
3	Crx (Crx)	15	2	690	86.81	84.7	84.9	85.7	**94.49**
4	Diabetes (Diab)	8	2	768	78.12	74.5	75.8	75.1	**88.42**
5	Glass (Glas)	9	7	214	76.1	73.9	70.1	74.4	**90.47**
6	Heart (Hrt)	13	2	270	83.7	81.9	82.2	82.6	**90.74**
7	Hepatitis (Hep)	19	2	155	81.9	81.8	80.5	79.4	**93.3**
8	Hypo (Hypo)	25	2	3163	**99.2**	98.9	98.4	98.1	98.89
9	Iono (Iono)	34	2	351	90	92.3	91.5	92.6	**96.57**
10	Iris (Iris	4	3	150	**95.3**	94.7	94.7	94	94.7
11	labor (lab)	16	2	57	79.3	86.3	89.7	84.7	**94**
12	Lymph (lym)	18	4	148	76.3	77.8	83.1	82.3	**91.42**
13	Sonar (Son)	60	2	208	79.8	77.5	79.4	79.3	**81.5**
14	TicTacToe (Tic)	9	2	958	99.4	**99.6**	99.2	98.6	99.26
15	Wine (Wine)	13	3	178	93.2	95	95	95.5	**98.24**
16	Zoo (Zoo)	16	7	101	93.1	96.8	97.1	95.1	**97**

6 Conclusion and Future Work

In this paper, we have proposed a novel approach for associative classification that generates class association rules from closed itemsets (ClosedCARs). We use a compact lattice based data structure for storing the ClosedCARs that helps in resolving the rule conflicts and in prediction phase. Experimental study reveals that the achieved classifiers are accurate. In future we intend to increase the understandability of the generated classifier by minimizing its size.

[3] Downloaded from http://www.comp.nus.edu.sg/~dm2

References

1. Agarwal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: 20th International Conference on Very Large Databases, pp. 487–499 (1994)
2. Antonie, M.L., Zaiane, O.R.: An Associative Classifier Based on Positive and Negative Rules. In: Proceedings of the 9th SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 64–69. ACM Press (2004)
3. Coenen, F., Leng, P.: Obtaining Best Parameter Values for Accurate Classification. In: Proceedings of 5th International Conference on Data Mining, pp. 549–552 (2005)
4. Deng, Z., Zheng, X.: Building Accurate Associative Classifier Based on Closed Itemsets and Certainty Factor. In: IEEE Third International Symposium on Intelligent Information Technology Application Workshop, pp. 141–144 (2009)
5. Gupta, A., Bhatnagar, V., Kumar, N.: Mining Closed Itemsets in Data Stream Using Formal Concept Analysis. In: Bach Pedersen, T., Mohania, M.K., Tjoa, A.M. (eds.) DAWAK 2010. LNCS, vol. 6263, pp. 285–296. Springer, Heidelberg (2010)
6. Heravi, M.J., Zaiane, O.R.: A Study on Interestingness Measures for Associative Classifiers. In: ACM Symposium on Applied Computing (2010)
7. Lan, Y., Chen, G., Wets, G.: Improving Associative Classification by Incorporating Novel Interestingness Measures. In: IEEE International Conference on e-Business Engineering (2005)
8. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class Association Rules. In: Proceedings of the International Conference on Data Mining, pp. 369–376 (2001)
9. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (1998)
10. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient Mining of Association Rules using Closed Itemset Lattices. Journal of Information Systems 24(1), 25–46 (1999)
11. Sun, Y., Wong, K.C., Wang, Y.: An Overview of Associative Classifiers. In: Proceedings of the International Conference on Data Mining, DMIN, pp. 138–143 (2006)
12. Thabtah, F.: A Review of Associative Classification Mining. The Knowledge Engineering Review 22, 37–65 (2007)
13. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
14. Yin, X., Han, J.: CPAR: Classification Based on Predictive Association Rules. In: Proceedings of SIAM Conference on Data Mining, pp. 331–335 (2003)
15. Zaki, M.J.: Generating Non-Redundant Association Rules. In: 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 34–43 (2000)
16. Zhang, X., Chen, G., Wei, Q.: Building a Highly-compact and Accurate Associative Classifier. Journal of Applied Intelligence 34, 74–86 (2011)

An Efficient Clustering Algorithm Based on Histogram Threshold

Shu-Ling Shieh[1,*], Tsu-Chun Lin[2], and Yu-Chin Szu[3]

[1] Department of Information Networking and System Administration, Ling Tung University, Taichung, Taiwan
ltcc63@teamail.ltu.edu.tw
[2] The Graduate Institute of Applied Information Technology, Ling Tung University, Taichung, Taiwan
rikuaono@hotmail.co.jp
[3] Department of Information Management, St. John's University, Taipei, Taiwan
belinda@mail.sju.edu.tw

Abstract. Clustering is the most important task in unsupervised learning and clustering validity is a major issue in cluster analysis. In this paper, a new strategy called Clustering Algorithm Based on Histogram Threshold (HTCA) is proposed to improve the execution time. The HTCA method combines a hierarchical clustering method and Otsu's method. Compared with traditional clustering algorithm, our proposed method would save at least ten several times of execution time without losing the accuracy. From the experiments, we find that the performance with regard to speed up the execution time of the HTCA is much better than traditional methods.

Keywords: Data mining, Clustering Algorithm, Histogram threshold.

1 Introduction

Data mining technologies are usually used for anthropology, social science, biology and the others [1]. Clustering technology is one of the data mining technologies. Clustering algorithm also moves forward, many methods have proposed in the past 50 years [2-6]. A part of clustering methods has proposed by combining two or three different method. Whether we can combine several methods in different area and develop a new clustering algorithm has become a issue that is worth to study.

In image bi-level threshold, Otsu's method is a popular algorithm [7]. The Otsu's used for image compression, significantly reduced execution time [8]. However, the clustering algorithms are generally applied in large scale data set. These data sets are as simple as image data. In this paper, we propose an efficient clustering method to improve the performance of clustering problem. The proposed clustering algorithm bases on hierarchical clustering method combine with refined Otsu's method, histogram construction technique [9] and optimization techniques [10]. The algorithm

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 32–39, 2012.

is significantly reduced the clustering execution time and lower error rate than general other clustering methods.

The remaining sections of this paper are organized as follows: the next section briefly presents related work concerning the Otsu's algorithm and k-means algorithm; in the third section, we present a detailed description of our algorithm and experimental results of the data set from UCI KDD Archive are provided in the fourth section. The conclusions are given in the last section.

2 Related Work

The clustering method is an important technology of data mining. The clustering technology is often used to discover the patterns and the hidden information in a data set. Clustering main purpose is to gather the same of characteristic and to separate the different of characteristic. The same objective is making similarity rate maximum within a cluster and difference rate maximum between others cluster.

K-means is a popular clustering method. The main steps are as follows [11]:

1. Select initial centers for each cluster. (Clusters number is K)
2. Assign each data vector to a cluster with its closest cluster center.
3. Compute new cluster center for each cluster. Repeat step 1 and step 2 until cluster membership stabilizes.

K-means method is working simply and effectively in most case. The initial centers selecting has different way, such as averagely choose or randomly choose in data set [12]. The selection of initial central vector will much affect the result of clustering. Therefore in this paper we would like to make comparison between these two choosing strategies.

Otsu's method is a popular threshold method on image processing. The main idea of Otsu's method is to use a gray level of histogram to search an optimal bi-section separation vector. The bi-section separation vector minimizes the within-class variance or maximizing between-class variance. Otsu's method algorithm is as follow:

Suppose an image is composed of R scales of gray levels, and it is consists of N pixels. The number of pixels at the *i*-th level is n_i, then $\mathrm{N} = \mathrm{n}_0 + n_1 + \ldots + n_{R-1}$. The probability of pixel with the *i*-th level in the image will be

$$\mathrm{p}_\mathrm{i} = \frac{n_i}{N} \quad and \quad \mathrm{p}_\mathrm{i} \geq 0, \quad \sum_{i=0}^{R} p_i = 1 \tag{1}$$

Now we separate this image into two classes of pixels based on gray levels. We would get two classes of pixel sets, $\mathrm{C}_1 = \{n_0, n_1, \ldots, n_L\}$ and $\mathrm{C}_2 = \{n_{L+1}, n_{L+2}, \ldots, n_{R-1}\}$, the probabilities pixels in of each class would be

$$\omega_1 = \sum_{i=0}^{L} p_i \quad and \quad \omega_2 = \sum_{i=L+1}^{R-1} p_i = 1 - \omega_1 \tag{2}$$

The means for classes C_1 and C_2 are

$$\mu_1 = \sum_{i=0}^{L} \frac{ip_i}{\omega_1}, \quad \mu_2 = \sum_{i=L+1}^{R} \frac{ip_i}{\omega_2} \quad and \quad \mu_T = \sum_{i=0}^{R-1} ip_i \tag{3}$$

Then we would get an between-class variance σ_B and the within-class variance σ_W. While

$$\sigma_B{}^2 = \omega_1(\mu_1 - \mu_T)^2 + \omega_2(\mu_2 - \mu_T)^2 \tag{4}$$

$$\sigma_W{}^2 = \omega_1\sigma_1{}^2 + \omega_2\sigma_2{}^2 \quad where \quad \sigma_1{}^2 = \sum_{i=0}^{L}(i-\mu_1)^2 \frac{p_i}{\omega_1}, \quad \sigma_2{}^2 = \sum_{i=L+1}^{R}(i-\mu_2)^2 \frac{p_i}{\omega_2} \tag{5}$$

Calculate the between-class variance repeatedly to obtain a maximum.

$$\sigma_B{}^2(L') = \max_{0 \le L \le (R-1)} \sigma_B{}^2(L) \tag{6}$$

Or compute the within-class variance repeatedly to obtain a minimum.

$$\sigma_W{}^2(L') = \min_{0 \le L \le (R-1)} \sigma_W{}^2(L) \tag{7}$$

The gray level L is the best threshold at this image.

Our proposed method has combined by hierarchical clustering with Otsu's threshold method. We propose an efficient clustering method to improve the performance of clustering problem. The algorithm is significantly reduced the clustering execution time and more accurately than the classical clustering algorithms.

3 Proposed Method

In this paper, a new strategy called Clustering Algorithm Based on Histogram Threshold (HTCA) is proposed for improving clustering performance by using Otsu's method to search the threshold by histogram. In the proposed HTCA algorithm is a hierarchical clustering method. The hierarchical clustering has two characteristic: will making a clustering tree to analyze each class how be clustered, and to apply this method to select threshold.

Suppose that there is a 2 dimensional data set with 150 data vectors with three known classes. Each step divide one class until the number of cluster is three ($k = 3$).

Otsu's method does threshold on gray level image. The method using the gray level histogram and calculate probability to select threshold gray level. For high dimensional data set, our HTCA method calculates the numbers of data vectors in every dimension and generates a histogram.

The main architecture of divisive hierarchical clustering algorithm will proceed k time of separating. The processes will at most generate $2k$-1 child classes, C_t, $t = 1,2,\ldots,2k-1$. While t is the index of the nodes in the tree. Let k is the number of groups after this separating process. Then this algorithm will create a binary tree to show clearly the separating status. Continue applying these steps and the result will be in k classes.

There are three main modules in HTCA. They are *Make_Histogram*, *Find_Splitting_Center* and *Assign_to_Cluster*. In the first step of HTCA, normalized data are assigned into the same class, which may be treated as root of a binary tree. The array N , $N = \{n_1, n_2, \ldots, n_{2k+1}\}$, is created based on the number of nodes in this tree. The initial root is the number of total records, and other nodes $n_2, n_3, \ldots, n_{2k-1} = 0$ in the tree. We use this array as the base of cutting criteria.

The separating process will be performed only when the number of data is greater than 0. They are those nodes with the property $n_t > 0$. When the cutting process is performed, the module *Make_Histogram* is called and a histogram H is built to compute the critical values based on the dimensions of the distribution of data. In this paper, the dimension data are normalized to be R distinct values. For example, if R=256 , then the range of data is separated into 256 regions, which represent the number of data vectors in each section, as shown in Fig. 1.

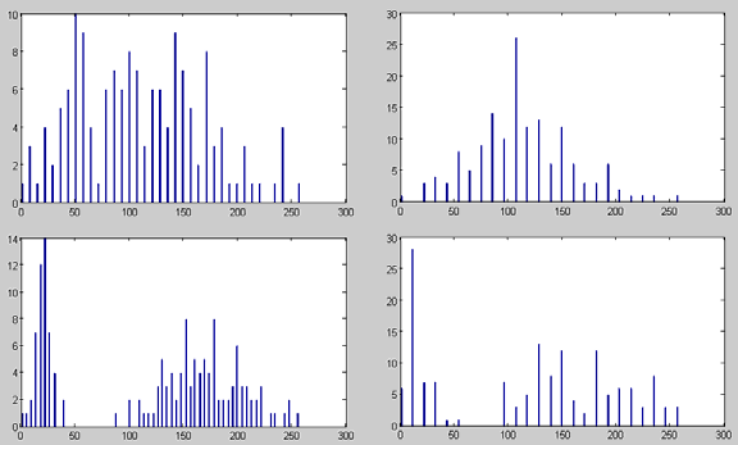


Fig. 1. Various dimensions' data vectors transform to the histogram on Iris data set ($R = 256$)

The values in x-axis represent the accumulated values of data vectors and the values in y-axis are the distribution of the normalized data in the range [1,256]. The left-upper diagram is the first dimension of the dataset. The right-upper one is the second dimension. The left-lower one is the third and the right lower one is the fourth dimension. From Fig.1 we get the distribution of vectors in the dataset and the result would be applied in the selection of cutting vector while applying the Otsu's method.

In the module *Find_Splitting_Center*, the Otsu's bi-level threshold algorithm is applied to find a cutting vector *L*. Then the medium of the parent class is calculated as m_1. Let $\frac{m_1+m_2}{2}=L$, i.e. $m_2=2L-m_1$. Then m_1 and m_2 are now the mediums of the two new child classes.

The procedure module *Assign_to_Cluster* uses the nearest adjacent distance criteria to assign the data vectors in the parent class to two child classes with the mediums m_1 and m_2 respectively, as shown in Fig. 2.

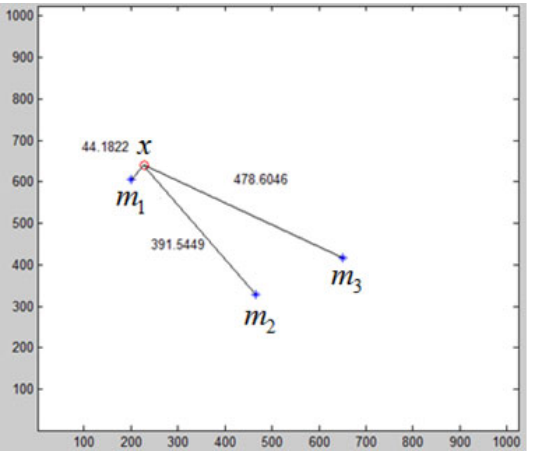


Fig. 2. The nearest neighbor method

Suppose that there is a data vector *x* with a Euclidian distance 44.1822; the distance from *x* to m_2 is 391.5449, and the distance from *x* to m_3 is 478.6046. The shortest distance among these three is the distance from *x* to m_1. Then this data vector *x* should be assigned to the same class with m_1.

HTCA algorithm is as follows:

The algorithm of HTCA•
(Input: $X=\{x_1,x_2,\ldots,x_m\}$, Output: $S=\{s_1,s_2,\ldots,s_k\}$)
Step 1: **(Initialization)**
$n_i=0, i=1,2,\ldots,2k-1$;
$C_1=X$;
$n_1=\mathrm{M}$;
$t=1$;
Step 2: **(Make dataset to Histogram)**
$H=$Make_Histogram(C_t);
Step 3: **(Find cluster center)**
if $n_t>0$ $[m_1,m_2]=$Find_Splitting_Center(C_t,H);

Step 4: **(Split dataset into subset)**
$[C_{2t}, C_{2t+1}]$ = Assign_to_Cluster(C_t, m_1, m_2);
Step 5: **(Iteration condition)**
$t = t + 1$;
if $t < k$ go to Step 2;
Step 6: **(Output clustering solution)**
for $i < 1$ to $2k - 1$
if $n_t > 0$ put C_t into S ;

Make_Histogram:
(Input: C_t , Output: H)
Make a histogram array H by C_t ;

Find_Splitting_Center:
(Input: C_t, H , Output: m_1, m_2)
Use Otsu's bi-level thresholding algorithm to calculate a best splitting-vector L .
Let m_1 is the center of C_t , and $m_2 = 2L - m_1$.

Assign_to_Cluster:
(Input: C_t, m_1, m_2 Output: C_{2t}, C_{2t+1})
Calculate the squared Euclidean distance for each C_t with m_1 and m_2 , that is, calculate the $d(x, m_1)$ and $d(x, m_2)$ for all $x \in C_t$.
If $d(x, m_1) <= d(x, m_2)$ put the data x into C_{2t} ;
Else if $d(x, m_2) < d(x, m_1)$ put the data x into C_{2t+1} ;

4 Experimental Results

All experiments were performed on the datasets from the University of California at Irvine – Machine Learning Repository [13] to illustrate the effectiveness of the proposed clustering algorithm. Two data sets were used to demonstrate the effectiveness of the proposed algorithm in the experiments. Iris data set has 150 data vectors with three known classes. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor).

WPBC is a dataset for the forecasting of breast cancer. There are 198 records of data vectors with two known classes. Eliminating one field of weakness, the dataset is distinguished as relapsing and non-relapsing ones.

The experiment compares these two methods in executed time and the rate of errors. The k-means use two strategies which are finding the medium vectors

correspondingly based on each section of the dataset and fining the medium vectors randomly. The latter method proceeds 10 times and generates an average. Using the execution time of HTCA as a basis, we find the execution time of K-means takes 13 to 16 times while proceeding WPBC dataset, as shown in table 1.

Table 1. Execution time comparison of HTCA, K-means and K-means(Random)

	HTCA	K-means	K-means(Random)
Iris	1	4.893406	6.958645
WPBC	1	16.45751	13.98443

The rate of error is defined as: 1. the number of data vectors that should be assigned into the same class but they are assigned into two different classes. 2. The number of data vectors that should be assigned into two different classes but they are assigned into the same class. The lower average of addition is, the better outcome is. This HTCA method not only raises the rate of clustering accuracy but also reduces the executing time, as shown in Table 2.

Table 2. The error rate of HTCA, K-means and K-means(Random)

	HTCA	K-means	K-means(Random)
Iris	12.54%	12.54%	12.71%
WPBC	42.45%	50.00%	48.20%

5 Conclusion

In this paper, we propose a new clustering algorithm named Clustering Algorithm Based on Histogram Threshold (HTCA) to speed up the execution time. The HTCA method combines a divided hierarchical clustering method and Otsu's method. Compared with traditional K-means algorithm, our proposed method would save at least ten several times of execution time without losing the accuracy. From the experiments, we find that the performance with regard to improved execution time of the HTCA is much better than traditional methods.

References

1. Hirschman, L., Park, J.C., Tsujii, J., Wong, L., Wu, C.H.: Accomplishments and challenges in literature data mining for biology. Bioinformatics 18, 1553–1561 (2002)
2. Berkhin, P.: Survey of clustering data mining techniques. Technique Report, Accrue Software (2002)
3. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31, 264–323 (1999)
4. Li, C., Biswas, G.: Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering 14, 673–690 (2002)

5. Rauber, A., Pampalk, E., Paralic, J.: Empirical evaluation of clustering algorithms. Journal of Information and Organizational Sciences, JIOS (2000)
6. Rui, X., Wunsch II, D.: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16, 645–678 (2005)
7. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics 9, 62–66 (1979)
8. Chang-Chin, H.: Efficient VQ Codebook Generation by Global/Local Clustering Algorithms (2009)
9. Patel, R., Shrawankar, U.N., Raghuwanshi, M.M.: Genetic Algorithm with Histogram Construction Technique. In: Proceedings of the 2009 Second International Conference on Emerging Trends in Engineering \& Technology, pp. 615–618. IEEE Computer Society (2009)
10. Sun, L., Lin, T.-C., Huang, H.-C., Liao, B.-Y., Pan, J.-S.: An Optimized Approach on Applying Genetic Algorithm to Adaptive Cluster Validity Index. In: Proceedings of the Third International Conference on International Information Hiding and Multimedia Signal Processing (IIH-MSP 2007), vol. 02, pp. 582–585. IEEE Computer Society (2007)
11. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data (1988)
12. Huang, Z.: Extensions of the K-means Algorithm for Clustering Large Data Sets with Categorical Values. In: Żytkow, J.M. (ed.) PKDD 1998. LNCS, vol. 1510, pp. 283–304. Springer, Heidelberg (1998)
13. Merz, C.J., Blake, C.L.: UCI repository of machine learning datasets. Department of Information and Computer Science. University of California, Irvine (1998)

A Resource Reuse Method in Cluster Sensor Networks in Ad Hoc Networks

Mary Wu[1], InTaek Leem[2], Jason J. Jung[1], and ChongGun Kim[1,*]

[1] Dept. of Computer Engineering, Yeungnam University, Korea
[2] Dept. of Game Creation, Deagu Mirae College, Korea
{Mrwu,j2jung}@ynu.ac.kr, itleem@empal.com, cgkim@yu.ac.kr

Abstract. Sensor nodes having the limited resource, energy efficiency is an important issue. Clustering on the sensor networks reduces the volume of inter-node communications and raises energy efficiency by transmitting the data collected from members by a cluster head via a sink node. But, due to radio frequency characteristics, interference and collision can occur between neighbor clusters, the resulted re-transmission is more energy consuming. The previous problems occurred between neighbor clusters can be resolved by assigning channels which do not overlap between neighbor clusters. In this paper, we propose a method which assigns and reuses channels which do not overlap between neighbor clusters in dynamic cluster sensor networks. The resource allocation model of the proposed method is analyzed by correctness and simplicity.

Keywords: Cluster sensor network, synchronization of inter-cluster, channel reuse, cluster topology matrix, resource allocation matrix.

1 Introduction

Wireless Sensor Networks collect data on the surrounding environment and can be applied to a variety of purposes such as intrusion detection in military areas, security areas, environmental monitoring of temperature and humidity. Sensor nodes aware around the symptoms, transmit measured data to a base station and the base station analyzes the data. The biggest limitation can be the limited resources of sensor nodes on the wireless sensor networks and many studies for efficient use of energy are actively underway to overcome this problem.[1-16]

Typically, neighboring sensor nodes collect similar information, so that the energy wasted due to duplicate transmission of similar information is large. From this perspective, in order to use energy efficiently, cluster methods on sensor networks have been studied for many. In clustering, each node belongs to a local cluster and a cluster head integrates the data collected from members of the cluster, and then transmits it to a sink node. This prevents duplicate transmission of similar information and make a low-power networking in the sensor networks.[8-16]

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 40–50, 2012.

On LEACH(Low-Energy Adaptive Clustering Hierarchy), a typical clustering protocol for sensor networks, cluster heads make the TDMA schedule of their cluster members and allocate it to each member node. Each node transmits data in its own transmission slot, operates in sleep mode in other times. This minimizes power dissipation due to idle listening.[8] However, the TDMA schedule of each cluster is generated independently, so that data transmission between a cluster head and members within a cluster causes collisions and interference for data transmission within neighbor clusters. In fig. 1, the radio range of a node affects in other clusters. In order to reduce this collisions and interference between the neighbor clusters, the cluster transmission control is required.

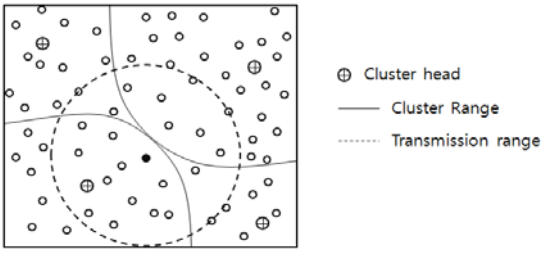
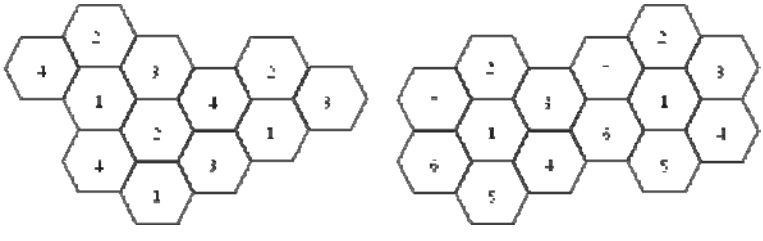


Fig. 1. Clustering and resource interference

Synchronization among clusters may give a chance to improve system performance. Timesharing or frequency allocation using a part of channels among neighbor cells is a method for inter-cluster synchronization. When the system is weak on the interference, increasing the frequency reuse factor reduces interference between cells, when the system is strong on the interference, reducing the frequency reuse factor makes clusters be allocated a wide channel bandwidth. Frequency reuse factor of 3, 4, 7, 9, 12, etc. is used.[17, 18]



a) Frequency reuse factor of 4 b) Frequency reuse factor of 7

Fig. 2. Frequency Reuse Pattern

In this study, a dynamic frequency allocation algorithm which does not overlap among neighbor clusters is proposed. This method uses a topology matrix which represents cluster topology and a resource allocation matrix which assigns a part of channels to cells for preventing overlap between neighbor clusters. Section 2 briefly

introduces for the role of head nodes on a cluster sensor network, section 3 introduces a topology matrix and a resource allocation matrix. 4 concludes conclusion.

2 Head Nodes in Clustering

Representative researches for the cluster-based routing protocol include LEACH, LEACH-C, HEED.[9,10] LEACH replaces cluster heads by random based on probability in order to evenly the energy consumption between nodes on the sensor network. At the start of each round, each node determines whether it acts as a header node with probability $P_i(t)$.

$$P_i(t) = \begin{cases} \dfrac{k}{N - k(r \mod \frac{N}{k})}, where \quad C_i(t) = 1, \\ 0 \qquad\qquad\qquad\quad , where \quad C_i(t) = 0. \end{cases} \tag{1}$$

$C_i(t)$ is an indication function, if the node was a cluster header during $r \mod \left(\frac{N}{k}\right)$ round, $C_i(t)$ is 0. If not, $C_i(t)$ is 1. In other words, If a node has been once a header during last $r \mod \left(\frac{N}{k}\right)$ rounds, the probability that the node can be a header again is 0. Above equation, i is the identifier of the node, t is the time, N is the total number of nodes, k is the number of clusters, r is round. In LEACH, member nodes frequently exchange information at the stage of the election of cluster heads. It consumes a lot of energy.[8]

In LEACH-C to complement this point, the cluster head is elected considering the energy state of all nodes. This approach needs the additional overhead to compare the amount of energy.[9]

HEED protocol algorithm is excellent in that node is only its own elements to elect a cluster head. The expression used in electing the cluster head uses the remaining energy of each node and uses the probability function (2).

$$CH_{prob} = C_{prob} X \frac{E_{residual}}{E_{\max}}. \tag{2}$$

$E_{\max}$ is the initial energy of the member nodes, $E_{resudual}$ is the remaining energy of the member nodes, C_{prob} is percentage of the cluster head for total network nodes.

If a node is determined as a cluster head, the elected cluster head broadcasts messages indicating it is elected as a cluster head. This message contains a number of cluster head nodes. Non-cluster head nodes which receive the message determine an

appropriate cluster head depending on signal strength and send a Join-Request to the cluster head to participate in the cluster. After this process, the entire network consists of multiple clusters.[10]

Cluster heads transfer a TDMA schedule for its cluster member nodes and in the assigned time slots, member nodes transmit data sensed to the header. In LEACH, TDMA for transmission control within a cluster is used. It minimizes transmission conflict within a cluster and makes efficient use of energy by increasing the sleep time of nodes.

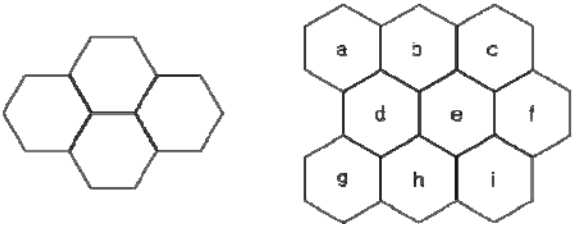
3 Topology Matrix and Resource Allocation Matrix

The proposed frequency channel assignment method uses a cluster topology matrix and a resource allocation matrix.

3.1 Topology Matrix

(1) Cell-based network representation

Topology matrix is generated as a hexagonal model based on clusters of sensor networks. The cluster topology of fig. 1 can be represented as the hexagonal model of fig. 3(a).



(a) Topology of fig. 1 (b) Topology of cluster 9

Fig. 3. Hexagonal model of cluster networks

For topology of fig. 3(b), a cell-based topology matrix (T) can be represented as the matrix of (3). The element '0' of the matrix (T) presents a non-existing area in cluster topology.

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} & t_{16} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} & t_{26} \\ t_{31} & t_{32} & t_{33} & t_{34} & t_{35} & t_{36} \end{bmatrix} = \begin{bmatrix} a & 0 & b & 0 & d & 0 \\ 0 & d & 0 & e & 0 & f \\ g & 0 & h & 0 & i & 0 \end{bmatrix}. \quad (3)$$

(2) Cluster adjacency matrix

The calculation for channel allocation is performed on a server and it can be the gateway of sensor networks. After a cluster setup phase, each cluster header transfers the information of neighbor clusters. The server generates the adjacency matrix (A) as

the matrix of (4) based on the neighbor information of clusters.[19-21] The adjacency sum matrix (AS) presents the sum of row elements in the adjacency matrix (A) of (4).

$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \end{array} \begin{array}{c} \begin{array}{ccccccccc} a & b & c & d & e & f & g & h & i \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \end{array}. \quad AS = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \end{array} \begin{bmatrix} 2 \\ 4 \\ 3 \\ 5 \\ 6 \\ 3 \\ 2 \\ 4 \\ 3 \end{bmatrix}. \tag{4}$$

(3) Cluster topology matrix

Topology matrix (T) of (3) which presents the topology of clusters is needed for non-overlapping channel allocation between neighbor clusters. It is created based on the adjacency matrix (A) of (4). The process of creating the topology matrix (T), based on the adjacency matrix (A) is presented as follows;

① The row which has the smallest sum of elements for each row in the adjacency matrix is selected. AS matrix shows the sum of elements for each row in (A) matrix and 'a', 'g' are the smallest value. The topology matrix of (3) shows that 'a' is selected as the first element t11 on a topology matrix.

② One of the row elements in the adjacency matrix for the first determined element 'a' is selected and is determined as t_{13} in the topology matrix. The row elements of 'a' in the adjacency matrix are 'b', 'd'. The 'b' is randomly selected and determined as t_{13} in fig. 4.

③ The common row element in the adjacency matrix for the current element, 'b' and the former element, 'a' is selected as the next element in the topology matrix. In fig. 4, the common row element in the adjacency matrix (A) for 'b' and 'a' is 'd'. Therefore, 'd' is selected as the next element of the topology matrix.

④ The position in the topology matrix of the selected element is determined as t_{22} according to the rule 1 of fig 5.

⑤ The process which selects the common row elements in the adjacency matrix for the current element and the former element in the topology matrix and determines the position of the topology matrix (T) is repeated until the topology matrix is completed.

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} & t_{16} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} & t_{26} \\ t_{31} & t_{32} & t_{33} & t_{34} & t_{35} & t_{36} \end{bmatrix} = \begin{bmatrix} a & 0 & b & 0 & c & 0 \\ 0 & d & 0 & e & 0 & f \\ g & 0 & h & 0 & i & 0 \end{bmatrix}$$

$t_{m(i-k)n(i-k)}$, $t_{m(i)n(i)}$, $t_{m(i+1)n(i+1)}$

Fig. 4. Topology matrix T of fig. 3b)

Table 1 shows the determination order of elements in the process of generation of the topology matrix of fig. 4 for the topology of fig. 3b). The element 'd' of the table 1 is the next element for the element 'a', 'b' of the topology matrix in fig. 4. Next, the common neighbor element for the element 'a', 'd' of the topology matrix is checked. There is no common neighbor element for the element 'a', 'd' of the topology matrix. Then, the common neighbor element for the element 'b', 'd' of the topology matrix is checked. The common neighbor element for the element 'b', 'd' of the topology matrix is the element 'e'. Therefore, the element 'e' is the next element for the element 'b', 'd' of the topology matrix in fig. 4. The position for the element 'e' in the topology matrix is determined according to the rule 3 of fig 5.

Table 1. The process of generation of the topology matrix for fig. 3 b)

The elements that are previously determined	The elements that are currently determined	The following elements
a	b	**d**
a, b	d	**e**
a, b, d	e	**c, h**
a, b, d, e	c	**f**
a, b, d, e	h	**g, i**

The position of the selected element is determined based on the position pattern of two elements which are previously determined and six patterns are appeared as fig. 5. In $t_{m(i)n(i)}$, 'm' presents a row id in the topology matrix, 'n' presents a column id in the topology matrix, 'i' presents the index for the order of position determination. $t_{m(i-k)n(i-k)}$ is the element that is previously determined, $t_{m(i)n(i)}$ is the element that is currently determined.

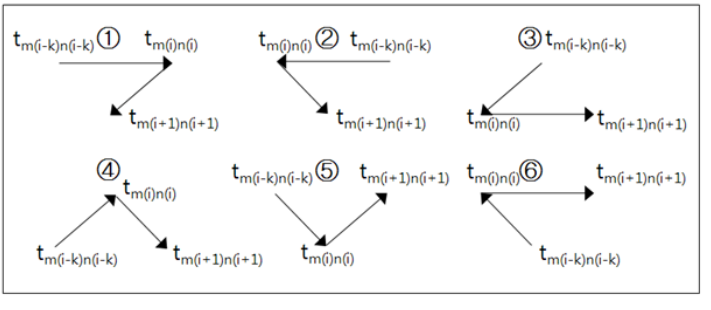


Fig. 5. The pattern of the position determination rule on a topology matrix

The pattern ① is presented as (5).

$$\text{if } (\ m_{(i-k)} == m_{(i)} \text{ and } n_{(i-k)} < n_{(i)}\ ), \text{ then } m_{(i+1)} = m_{(i)} + 1,\ n_{(i+1)} = n_{(i)} - 1 \tag{5}$$

Based on both the element previously determined 'a', t_{11} and the element currently determined 'b', t_{13}, the position in the topology matrix of 'd' is determined as t_{22} by function (5).

Table 2. The position determination rule on a topology matrix

Pattern No.	Rules	Pattern No.	Rules
①	if ($m_{(i-k)} == m_{(i)}$ and n(i-k) < $n_{(i)}$) then $m_{(i+1)} = m_{(i)} +1$, $n_{(i+1)} = n_{(i)} -1$	④	if ($m_{(i-k)} > m_{(i)}$ and $n_{(i-k)} < n_{(i)}$) then $m_{(i+1)} = m_{(i)} +1$, $n_{(i+1)} = n_{(i)} +1$
②	if ($m_{(i-k)} == m_{(i)}$ and $n_{(i-k)} > n_{(i)}$) then $m_{(i+1)} = m_{(i)} +1$, $n_{(i+1)} = n_{(i)} +1$	⑤	if ($m_{(i-k)} < m_{(i)}$ and $n_{(i-k)} < n_{(i)}$) then $m_{(i+1)} = m_{(i)} - 1$, $n_{(i+1)} = n_{(i)} +1$
③	if ($m_{(i-k)} < m_{(i)}$ and $n_{(i-k)} > n_{(i)}$) then $m_{(i+1)} = m_{(i)}$, $n_{(i+1)} = n_{(i)} +2$	⑤b	if ($m_{(i-k)} < m_{(i)}$ and $n_{(i-k)} < n_{(i)}$) then $m_{(i+1)} = m_{(i)}$, $n_{(i+1)} = n_{(i)} - 2$
③b	if ($m_{(i-k)} < m_{(i)}$ and $n_{(i-k)} > n_{(i)}$) then $m_{(i+1)} = m_{(i)} -1$, $n_{(i+1)} = n_{(i)} -1$	⑥	if ($m_{(i-k)} > m_{(i)}$ and $n_{(i-k)} > n_{(i)}$) then $m_{(i+1)} = m_{(i)}$, $n_{(i+1)} = n_{(i)} +2$

After applying sequential from pattern ① to pattern ⑥ of table 2 for the element currently determined and the element previously determined, a matched rule is used for the position determination in a topology matrix. The rule ③b is applied in case that the position which is determined by ③ is already used by another element. The rule ⑤b is also identical. Fig. 6 shows the rules which are applied in the process of creating the topology matrix T for the topology of fig. 3(b).

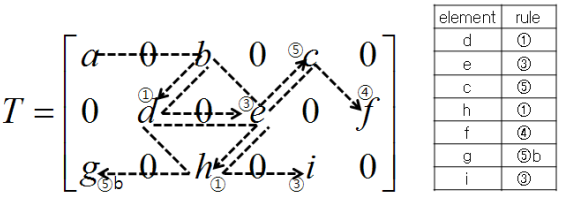
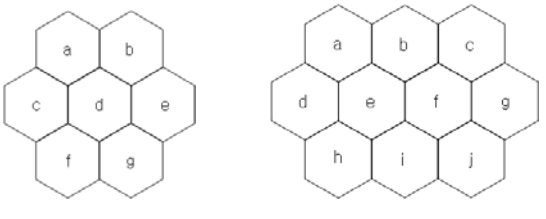


Fig. 6. Topology matrix calculation process and results

(4) Topology matrix for various hexagonal cluster topologies



a) Topology for cluster 7 b) Topology for cluster 10

Fig. 7. Topology for neighbor cluster of which the number is more than 3

For cluster topology of fig. 7(a), the adjacency matrix is shown as the matrix (6). AS matrix presents the sum of row elements in the adjacency matrix A. A topology matrix is generated based on the adjacency matrix. In the topology matrix of fig. 8 for fig. 7a, after the positions of elements 'a', 'b', 'd' are determined, the common neighbor element for the element 'a', 'd' of the topology matrix is checked. The common neighbor element for the element 'a', 'd' of the topology matrix is the element 'c'. The position of 'c' is calculated as t_{20} depending on the position determination rule ⑤b. But, t_{20} does not exist in the matrix. In this case, all the elements on the topology matrix are shifted in the right direction, and the rest of the process for matrix creation is continued. Fig. 8 shows the example that the elements are shifted to the right.

$$A = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \end{array} \begin{array}{c} \begin{array}{ccccccc} a & b & c & d & e & f & g \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}, \quad AS = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \end{array} \begin{bmatrix} 3 \\ 3 \\ 3 \\ 6 \\ 3 \\ 3 \\ 3 \end{bmatrix}. \tag{6}$$

$$T = \begin{bmatrix} a & 0 & b \\ 0 & d & 0 \end{bmatrix} \Rightarrow T = \begin{bmatrix} 0 & a & 0 & b & 0 \\ c & 0 & d & 0 & e \\ 0 & f & 0 & g & 0 \end{bmatrix}$$

Fig. 8. Example of the elements on the topology matrix shifted in the right direction

3.2 Resource Allocation Matrix

Resource assignment matrix (RA) is created based on the topology matrix (T) to assign frequency channels which do not overlap between neighbor clusters. The function (7) shows the resource allocation matrix with frequency reuse factor 3 for fig. 3(b). The '0', '1', '2' elements of the resource allocation matrix mean the identifier of frequency channels assigned to the cluster, and '∞' presents the element for a non-existing cluster.

$$RA = \begin{bmatrix} 1 & \infty & 0 & \infty & 2 & \infty \\ \infty & 2 & \infty & 1 & \infty & 0 \\ 1 & \infty & 0 & \infty & 2 & \infty \end{bmatrix},$$
$$where \quad ra_{ij} = \begin{cases} (3i + j) \ \% \ 3, & for\, t_{ij} \in T, \ if\, t_{ij} \neq 0, \\ \infty, & otherwise. \end{cases} \tag{7}$$

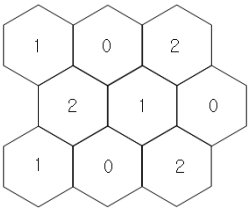


Fig. 9. The result of resource allocation on the cluster network (frequency reuse factor=3)

Fig 9 shows the result of frequency channel assigned to each cluster by using resource allocation matrix with reuse factor 3. A frequency channel which is not overlapping with neighboring clusters may be assigned to each cluster.

Due to the nature of FDMA, when the system is weak on the interference, increasing the frequency reuse factor reduces interference between cells, when the system is strong on the interference, reducing the frequency reuse factor makes clusters be allocated a wide channel bandwidth.

Resource allocation matrix according to the frequency reuse factor is generated by the following general function (8).

$$ra_{ij} = \begin{cases} (n \times i + j) \ \% \ n, & for\ t_{ij} \in T,\ if\ t_{ij} \neq 0 \\ \infty, & otherwise \end{cases} \tag{8}$$

When a neighbor relationship is constituted only with the hexagonal model, a topology matrix represents the ideal topology, and non-overlapping frequency channel may be assigned to each cluster by using allocation resource matrix

4 Conclusions

For energy efficiency, the synchronization of inter-cluster in the cluster-based sensor networks is important. A matrix based method that can efficiently provide non-overlapping frequency channels between neighbor clusters is proposed. The proposed method shows successful allocating of non-overlapping frequency channel for hexagonal neighbor topologies. The method can be easily applied in practical systems. As the future work, the channel allocation for non-hexagonal neighbor topologies has to be studied.

Acknowledgements. This research was supported by the Yeungnam University research grants 2011. This research was also supported by the Ministry of Education, Science Technology (MEST) and Korea Institute for Advancement of Technology (KIAT) through the Human Resource Training Project for Regional.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38 (2002)
2. Hak, L.S., Hwan, K.D., Jae, Y.J.: Ubiquitous sensor network technic trend. Korea Internet Information Institute Paper, 97–107 (2005)
3. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-Efficient Communication Protocol for Wireless Microsensor Networks. In: Proceedings of the Hwaii International Conference on System Science, pp. 1–10 (January 2000)
4. Ye, W., Heidemann, J., Estrin, D.: An Energy-Efficient MAC Protocol for Wireless Sensor Networks. In: Proceedings of the 21st IEEE INFOCOM, vol. 3, pp. 1567–1576 (June 2002)
5. Dam, T., Langendoen, K.: An Adaptive Energy-Efficient MAC Protocol for Wireless Sensor Networks. In: Proceedings of the 1st ACM Conference on Embedded Networked Sensor Systems, pp. 171–180 (November 2003)
6. Rhee, I., Warrier, A., Aia, M., Min, J.: Z-MAC: a Hybrid MAC for Wireless Sensor Networks. In: Proceedings of the 3rd ACM Conference on Embeded Networked Sensor Systems, pp. 90–101 (November 2005)
7. Degesys, J., Rose, I., Patel, A., Nagpal, R.: DESYNC: Self-Organizing Desynchronization and TDMA on Wireless Sensor Networks. In: Proceedings of the 6th International Symposium on Information Processing in Sensor Networks, pp. 11–20 (April 2007)
8. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-Efficient Communication Protocol for Wireless Microsensor Networks. In: Proceedings of the 3rd Hawaii International Conference on System Sciences (2000)
9. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transactions on Wireless Communication (2002)
10. Younis, O., Fahmy, S.: HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad-hoc Sensor Networks. IEEE Transaction on Mobile Computing (October 2004)
11. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: A stable election protocol for clustered heterogenous wireless sensor networks. In: International Workshop on SANPA, Boston, vol. (4), pp. 660–670 (2004)
12. Ying, L., Haibin, Y.: Energy Adaptive Cluster-Head Selection for Wireless Sensor Networks. In: Proceedings of the 6th International Conference on Parallel and Distributed Computing Applications and Technologies, pp. 634–638 (December 2005)
13. Vlajic, N., Xia, D.: Wireless Sensor Networks: To Cluster or Not To Cluster? In: International Symposium on a World of Wireless (2006)
14. Lin, T.C., Huang, H.C., Liao, B.Y., Pan, J.S.: An Optimized Approach on Applying Genetic Algorithm to Adaptive Cluster Validity Index. International Journal of Computer Sciences and Engineering Systems 1(4), 253–257 (2007)
15. Kim, J.S., Lee, J.H., Rim, K.W.: 3DE: Selective Cluster Head Selection scheme for Energy Efficiency in Wireless Sensor Networks. In: The 2nd ACM International Conference on Pervasive Technologies Related to Assistive Environments (2009)
16. Hong, T.-P., Wu, C.-H.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. Journal of Information Hiding and Multimedia Signal Processing 2(2), 173–184 (2011)
17. Marsh, G.W., Kahn, J.M.: Channel Reuse Strategies for Indoor Infrared Wireless Communications. IEEE Transactions on Communications 45(10) (October 1997)

18. Wang, X., Berger, T.: Spatial channel reuse in wireless sensor networks. Journal of Wireless Networks 14(2) (March 2008)
19. Wu, M., Kim, S., Kim, C.: A Routing Method Based on Cost Matrix in Ad Hoc Networks. In: Nguyen, N.T., Katarzyniak, R., Chen, S.-M. (eds.) Advances in Intelligent Information and Database Systems. SCI, vol. 283, pp. 337–347. Springer, Heidelberg (2010)
20. Wu, M., Kim, C.: A cost matrix agent for shortest path routing in ad hoc networks. Journal of Network and Computer Applications 33, 646–652 (2010)
21. Wu, M., Kim, C., Jung, J.J.: Leader Election Based on Centrality and Connectivity Measurements in Ad Hoc Networks. In: Jędrzejowicz, P., Nguyen, N.T., Howlet, R.J., Jain, L.C. (eds.) KES-AMSTA 2010, Part I. LNCS, vol. 6070, pp. 401–410. Springer, Heidelberg (2010)

Generation of Tag-Based User Profiles for Clustering Users in a Social Music Site

Hyon Hee Kim, Jinnam Jo, and Donggeon Kim

Department of Statistics and Information Science, Dongduk Women's University
23-1 Wolgok-Dong, Sungbuk-Gu, Seoul, South Korea
{heekim,jinnam,dongg}@dongduk.ac.kr

Abstract. Collaborative tagging has become increasingly popular as a powerful tool for a user to present his opinion about web resources. In this paper, we propose a method to generate tag-based profiles for clustering users in a social music site. To evaluate our approach, a data set of 1000 users was collected from last.fm, and our approach was compared with conventional track-based profiles. The K-Means clustering algorithm is executed on both user profiles for clustering users with similar musical taste. The test of statistical hypotheses of inter-cluster distances is used to check clustering validity. Our experiment clearly shows that tag-based profiles are more efficient than track-based profiles in clustering users with similar musical tastes.

Keywords: collaborative tagging, tag-based profiles, clustering users, social music sites.

1 Introduction

Collaborative tagging is a powerful tool which enables a user to organize contents for future navigation, filtering, or searching by his own way. In a collaborative tagging system, a user adds a keyword called a tag, which is freely chosen by himself to web resources. Since a tag does not have any pre-defined term or hierarchies of a term, a set of tags represents a user's preferences and interests about the web resources. Therefore, their use for personalized recommendation services in a social web site has attracted much attention.

Recently, personalized recommendation services have been increasingly popular in social music sites such as last.fm and Pandora. Those sites recommend other users with similar musical tastes, or music items and artists that the group of the users with similar musical tastes prefers. To provide personalized recommendation successfully, it is imperative to create users' profiles considering their musical preferences and to cluster users based on their profiles. After clustering users, collaborative recommendations which recommend items which people with similar tastes and preferences liked in the past can be used [1].

There are largely two approaches to catch a user's musical preferences in social music sites. One is based on track information that the user has listened to, and the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 51–61, 2012.

other is based on a user's rating about music items. In the case of using track information, it takes time enough for a user to listen to his favorite music items. In addition, the user does not listen to all of his favorite music items through the social music sites. On the other hand, rating systems allow a user to rate the music items easily better. However, the rating information that is represented by the number, usually from 1 to 5, does not reflect users' musical preference directly.

From this point of view, it is imperative to generate user profiles considering users' tagging information. A user can add tags to his favorite music items without listening to the music and represent his opinion more flexibly using keywords rather than a conventional rating system. Such features of collaborative tagging sometimes lead to the semantic ambiguity, such as polysemy, i.e., a single word contains multiple meanings, and synonymy, i.e., different words represents the same meaning [2]. Therefore, resolving the semantic ambiguity of tags in the tag-based user profiling is required.

In this paper, we present a method of generating tag-based profiles for clustering users in a social music site. First, to resolve semantic ambiguity of tags, we developed UniTag ontology which provides three types of reasoning rules for resolving semantic ambiguity. Second, we propose a method to generate a user's profile based on the tags and tracks. Third, to evaluate efficiency of our approach, we execute k-means clustering algorithm on both the tag-based profiles and the track-based profiles. Data sets are crawled from last.fm, which is the well known social music site, about 1000 users. Our experimental result shows that the tag-based approach clusters users more appropriately than the track-based approach.

The remainder of this paper is organized as follows. In Section 2, we mention related work, and In Section3, we discuss the overview of UniTag ontology. Section 4 explains generation of tag-based profiles and track-based profiles. Section 5 provides the experimental evaluation, and finally, Section 6 gives concluding remarks.

2 Related Work

Firan et al. [3] propose tag-based profiles, which collect tags together with corresponding scores representing the user's interest as well as track-base profiles which collect music tracks with associated preference scores describing users' musical tastes. Their experimental results show that the tag-based music recommendation improves result's quality significantly. Cai and Li [4] also suggest a tag-based user profile and a resource profile, and apply them to personalized search. However, the approaches mentioned above do not consider semantic ambiguity of tags.

Durao and Dolog [5] provide personalized recommendations by calculation of a basic similarity between tags. Further, they extend the calculus with additional factors such as tag popularity, tag representativeness and affinity between users and tags. To extract semantic relatedness between two tags, they used WordNet dictionary and ontologies from open linked data. Hierarchical agglomerative clustering of tags is used for personalized recommendations [6]. They show that clusters of tags can be effectively used as a means to ascertain the user's interest as well as to determine the topic of a resource.

Foafing the music [7] and MusicBox [8] are personalized music recommendation systems. The system uses FOAF profile, which provides a framework to represent information about people, their interests, relationships between them and social connections [7], and filters music related information from RSS based on the FOAF profile. The FOAF profile provides a user's psychological factors like personality, demographic preferences, socio-economics, situation, and musical preferences explicitly. In MusicBox [8], modeling of social tagging data with three-order tensors capturing correlations between users, tags, and music items is proposed. The system shows that the proposed method improves the recommendation quality.

3 Resolving Semantic Ambiguity of Tags

In this Section, we present the overview of the UniTag Ontology for a social music site and then explain three types of reasoning rules to resolve semantic ambiguity of tags and other type of reasoning rules to classify music items using tags in detail.

3.1 Overview of the UniTag

In tag-based recommendation systems, the tag ontology is used to resolve semantic ambiguity of tags [2, 9]. [9] suggests the meaning of a tag, which a user adds by looking for MOAT ontology. The user finds terms registered in the MOAT ontology, and replace his tag with the registered term. This approach guaranties semantic clarity of tags, but it restricts a user's freedom of expression, which is the strength of collaborative tagging. We developed a method to create tag ontology called TagSES based on the set of tags which users add in our former study [2]. TagSES extracts semantic information from tags which are chosen freely without any restriction.

In this study, we take the same approach to generate the UniTag ontology for a social music site. In the case of social music sites, considering that 68% of tags are related to genres, UniTag maps tags which are represented by different terms onto the representative genre tags. There exist diverse classifications of musical genres [10]. Therefore, to define the representative genre tags, we selected top ten frequently used genre tags, i.e., rock, hiphop, electronic, metal, jazz, rap, funk, folk, blues, and reggae.

In general, tag ontologies are represented by a ternary relation composed of three common entities: users, tags, and resources [11]. The UniTag ontology is based on the ternary relation, but its concept is augmented with representative tag to infer semantic agreement of a set of tags. Classes and properties of the UniTag ontology are explained as follows.

uniTag:Users explains users for a social music site and has two subcalsses, *uniTag:User* and *uniTag:UserGroup*. A user might be an instance of *uniTag:User* class, while a group of users with similar musical tastes might be an instance of *uniTag:UserGroup* class. *uniTag:Users* has a property *uniTag:memberOf*. *uniTag:memberOf* property represents an instance of *uniTag:User* class is member of *uniTag:UserGroup* class.

uniTag:Tags models a set of tags in a social music site, and has two subclasses *uniTag:Tag* and *uniTag:RTags* subclasses. *uniTag:Tag* class represents each tag which users add, while *uniTag:RTags* represents representative musical genre tag defined in this study. A tag in *uniTag:Tag* can have prefix, which represents country and types of musical genre. For example, French rock, folk rock, progressive rock, etc. have prefixes, French, folk, and progressive and the prefix is represented by *uniTag:tagPrefix* properties. The remaining term, i.e, rock, is represented by *uniTag:subTag* properties. If there is no prefix in a tag, then *uniTag:tagPrefix* has null value and *uniTag:subTag* has the tag as value. Tags in *uniTag:Tags* are mapped onto representative tags in *uniTag:RTags* according to reasoning rules, which will be explained in Section 3.2.

uniTag:RTags class has ten instances of representative musical genres, i.e., rock, hiphop, electronic, metal, jazz, rap, funk, folk, blues, and reggae. Those musical genres are chosen by analyzing tags of social musical sites. *uniTag:tagVariation* property represents diverse expressions of the representative musical genres. For example, hip-hop and hip hop are assigned to *uniTag:tagVariation* of hiphop in *uniTag:RTags*. *uniTag:classifiedAs* property classifies each tag in *uniTag:Tag* class as *uniTag:RTags*. *uniTag:isKindOf* property represents a tag in *unTag:Tag* is sub genre of other tag, and is transitive. *uniTag:istheSameAs* property represents diverse expressions of a tag which belongs to a representative tag.

uniTag:Resource models music items which users have listened to in a social web site. Users can listen to the music items or add tags, and *uniTag:hasListened* and *uniTag:hasAdded* property are used. *uniTag:belongsTo* property represents a musical genre which a musical item belongs to.

3.2 Reasoning Rules

To map a tag which a user adds onto a representative tag, three types of rules are defined as follows: reasoning rules using prefixes of tags, reasoning rules using expert knowledge, and reasoning rules using synonym. Additionally, one more type of rules is required to classify music items into ten genre categories.

Rules for Reasoning Prefix: After analyzing a set of tags in last.fm, we realized that there exist four types of prefixes ahead of a genre tag: prefix for nationality such as French rock, English rock, and Spanish rock, prefix for combination of other musical genre such as folk rock, punk rock, prefix for new sub genre such as progressive rock, alternative rock, etc., and finally prefix for general use of genre such as indie rock, post rock, etc. All of those tags can be mapped onto the representative genre tag, i.e., rock. The following rules are defined for reasoning prefixes.

Rule1. If a prefix of a tag is an instance of *uniTag:Prefix* class and the remaining term of the tag is an instance of *uniTag:RTags* class, then the tag is classified as an instance of *uniTag:RTags* class.

```
(Tag (?t) ∧ tagPrefix (?t, ?p) ∧ Prefix(?p) ∧
subTag(?t, ?s) ∧ RTags(?s)) → classifiedAs(?t, ?s)
```

Rule2. If a prefix of a tag is an instance of *uniTag:RTags* class and the remaining term is an instance of *uniTag:RTags* class, then the tag is classified as an instance of *uniTag:RTags*.

```
(Tag (?t) ∧ tagPrefix ?t, ?p) ∧ RTags (?p) ∧ subTag
(?t,?s) → classifiedAs (?t, ?s)
```

Rules for Expert Knowledge: The rules map a tag which represents a musical genre onto a representative tag of the musical genre based on the expert knowledge of the musical domain. For example, the soul music is a kind of the rhythm and blues music, and the rhythm and blues music is a kind of blues music. Since the "is a kind of" relation is transitive, the soul music is classified as the blues music. Also, diverse expressions of a musical genre are inferred based on the rules.

Rule3. If a tag represents a sub genre of a genre A and the genre A is also sub genre of other genre B, then the tag belongs to the genre B.

```
(Tag (?t) ∧ isKindof (?t, ?A) ∧ isKindof (?A,?B)
→ isKindof (?t, ?B)
```

Rules for Reasoning Synonym: The rules map diverse expressions of a musical genre onto a representative musical genre. For example, hip-hop and hiphop are mapped onto the representative tag hip hop.

Rule4. If a tag A is a variation of a representative tag R and the tag is the same as other tag B, then the tag B is also a variation of the representative tag R.

```
(Tag (?t) ∧ tagVariation (?t, ?R) ∧ istheSameAs (?t,?s)
→ tagVariation (?s, ?R)
```

Rules for Classifying Music Items: The rules classify music items into ten musical genres based on tags which music items have. Different from the above three rules, the rule is used for generating track-based profiles.

Rule5. If a tag t is added to music item r, and the tag t is consistent with a representative genre g, then the music item belongs to the musical genre g.

```
 (Tag (?t) ∧ Resource (?r) ∧ hasAdded(?r, ?t) ∧
istheSameAs (?t,?g) → belongsTo (?t, ?g)
```

4 Generating User Profiles

After resolving semantic ambiguity of tags using the reasoning rules explained in Section 3.2, user profiles are generated based on the set of tags. In this Section,

we describe how a user profile is generated in detail. To compare the proposed tag-based profile with the conventional approach, a track-based profile is also generated. The main difference between the conventional track-based profile and our track-based profile is that our track-based profile uses tags to classify music items. Since a user tends to listen to diverse genres of music, it has some difficulty to classify all of music items into representative musical genre. And therefore, tags which are added to the music items are used for this purpose. Let us take a closer look at the method of generating tag-based user profiles and track-based user profiles.

4.1 Tag-Based User Profiles

If a user adds tags to specific music items often, he can be considered to have musical tastes for the items. A set of tags added is used to generate a tag-based user profile. Therefore, the tag-based user profile is represented by frequency of ten representative tags added in the vector space shown as $P_u = (T_{u,1}, T_{u,2}, .., T_{u,k},\ldots, T_{u,10})$, where $T_{u,k}$ is the frequency of the tag k which the user u adds. An example of tag-based profiles of five users is shown in Table 1.

Table 1. An example of tag-based profiles

	rock	hiphop	electronic	metal	jazz	rap	funk	folk	blues	reggae
user1	6	2	2	3	2	4	3	1	1	1
user2	5	0	0	0	0	0	0	0	1	0
user3	2	2	1	1	1	1	2	0	0	1
user4	10	1	0	1	2	0	2	3	3	1
user5	1	4	0	0	0	4	1	0	0	0

As shown in Table 1, the tag-based profiles reflect users' musical preference concretely. For example, both user1 and user2 prefer rock music. While user1 is interested in diverse musical genre, user2 is extensively interested in rock genre. The algorithm for generating a tag-based profile is shown as follows.

Algorithm 1. Generation of A Tag-based Profile

Input: set of Representative tags T_r, set of a user's tag T_u

Output: set of frequencey for each representative tag of the user FT_r

```
var   RTags[] = {rock, hiphop, electronic, metal, jazz,
rap, funk, folk, blues, reggae}
var tagFrequency[] = { }, tempFrequency [] = { }
var RTag = null
while ∃next tag t in Tu do
       RTag = FindRTag (t)
   If Rtag == RTags [i] then
       {   tempFrequency[i] = tempFrequency[i] + 1
```

```
            tagFrequency [i] = tempFrequency [i] }
   else
            tagFrequency [i] = tempFrequency [i]
endwhile
```

The function **FindRTag**(t) finds a representative tag of a tag t. If the tag t is consistent with classifiedAs (?t, ?s), then the function returns *s* as a value of variable *RTag*. If the *RTag* is the same as one of array of RTags, then the frequency of the representative tag is increased. A set of tags of a user are examined, and as a result, a user's tag-based profile is generated.

4.2 Track-Based User Profiles

A track-based user profile is generated using the number of music items which a user has listened to. To classify music items into ten representative genre tags, tags which are added to music items are used. Therefore, a track-based user profile is represented by the number of music items belonging to a specific musical genre in a vector space shown as $P_u = (N_{u,1}, N_{u,2}, .., N_{u,k},\ldots, N_{u,10})$. $N_{u,k}$ is the number of music items which a user has listen belonging to musical genre k. Table 2 shows an example of track-based profiles of five users. In comparison with tag-based profiles, values of each category are somewhat larger than tag-based profiles, and therefore standardization is necessary to compare the two types of profiles.

Table 2. An example of track-based profiles

	rock	hiphop	electronic	metal	jazz	rap	funk	folk	blues	reggae
User1	65	176	5	4	0	168	0	3	0	0
User2	411	8	11	109	3	5	8	1	0	0
User3	157	7	11	10	6	2	1	39	4	2
User4	257	20	9	18	2	5	0	9	0	0
User5	110	277	15	8	6	85	10	3	2	7

The algorithm for generating a track-based profile is shown as follows.

Algorithm 2. Generation of A Track-based Profile

Input: set of tracks of a usr TR_u, set of Representative tags T_r

Output: set of number of a user's tracks for each representative musical genre T_n

```
var RTags[] = {rock, hiphop, electronic, metal, jazz,
rap, funk, folk, blues, reggae}
var numTrack[ ] = { }, tempnumTrack [ ] = { }
var RTrack = null
while ∃next tag t in Tu do
   RTrack = FindGenre (t)
```

```
   If Rtrack == RTags [i] then
       {   tempnumTrack [i] = tempnumTrack[i] + 1
           numTrack[i] = tempnumTrack [i] }
   else
           numTrack [i] = tempnumTrack [i]
endwhile
```

The function **FindGenre**(t) finds a representative genre which each music item belongs to. If the tag t is consistent with belongsTo (?t, ?s), then the function returns *s* as a value of the variable *RTrack*. If the *RTrack* is the same as one of array of *RTags*, then the number of tracks which a user has listened to in the genre is increased.

5 Experimental Evaluation

To evaluate the effectiveness of our method, K-Means clustering algorithm [12] with two types of user profiles on last.fm data set is executed. 1,000 user information, tags which the users add, music items which the users have listened to and tags which are added to the music items are collected. Six clusters are generated in each user profile, and inter-cluster distances are estimated. To check clustering validity, the statistical hypothesis testing is used.

5.1 Clustering Users

First of all, our data set shows musical genre of rock has extensive preference in both tag-based profiles and track-based profiles. In addition, the frequency of tag usage in tag-based profiles and the number of music items in track-based profiles are different. Therefore, standardization is executed on both user profiles. To cluster users with similar musical tastes, K-Means clustering algorithm is selected because of its simplicity and efficiency. To obtain good a quality clustering using K-Means algorithm, the number of clusters, k should be carefully chosen. For this purpose, canopy clustering, a fast approximate clustering technique [13] is executed in advance. As a result, six centroid points are chosen, and six clusters are created based on both user profiles. The centers of six clusters from standardized profiles are shown in Table 3 and Table 4.

Table 3. Values of Centers of Tag-based Profiles

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Cluster1	0.241	1.472	0.626	0.130	1.267	1.621	2.168	0.274	1.078	0.381
Cluster2	2.171	0.032	0.517	3.052	0.011	-0.030	0.328	1.533	1.245	0.162
Cluster3	-0.206	-0.273	-0.517	-0.178	-0.180	-0.294	-0.233	-0.171	-0.204	-0.136
Cluster4	-0.341	0.660	-0.459	-0.284	-0.208	1.178	-0.179	-0.321	-0.166	0.273
Cluster5	-0.074	-0.155	1.320	-0.230	-0.115	-0.261	-0.209	-0.070	-0.172	-0.071
Cluster6	2.815	7.640	5.168	-0.136	9.254	6.135	7.000	4.286	4.421	5.254

Table 4. Values of Centers of Track-based Profiles

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Cluster1	-0.411	0.495	0.406	-0.338	1.565	0.131	1.632	-0.135	0.147	0.812
Cluster2	0.200	-0.444	0.007	-0.341	0.907	-0.468	-0.288	2.617	1.097	0.020
Cluster3	-0.897	1.651	-0.539	-0.442	-0.213	1.836	0.059	-0.507	-0.415	0.034
Cluster4	1.925	-0.590	-0.404	0.852	-0.264	-0.491	0.655	-0.002	2.850	-0.108
Cluster5	0.914	-0.557	-0.216	0.794	-0.296	-0.511	-0.297	0.014	-0.157	-0.147
Cluster6	-0.472	-0.327	0.380	-0.373	-0.184	-0.371	-0.241	-0.205	-0.300	-0.093

5.2 Evaluation of Results

Inter-cluster distances are estimated to check cluster validity. Given all centroid points, the distances between all pairs of centroids are calculated using the cosine distance measure. If the clusters have centroids that are too close, it indicates distinctions between cluster members are hard to support. That is, the centroids of good clusters will be far from each other.

Since both user profiles are grouped into six clusters, 15 inter-cluster distances are estimated, respectively. If the mean of the inter-cluster distances of tag-based profiles are greater than that of track-based profiles, then we may conclude that the clusters of the tag-based profiles are grouped more efficiently than the clusters of the track-based profiles.

We used t-test to compare the means of the inter-cluster distances of tag-based profiles and track-based profiles. The test result is shown in Table 5. The mean of inter-cluster distances of tag-based profiles was significantly greater than that of track-based profiles indicating the centroid points generated by tag-based profiles are more well-separated than the centroid points by track-based profiles on the average. Therefore, we conclude that the quality of clustering users based on the proposed tag-based profiles is better than the conventional approach.

Table 5. T-test result for the means of inter-cluster distances

	N	Mean	Std Dev	t	p-value
Tag-based profiles	15	0.8325	0.6834	2.55	0.0165
Track-based profiles	15	0.3785	0.0885		

6 Conclusions and Future Work

In this paper, we have presented a method to generate tag-based profiles for clustering users with similar musical tastes in a social music site. In collaborative tagging, since a user chooses tags freely without any limitation such as preserved keywords, set of tags reflects users' musical preference directly. Therefore, tag-based user profiles catch their musical tastes better than the conventional rating system which is represented by numbers. In addition, tag-based profiles can overcome limitations of

the conventional track-based profiles, which need time for users to have listened to music items.

To evaluate our approach, the track-based profiles are also generated. There is difference between our track-based profiles and the conventional track-based profiles. Tags which are added to the music items are also used to classify musical items as pre-defined ten musical genres. K-Means clustering algorithm is executed on both user profiles and six clusters are generated for each profile. To check cluster validity, the statistical verification approach is conducted using inter-cluster distances. Our experimental result shows that inter-cluster distances of tag-based profiles are longer than track-based profiles, which indicates quality of clusters of tag-based profiles are better than one of track-based profiles.

Clustering algorithms are widely applied to diverse applications such as mobile ad hoc network and wireless sensor networks, and adapted to those applications [14]. As a future work, we are considering using an weighted clustering algorithms to cluster users with similar tastes, and developing a music recommendation algorithm based on the proposed tag-based user profiles. Also, evaluation model of user clustering considering other variables such as intra-cluster distances is also being considered.

Acknowledgments. This was supported by the Dongduk Women's University grant.

References

1. Adomavicius, A., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering 17(8), 743–749 (2005)
2. Kim, H.H.: A Personalized Recommendation Method Using a Tagging Ontology for a Social E-Learning System. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS, vol. 6591, pp. 357–366. Springer, Heidelberg (2011)
3. Firan, C.S., Nejdl, W., Paiu, R.: The Benefit of Using Tag-Based Profiles. In: LA-Web 2007, Snatiago de Chile (2007)
4. Cai, Y., Li, Q.: Personalized Search by Tag-based User Profile and Resource Profile in Collaborative Tagging Systems. In: Proc. of the ACM International Conference on Information and Knowledge Management (CIKM 2010), New York, USA (2010)
5. Durao, F., Dolog, P.: Extending a Hybrid Tag-Based Recommender System with Personalization. In: Proc. 2010 ACM Symposium on Applied Computing (SAC 2010), Sierre, Switzerland, pp. 1723–1727 (2010)
6. Shepitsen, A., et al.: Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering. In: Proc. of ACM International Conference on Recommender Systems (RecSys 2008), Lausanne, Switzerland (2008)
7. Celma, O., Ramirez, M., Herrera, P.: Foafing the Music: a Music Recommendation System Based on RSS Feeds and User Preferences. In: Proc. of International Conference on Music Information Retrieval (ISMIR 2005), Londun, UK (2005)
8. Nanopoulos, A., et al.: MusicBox: Personalized Music Recommendation based on Cubic Analysis of Social Tags. IEEE Trans. on Audio, Speech and Language Proceeding 18(2), 1–7 (2010)

9. Passant, A., Laublet, P.: Meaning of A Tag: A collaborative Approach to Bridge the Gap Between Tagging and Linked Data. In: Proc. of the Linked Data on the Web Workshop, Beijing, China (2008)
10. Aucouturier, J., Pachet, F.: Representing Musical Genre: A State of Art. Journal of New Music Research 32(1), 83–93 (2003)
11. Gruber, T.: Ontology of Folksonomy: A Mash-up of Apples and Oranges. Int. J. on Semantic Web & Information Systems 3(2), 1–11 (2007)
12. Lloyd, S.P.: Least Squres Quantization in PCM. IEEE Trans. Information Theory 23, 128–137 (1982)
13. McCallim, A., Nigam, K., Ungar, L.H.: Efficient Clustering of High-Dimensional Data Set with Application to Reference Matching. In: Proc. of ACM SIGKDD (KDD 2000), Boston, USA, pp. 169–178 (2000)
14. Hong, T.P., Wu, C.H.: An Improved Weighted Clustering Algorithm for Determination of Application Nodes in Heterogeneous Sensor Networks. Journal of Information Hiding and Multimedia Signal Processing 2(2), 173–184 (2011)

A Proposed IPC-Based Clustering and Applied to Technology Strategy Formulation

Tzu-Fu Chiu[1], Chao-Fu Hong[2], and Yu-Ting Chiu[3]

[1] Department of Industrial Management and Enterprise Information, Aletheia University, Taiwan, R.O.C.
chiu@mail.au.edu.tw
[2] Department of Information Management, Aletheia University, Taiwan, R.O.C.
cfhong@mail.au.edu.tw
[3] Department of Information Management, National Central University, Taiwan, R.O.C.
gloria@mgt.ncu.edu.tw

Abstract. In order to aggregate the professional knowledge of examiners (of patent office) in the IPC code assignment and the innovative information within the patent documents, an IPC-based clustering is proposed for formulating the technology strategy. Technology strategy represents managers' efforts to think systematically about the role of technology in decisions affecting the long-term success of the organization. IPC-based clustering is utilized to generate the technical categories via the IPC and Abstract fields, while link analysis is adopted to generate the relation types for the whole dataset via the Abstract, Issue Date, and Assignee Company fields. During experiment, the technical categories have been identified using IPC-based clustering, and the technology strategies for significant companies have been formulated through link analysis. Finally, the technical categories and technology strategies will be provided to the managers and stakeholders for assisting their decision making.

Keywords: technology strategy, IPC-based clustering, link analysis, thin-film solar cell, patent data.

1 Introduction

Solar cell (especially thin-film solar cell), one of green energies, is growing at a fast pace with its long-lasting and non-polluting natures. It is worthwhile for researchers and practitioners to devote their efforts to explore the technology strategy via patent data. In patent database, the IPC codes are provided by the examiners of patent office [1] and contain the professional knowledge of the experienced examiners. It would be reasonable for a research to base on the IPC code and the term vectors of Abstract to classify the patents into a certain number of categories. Afterward, link analysis is adopted to discover the relations between IPC category and company. Consequently, the technical categories will be obtained and the technology strategy of thin-film solar cell will be formulated for assisting the R&D planning for managers and companies.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 62–72, 2012.

2 Related Work

As this study is attempted to suggest the technology strategy for companies and stakeholders via patent data, a research framework is required and can be built using methods of data mining and cluster analysis. Due to the collected patent data originating from various companies and spreading over a period of time (2000 to 2009), IPC-based clustering is employed to generate the technical categories and link analysis is adopted to discover the relations between technical categories and companies (as well as years). Subsequently, the research framework will be applied to formulate the technology strategy on thin-film solar cell. This study is also based on the idea of authors' previous research [2]. Therefore, the related areas of this study would be technology strategy, thin-film solar cell, IPC-based clustering, and link analysis, which will be described briefly in the following subsections.

2.1 Technology Strategy

Technology strategy represents managers' efforts to think systematically about the role of technology in decisions affecting the long-term success of the organization [3]. Strategic planning is an organization's process of defining its strategy, or direction, and making decisions on allocating its resources to pursue this strategy, including its capital and people [4]. While formulating technology strategy, an organization needs to review its core technologies, position itself relative to technology development, and respond to the technology standards. Technological pioneering strategy can be classified as: market pioneers, quick-followers, late-entrants, and strategic alliances [3]. Strategic choices can be also divided into two large categories: business-level strategies and corporate-level strategies [5]. In this study, a research framework of IPC-based clustering and link analysis will be constructed for technology strategy formulation on thin-film solar cell.

2.2 Thin-Film Solar Cell

Solar cell, a sort of green energy, is clean, renewable, and good for protecting our environment. It can be mainly divided into two categories (according to the light absorbing material): crystalline silicon (in a wafer form) and thin films (of other materials) [6]. A thin-film solar cell (TFSC), also called a thin-film photovoltaic cell (TFPV), is made by depositing one or more thin layers (i.e., thin film) of photovoltaic material on a substrate [7]. The most common materials of TFSC are amorphous silicon and polycrystalline materials (such as: CdTe, CIS, and CIGS) [6]. In 2009, the photovoltaic industry production increased by more than 50% (average for last decade: more than 40%) and reached a world-wide production volume of 11.5 GWp of photovoltaic modules, whereas the thin film segment grew faster than the overall PV market [8]. Therefore, thin film would be appropriate for academic and practical researchers to contribute efforts to explore and formulate the technology strategy for companies and industry.

2.3 IPC-Based Clustering

An IPC (International Patent Classification) is a classification derived from the International Patent Classification System (supported by WIPO) which provides a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain [9]. IPC classifies technological fields into five hierarchical levels: section, class, subclass, main group and sub-group, containing about 70,000 categories [10]. The IPC codes of every patent are assigned by the examiners of the national patent office and contain the professional knowledge of the experienced examiners [11]. Therefore, it would be reasonable for a research to base on the IPC code and the term vectors of Abstract to cluster the patents into a number of categories. The IPC codes have been applied for assisting patent retrieval in some researches [11, 12].

In this study, the mean vector of the patents comprised in the same IPC code group will be utilized as a centroid for clustering to divide the whole dataset of patents into a certain number of categories.

2.4 Link Analysis

Link analysis is a collection of techniques that operate on data that can be represented as nodes and links [13]. A node represents an entity such as a person, a document, or a bank account. A link represents a relationship between two entities such as a reference relationship between two documents, or a transaction between two bank accounts. The focus of link analysis is to analyze the relationships between entities. The areas related to link analysis are: social network analysis, search engines, viral marketing, law enforcement, and fraud detection [13]. Additionally, in social network analysis, the degree centrality of a node can be measured as in Equation (1), where $a(P_i, P_k) = 1$ if and only if P_i and P_k are connected by a link (0 otherwise) and n is the number of all nodes [14]. In data mining, the strength of a relationship (i.e., the similarity between nodes) can be measured as in Equation (2), where Q and D are the term frequency vectors and n is the vocabulary size [15].

$$C_D(P_k) = \sum_{i=1}^{n} a(P_i, P_k) \Big/ (n-1) \tag{1}$$

$$Sim(Q, D) = \sum_{i=1}^{n} (Q_i \times D_i) \Big/ \sqrt{\sum_{i=1}^{n} (Q_i)^2} \times \sqrt{\sum_{i=1}^{n} (D_i)^2} \tag{2}$$

In this study, link analysis will be employed to generate the linkages between IPC category and year (/company) for relation recognition.

3 A Research Design for Strategy Formulation

A research framework for technology strategy formulation, based on IPC-based clustering and link analysis, has been constructed as shown in Fig. 1. It consists of four phases: data preparation, IPC-based clustering, link analysis, and new findings; and will be described in the following subsections.

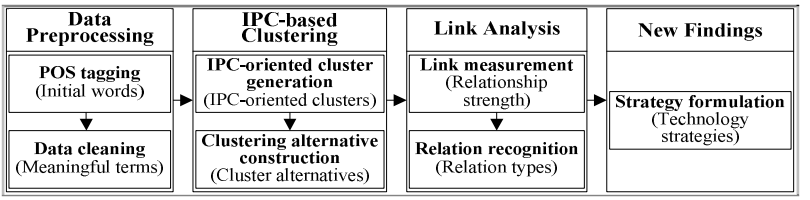


Fig. 1. A research framework for technology strategy formulation

3.1 Data Preprocessing

In first phase, the patent data of thin-film solar cell (during a certain period of time) will be downloaded from the USPTO [16]. For considering an essential part to represent a patent document, the Abstract, Assignee, and Company fields are selected as the objects for this study. Afterward, two processes, POS tagging and data cleaning, will be executed to clean up the textual data of the Abstract field.

(1) POS Tagging: An English POS tagger (i.e., a Part-Of-Speech tagger for English) from the Stanford Natural Language Processing Group [17] will be employed to perform word segmenting and labeling on the patents (i.e., the abstract field). Then, a list of proper morphological features of words needs to be decided for sifting out the initial words.

(2) Data Cleaning: Upon these initial words, files of n-grams, synonyms, and stop words will be built so as to combine relevant words into compound terms, to aggregate synonymous words, and to eliminate less meaningful words. Consequently, the meaningful terms will be obtained from this process.

3.2 IPC-Based Clustering

Second phase is used to describe the IPC-based clustering proposed by this research, including IPC-oriented cluster generation and clustering alternative construction. The former is used to generate the clusters via the centroids and term vectors. The latter is designed to construct and evaluate the alternatives for selecting the appropriate one.

(1) IPC-oriented Cluster Generation: Firstly, the patents with the same IPC code will be put together to form an IPC code group, if a patent which has more than one IPC code will be assigned to multiple groups. Secondly, patents in an IPC code group will then be used to calculate the group centroid M_i via the term vectors of the Abstract field of patents using Equation (3) [18] where x is a term vector and C_i is the ith group.

$$M_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{3}$$

Lastly, the centroids and the term vectors of patents will be utilized to produce the IPC-oriented clusters. That is, the whole dataset of patents will be distributed into a certain number of clusters using the Euclidean distance measure as in Equation (4)

[24] where X_i is the term vector of $_i$th patent. A patent will be assigned to a specific IPC code cluster, while the shortest distance $D(X_i, M_j)$ exists.

$$D(X_i, M_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - m_{jk})^2} \tag{4}$$

(2) Clustering Alternative Construction: At the beginning, the clustering alternative will be constructed from containing a small number of clusters (e.g., 4 clusters) to containing a large number of clusters. Secondly, the accuracy of the clustering alternatives will be evaluated via the *F* score ($F = 2 / ((1 / Precision) + (1 / Recall))$ [19]. Lastly, an original or adjusted alternative with the higher *F* score will be selected as the appropriate clustering result. In the selected alternative, each cluster is regarded as an IPC-oriented category (or technical category).

3.3 Link Analysis

Third phase is designed to perform the link analysis for measuring the relationship strength between entities, so as to obtain the relations between Assignee Company (i.e., company) and IPC-oriented categories as well as between Assignee Company and Issue Year (i.e., year).

(1) Link Measurement: In order to measure the relationship strength, the linkages between companies and IPC-oriented categories (as well as years) will be calculated via Equation (2) so as to provide a basis for the next relation recognition step.

(2) Relation Recognition: According to the linkage values between companies and categories (L1) as well as between companies and years (L2), nine kinds of relations are likely found: a strong linkage L1 (between companies and categories: not less than 5) and a strong linkage L2 (between companies and years: not less than 5) recognizing a relation type A; a strong L1 and a medium L2 (between 3 and 4) recognizing a type B; a strong L1 and a weak L2 (between 1 and 2) recognizing a type C; a medium L1 (between 3 and 4) and a strong L2 recognizing a type D; both medium L1 and L2 recognizing a type E; a medium L1 and a weak L2 recognizing a type F; a weak L1 (between 1 and 2) and a strong L2 recognizing a type G; a weak L1 and a medium L2 recognizing a type H; finally both weak L1 and L2 recognizing a type I.

3.4 New Findings

Last phase is intended to generate the technology strategy based on the relation types (Type A to I) which are derived from the relations "between companies and technical categories" (L1) and "between companies and years" (L2).

Strategy Formulation: In accordance with relation types (Type A to I) between companies and categories (as well as years), the technology strategy will be formulated as follows: (i) type A or B (with 'strong L1' and 'strong or medium L2') forming the market-pioneer strategy, (ii) type D or E (with 'medium L1' and 'strong

or medium L2') forming the strategic-alliance strategy, (iii) type G or H (with 'weak L1' and 'strong or medium L2') forming the quick-follower strategy, (iv) type C or F (with 'strong or medium L1' and weak 'L2') forming the late-entrant strategy, and (v) type I (with both weak L1 and L2) forming the niche-market strategy. The formulated strategies will be helpful for the managers and stakeholders to assist their decision making.

4 Experimental Results and Explanation

The experiment has been implemented according to the research framework. The experimental results will be explained in the following five subsections: result of data preprocessing, result of IPC code group and IPC-coded centroid, result of IPC-based clustering, result of link analysis, and result of new findings.

4.1 Result of Data Preprocessing

As the aim of this study is to formulate the technology strategy on thin-film solar cell, 160 patent records (mainly the Abstract, IPC, Issue Date, and Country fields) were collected from USPTO during 2000 to 2009, using key words: "'thin film' and ('solar cell' or 'solar cells' or 'photovoltaic cell' or 'photovoltaic cells' or 'PV cell' or 'PV cells')" on "title field or abstract field". Afterward, the POS tagger was triggered and the data cleaning process was executed to do the data preprocessing. Consequently, the Abstract data were cleaned up and the meaningful terms were obtained.

4.2 Result of IPC Code Group and IPC-Coded Centroid

According to the IPC field, the number of IPC code groups (down to the fifth level) in 160 patents were 190, as many patents contained more than one IPC code, for example, Patent 06420643 even contained 14 codes. But there were up to 115 groups consisting of only one patent. If the threshold of the number of comprising patents was set to 5, there were 31 leading groups including the first group H01L031/18 (consisting of 48 patents), the second group H01L021/02 (27 patents), and down to the 31st group H01L031/0236 (5 patents), as in Fig. 2.

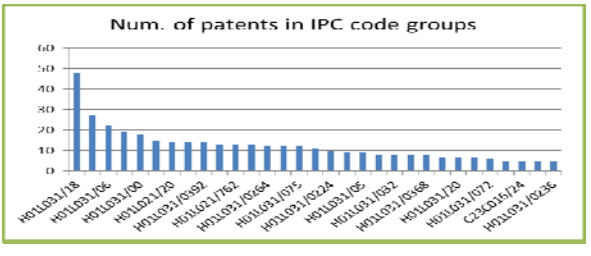


Fig. 2. The number of patents in IPC code groups

The patents contained in each IPC code group were used to generate the IPC-coded centroid for that group via Equation (3). These IPC-coded centroids would be utilized to produce the IPC-oriented clusters afterward.

4.3 Result of IPC-Based Clustering

The clustering alternatives contained the ones from including 4 leading groups, to 5 leading groups, …, till 31 leading groups via Equation (4), using the IPC-coded centroids and the term vectors of Abstract data. In the first alternative, the number of patents distributed into 4 groups was: 118 in H01L031/18, 11 in H01L021/02, 19 in H01L031/06, and 12 in H01L031/036. The accuracy (i.e., average *F* score) of this alternative was 0.4514. Each IPC code group with its distributed patents was regarded as an IPC-oriented cluster. The other alternatives (from including 5 to 31 groups) were then constructed successively. Some of the clustering alternatives with their including clusters and accuracies were calculated and summarized as in Table 1.

Table 1. A summary of clustering alternatives with their including clusters and accuracies

Alternative	Num. of clusters	Num. of patents in cluster	Accuracy
1	4	118, 11, 19, 12	0.4514
2	5	93, 11, 15, 10, 31	0.5109
3	6	87, 11, 15, 10, 29, 8	0.5213
4	7	84, 5, 15, 8, 29, 8, 11	0.514
5	8	77, 5, 12, 8, 26, 8, 11, 13	0.5055
6	9	77, 5, 12, 1, 26, 8, 11, 13, 7	0.5052
7	10	73, 5, 12, 1, 24, 8, 11, 10, 7, 9	0.4988
8	15	55, 2, 2, 1, 24, 8, 0, 7, 10, 9, 10, 0, 10, 10, 12	0.5466
9	20	48, 2, 2, 0, 16, 8, 0, 5, 8, 5, 9, 0, 10, 10, 11, 8, 1, 6, 8, 3	0.5251
10	25	47, 2, 2, 0, 16, 7, 0, 5, 6, 5, 5, 0, 5, 10, 1, 8, 0, 6, 7, 3, 10, 2, 2, 5, 6	0.4992
11	31	42, 2, 0, 0, 16, 7, 0, 2, 6, 5, 5, 0, 4, 7, 1, 8, 0, 3, 6, 3, 10, 2, 0, 0, 6, 5, 4, 5, 5, 2, 4	0.5192
adjusted	9	70, 23, 15, 12, 10, 10, 7, 7, 6	0.5617

According to Table 1, the accuracies of most alternatives varied from 0.50 to 0.53. An adjusted method was suggested which selected the leading clusters from the alternative 11 (with 31 clusters) by setting the threshold of the number of comprising patents to 6, so as to increase the accuracy to 0.5617. After the IPC-oriented cluster generation, the adjusted alternative contained nine clusters: H01L031/18, H01L031/00, H01L021/00, H01L021/20, H01L031/052, H01L031/048, H01L031/04, H01L031/0336, and H01L031/20, and regarded as the IPC-oriented categories.

4.4 Result of Link Analysis

Using link analysis, the linkage strength between 13 significant companies and IPC-oriented categories (as well as years) were calculated and shown in Table 2 (for categories) and Table 3 (for years), where the items in boldface were the ones over the threshold setting 0.15.

(1) Link Measurement: The linkage strength between 13 significant companies and IPC-oriented categories (as well as years) were calculated and shown in Table 2 (for categories) and Table 3 (for years), where the items in boldface were the ones over the threshold setting 0.15.

Table 2. The linkage strength between companies and categories

Company	H01L0 31/18	H01L0 31/00	H01L0 31/048	H01L0 31/052	H01L0 21/20	H01L0 31/0336	H01L0 21/00	H01L0 31/04	H01L0 31/20
Canon (JP)	**0.22**	0.04	0.09	0	**0.15**	0	0	0	0.05
Sharp (JP)	**0.21**	0.12	0	0	0	0	0	0.09	0
Kaneka (JP)	0.05	0	**0.19**	**0.18**	0	0	0	0	**0.17**
Sony (JP)	0.03	0	**0.45**	0	0	0	0	0	0
Matsushita (JP)	0.03	0	0	0	**0.18**	0.13	0	**0.17**	0
Midwest R.I. (US)	0.09	0	0	0.13	0	0	0	0	0.11
AstroPower (US)	0.03	**0.30**	0	0	0	0	0	0	0
Trustees P.U. (US)	0.06	0	0	0.14	0	0	**0.25**	0	0
Angewandte (DE)	0.06	0	0	0	0	0	0	0	0.14
ANTEC S. (DE)	0.10	0	0	0	0	0	0	0	0
Luch (US)	0	**0.33**	0	0	0	0	0	0	0
Seiko E. (JP)	0	0	0	0	**0.38**	0	0	0	0
Showa S. (JP)	0	0	0	0	0	**0.75**	0	0	0

Table 3. The linkage strength between companies and years

Company	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Canon (JP)	0.12	0.10	0.12	**0.19**	0.10	0	0	0	0	0
Sharp (JP)	0	0.08	0	**0.18**	**0.20**	0	0.07	0.09	0	0
Kaneka (JP)	0	0.13	**0.16**	0.06	0	0.10	0.08	0	0	0
Sony (JP)	0.05	**0.15**	0.06	0	0.08	0	0	0	0	0
Matsushita (JP)	0.12	0.05	0	0.07	0	**0.17**	0	0	0	0
Midwest R.I. (US)	0.06	0.05	0	0	0	0	**0.25**	**0.17**	0	0
AstroPower (US)	0.06	0.05	0.13	0	0	0	0	0	0	0
Trustees P.U. (US)	0	0	0.06	0	0.09	0	0.13	0	**0.25**	0
Angewandte (DE)	0.13	0.05	0	0	0	0	0	0	0	0
ANTEC S. (DE)	0	0.05	0.07	0.08	0	0	0	0	0	0
Luch (US)	0	0.05	0.07	0	0	0	0	0	0	**0.33**
Seiko E. (JP)	0.06	0	0.07	0	0.10	0	0	0	0	0
Showa S. (JP)	**0.21**	0	0	0	0	0	0	0	0	0

(2) Relation Recognition: According to the above Table 2 and 3, the relation type for company Canon (JP), as an example, was: type A (as both L1 and L2 are 5). Subsequently, the relation types between companies and categories (as well as years) were summarized below in Table 4 and then used to formulate the technology strategy.

Table 4. The relation types between companies and categories (/years)

Company	Categories	Years	Relation type
Canon (JP)	**H01L031/18**, H01L031/00, H01L031/048, **H01L021/20**, H01L031/20	2000, 2001, 2002, **2003**, 2004	A
Sharp (JP)	**H01L031/18**, H01L031/00, H01L031/04	2001, **2003**, **2004**, 2006, 2007	D
Kaneka (JP)	H01L031/18, **H01L031/048**, **H01L031/052**, H01L031/20	2001, **2002**, 2003, 2005, 2006	D
Sony (JP)	H01L031/18, **H01L031/048**	2000, **2001**, 2002, 2004	H
Matsushita (JP)	H01L031/18, **H01L021/20**, H01L031/0336, **H01L031/04**	2000, 2001, 2003, **2005**	E
Midwest R.I. (US)	H01L031/18, H01L031/052, H01L031/20	2000, 2001, **2006**, **2007**	E
AstroPower (US)	H01L031/18, **H01L031/00**	2000, 2001, 2002	H
Trustees P.U. (US)	H01L031/18, H01L031/052, **H01L021/00**	2002, 2004, 2006, **2008**	E
Angewandte (DE)	H01L031/18, H01L031/20	2000, 2001	I
ANTEC S. (DE)	H01L031/18	2001, 2002, 2003	H
Luch (US)	**H01L031/00**	2001, 2002, **2009**	H
Seiko E. (JP)	**H01L021/20**	2000, 2002, 2004	H
Showa S. (JP)	**H01L031/0336**	**2000**	I

4.5 Result of New Findings

Technology Strategy Formulation: According to the above summarized Table 4, the technical categories and significant companies had been identified, and the relation

types between companies and categories (/years) had been recognized. Subsequently, the technology strategies for companies would be formulated and described as follows.

(a) Business-Level Strategies: From the above Table 4, it would be a reasonable strategic choice for company Canon (JP) to emphasize on H01L031/18, H01L031/00, H01L031/048, H01L021/20, and H01L031/20 (especially on H01L031/18 and H01L021/20) categories. Successively, it would be an acceptable strategic choice for Sharp (JP) on H01L031/18, H01L031/00, and H01L031/04 (especially on H01L031/18); for Kaneka (JP) on H01L031/18, H01L031/048, H01L031/052, and H01L031/20 (especially on H01L031/048 and H01L031/052); …; and finally for Showa S. (JP) on H01L031/0336.

(b) Corporate-Level Strategies: Referring to Table 4, based on the 'Relation type' column, company Canon (JP) would be capable to pick up the "market-pioneer" strategy to be a leader (especially on H01L031/18 and H01L021/20 categories) as it possessed both strong linkages of category and year (type A). It would be suitable for companies Sharp (JP) and Kaneka (JP) to select the "strategic-alliance" strategy to integrate their strength together (especially on H01L031/18, H01L031/048, and H01L031/052 categories) to compete with the leader, because they possessed the medium linkage of category and strong linkage of year (type D). For companies Matsushita (JP), Midwest R.I. (US), and Trustees P.U. (US), it would be also proper for them to choose the "strategic-alliance" strategy to integrate their merits together (especially on H01L021/20, H01L031/04, and H01L021/00 categories) to compete with others, because they possessed both medium linkages of category and year (type E). For companies Sony (JP), AstroPower (US), ANTEC S. (DE), Luch (US), and Seiko E. (JP), it would be appropriate to take the "quick-follower" strategy to follow the forerunners (especially on their focused categories), as they possessed the weak linkage of category and medium linkage of year (type H). Lastly, it would be proper for companies Angewandte (DE) and Showa S. (JP) to select the "niche-market" strategy to emphasize on the small technological area (especially on their focused categories), because they possessed both weak linkages of category and year (type I).

In addition, companies Trustees P.U. (US) and Luch (US) held the most recent patent documents (2008 and 2009) and would be suitable to put more resources into their focused categories H01L021/00 and H01L031/00 respectively.

5 Conclusions

The research framework of IPC-based clustering for formulating the technology strategy on thin-film solar cell has been formed. The experiment was performed and the experimental results were obtained. The clustering alternatives and relation types were generated. The technical categories were: H01L031/18, H01L031/00, H01L021/00, H01L021/20, H01L031/052, H01L031/048, H01L031/04, H01L031/0336, and H01L031/20. The significant companies were: Canon, Sharp, Kaneka, Sony,

Matsushita, Midwest R.I., AstroPower, Trustees P.U., Angewandte, ANTEC S., Luch, Seiko E., and Showa S. The business-level and corporate-level strategies for companies were also formulated. For examples, company Canon (JP) would be capable to pick up the "market-pioneer" strategy to strive to be a leader (especially on H01L031/18 and H01L021/20 categories); companies Sharp (JP) and Kaneka (JP) would be suitable to select the "strategic-alliance" strategy to integrate their strength together (especially on H01L031/18, H01L031/048, and H01L031/052 categories) to compete with the leader. The suggested strategies would be provided to the managers and stakeholders for assisting their decision making.

In the future work, the proposed IPC-based clustering will be explained in more detail. The other aspects of company information (e.g., the public announcement, open product information, and financial reports) may be included in strategy formulation so as to enhance the validity of the proposed method.

Acknowledgments. This research was supported by the National Science Council of the Republic of China under the Grants NSC 99-2410-H-156-014 and NSC 99-2632-H-156-001-MY3.

References

1. Intellectual Property Office, Patent classifications (March 15, 2011), http://www.ipo.gov.uk/pro-types/pro-patent/p-class.htm
2. Chiu, T.-F., Hong, C.-F., Chiu, Y.-T.: To Propose Strategic Suggestions for Companies via IPC Classification and Association Analysis. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS, vol. 6591, pp. 218–227. Springer, Heidelberg (2011)
3. Floyd, S.W., Wolf, C.: Technology Strategy. In: Narayanan, V.K., O'Connor (eds.) Encyclopedia of Technology and Innovation Management, pp. 37–45. John Wiley & Sons (2010)
4. Wikipedia, Strategic planning (October 15, 2010), http://en.wikipedia.org/wiki/Strategic_planning
5. Barney, J.B., Hesterly, W.S.: Strategic Management and Competitive Advantage: Concepts and Cases. Prentice Hall (2010)
6. Solarbuzz, Solar Cell Technologies (October 20, 2010), http://www.solarbuzz.com/technologies.htm
7. Wikipedia, Thin film solar cell (October 20, 2010), http://en.wikipedia.org/wiki/Thin_film_solar_cell
8. Jager-Waldau, A.: PV Status Report 2008: Research, Solar Cell Production and Market Implementation of Photovoltaics, JRC Scientific and Technical Reports (2010)
9. WIPO, Preface to the International Patent Classification (IPC) (October 30, 2010), http://www.wipo.int/classifications/ipc/en/general/preface.html
10. Sakata, J., Suzuki, K., Hosoya, J.: The analysis of research and development efficiency in Japanese companies in the field of fuel cells using patent data. R&D Management 39(3), 291–304 (2009)
11. Kang, I.S., Na, S.H., Kim, J., Lee, J.H.: Cluster-based Patent Retrieval. Information Processing & Management 43(5), 1173–1182 (2007)

12. Chen, Y.L., Chiu, Y.T.: An IPC-based Vector Space Model for Patent Retrieval. Information Processing & Management 47(3), 309–322 (2011)
13. Maimon, O., Rokach, L. (eds.): Data Mining and Knowledge Discovery Handbook, 2nd edn. Springer, Heidelberg (2010)
14. Freeman, L.C.: Centrality in Social Networks: Conceptual Clarification. Social Networks 1, 215–239 (1979)
15. Lee, D.L., Chuang, H., Seamons, K.: Document Ranking and the Vector-Space Model. IEEE Software 14(2), 67–75 (1997)
16. USPTO (2010) USPTO: the United States Patent and Trademark Office (July 14, 2010), http://www.uspto.gov/
17. Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger (October 15, 2009), http://nlp.stanford.edu/software/tagger.shtml
18. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2007)
19. Hotho, A., Nürnberger, A., Paaß, G.: A brief survey of text mining. LDV Forum - GLDV Journal for Language Technology and Computational Linguistics 20(1), 19–62 (2005)

Cluster Control Management as Cluster Middleware

Rajermani Thinakaran[1] and Elankovan Sundararajan[2]

[1] Computing Department, School of Science and Technology, Nilai Univesity College, Negeri Sembilan, Malaysia
[2] Industrial Computing Programme, School of Information Technology Faculty of Information Science and Technology National University of Malaysia, 43600 UKM, Bangi Selangor, Malaysia
rajermani@nilai.edu.my, elan@ftsm.ukm.my

Abstract. Cluster Network technologies have been evolving for the past decades and still gaining a lot of momentum for several reasons. These reasons include the benefits of deploying commodity, off-the-shelf hardware (high-power PCs at low prices), using inexpensive high-speed networking such as fast Ethernet, as well as the resulting benefits of using Linux. One of the major difficulty encounters by cluster administrator in the exploitation of Cluster Networks is handling common tasks, such as setting up consistent software installation or removing on all the nodes or particular node and listing of files or processes will require a lot of time and effort and may affect productivity of the cluster. Even though there are numerous systems available which is known as Cluster Middleware (CM), most of it is developed specifically for in-house use such as scientific research or commercial purpose. Some of them mainly design for job execution rather than node management. To mitigate this problem in the cluster network environment, Cluster Control Management (CCM) system has been developed as a solution for this problem. CCM presently contains six tools, and is extensible to allow more tools to be added. All this six tools are designed for remote cluster administration which includes Switch Manager, Program Manager, Report Manager, User Manager and Administrator Manager. In this paper, we discuss the architecture and the design for CCM using a prototype called UKM C^2M deploys using open-source software (OSS).

1 Introduction

Parallel computing has seen many changes since the days of the highly expensive and proprietary super computers. Changes and improvements in performance have also been seen in the area of mainframe computing for many environments. But these compute environments may not be the most cost effective and flexible solution for a problem. Over the past decade, cluster technologies have been developed that allow multiple low cost computers to work in a coordinated fashion to process jobs have emerged. Their taxonomy is based on how their processors, memory, and interconnect are laid out. One of common system is cluster computer.

A cluster computer is a type of parallel or distributed computer system, which consists of a collection of inter-connected stand-alone computers or node working

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 73–82, 2012.

together as a single integrated computing resource. The nodes on a cluster are networked in a tightly-coupled fashion where they are all on the same subnet of the same domain, often networked with very high bandwidth connections. The nodes are homogeneous; they all use the same hardware, run the same software, and are generally configured identically. Each node in a cluster is a dedicated resource and generally only the cluster Middleware run on a cluster node [1], [2].

Cluster middleware is generally considered as the layer of software sandwiched between the operating system and applications in a cluster environment. Without a proper cluster middleware, a cluster is just a pile of hardware. Having a proper cluster middleware can make the difference between 20% to 99% utilization nodes in the cluster environment. It has been around since 1960's and provides various services in cluster environment and development. [1], [3].

Clusters can be built with any number of nodes because of their loosely coupled architecture. To monitor and manage all the nodes in the cluster is quite tedious. For instance, handling common tasks such as listing of files or processes, and installing software packages on all or specific nodes in the cluster will require a lot of time and effort by cluster administrators and will definitely affect productivity. An Interactive monitoring and management Cluster Middleware (CM) are becoming a necessity.

Although numerous systems are available in CM, most of it are developed for specific purpose either for in-house research uses [5], [6], [7], [8], [9] which is not available for public or proprietary systems [12], [13]. Some of them which are freely available do not meet the requirements needed at the node management and administration level [11].

Therefore, a system called UKM C^2M has been developed. The system provides simple and effective way for cluster administrator to easily manage cluster computer. It consists of a rich set of supporting modules on top of open source software. It allows administrators to specify and install a common software stack on all cluster nodes, and enable centralized control and diagnostics of components with minimal effort.

This paper is organized as follows. Section 2, discusses related work on existing systems. In section 3, we discuss about Cluster Control Management (CCM). Section 4 describes the UKM C^2M architecture, tools and functions while section 5, UKM C^2M prototype. Lastly, section 6 conclusions.

2 Related Work

2.1 Cluster Computer

The recent popularity and the availability of powerful microcomputers and high-speed networks as low-cost commodity components is changing the way we do computing. The idea of using a cluster of computers for running computationally intensive tasks appears to have been conceived by Maurice Wilkes in the late 1970s [3]. This technology opportunity lead to the possibility of using networks of computers for high performance computing, popularly called as cluster computing.

Cluster computers are a set of commodity PC's (nodes) dedicated to a network designed to capture their cumulative processing power for running parallel processing applications. It was specifically designed to take large programs and sets of data and subdivide them into component parts, thereby allowing the individual nodes of the cluster to process their own individual "task" of the program [1]. There are many cluster configurations, but a simple architecture such as the one shown in Fig. 1 is used to visualize the basic concept.

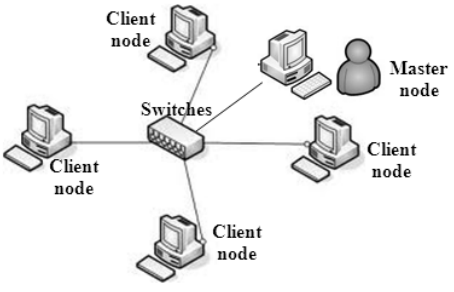


Fig. 1. Typical architecture of a cluster computer

The basic building blocks of clusters are broken down into multiple categories: the cluster nodes, cluster operating system, network switching hardware and the node/switch interconnect Fig 2. Significant advances have been accomplished over the past one decade to improve the performance of both the compute nodes as well as the underlying switching infrastructure.

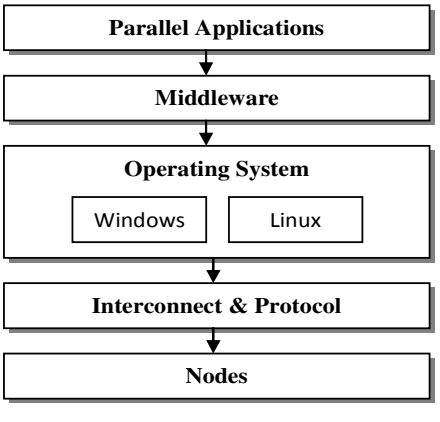


Fig. 2. Cluster components

2.2 Cluster Middleware

Cluster Middleware is a layer that resides between operating system and the applications, is one of the common mechanisms used in clusters which interacts the

user with the operating system. Middleware provides various services required by an application to function correctly. In Fig. 3 Cluster Middleware consist of Job Management System, Cluster Management System, Parallel Execution Libraries and Global Process Space.

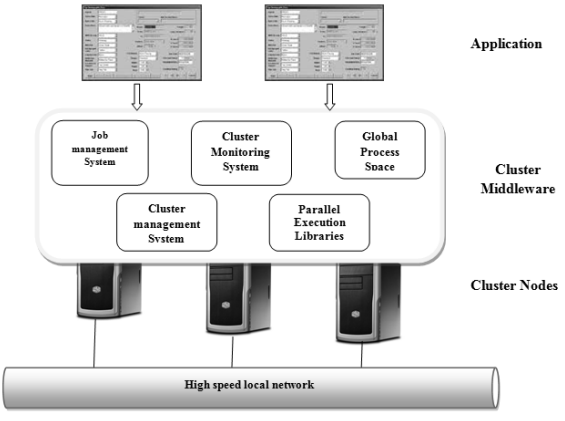


Fig. 3. Cluster Middleware layers

Job Management System (JMS), this system responsible for cluster user's on job control, monitor jobs in progress, and fetch job results. Administrator also use JMS to implement various usage policies, such as fair sharing processing resources, limiting resource consumption and making advance reservation.

While Cluster Management System used to manage nodes in a cluster. It is used by the administrator for cluster configuration. One of the most important tools is automatic installer such as installing software stack on remote nodes.

Cluster Monitoring System collects various information concerning nodes in the cluster such as system load, amount of free memory, amount of free disk, active processes, etc. The collected information is used by cluster administrator to identify potential problems within the cluster or to record cluster utilization. While, cluster users may use these data to analyze and debug execution of their jobs. Furthermore, information about nodes is used by JMS in the process of job scheduling.

Parallel Execution Libraries provide application developers methods for achieving application parallelization and synchronization among different tasks (processes) executing on the same or different computer. Most common parallel libraries are MPI and PVM. Various parallel programming languages can also be used for parallel application development.

Global Process Space provides control of all running processes in any of the node within the cluster. To use this system, all processes on the nodes must have unique identifiers at the operating system level.

2.3 Opens Source Software

UKM C^2M as CCM implemented using open-source software (OSS). OSS is any computer software distributed under a license which is required to have its source code freely available, end-users has the right to modify and redistribute the software, as well as the right to package and sell the software. It is free distribution, while source code freely available. This development approach has helped produce reliable, high quality software quickly and inexpensively. Moreover free software can be developed in accord with purely technical requirements. It does not require thinking about commercial pressure that often degrades the quality of the software [11].

LAMP is an acronym for a solution stack of free, open source software, originally coined from the first letters of Linux (operating system), Apache HTTP Server, MySQL (database software) and PHP as a side-sever programming, while SSH as connectivity tools and MRTG as cluster monitoring tools are used as a principal components to build UKM C^2M.

Ubuntu 10.04 Server Edition (Lucid Lynx), 11th release based on the Debian GNU/Linux distribution provides an up-to-date, stable operating system with a strong focus on usability and ease of installation. A web statistics [16] in January 2011, says that Ubuntu's share of Linux usage is between 40 to 50%.

The **Apache HTTP Server** is web server software notable for playing a key role in the initial growth of the World Wide Web which is developed by Rob McCool at the National Center for Supercomputing Applications, University of Illinois, Urbana-Champaign. This application is available for a wide variety of operating systems, including Linux. Since April 1996 Apache has been the most popular HTTP server software in use. As of May 2011 Apache served over 63% of all websites. It supports a variety of features include server-side programming such as PHP.

PHP is a general-purpose server-side scripting language embedded into the HTML source document and interpreted by a web server with a PHP processor module, which generates the web page document. PHP can be deployed on most web servers, many operating systems and platforms, and can be used with RDBMS. It is available free of charge, and the PHP Group provides the complete source code for users to build, customize and extend for their own use.

Secure Shell (SSH) is a network protocol that allows data to be exchanged using a secure channel between two networked devices. SSH is a protocol that can be used for many applications across many platforms including Linux and primarily used to access shell accounts.

MRTG (Multi Router Traffic Grapher), is a free software for monitoring nodes on cluster network and works under Ubuntu. By integrating MRTG with C2M, it allows an administrator to observe live visual traffic load (i.e. system load and login sessions) on nodes over a specific period of time in graphical form. This data can be useful in tracking down and troubleshooting nodes. MRTG utilizes database (MySQL) in order to generate HTML web pages and graphs containing GIF images via a web browser interface.

2.4 Cluster Management System

The challenge of managing, monitoring and maintaining nodes across cluster environment requires sophisticated, effective, and easy to use tools. Since 1997 when the CplantTM [10] project began, the interest in efficient system management, administration and monitoring tools for clusters has been explored and developed by the researches. Many solutions have been developed by various groups [5] - [9] and vendors [12], [13].

- **C3(Cluster Command and Control):** Is a suite of cluster tools developed at Oak Ridge National Laboratory. The tools cover cluster-wide command execution, file distribution and gathering, process termination, remote shutdown and restart, and system image updates. The C3 tool suite makes use of native UNIX commands modified to serve as parallel commands [5].
- **(Managing Multiple Multi-user Clusters):** A Java-based tool suite called developed by Al Geist and Jens Schwidder at Oak Ridge National Laboratory [6] .Contains six tools which are Multi-User Reservation tool, Job Submission tool , Cluster Monitor tool, Install Software, an Administrative tool and Partition tool.
- **SMILE:** Scalable Multicomputer Implementation using Low-cost Equipment Beowulf Cluster developed at Computer and Network Systems Research Laboratory, Department of Computer Engineering, Kasetsart University [7]. This system is currently used as a platform for parallel processing research, and computational science application development. It is primarily designed to monitor and manage resources in a cluster. Its notable feature for managing resources in a cluster are shutting down or booting up any nodes. Remote login and remote execute command to any nodes. Submit parallel command that execute cluster wide and get the result back at management point. Browse system configuration of any node, query important statistic such as CPU, memory, I/O, and network usage from any node or from the whole cluster are the features for monitoring level.
- **CluMSy:** Was built and tested on a cluster of four nodes connected via a local area network (LAN), one node runs Red Hat Linux 8.0 and the others run Red Hat Linux 9.0 [8]. CluMSy are classified into three categories: (1) cluster-level file management, (2) cluster-level process management, and (3) cluster level account management commands. Aside from monitoring and management functions that incorporate classical UNIX commands, CluMSy also assists users in the preparation and termination of parallel program executions.
- **Sun Cluster Software:** Created by Sun Microsystems, a subsidiary of Oracle Corporation used to improve the availability of software services such as databases, file sharing on a network, electronic commerce websites, etc [12].
- **RemoteNet:** RemoteNet created by ManualEnds Tehnology, is an automation control program that lets you work on another computer remotely (or a group of computers - the cluster system), for example, File Transfer (Download / Upload files) between two PCs, Real Time Screen View to remote client computer, Chat room, Broadcasting message to client computers, and Voice conversations (Voice over IP, VOIP) enable two users to communicate with each other using microphones and speakers. The remote computer can be anywhere on the Internet or in local network. etc [13].

3 Cluster Control Management

The overview of existing system architectures [5] – [8] which were presented in the previous sections, Cluster Management System and Cluster Monitoring System have been integrated together. It is clear that both systems (Cluster Management System and Cluster Monitoring System) have different features. As for that, a new term "Cluster Control Management (CCM)" as new cluster middleware was introduced.

4 UKM C^{2}M: System Architecture

Fig. 4 shows the architecture of UKM C^{2}M. The cluster consists of a number of independent homogenous Linux nodes interconnected by a network. These nodes can be logically viewed as a single-unified resource due to the CCM middleware.

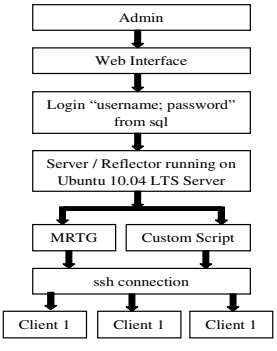


Fig. 4. UKM C^{2}M System Architecture

UKM C^{2}M is designed for simple, low-cost management of distributed and clustered computing environments. After some detailed study on the available systems, six tools was identified. The tools are (1) Switch Manager, (2) Program Manager, (3) Monitoring Manager, (4) User Manager, (5) Admin Manager and (6) Report Manager. It simplifies cluster administrator works by providing management from a single point-of-control.

- **Switch Manager:** Define Provides cluster configuration service has a self introspection mechanism which is automatically carried out by this system according to the cluster administrator settings. These modules provide service for shutting down and rebooting the target or all the nodes from master node. It also display all active and none active nodes in graphic form.
- **Program Manager:** To install new software on selected or all the client nodes is according to date and time setting. Software installation can be performed from master node, eternal storage such as CD, DVD or downloaded from the Internet. Uninstall software on selected or all the client nodes according to the date setting also can be carried out.

- **Admin Manager:** Setting for administrator authentication at its local or remote login into UKM C^2M system and delete account with administrators privilege. After successful authentication administrator can start working with UKM C^2M.When the user administrator a new job, the system opens Secure Shell (ssh) connection to the desired node and executes corresponding action.
- **User Manager:** Create and remove user account at selected node or all nodes at the cluster network level. To start working with UKM C^2M, user should supply a login name and a password. UKM C^2M has a stand-alone access list, which is checked for existence of the provided user name and password.
- **Monitor Manager:** Query important statistics usage on physical resources, such as CPU, memory, swap, disk I/O and network I/O of nodes, switch and storage, which are fundamental for many runtime environments. MRTG (Multi Router Traffic Graph) had been chosen for the monitoring purpose due to flexibility of adapting this tool into UKM C^2M system.
- **Monitor Manager:** Query important statistics usage on physical resources, such as CPU, memory, swap, disk I/O and network I/O of nodes, switch and storage, which are fundamental for many runtime environments. MRTG (Multi Router Traffic Graph) had been chosen for the monitoring purpose due to flexibility of adapting this tool into UKM C^2M system.
- **Report Manager:** Collects, keeps track and generate node information statistic report continuously for later retrieval. The reports saved in storage device as text file or printed as hard copies. These include **Switch Manager Report** – where it provides cluster-wide configuration information report consists of shutting/reboot date, client name (node), user name and administrator incharge. Also include current active node with the user name. **Program Manager Report** – generates details such as application name, version, size, install/uninstall date and administrator incharge. **User Manager Report** – produces details such as user name, date and time log in and log out. **Admin Manager Report** – produces details such as accounts with administrator privilege, name, date and time log in and out and job process.

5 UKM C2M: Prototype Developed Used Open Source Software

The prototype implementation of UKM C^2M was built and tested on a cluster of four (4) nodes connected via a local area network (LAN) runs Ubuntu 10.04 LTS Server. Cluster administrator connects to the UKM C^2M agent with a web browser. After initial authentication and authorization, administrator can define and execute actions for the services requested from the nodes user. The actions are then stored in the database and middleware agent takes care of dispatching them to the individual machines. On the cluster nodes middleware agent also is responsible for executing the actions, following up their progress and in case of a need, performs maintenance and reports on the progress of the action.

Benefits for typical administrator (users) are where many of the cluster management systems have a lot of command line options, scriptable parameters, and different tools, which are very difficult to learn for the typical user. UKM C^{2}M WEB based interface simplifies job submission and management while WEB interface does not require any additional software and can be used from any HTML compliant browser.

Fig. 5. UKM C^{2}M: Administrative Login Menu and Main Menu

6 Conclusion

The challenge of maintaining, monitoring and administrating independent homogenous nodes across the computer cluster requires an effective and easy to use system. To overcome this, UKM C^{2}M as Cluster Control Management as a new middleware had been developed to manage, monitor and administer nodes in a cluster networks. It allows a single system administrator to manage nodes in the clusters at the same time. UKM C^{2}M is being developed using open source software and six tools are presently incorporated into the suite which are Program Manager, System Manager, User Manager, Admin Manager, Monitor Manager and Report Manager. With this system it is hoped that the productivity of cluster administrator will increase.

References

1. Sloan, J.D.: High Performance Linux Cluster with Oscar, Rocks, openMosix and MPI. O'Reilly Media Inc. (2004)
2. Baker, M., Apon, A., Buyya, R., Jin, H.: Cluster computing and applications. In: Kent, A., Williams, J. (eds.) Encyclopedia of Computer Science and Technology, New York, pp. 87–125 (2002)

3. Buyya, R.: High performance cluster computer: System and architectures. Prentice Hall PTR, United States of America (1999)
4. Needham, R.M., Herbert, A.J.: The Cambridge distributed computing systems. Addison-Wesley, MA (1982)
5. Cluster Command and Control (C3) Tool Suite, http://www.csm.ornl.gov/torc/C3/Papers/pdcp-v2.0.pdf
6. Managing Multiple Multi-user PC Clusters, http://www.csm.ornl.gov/torcpress
7. Uthayopas, P., Phaisithbenchapol, S., Chongbarirux, K.: Building a Resources Monitoring System for SMILE Beowulf Cluster. In: Proc. HPC ASIA 1999 (1999)
8. Jane, E., Chua, E.: CluMSy: A Middleware for Cluster Management. Philippine Computing Journal 1(1) (March 2006)
9. ChameleonII: Laboratory for Communication Networks. Royal Institute of Technology (KTH), Stockholm, Sweden (April 2007)
10. Brightwell, R.B., Fisk, L.A., Greenberg, D.S., Hudson, T.B., Levenhagen, M.J., Maccabe, A.B., Riesen, R.E.: Massively Parallel Computing Using Commodity Components. Parallel Computing 26(2-3), 243–266 (2000)
11. http://www.opensource.org/docs/definition.php
12. Solaris, http://www.oracle.com/us/products/servers-storage/solaris/cluster-067314.html
13. RemoteNet, http://www.manualends.com/Products/Remote/Me-CRME/MECRME.htm

A Novel Nonparametric Approach for Saliency Detection Using Multiple Features

Xin He, Huiyun Jing, Qi Han, and Xiamu Niu

Department of Computer Science and Technology, Harbin Institute of Technology, No.92, West Da-Zhi Street, Harbin, China
xm.niu@hit.edu.cn

Abstract. This paper presents a novel saliency detection approach using multiple features. There are three types of features to be extracted from a local region around each pixel, including intensity, color and orientation. Principal Component Analysis(PCA) is employed to reduce the dimension of the generated feature vector and kernel density estimation is used to measure saliency. We compare our method with five classical methods on a publicly available data set. Experiments on human eye fixation data demonstrate that our method performs better than other methods.

Keywords: Saliency detection, Multiple features, PCA, Kernel density estimation.

1 Introduction

In the computer vision field, research on saliency detection has attracted much more interest. Visual saliency plays an important role that distinguishes important or interesting parts from their surrounding and drives our perceptual attention. Saliency detection is indispensable in many theories of visual attention and has been widely applied in many areas such as object detection[1], image cropping[2], image browsing[3], and image/video compression[4].

There are two types of saliency detection. The first type is bottom-up saliency detection, which is fast, stimulus-driven and independent of the knowledge. The second type is top-down saliency detection, which is show, goal-oriented and requires the prior knowledge.

Over the last two decades, many studies have focused on the bottom-up saliency detection. The most important and influential model is proposed by Itti et al.[5]. Based on feature integration theory[6], Itti's saliency model combines three feature channels including intensity, color and orientation into the final saliency map. Bruce and Tsotsos[7] proposed a overt attention model based on the principle of information maximization sampled from scene. Shannon's self-information measure $-\log(p(x))$ is used to measure saliency, where x is a local feature vector in the image. The local features are derived from independent component analysis(ICA), where the ICA basis set was learned from a set of 360000 $7 \times 7 \times 3$ image patches from 3600 natural images. Zhang et al.[8] proposed a

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 83–90, 2012.

similar approach based on a Bayesian framework from which bottom-up saliency emerges as the self-information of visual features. Their measure of saliency is derived from natural image statistics, obtained in advance from a collection of natural image. Gao et al.[9] proposed a discriminant center-surround hypothesis architecture formulated as a classification problem that measures saliency by expected probability of error between feature responses in center and surround. The saliency of each location is equal to the discriminant power of a set of features and the saliency detector is derived from various stimulus modalities, including intensity, color, orientation and motion. Seo and Milanfar[10] proposed a nonparametric saliency detection approach based on local steering kernel(LSK) features and utilized self-resemblance mechanism to compute saliency map, where each pixel indicates the statistical likelihood of saliency of a feature matrix given its surrounding feature matrices. The local steering kernel feature robustly obtains the structure information of local region of image by analyzing the pixel value differences based on estimated gradients.

The bottom-up saliency detection can be divided into two categories, according to whether the input image is decomposed into some feature maps. The methods proposed by Itti et al.[5] and Gao et al.[9] decompose the input image into three feature maps and obtain the feature saliency maps computed by different saliency computational schemes. And linear combinations are used to combine three features saliency maps for the final saliency map. The methods belong to the first category. On the contrary, the second category is that the saliency map is computed from a single feature, like Bruce et al.[7], Zhang et al.[8](ICA filter) and Seo et al.[10].

In this paper, we present a novel nonparametric approach to measure saliency of natural color image by computing kernel density estimation of the feature vector, which combines multiple features into a single feature. We describe our method in Section 3. In Section 4, we compare our method with five methods and conclusions are given in Section 5.

2 Related Works

In this section, we briefly introduce the two previous models, Itti's model[5] and Bruce's model[7]. As referred in Section 1, Itti's model belongs to the category of combination of decomposed multiple features map and Bruce's model belongs to the category of a single feature for measuring saliency.

2.1 Itti's Model

Fig. 2 shows the general architecture of Itti's model. Itti's model firstly decompose the input image into three feature channels, including color, intensity, and orientation. Dyadic gaussian pyramids, linear center-surround differences and normalization operations are used to produce feature map. The main steps of the process are represented as follows.

$$\mathcal{I}(c,s) = |I(c) \ominus I(s)| \tag{1}$$

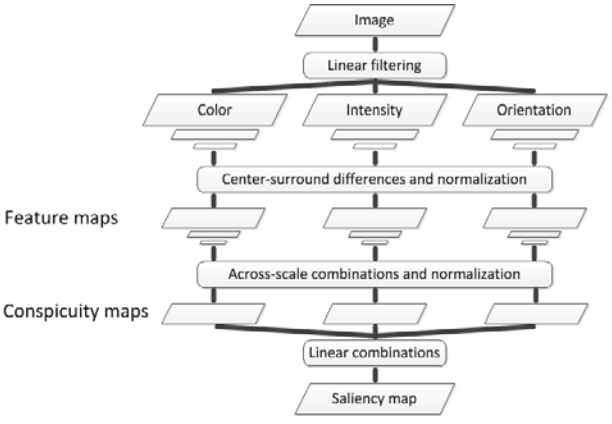


Fig. 1. General architecture of Itti's model

$$\begin{aligned} \mathcal{RG}(c,s) &= |(R(c) - G(c)) \ominus (G(s) - R(s))| \\ \mathcal{BY}(c,s) &= |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \end{aligned} \tag{2}$$

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)| \tag{3}$$

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}(\mathcal{I}(c,s)) \tag{4}$$

$$\overline{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{5}$$

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta))) \tag{6}$$

where $c (c \in \{2,3,4\})$ and $s (s \in \{5,6,7,8\})$ are spatial scales created using dyadic gaussian pyramids. $\ominus$ denotes the across-scale difference. $\mathcal{N}$ is normalization operator. $\mathcal{I}$, $\mathcal{C}$ and $\mathcal{O}$ are feature maps, which are combined into conspicuity maps $\overline{\mathcal{I}}$, $\overline{\mathcal{C}}$ and $\overline{\mathcal{O}}$.

The final saliency map $\mathcal{S}$ can be obtained by normalizing and summing the conspicuity maps as follows.

$$\mathcal{S} = \frac{1}{3}(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})) \tag{7}$$

2.2 Bruce's Model

Bruce's model measures saliency as the maximum information of local image patches. Firstly, a sparse basis is firstly learned via ICA(Independent Component Analysis) on a large sample of small RGB patches from a set of 3600 natural images. Then for the input image, ICA coefficients are extracted from

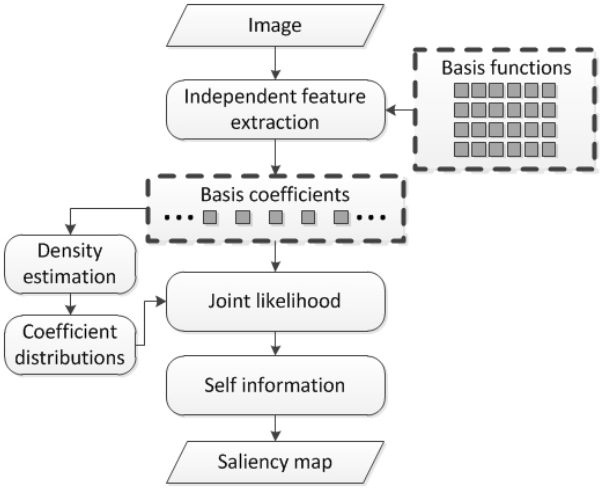


Fig. 2. The framework of Bruce's model

local neighborhoods around each pixel as a feature vector. Based on the probability density estimate, the distribution of values for any single coefficient can be obtained by a set of ICA coefficients for every local neighborhood. The final saliency is computed as Shannon's measure of self-information from the product of all the individual likelihoods corresponding to a particular local region. The framework of Bruce's model is shown in Fig. 2.

3 The Proposed Saliency Detection Method

Itti's model only simply combines the three conspicuity maps into the final saliency map through linear combinations. Global spatial distribution of features aren't considered, which implies that the features of lower probability attract more attention. Bruce's model only considers the ICA features and ignores other features. We try to combines three features into a single feature and measure saliency using gaussian kernel density estimation to compute the probability of the feature.

Similar to Itti's model, the initial image are performed by decomposition, gaussian pyramids, linear center-surround and normalization operations. Different to Itti's model, only four spatial scales are used in gaussian pyramids for simplicity in our method.

$$\mathcal{I} = |I(c) \ominus I(s)| \tag{8}$$

$$\begin{aligned} \mathcal{RG} &= |(R(c) - G(c)) \ominus (G(s) - R(s))| \\ \mathcal{BY} &= |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \end{aligned} \tag{9}$$

$$\mathcal{O} = |O(c, \theta) \ominus O(s, \theta)| \tag{10}$$

where $c \in \{2, 3\}$, $s \in \{5, 6\}$ and $\theta \in \{0°, 45°, 90°, 135°\}$.

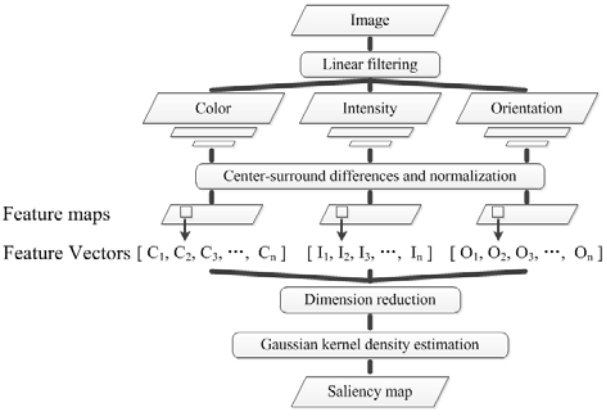


Fig. 3. The framework of our method

For each feature map, we extract a small region(3×3) around each pixel as the feature vector of the pixel and combine the three types of features into a single feature vector. Generally speaking, the data of the feature vector is high dimension. So PCA(Principal Component analysis) is used for dimension reduction.

$$\mathcal{F}(i) = PCA(\mathcal{I}(r(i)), \mathcal{RG}(r(i)), \mathcal{BY}(r(i)), \mathcal{O}(r(i))) \tag{11}$$

where $r(i)$ denotes a small region(3×3) around pixel i.

In order to calculate the probability of each integrated feature, we use a gaussian kernel density estimation and the saliency of each pixel can be measured as follows.

$$\mathcal{S}(i) = \mathcal{N}(\frac{1}{p(\mathcal{F}(i))}) \tag{12}$$

where $\mathcal{N}(\cdot)$ is a normalization function, normalizing the value to $[0, 1]$.

4 Experimental Results

We perform experiments on human eye fixation data from natural images. The dataset is provided by Bruce and Tsotsos[7] as the benchmark dataset for

Table 1. Experimental Result

Model	KL(SE)	AUC(SE)
Itti et al.[5]	0.1130(0.0011)	0.6146(0.0008)
Bruce and Tsotsos[7]	0.2029(0.0017)	0.6727(0.0008)
Gao et al.[9]	0.1535(0.0016)	0.6395(0.0007)
Zhang et al.(DoG filters)[8]	0.1723(0.0012)	0.6570(0.0007)
Zhang et al.(ICA filters)[8]	0.2097(0.0016)	0.6682(0.0008)
Our Method	0.2608(0.0024)	0.6597(0.0008)

comparing human eye predictions between methods. The dataset contains eye fixation data from 20 subjects for a total of 120 natural images. The Kullback-Leibler (KL) divergence and the area under receiver operating characteristic(AUC) were computed as performance metrics. A high value of these two metrics means better performance. Zhang et al.[8] noted that the original KL divergence and ROC

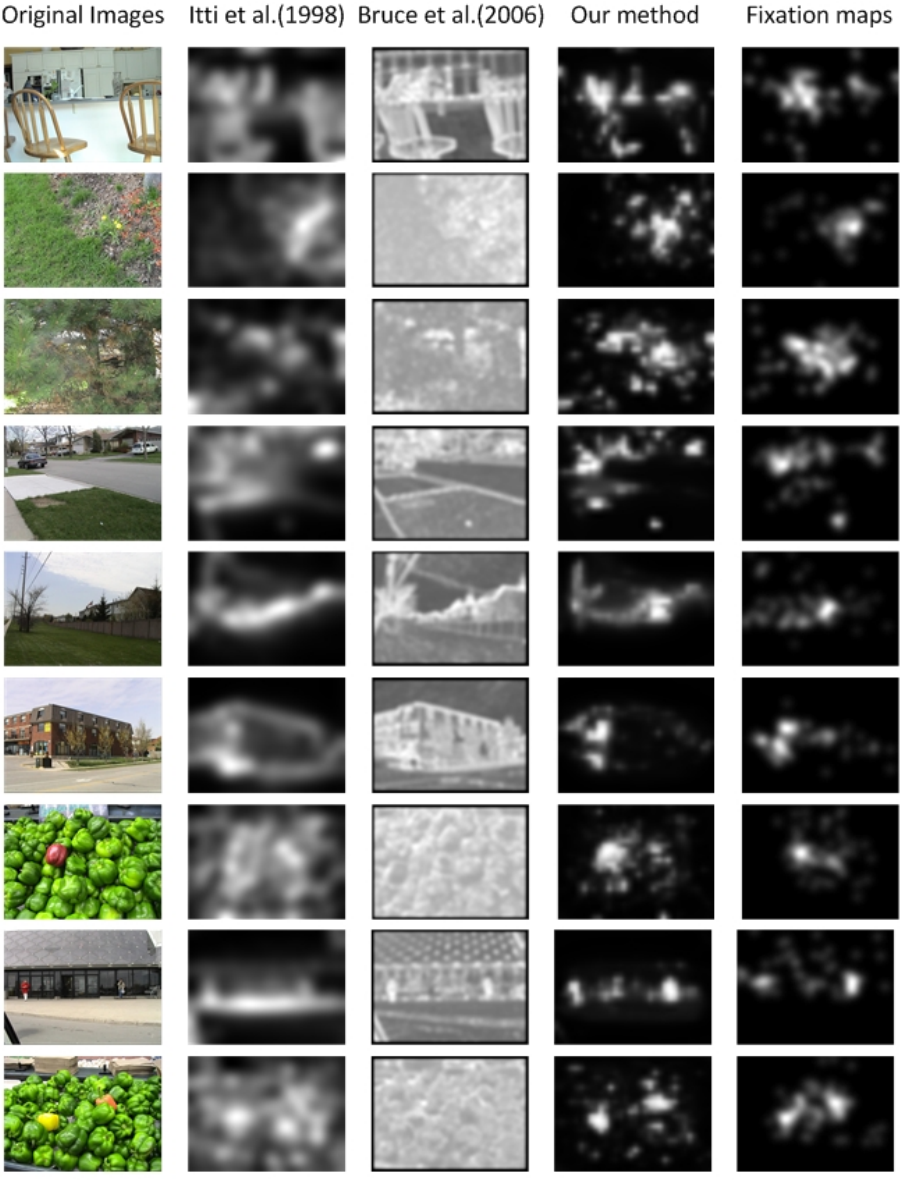


Fig. 4. Examples of saliency map

area measurement are corrupted by an edge effect which yielding artificially high results. In order to eliminate border effects, we adopt the same procedure described by Zhang et al.[8] to compute KL divergence and ROC area. Our method is compared with five classical methods, including Itti et al.[5], Bruce et al.[7],Gao et al.[9], and Zhang et al.[8]. As we can see in Table 1, our method performs a little worse than Bruce and Tsotsos' method on this dataset by the ROC metric, but significantly better than the five classical methods by the KL metric.

Fig. 4 provides some visual comparisons of our method with Itti et al.[5] and Bruce et al.[7]. Visually, our method is more consistent with human eye fixation data.

5 Conclusion

In this paper, we present a novel saliency detection approach using multiple features. Based on the analysis of Itti's model and Bruce's model, we extract three types of features from a local region around each pixel. The features of intensity, color and orientation are combined into a single feature vector and Principal Component Analysis(PCA) is employed to reduce the dimension of the generated feature vector. The saliency of each pixel is determined by computing the kernel density estimation of the features after dimension reduction. We compare our method with five classical methods on visual fixation data. Our method performs a little worse than Bruce' method by the ROC metric, but significantly better than the five classical methods by the KL metric. In future work, we will incorporate more features(e.g. texture feature[11], motion feature[12]) to detect saliency.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (60832010, 61100187), the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF. 2010046) and the China Postdoctoral Science Foundation (2011M500666).

References

1. Papageorgiou, C., Poggio, T.: A trainable system for object detection. International Journal of Computer Vision 38(1), 15–33 (2000)
2. Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic thumbnail cropping and its effectiveness. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, Vancouver, Canada, pp. 95–104 (October 2003)
3. Rother, C., Bordeaux, L., Hamadi, Y., Blake, A.: AutoCollage. ACM Transactions on Graphics 25, 847–852 (2006)
4. Itti, L.: Automatic foveation for video compressing using a neurobiological model of visual attention. IEEE Transactions on Image Processing 13(10), 1304–1318 (2004)
5. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)

6. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–138 (1980)
7. Bruce, N.D.B., Tsotsos, J.K.: Saliency based on information maximization. Advances in Neural Information Processing Systems 18, 155–162 (2006)
8. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: a bayesian framework for saliency using natural statistics. Journal of Vision 8(7), 32–51 (2008)
9. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. Journal of Vision 8(7), 13–31 (2008)
10. Seo, H.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision 9(12), 15–41 (2009)
11. Kaganami, H.G., Ali, S.K., Zou, B.: Optimal Approach for Texture Analysis and Classification based on Wavelet Transform and Neural Network. Journal of Information Hiding and Multimedia Signal Processing 2(1), 33–40 (2011)
12. Liu, P., Jia, K.: Research and Optimization of Low-Complexity Motion Estimation Method Based on Visual Perception. Journal of Information Hiding and Multimedia Signal Processing 2(3), 33–40 (2011)

Motion Vector Based Information Hiding Algorithm for H.264/AVC against Motion Vector Steganalysis

Huiyun Jing, Xin He, Qi Han, and Xiamu Niu

Department of Computer Science and Technology, Harbin Institute of Technology, No.92, West Da-Zhi Street, Harbin, China

Abstract. In this paper, we present a novel motion vector based information hiding algorithm for H.264/AVC against motion vector steganalysis. For resisting motion vector steganalysis, the underlying statistics of the motion vectors used by motion vector steganalysis are remained during the secret information embedding process. We choose one part of motion vectors to be modified for restoring the statistics used by motion vector steganalysis. In order to guarantee imperceptibility, secret information is embedded by modulating the best search points of other part of motion vectors during the quarter-pixel motion estimation process. Experimental results show that the proposed algorithm effectively resist motion vector steganalysis, while having good video quality.

Keywords: Video information hiding, Motion vector(MV), H.264/AVC, motion vectors steganalysis counteraction.

1 Introduction

In the information security field, information hiding has attracted increasing attention as an effective and important complement to the traditional encryption method. For its unique characteristics in security, imperceptibility, robustness, etc, information hiding has been widely used in copyright protection and covert communication. During recent years, much research effort has been made to information hiding algorithm [1,2,3]. Among the various carriers of information hiding, video is the most widely used one. This paper focuses on the research of video information hiding algorithm in the compressed domain based on H.264/AVC standard.

In existing compressed domain based video information hiding algorithms, DCT coefficients [4,5] and motion vectors [6,7,8,9,10] are often chosen for embedding secret information. However, much fewer nonzero DCT coefficients are available for embedding than motion vectors in low bit-rate video. Because the motion vector based video information hiding algorithms have the advantage of much less degradation of the reconstructed image quality, they have drawn more and more researchers attention [6,7,8,9,10].

Video information hiding algorithms based on motion vectors can be classified into two categories: 1. secret information is embedded by directly altering the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 91–98, 2012.

motion vectors[6,7,8]; 2. secret information embedding is performed during the motion estimation process[9,10]. As the secret information is embedded during the motion estimation process, the altered motion vectors are selected from the sub-optimum motion vectors during motion estimation process. So, the second category of algorithms leads to much less degradation of the reconstructed image quality than the first category. However these existing video hiding algorithms based on motion vector have no ability against motion vector steganalysis[11]. Motion vector steganalysis employs the difference between probability mass function of the difference between adjacent motion vectors in the stego-video and in the cover video to determine the existence of the hidden message.

In this paper, we present a novel information hiding algorithm against motion vector steganalysis with guaranteeing imperceptibility. We try to make the probability mass function of stego-video remain the same as the probability mass function in cover video, which results in the invalidation of motion vector steganalysis. Moreover, secret information is embedded by changing the best motion vectors to the sub-optimum motion vectors during the quarter-pixel motion estimation process in H.264/AVC.

2 Motion Vector Steganalysis

Su et al.[11] proposed a video steganalysis algorithm against motion vector steganography[6,7,8,9,10]. Utilizing the phenomenon that the embedding process ruins the spatial and temporal correlation of the motion vector field, they extracts the statistical characteristics reflecting the phenomenon and employs the feature classification technique (support vector machine) to determine the existence of the hidden message.

Su et al.[11] modeled the existing motion vector embedding process as adding an independent noise signal to the horizontal and vertical components of the motion vectors. So, the horizontal component X_i^h of the motion vectors for the ith macroblock in the stego-video can be represented as follows

$$X_i^h = S_i^h + \eta_i^h, i = 1, \cdots, N \tag{1}$$

where S_i^h are the horizontal component of the motion vectors in the ith macroblock of the cover video, η_i^h represents the secret data embedded into the corresponding horizontal component of the motion vectors, and N is the number of the macroblocks in one video frame.

Su et al.[11] define a differential operator ∇ to calculate the difference between the motion vectors of the adjacent macroblocks in the same frame. Let $h_\triangle[n]$, $h_s[n]$ and $h_x[n]$ respectively be the probability mass function of $\nabla\eta_i^h$, ∇S^h and ∇X^h. $h_\triangle[n]$ introduces aliasing effect into $h_x[n]$. The aliasing degrees of the probability mass function $h_x[n]$ are defined as

$$\begin{aligned} E_1 &= \frac{h[k-1] + h[k+1]}{2} - ah[k] \\ E_2 &= \frac{h[-k+1] + h[-k-1]}{2} - ah[-k] \end{aligned} \tag{2}$$

where k the amplitude of the motion vector modification, usually $k = 1$, and α is an adaptive factor to adjust precision of detection. Due to the aliasing effect introduced by $h_{\triangle}[n]$, the values of E_1 and E_2 computed from the cover video are usually larger than the values computed from the stego-video.

Furthermore, the aliasing degrees of the probability mass function can be also defined in the frequency domain. The characteristic function center of mass (COM) is introduced as a measure to quantify the alteration degrees.

$$C(H[m]) = \frac{\sum_{i=0}^{T} i|H[i]|}{\sum_{i=0}^{T} |H[i]|} \tag{3}$$

where $T = N/2$, $H_s[m]$, $H_x[m]$ and $H_{\triangle}[m]$ are obtained by taking N-element Discrete Fourier Transform of the probability mass functions $h_s[n]$, $h_x[n]$ and $h_{\triangle}[n]$ respectively.

All the above processes can be repeated for the vertical component of the motion vectors in the spatial domain. The three statistics E_1, E_2 and COM of vertical and horizontal directions in the temporal domain can be also obtained. All these statistics calculated from the spatial and temporal domains yield 12 statistics, which form the feature vector for training the SVM to tell the stego-videos from cover videos.

We employ the steganalysis method to test the performance of the video information hiding algorithm proposed by Qiu et al.[10] against motion vector steganalysis. The classification and error rates, computed by only choosing 1/3 of motion vectors to be embedded, are listed in Table 1. The experimental results show that the the video information hiding algorithm proposed by Qiu et al. has no ability against motion vector steganalysis. Positive detection (PD) indicates classifying the stego videos correctly, while Negative detection (ND) denotes classifying the nonstego videos correctly. False positive (FP) represents classifying the nonstego videos as stego videos. False negative (FN) indicates classifying the stego videos as nonstego videos.

Table 1. Steganalysis counteraction ability of Qiu's method

Video	PD	ND	Classification rate	FP	FN	Error rate
football	94.19%	45.95%	70.07%	54.05%	5.81%	29.93%
stefan	91.28%	75.84%	83.56%	24.16%	8.72%	16.44%
bus	97.99%	67.11%	82.55%	32.89%	2.01%	17.45%

3 Proposed Motion Vector Based H.264/AVC Video Information Hiding Algorithm

We perform the same information embedding procedure as the algorithm proposed by Qiu et al.[11], which embeds secret information during the quarter-pixel motion estimation process in H.264/AVC. All the statistics used by the motion vector steganalysis are relevant to the probability mass function of the difference between adjacent motion vectors. We try to keep the probability mass function

of the difference between adjacent motion vectors in stego-video same as the probability mass function in cover video, which is different form the algorithm proposed by Qiu et al.. Choosing one part of motion vectors to be modified for restoring the probability mass function in stego-video and choosing other parts for embedding secret information are the key point of the ability against motion vectors steganalysis.

3.1 Probability Mass Function Restoration

Secret information embedding process ruins the spatial and temporal correlation of the motion vector field in the video sequence, which leads to the aliasing effect in the probability mass function of the difference between adjacent motion vectors in stego-video. Su et al.[11] employs the E_1, E_2 and COM statistics to compute the aliasing degree of the probability mass function. These statistics E_1 and E_2 are only relevant to $h[k-1]$, $h[k+1]$, $h[k]$, $h[-k+1]$,$h[-k-1]$, $h[-k]$. When $k=1$, E_1 and E_2 are relevant to $h[-2]$, $h[-1]$, $h[0]$, $h[1]$, $h[2]$. The statistics COM are only relevant to H[0],...,H[i],...,H[N/2], where $H[i]$ is N-element Discrete Fourier Transform of the probability mass functions $h[i]$.

The probability mass function of the difference between adjacent motion vectors in cover video obeys spiculate super-Gaussian distribution with zero peaks. However, the probability mass function of stego video obeys Gaussian distribution with zero peaks. The difference of these two kinds distribution is primarily reflected in the difference of values of $h[-2]$, $h[-1]$, $h[0]$, $h[1]$ and $h[2]$. $h_x[0]$ is much smaller than $h_s[0]$, while $h_x[i]$ is larger than $h_s[i]$ $(i=\pm 2,\pm 1)$. The difference between $h_x[j]$ and $h_s[j]$ are little, where $j=\pm 3,\ldots \pm n$.

Based on the above analysis, it is feasible that trying to keep the values of $h_x[-2]$, $h_x[-1]$, $h_x[0]$, $h_x[1]$, $h_x[2]$ the same as $h_s[-2]$, $h_s[-1]$, $h_s[0]$, $h_s[1]$, $h_s[2]$ for the invalidation of SVM classifier. Then the algorithm obtains the ability against motion vector steganalysis. During the secret information embedding procedure, we choose the some of motion vectors X_{i+1} satisfying $\nabla X_i = X_i - X_{i+1} = \pm 2$ or $\nabla X_i = X_i - X_{i+1} = \pm 1$ to be modified. The modified motion vectors $X_{i+1}^{'}$ satisfy $\nabla X_i = X_i - X_{i+1} = 0$.

Because the final modified motion vectors should be selected from these quarter-pixel locations $\{1,2,3,4,A,5,6,7,8\}$(Fig. 1) in order to curb the bit-rate increase, not all the chosen motion vectors X_{i+1} satisfying $\nabla X_i = X_i - X_{i+1} = \pm 2$ or $\nabla X_i = X_i - X_{i+1} = \pm 1$ can be modified to satisfy $\nabla X_i = X_i - X_{i+1} = 0$.

Given the chosen motion vector X_{k+1} satisfying $\nabla X_k = X_k - X_{k+1} = 2$, only the motion vector X_{k+1} pointed at one of these locations $\{1,4,6\}$(Fig. 1) can be shifted to one of these new locations $\{3,5,8\}$(Fig. 1) for satisfying $\nabla X_k = X_k - X_{k+1} = 0$. And the final modified motion vectors are chosen from these locations $\{3,5,8\}$(Fig. 1) by following the Lagrangian optimization rule which is complied by H.264/AVC standard. When the the chosen motion vector X_{k+1} satisfies $\nabla X_k = X_k - X_{k+1} = -2$, only the motion vector X_{k+1} pointed at one of these locations $\{3,5,8\}$(Fig. 1) can be shifted to one of these new locations $\{1,4,6\}$(Fig. 1) for satisfying $\nabla X_k = X_k - X_{k+1} = 0$.

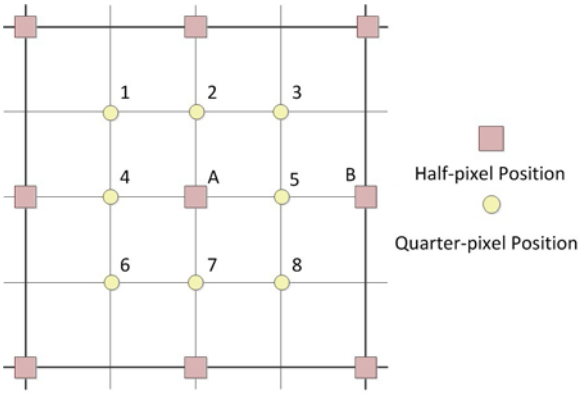


Fig. 1. Fractional motion vector search positions

The above analysis can be repeated on the case of X_{k+1} satisfying $\nabla X_k = X_k - X_{k+1} = \pm 1$. While the the chosen motion vector X_{k+1} satisfies $\nabla X_k = X_k - X_{k+1} = 1$, the motion vector X_{k+1} only pointed at one of these locations $\{1, 4, 6\} \bigcup \{2, A, 7\}$ (Fig. 1) can be shifted to one of these new locations $\{2, A, 7\} \bigcup \{3, 5, 8\}$(Fig. 1) for satisfying $\nabla X_k = X_k - X_{k+1} = 0$.

After the restoration procedure, there is only little difference between the probability mass function of modified motion vectors in the stego-video and the motion vectors in the cover video. Thus, the SVM classifier can not tell the difference between stego-videos and cover-videos. Therefore, the SVM classifier can not classify stego-videos correctly.

3.2 Embedding Scheme

We only choose 1/3 of motion vectors to be embedded secret information. While we choose 1/3 of motion vectors to be modified for restoring the probability mass function in the spatial domain, we choose 1/3 of motion vectors to be modified for restoring the probability mass function in the temporal domain. Let N be the number of the inter-coded blocks. The secret information is embedded into the H.264/AVC compressed video stream by going through following steps:

1. If $N\%3 = 0$, we respectively embed the secret information bits w_n and w_{n+1} and into the horizontal and vertical components of the motion vector in the N inter-coded blocks;
2. If $N\%3 = 1$, we modify the horizontal and vertical components of the motion vectors in the N inter-coded blocks for restoring the probability mass function in the spatial domain.
3. If $N\%3 = 2$, we modify the horizontal and vertical components of the motion vectors in the N inter-coded blocks for restoring the probability mass function in the temporal domain.

4 Experimental Results

The proposed algorithm has been simulated on CIF sequences. The tests were performed using the following H.264 configuration CABAC, QP=30, 25 frames/sec and with a GOP structure of IPPPPPPPP. The performance of the video information hiding algorithms was evaluated with the metrics of subjective testing, Peak-Signal-to-Noise-Ratio (PSNR), and the ability against motion vector steganalysis. Experimental results show that the the proposed algorithm effectively resist motion vector steganalysis, while having good video quality.

4.1 Subjective Testing

For making subjective testing, we compared the original video frames and the corresponding stego video frames. Fig. 2 is the comparison between the original video frame and the corresponding stego-video frame with hidden information. As can be seen in Fig. 2, there is no perceptual difference between them.

Fig. 2. Comparison of images extracted from cover video and stego-video

4.2 Peak-Signal-to-Noise-Ratio (PSNR) Testing

Peak Signal-to-Noise Ratio(PSNR) is commonly used as a measure of quality degradation of reconstructed video images after embedding information. Table 2 shows the average PSNR values of the stego-videos with our algorithm and Qiu's algorithm. As can be seen in Table 2, our algorithm get the comparable results with Qiu's algorithm.

Table 2. PSNR comparison

Video	PSNR(Qiu's algorithm)	PSNR(Our algorithm)
football	35.710db	35.636db
stefan	34.512db	34.384db
bus	34.488db	34.346db

4.3 Ability against Motion Vector Steganalysis

Here, we test the ability against motion vector steganalysis. Table 3 shows that our algorithm has the ability against motion vector steganalysis, when choosing 1/3 of motion vectors to be embedded. And then, we compare the ability against motion vector steganalysis of Qiu's algorithm and our algorithm. Table 4 shows that our algorithm performs more better than Qiu's algorithm.

Table 3. Steganalysis conteraction ability of our algorithm

Video	PD	ND	Classification rate	FP	FN	Error rate
football	48.84%	57.92%	53.38%	42.08%	51.16%	46.62%
stefan	34.83%	67.11%	45.97%	32.89%	75.17%	54.03%
bus	12.75%	89.93%	51.34%	10.07%	9.06%	48.66%

Table 4. Comparison of steganalysis conteraction ability

Video	Classification rate Qiu's algorithm	Classification rate Our algorithm
football	70.07%	53.38%
stefan	83.56%	45.97%
bus	82.55%	51.34%

5 Conclusions and Future Work

This paper presents a novel motion vectors based information hiding algorithm for H.264/AVC. The proposed algorithm embeds information by changing the parity of one part of motion vector prediction errors during the quarter-pixel motion estimation process. And choosing the other parts of motion vectors to be modified for restoring the probability mass function of the difference between adjacent motion vectors in stego-video makes the algorithm get the ability against motion vectors steganalysis. Experimental results show that the proposed algorithm effectively resist motion vector steganalysis, at the same time cause unnoticeable quality degradation.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (60832010, 61100187), the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF. 2010046) and the China Postdoctoral Science Foundation (2011M500666).

References

1. Medeni, M.O., Souidi, E.M.: A novel steganographic protocol from error-correcting codes. Journal of Information Hiding and Multimedia Signal Processing 1(4), 337–343 (2010)
2. Kermanidis, K.L.: Capacity-rich knowledge-poor linguistic steganography. Journal of Information Hiding and Multimedia Signal Processing 2(3), 247–258 (2011)
3. Huang, X., Abe, Y., Echizen, I.: Capacity adaptive synchronized acoustic steganography scheme. Journal of Information Hiding and Multimedia Signal Processing 1(2), 72–90 (2010)
4. Hsu, C., Wu, J.: Dct-based watermarking for video. IEEE Transactions on Consumer Electronics 44(1), 206–216 (1998)
5. Wang, Y., Pearmain, A.: Blind mpeg-2 video watermarking robust against geometric attacks: a set of approaches in dct domain. IEEE Transactions on Image Processing 15(6), 1536–1543 (2006)
6. Jordan, F., Kutter, M., Ebrahimi, T.: Proposal of a watermarking technique for hiding/retrieving data in compressed and decompressed video. ISO/IEC document, JTC1/SC29/WG11 MPEG97/ M 2281, 27–31 (1997)
7. Zhao, Z., Yu, N., Li, X.: A novel video watermarking scheme in compression domain based on fast motion estimation. In: International Conference on Communication Technology, vol. 2, pp. 1878–1882 (2003)
8. Kung, C., Jeng, J., Lee, Y., Hsiao, H., Cheng, W.: Video watermarking using motion vector. In: International Conference on Computer Vision, Graphics and Image, pp. 547–551 (2003)
9. Song, J., Liu, K.: A data embedding scheme for h. 263 compatible video coding. In: IEEE International Symposium on Circuits and Systems, vol. 4, pp. 390–393 (1999)
10. Qiu, G., Marziliano, P., Ho, A., He, D., Sun, Q.: A hybrid watermarking scheme for h. 264/avc video. In: International Conference on Pattern Recognition, vol. 4, pp. 865–868 (2004)
11. Su, Y., Zhang, C., Zhang, C.: A video steganalytic algorithm against motion-vector-based steganography. Signal Processing (2011)

A Novel Coding Method for Multiple System Barcode Based on QR Code

Xiamu Niu, Zhongwei Shuai, Yongqiang Lin, and Xuehu Yan

Information Countermeasure Technique Research Institute,
Harbin Institute of Technology, Harbin, China
{xiamu.niu,zhongwei.shuai,yongqiang.lin,
xuehu.yan}@ict.hit.edu.cn

Abstract. An encoding and decoding method for Multiple System Barcode (MSB) based on QR code is designed in the paper through fusing several Black and White Barcodes (BWB) representing different color planes. The QR code standard is compatible and could be extended to MSB. MSB's capacity is increased in the method. The method can be applied to other standard barcode and large-capacity data storage and transmission. The experimental results are encouraging and demonstrate that our method is effective which is easy to understand, implement and extend.

Keywords: Barcode, Multiple System Barcode, QR code, Fuse.

1 Introduction

Barcode has been widely used in commercial distribution, storage, library and information, postal, railway, transportation, production and automation management from its inception in the early 1970s because it's rapid, accurate, low-cost and highly reliable. One-dimensional barcode could provide information with a set regularly arranged by bar, empty and the corresponding characters. The bar is dark and empty which can be read by the barcode scan devices. The characters are read directly by persons composed by a group of Arabic numerals. An example of one-dimensional barcode is shown in Fig.1(a) where the group of bar, empty and the corresponding characters provide the same information. Two-dimensional barcode generated to meet some practical application requirements one-dimensional bar code can't generally refer to black and white two-dimensional barcode.

Black and white two-dimensional barcode can represent characters, images and other information portably without database supporting with the main feature information could be represented in the horizontal and vertical directions. Black and white two-dimensional barcode[1] has the advantages of high capacity, wide encoding range, high reliability, high error-correcting ability, representing character, image information and so on in addition to the basic characteristics of one-dimensional bar code. In nowadays the following two-dimensional barcodes are used widely: PDF417 code [2] [3] and QR Code [4]. PDF417 code is shown in Fig.1(b) and QR Code is

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 99–108, 2012.

shown in Fig.1 (c). PDF417 code is developed by Symbol of the United States. QR Code is developed by Denso Corporation of Japan in September 1994. QR Code a two-dimensional matrix barcode is so widely used for its speed reading, round reading, multiple-format data representing, efficient Chinese characters representing and so on that the method is realized with QR Code as an example to ensure working and encoding and decoding efficiency of MSB.

As shown in Fig.2 traditionally barcode is called BWB or binary barcode. The barcode with different gray scale or colors is called MSB that could be called gray MSB or color MSB according to the different forms. The barcode with different gray scale or color distinction improves the data-storage capacity, but it's one-dimensional

(a) One-dimensional barcode (b) PDF417 code (c) QR Code

Fig. 1. Examples of barcode

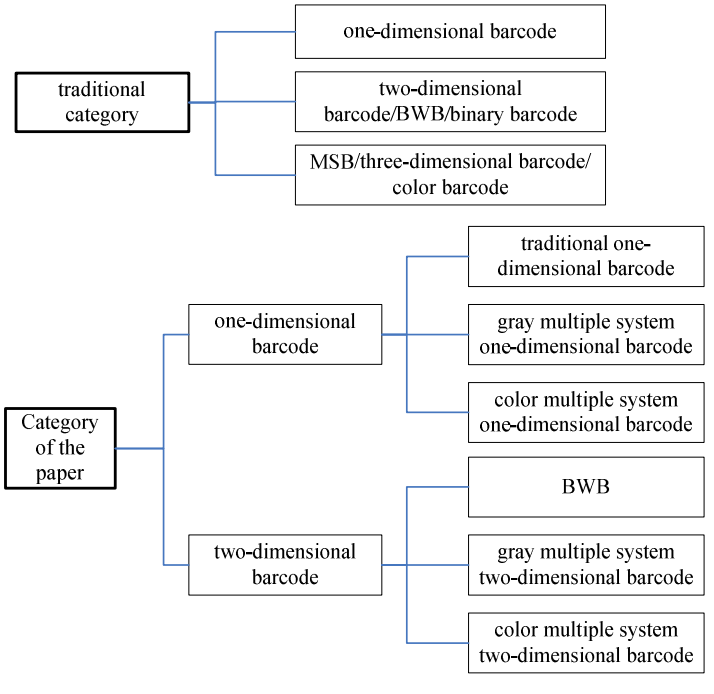


Fig. 2. Different categories between tradition and the paper according to dimension

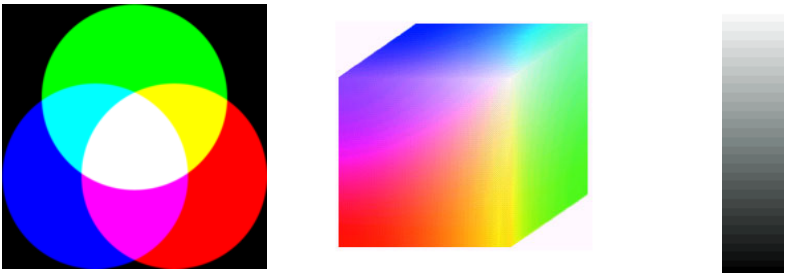
or two-dimensional, so it is defined as MSB not three-dimensional or multi-dimensional barcode. The MSB has multiple system one-dimensional barcode or multiple system two-dimensional barcode based on the one-dimensional barcode and two-dimensional barcode, also gray multiple system one-dimensional barcode, color multiple system one-dimensional barcode, gray multiple system two-dimensional barcode and color multiple system two-dimensional barcode according to the gray scale or color distinction. Data-storage capacity of MSB is larger than BWB that doesn't change the code size. For example, storage capacity of multiple system two-dimensional gray barcode with 8 kinds of gray or 8 colors is 4 times black and white two-dimensional barcode.

Research achievements of barcode technology have mainly concentrated in its capacity, reading effect and the application of some new coding method [5]. MSB has been studied a lot[6][7][8][9][10][11][12][13][14], but little applied. The concept of three-dimensional barcode has been put forward in [10], in which barcode coding capacity, error control and identification technology is researched. Three-dimensional barcode concept using color as the third dimension has been proposed in paper [13] that extends binary barcode to MSB through increasing the color-information types and researched coding theory of the MSB. The binary code has been extended to MSB in paper [14] through increasing different colors. The color-information types in every part based on the traditional one-dimensional barcode, the character set capacity has been substantially expanded, thus the one-dimensional barcode application range has been further expanded with the information density unchanged. However, firstly, the researches above could not load a large amount of information since individual barcode capacity is small generally. Secondly, existing standard barcode is not compatible and could not be extended to MSB directly because encoding and decoding is in a plane, coding methods all are using color to represent a certain number of information bits, and new codec design and program is needed in realizing. A novel encoding and decoding method for MSB is proposed in the paper. The standard barcode is compatible and could be extended to MSB. The barcode capacity is increased through the method.

2 Coding Method for MSB through Fusing Several QR Code

The gray scale or color space is introduced to expand the capacity for MSB on the basis of BWB. The MSB is categorized as multiple system one-dimensional barcode or multiple system two-dimensional barcode. The MSB can greatly improve the storage capacity compared with BWB. The main idea of the method for MSB is fusing vertically to be multiple system for existing code. Multiple system two-dimensional barcode based on QR code and the color space will be introduced in the followings, and principle and technique of gray scale is similar.

The color should be easy to distinguish in order to better recognize color barcode for image acquisition equipments. As shown in Fig.3(a), the RGB three planes are fused together to generate 8 colors. Taking 8 colors as an example in multiple system two-dimensional color barcode, principally it's better to choose 8 points as the using



(a)RGB colors fusion (b)The RGB color space (c)Gray scale

Fig. 3. Different RGB colors and the gray scale

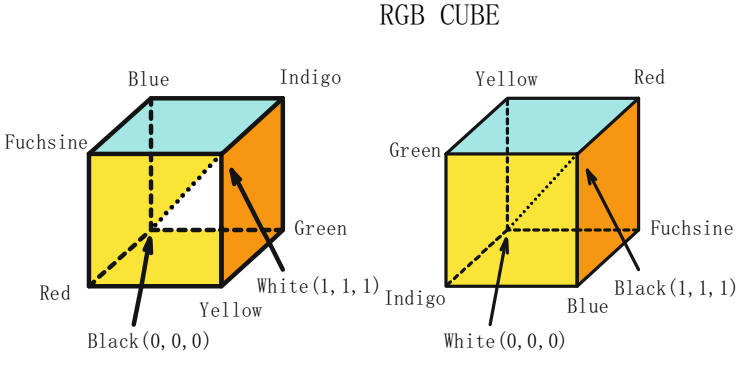


Fig. 4. The vertex colors of RGB color space

colors that have maximum Euclidean distance in the RGB color space. As shown in Fig.3(b) and Fig.4, these 8 colors chosen are red, green, blue, black, white, fuchsine, indigo and yellow, those have strong contrast, distinguish easily. Appropriate colors [15] could be chosen according to the actual situation for other applications to be distinguished easily and be resistance-distortion, etc.

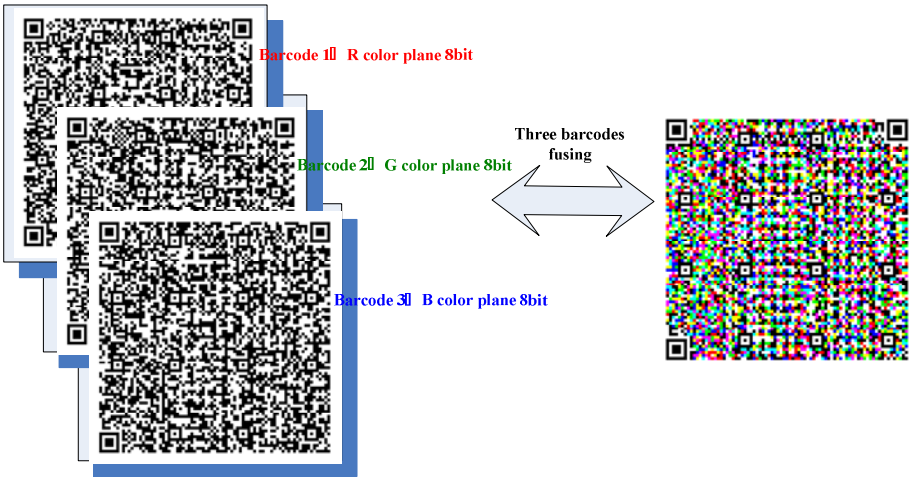
The gray scale of gray MSB based on QR code theoretically should be divided uniformly in 0-255 gray value range according to the needs of the gray categorized numbers to determine the gray value, as shown in Fig.3 (c). Barcode-load characteristics should be considered in practical applications, adjacent gradation spacing of gray-scale range can be slightly closer for easy identification, and farther for not easy. Gray MSB based on QR code and color MSB based on QR code are shown in Fig.5 (a) and (b).

2.1 Coding Method for Color MSB Based on QR Code

The method for color MSB based on QR code is shown in Fig.6. 8 colors multiple system two-dimensional barcode could be generated through fusing three color planes using three BWBs representing the R/G/B color planes.

(a)Gray MSB (b)Color MSB

Fig. 5. Examples of gray MSB and color MSB based on QR code



Detail Code Fusion Examples

R:	0	1	1	0	1	0	1	1	0	…
G:	1	0	0	1	0	1	0	0	1	…
B:	1	0	1	0	1	0	0	0	1	…
MSB	011	100	101	010	101	010	100	100	011	…

Fig. 6. Fusing RGB barcodes into a color barcode

8 colors MSB could be generated through combination displaying of 3 QR code encoding results using BWB technology when encoding. The 3 BWBs could be obtained through dividing the color MSB into three planes based on R/G/B color planes when decoding, further MSB decoding work could be completed through decoding three BWBs using BWB technology, and join the decoding results together.

The advantage of this method is full using coding resources of black and white two-dimensional barcode, easy implementation and strong expansibility.

This method can also be used for one-dimensional barcode, but the specific method of one-dimensional barcode isn't expatiated for small capacity and not wide application in the paper.

As shown in Fig.6, three planes respectively represent a R/G/B color value, three planes are respectively encoded in the form of QR code, and encoding result of each module is 0 or 1, 0 means white (simple) and 1 black (dark). For the QR Code of three planes, encoding results of each module of the octal barcode are obtained through fusing each plane coding results. For example, for a module, R plane encoding result is 0, G plane encoding result is 1, B plane encoding result is 1, and the coding result of this module of octal two-dimensional barcode is 011, and then displaying the color according to the 011 corresponding color. When decoding convert the acquisition color to 011 according to the corresponding relation, peel off three planes respectively and decode, then join the decoding data together.

Corresponding relations between fused data and 8 colors are shown in Table 1.

Table 1. Corresponding relations between code (fused data) and 8 colors

Code	RGB	Color
000	(255,255,255)	White
001	(255,255, 0)	Yellow
010	(255, 0,255)	Fuchsine
011	(255,0,0)	Red
100	(0,255,255)	Indigo
101	(0,255,0)	Green
110	(0,0,255)	Blue
111	(0,0,0)	Black

Table 2. Corresponding relations between fused data and 16 colors

Code	RGB	Color
0000	(255,255,255)	White
0001	(128,128,255)	
0010	(255,128,128)	
0011	(128,255,128)	
0100	(128,128,128)	
0101	(255,0,255)	Fuchsine
0110	(255,255,0)	Yellow
0111	(0, 255, 255)	Indigo
1000	(0,128,128)	
1001	(128,0,128)	
1010	(128,128,0)	
1011	(255,0,0)	Red
1100	(0,255,0)	Green
1101	(0,0,255)	Blue
1110	(192,192,192)	
1111	(0,0,0)	black

2.2 Extension of the Method

The technology above that is obtaining one MSB through 3 QR codes representing the R/G/B color plane sacking together, can be extended to any two-dimensional planes fusing obtaining one MSB through standard specifying each codec and the

corresponding color not through R/G/B planes which is a special case. For example, fusing 8 two-dimensional barcodes can produce 2 to the power of 8 or 256 colors with capacity promoting 32 times fusing 3 two-dimensional planes with 8 kinds of colors.

This method also can be extended based on gray, MSB obtained by fusing 8 two-dimensional barcode can be represented using the 256 kinds of gray.

Noting that, the barcode scan device identification capability requirement increases as the number of colors or gray scale increases.

Tacking fusing 4 QR codes as an example, codec and the colors can be corresponded as shown in Table 2 also can be changed based on actual need.

3 Experimental Results and Analysis

In the realization, the parameters of format information, version, error correction level and mask pattern reference could be fixed if the same to improve the decoding efficiency in the using process of MSB based on QR Code. The location center point of barcode could be fixed if the same for each barcode, so it need not the detection and location for each barcode. It also can improve the decoding efficiency.

Some parameters decoded for the plane could be fixed to solve the decoding failure question for other planes since color distortion caused by the image acquisition device recognition ability (refer to decoded plane). Parameters could be fixed mainly include: timing pattern center point coordinates, alignment pattern center point coordinates, format information, version, error correction level, mask pattern reference, etc. The reason is that the same parameters could be used for decoding to improve decoding efficiency obtained by fusing three QR Codes since coordinates of timing pattern, alignment pattern, .etc for the three planes are identical for 8 colors MSB. In addition, for different planes error correction level and other parameters can be different except version size, such flexibility is further enhanced.

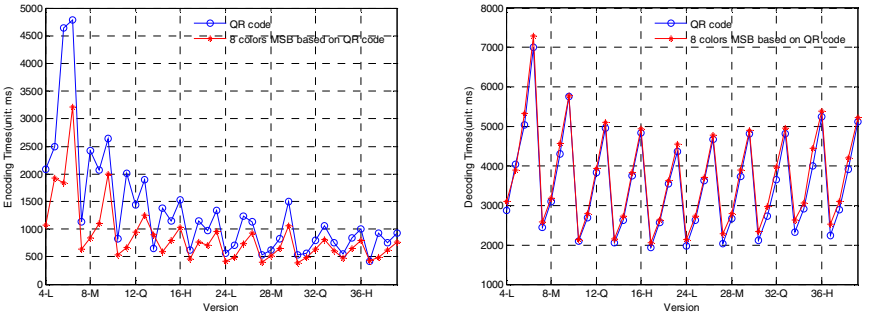
The experimental environment and configuration parameters are shown in table 3.

Table 3. Experimental environment and configuration parameters

Hardware	Software	Versions	Error correction levels	Size of test cases
Dell Precision WorkStation T5500 Tower, 2.93G memory, Xeon (Intel Xeon) E5504 @ 2.00GHz	Windows XP, VC 6.0	4, 8, 12, 16, 20, 24, 28, 32, 36, 40	L, M, Q, H	37392 bytes

The encoding and decoding time, capacity and error correction level are compared between normal QR code and 8 colors MSB based on QR code.

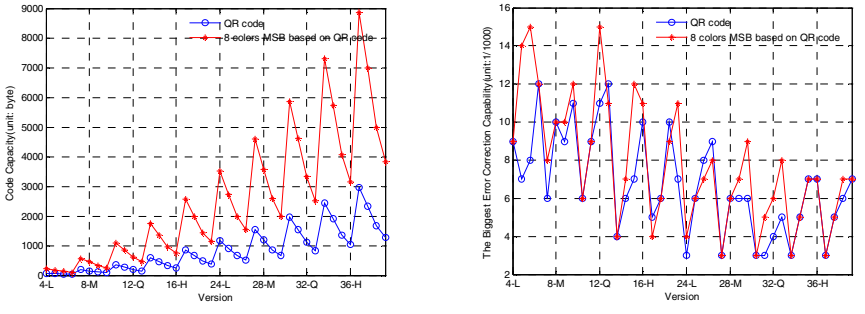
The encoding and decoding times (unit: ms) of test cases are shown in Fig.7 (a) and (b). The result is consistent with designed theory for the decoding times are similar for the same data volume, and the encoding time of 8 colors MSB based on QR code is slightly less that may because of copying less frequently.



(a) Encoding Times (b) Decoding Times

Fig. 7. The comparison of coding times between QR code and 8 colors MSB based on QR code (unit: ms)

The barcode capacity (unit: byte) can be obtained as shown in Fig.8 (a) according to a barcode number of encoding the test cases. From the figure it is consistent with the designed theory that the capacity of 8 colors MSB based on QR code is 3 times QR code.



(a) Code Capacity (b) The Biggest Error Correction Capability

Fig. 8. The comparison of code capacity and the biggest error correction capability between QR code and 8 colors MSB based on QR code

The error correction levels are shown in Fig.8 (b) through adding a certain proportion of errors (unit: 1/1000) from the biggest error correction capability among 100 tests. From the figure, though the data distribution is a little random, overall the error correcting capability is consistent with the designed theory.

4 Conclusions and Feature Work

This paper presents a novel encoding and decoding method for the color/gray MSB based on QR code. It's easy to understand, easy to implement, easy to expand from barcode standard to MSB not changing the coding and decoding method. It also can be applied to store and transmit large amounts of data. MSB is a new and valuable research direction, some theory and practice are researched in this paper. Further research will be focused on the realization of other standard formats of the MSB, as well as continuous MSB.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (60832010, 61100187) and the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF. 2010046).

References

1. Tingting, H.: The Research on the Technology of QR Code Recognition, MS Thesis of Central South University (2008)
2. Niu, X.-M., Huang, W.-J., Wu, D., Zhang, H.: Information Hiding Technique Based on 2D Barcode. Acta Scientiarum Naturalium Universitatis Sunyatseni. 43(suppl.2), 21–25 (2004)
3. ISO/IEC 15438-2006. International Organization for Standardization: Information Technology-Automatic Identification and Data Capture Techniques-PDF417 Bar Code Symbology Sepcification (2006)
4. ISO/IEC 18004-2000. International Organization for Standardization: Information Technology—Automatic Identification and Data Capture Techniques—Bar Code Symbology—QR Code (2000)
5. Automatic Identification Manufacturers. Automatic Identification and Data Capture Techniques-an Overview [EB/OL] (2009), http://www.aimglobal.org/technologies/ai-dc_overview.asp
6. Cheong, C., Kim, D.C., et al.: Color Classification using Quiet Zone Information for Color-Based Image Code Recognition [TP], IEICE Technical Report, PRMU2006-139, pp. 55–60 (2006)
7. Ahn, J., Cheong, C., et al.: Augmented Color Recognition by Applying Erasure Capability of Reed-Solomon Algorithm. In: Proceedings of Advances in Systems, Computing Sciences and Software Engineering, pp. 107–112 (2006)
8. Parikh, D., Jancke, G.: Localization and Segmentation of A 2D High Capacity Color Barcode. In: IEEE workshop on Applications of Computer Vision, pp. 1–6 (2008)
9. Jancke, G.: System and Method for Encoding High Density Geometric Symbol Set. U.S. 0285761 A1 (2005)
10. Liu, N.-Z., Yang, J.-Y.: Encoding Theory and Design of Three-Dimensional Bar Code. Chinese Journal of Computers 30(4), 686–692 (2007)
11. Pei, S., Wu, B., et al.: Codec System Design for Color Barcode Symbols. In: Proceedings of IEEE International Conference on Computer Information Technology (CIT), pp. 539–544 (July 2008)

12. Songwen, P., Baifeng, W.: Encoding Theory and System Design of Multi-dimensional Matrix Barcode. Journal of Image and Graphics 21(7), 1018–1024 (2009)
13. Guo, W., Zhao, S.: Encoding Theory and Feasibility Studies of Three Dimensional Multiple System Barcode. Journal of Beijing Technology and Business University(Natural Science Edition) 27(6), 49–53 (2009)
14. Meng, X., Liu, X.-D.: Encoding Study of 1-dimensional Color Multiple System Barcode. Journal of Zhejiang University(Engineering Science) 38(5), 55–561 (2004)
15. Wang, H.-S., Liu, X.-D.: Print and Recognition Research of Three-Dimensional Barcode. Computer Engineering and Design 32(1), 370–373 (2011)

A Research on Behavior of Sleepy Lizards Based on KNN Algorithm

Xiaolv Guo[1], Shu-Chuan Chu[2], Lin-Lin Tang[1,*], John F. Roddick[2], and Jeng-Shyang Pan[1]

[1] Harbin Institute of Technology Shenzhen Graduate School, 518055 Xili, Nanshan, Shenzhen, China
[2] School of Computer Science, Engineering and Mathematics Flinders University of South Australia GPO Box 2100, Adelaide, South Australia 5001
linlintang2009@gmail.com

Abstract. A new research method for the sleepy lizards based on the KNN algorithm and the traditional social network algorithms is proposed in this paper. The famous paired habit of sleepy lizards is verified here based on our proposed algorithm. In addition, some common population characteristics of the lizards are also introduced by using the traditional social net work algorithms. Good performance of the experimental results shows efficiency of the new research method.

Keywords: Social Network Analysis (SNA), K-Nearest Neighbor (KNN) Algorithm, Sleep Lizard.

1 Introduction

The social network analysis research work which is evolved from psychology, sociology, anthropology and mathematics rises in the 1930s. It is initially used for studying the real relationship among people in society. And it has been continuously developed, especially in the 1930s to the late 1970s, Harvard University and the University of Manchester, have made a significant contribution to social network analysis [1]. The Harvard researches drew together algebraic models of groups using set theory and multidimensional scaling to establish concepts such as the strength and distance of connections [2].

The basic idea of social network is to establish and analysize the relationship among the members of the social group for finding some potential relation between them which will be used in some special research of social group. There are several classical research methods in the social network analysis research area, such as the Sociogram algorithm, the school of Manchester algorithms and the Structure Hole theory. Each of them has gained a lot of research results which give a great push to the development of the social network analysis.

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 109–118, 2012.

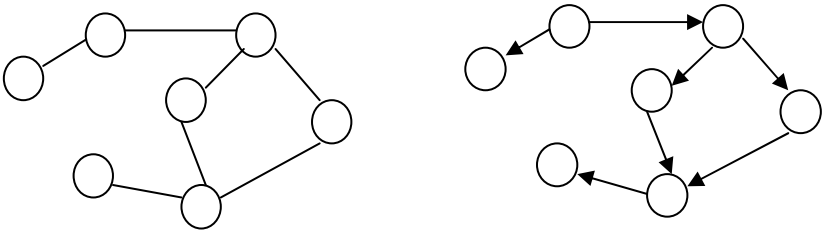
On the other side, the KNN algorithm [3] which was introduced by Cover and Hart in 1968 is often used for classifying. The basic idea of KNN is to find the K nearest members in the training set. Then the classification of the given member X can be determined by category of these nearest members. The common rule is that X has the same category of the most of the K nearest members. The most commonly used distance measuring criterion is the classical Euclidean distance.

The whole paper is organized as follows. The basic knowledge of the social network will be introduced in section 2. And the simple introduction of KNN algorithm will also be proposed in this section. Our testing sample will be shown in section 3. The proposed KNN based algorithm and the simulation results are shown in section 4. And the conclusion section 5.

2 Related Knowledge

2.1 Social Network

Actually, the SNA can be described as a research study for structures which consist of at least two social entities (usually more) and the links among them. There are many kinds of relationships which are used for linking, such as the kinship relations, social roles, actions, affective, material exchanges and common behaviors. The graph method and the matrix method are usually used for expressing those structures. The following figure 1 gives the general idea of such graphs.



(a) (b)

Fig. 1. (a) Undirected graph, (b) Directed graph

Another important method for SNA is the matrix method. The most commonly used matrixes are the adjacency matrix, incidence matrix, distance matrix and the valued matrix. One of the most commonly used and simplest is the adjacency matrix. The adjacency matrix is a $n \times n$ matrix for a given graph G with n nodes. If there is a link between node i and node j, the value in the place (i, j) will be set 1. It is similar between the graph method and the matrix method in the classification. Matrix method can also be divided into the undirected and the directed matrix methods based on the different relationship between the different models as shown in table 1.

Table 1. Adjacency Matrix Example

(1) undirected matrix

	a	b	c	d	e	f	g
a	—	1	0	1	0	0	0
b	1	—	1	0	0	1	0
c	0	1	—	0	0	0	1
d	1	0	0	—	0	0	0
e	0	0	0	0	—	1	0
f	0	1	0	0	1	—	1
g	0	0	1	0	0	1	—

(2) directed matrix

	a	b	c	d	e	f	g
a	—	0	1	0	0	0	1
b	0	—	0	0	1	0	0
c	0	0	—	1	1	1	0
d	0	0	0	—	0	0	0
e	0	0	0	0	—	0	0
f	0	0	0	1	0	—	0
g	0	0	0	0	0	0	—

Generally speaking, the undirected matrix is a symmetric matrix and the directed matrix is a asymmetric matrix. And there are some other matrix methods such as the adjacent matrix with weight.

There are many useful definitions for the SNA. Degree is the adjacent number of the actor. Geodesics are the shortest path between the two nodes. Distance is the length of the Geodesics. Diameter is the length of the longest Geodesics. Connected graph is a graph that each pair of the nodes in it can be connected. And the Disconnected graph's definition is on the contrary. Size is the number of a group. Density is the tightness of among the different nodes. Cohesion is the strength between the actors in the group. Centrality gives the location information of the actors.

The most useful two definitions which are used for analyzing the group behavior character are the degree centrality analysis and the middle centrality analysis. The definition is shown below.

$$C_{AD}(i) = d(n_i) = \sum_j x_{ij} = \sum_i x_{ji} \tag{1}$$

$$C_{AD}(i) = d(n_i) = \sum_j x_{ij} + \sum_i x_{ji} \tag{2}$$

Here, the formula (1) and (2) represent the absolute undirected and directed degree centrality respectively. And $x_{i,j}$ is the path number from other points to the point n_i .

$x_{j,i}$ is the path number from other points to the point n_i. The absolute and relative betweenness centrality is shown below.

$$C_{AB}(i) = \sum_{j}^{n} \sum_{k}^{n} b_{jk}(i), j \neq k \neq i, j < k \tag{3}$$

$$C_{RB}(i) = \frac{2C_{AB}(i)}{(n-1)(n-2)} \tag{4}$$

$$b_{jk}(i) = \frac{g_{jk}(i)}{g_{jk}} \tag{5}$$

Here, g_{jk} is the geodesics number between j and k. $g_{jk}(i)$ is the number of geodesics of the third point i which is on the geodesics between j and k. And $b_{jk}(i)$ can been seen as the probability of the point i for controlling the point k and j. n is the points number of the group.

2.2 KNN Algorithm and Its Improvement

The Euclidean distance shown in the following formula (6) is often used in KNN. Here, x and c_t are two vectors.

$$d(x, c_t) = \sqrt{\sum_{i=1}^{n} (x_i - c_{ti})^2} \tag{6}$$

The basic idea of KNN algorithm is shown in the following figure 2. If the K value in the KNN algorithm equals to 3, then the testing point will have the same classification with the round points. And if the K value is 5, then the testing point will belong to the triangle set.

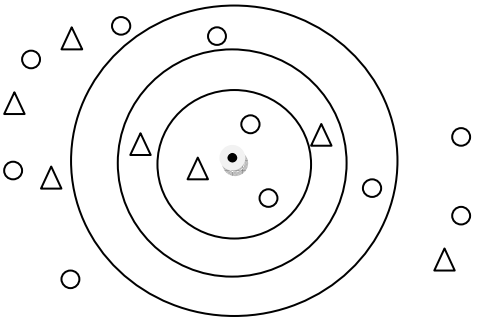


Fig. 2. KNN algorithm graph

In fact, there are a lot of different improvements for the traditional KNN algorithm, such as the WKPDS algorithm which was proposed by Hwang and Wen [4], ENNS (Equal-Average Nearest Neighbor Search) algorithm which was introduced in reference [5] by Pan in 2004, EENNS algorithm which was shown in reference [6] by Qiao and the EEENNS algorithm given in reference [7] by Lu. In this paper, we use the ENNS algorithm in our experiment. The basic theory in this algorithm can be described as follows:

Let the sum of the vector x as the following formula (7) shows.

$$S_x = \sum_{i=1}^{N} x_i \tag{7}$$

As we all know that $N \cdot D\left(X, C_j\right) \geq \left(S_X - S_{C_j}\right)^2$. Here N is the dimension of the vector space. X and C_j and the wavelet coefficients vector for x and c_j. If there is following inequality (8) is satisfied, then there is the formation of the formula (9).

$$\left|S_X - S_{C_j}\right| \geq \sqrt{N \cdot D\left(X, C_i\right)} \tag{8}$$

$$D\left(X, C_j\right) \geq D\left(X, C_i\right) \tag{9}$$

And if the formula (10) is satisfied, the inequality (11) will also be satisfied. Here, the

$$| C_x - C_{c_j} | \geq \sqrt{\frac{N}{2^n} D\left(X, C_i\right)} \tag{10}$$

$$D\left(X, C_j\right) \geq D\left(X, C_i\right) \tag{11}$$

The steps of such an algorithm are proposed as follows:

Step 1. Apply the wavelet transform [8] onto the training set to get the wavelet coefficient vector. And so does the testing vector. Then, the transformed vector should be sorted by comparing the approach coefficients from small to large.

Step 2. Compare the testing vector with the training vectors and select the point which has the least absolute value between them as the initially center vector. Then we choose K vectors with the center vector that has the least absolute difference value between the testing vector and it. And the initial distance is chosen as the largest distance between our testing vector and these training K vectors shown in the following formula (12).

$$D_{\min} = D\left(X, C_i\right) \tag{12}$$

Step 3. For the vectors before the K-vector block, check the inequality (13) shown below to judge if it should be in the K-vector block. And for the vectors behind the K-vector block, check the inequality (14). If they follow the rules respectively, then the K vectors are the aim group. The searching process comes to an end.

$$C_{c_j} \le C_x - \sqrt{\frac{N}{2^n} D_{\min}} \tag{13}$$

$$C_{c_j} \ge C_x + \sqrt{\frac{N}{2^n} D_{\min}} \tag{14}$$

Step 4. If the vectors don't follow the rules above respectively, then the distance between the testing vector and them should be compared with the initial largest distance $D_{\min}$ to decide if the vector should be changed into the K-vector block. The comparison algorithm is chosen as the WKPDS algorithm. Then the steps 3 and 4 are repeated until all the vectors in the training set have been checked.

3 Introducing and Preprocessing for Our Sleepy Lizard Samples

3.1 Basic Sample Information Introduction

Professor Stephan T. Leu and his colleagues in Flinders University got the living data about the 55 sleepy lizards that live in one area near the Bundey Bore Station after three months observation from Sept. 15th, 2009 to Dec. 15th 2009. There are three prime contents in this record: body temperature in every two minutes, step number in two minutes and the location information in every ten minutes. We take the location information as an example which is shown in the following figure 3.

991	976	9	20	17	30
974	970	9	21	9	30
977	977	9	21	9	40
971	971	9	21	9	50
978	973	9	21	11	40
986	989	9	21	12	40
982	978	9	21	13	20
981	972	9	21	13	30
983	984	9	21	13	40
980	978	9	21	13	50
982	980	9	21	14	30
984	992	9	21	14	40
988	980	9	21	14	50
989	981	9	22	8	10
984	978	9	22	8	20
979	979	9	22	8	30

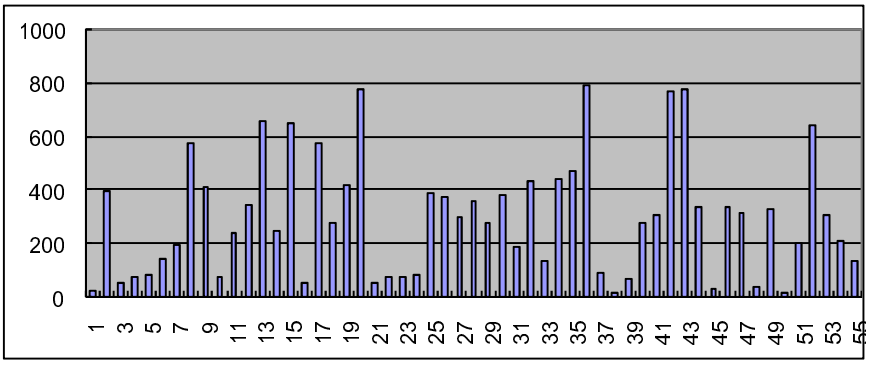
Fig. 3. Location information after reprocess

The first two columns in the above figure 3 are the X-coordinate values and the Y-coordinate values after some reprocess respectively. Considering the living property of sleepy lizards and the accuracy of our equipments, 14 meters are chosen for the threshold to judge the encounter area for two sleepy lizards. If we take the encounter matter as a basic information transmission relationship, then the communication relation between different sleepy lizards can be got as follows.

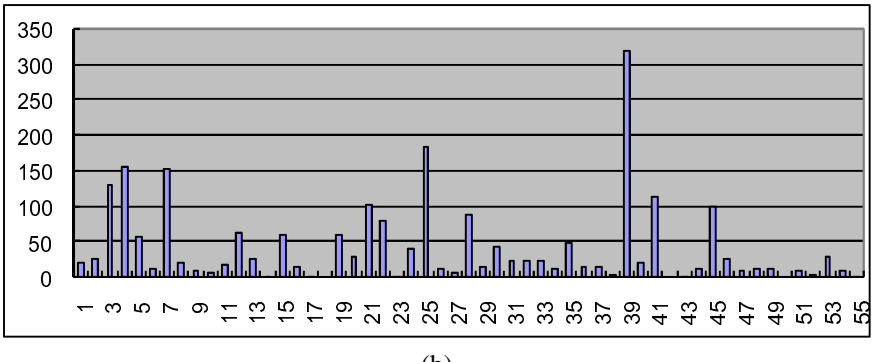
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	2	0	15	0	25	0	0	0	11	0	0	0	0	0	1	0	0	2	0	2	2
3	0	0	0	8	0	0	0	5	0	0	0	1	0	0	0	2	3	0	3	1	0	0	0	0	5
4	0	0	8	0	0	1	1	4	0	0	0	0	0	0	0	0	3	0	0	5	4	0	0	1	5
5	0	2	0	0	0	0	1	0	2	2	0	2	0	0	0	0	0	0	0	0	0	3	0	2	3
6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
7	0	15	0	1	1	0	0	0	2	21	0	0	11	0	0	0	0	0	8	0	0	13	0	20	8
8	0	0	5	4	0	0	0	0	0	0	0	5	0	4	0	0	538	0	0	0	0	0	0	3	9
9	0	25	0	0	2	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0
10	0	0	0	0	2	0	21	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	17	13
11	13	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
12	0	0	1	0	2	0	0	5	0	2	0	0	0	218	0	0	7	0	2	0	0	1	0	0	39
13	0	11	0	0	0	0	11	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3
14	0	0	0	0	0	0	0	4	0	0	0	218	0	0	0	0	0	0	2	0	0	0	0	0	15
15	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0
17	0	0	3	3	0	0	0	538	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	3	12
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0
19	0	1	3	0	0	0	8	0	0	0	0	2	0	2	0	0	0	0	0	0	0	10	0	6	14
20	0	0	1	5	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	19	0	0	0
21	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	1	0	0	0	0	0
22	0	2	0	0	3	0	13	0	3	0	0	1	0	0	0	0	0	0	10	0	0	0	0	0	3
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	19	0	0	0	0	0
24	0	2	0	1	2	0	20	3	0	17	0	0	2	0	0	0	3	0	6	0	0	0	0	0	10
25	0	2	5	5	3	0	8	9	0	13	0	39	3	15	0	0	12	0	14	0	0	3	0	10	0

Fig. 4. Part of the communication matrix between sleepy lizards for our samples

The classical degree centrality analysis and the betweenness centrality analysis graphs are shown in the following figure 5.



(a)



(b)

Fig. 5. Degree centrality analysis graph (a) and betweenness centrality analysis graph (b)

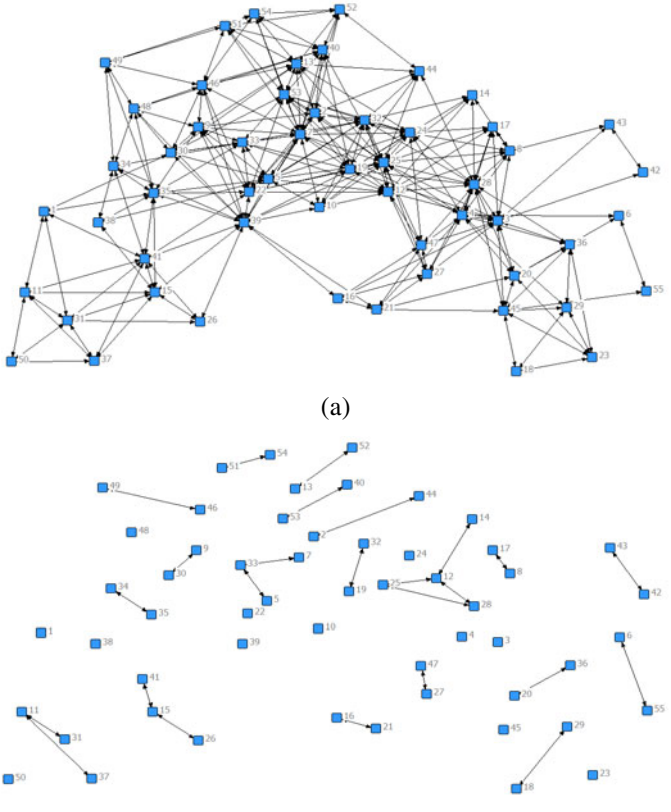
The mean purpose of such graphs is to find the different social status and different resource controlling ability for every sleepy lizard in the samples.

As we can see from the above figure 5, communication strength for every sleepy lizard and places they take during the communication are shown clearly. For example, though the 39th sleepy lizard's degree centrality is low which means communication strength is weak, the betweenness centrality is high which means the place it takes is very important.

3.2 Paired Living Analysis

The famous paired living property of sleepy lizards can also be studied by a relation graph based on the communication definition mentioned above which is shown in the following figure 6 (a). And when the threshold value is large enough, the paired living character can also be verified in figure 6 (b).

As we can see from the above figure 6, the famous paired living character can be verified from the communication relationship.



(a)

(b)

Fig. 6. Communication graph (a) and communication graph under threshold value 30

4 Our Proposed Method

We introduce the location information based KNN method to study the intimacy and even to verify the paired living character in this paper. The reason for us to do so is based on the pre-analysis of this kind of animal whose motion range is relative fixed.

The steps of our proposed method basically follow the ENNS algorithm mentioned above. And the location information which is a two demotions vector is used for one sleepy lizard sample. The experimental result example for a sample 42th sleepy lizard's K nearest friends can be shown in the following figure 7.

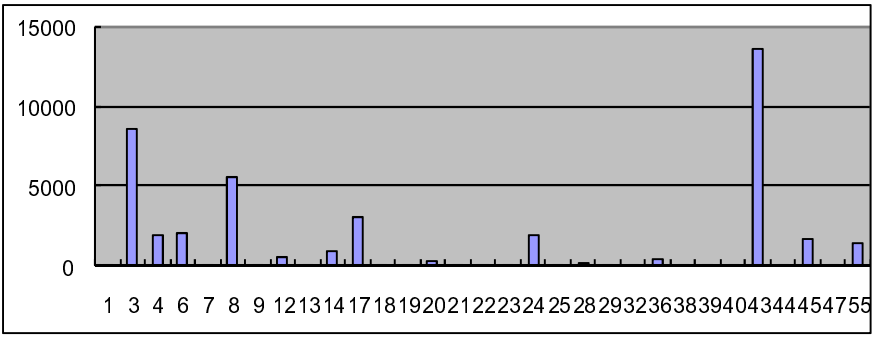


Fig. 7. 3-nearest frequency for the 42th sample based on KNN algorithm

As we can see that the 43th is the most 'nearest' for the 42th sample. And the 3-nearest frequency simulation results for the 43th shown in figure 8 can help us to verify the paired living character.

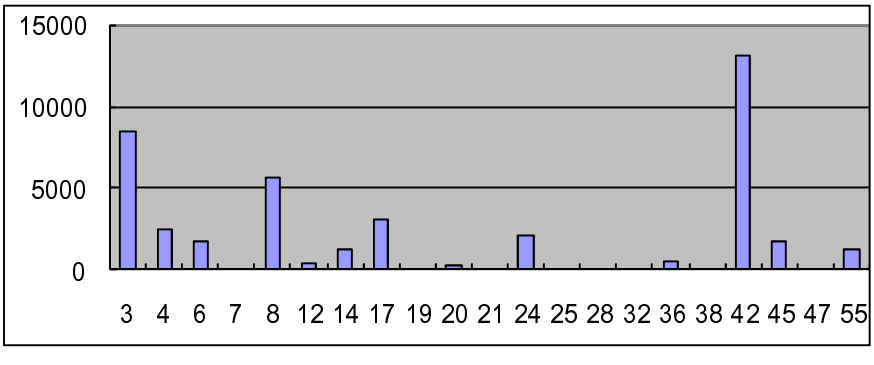


Fig. 8. 3-nearest frequency for the 43th sample based on KNN algorithm

5 Conclusion

This paper proposed a new research perspective for a small sleepy lizard group. By introducing the traditional KNN algorithm, a good performance for the relationship

between the samples can be given. Even the famous paired living character also can be verified. To make full use of the data base and find more new living habits of such a small group sleepy lizards is our future work.

References

1. Granovetter, M.: Getting a Job. A Study of Contacts and Careers. Harvard Univ. Press (1974)
2. Granovertter, M.: The Strength of Weak Ties. Am. J. Sociology 78(6), 1360–1380 (1973)
3. Kuncheva, L.: Fitness Function in Editing KNN Reference Set by Genetic Algorithms 30(6), 1041–1049 (1997)
4. Hwang, W.J., Wen, K.W.: Fast KNN Classification Algorithm Based on Partial Distance Search. IEEE Trans. Computers 24(7), 750–753 (1975)
5. Pan, J.S., Lu, Z.L., Sheng, H.S.: An Efficient Encoding Algorithm for Vector Quantization based on Subvector Technique. IEEE Trans. on Image Processing 12(3), 265–270 (2003)
6. Pan, J.S., Qiao, Y.L., Sun, S.H.: A Fast K Nearest Neighbors Classification Algorithm. IEICE Trans. Fundamentals E87-A(4), 961–963 (2004)
7. Lu, Z.M., Sun, S.H.: Equal-average Equal-variance Equal-norm Nearest Neighbor Search Algorithm for Vector Quantization. IEICE Trans. Inf. & Syst. 86(3), 660–663 (2003)
8. Hassana, G.K., Shaker, K.A., Zou, B.J.: Optimal Approach for Texture Analysis and Classification based on Wavelet Transform and Neural Network. Journal of Information Hiding and Multimedia Signal Processing 2(1), 33–40 (2011)

Directional Discriminant Analysis Based on Nearest Feature Line

Lijun Yan[1,2,*], Shu-Chuan Chu[3], John F. Roddick[3], and Jeng-Shyang Pan[4]

[1] Intelligent Control Research Center, Guangzhou Institute of Adavanced Technology, Chinese Academy of Sciences
IAT, GZ&CAS No.1121, Haibin Rd, Guangzhou, China
[2] Department of Automatic Control and Test, Harbin Institute of Technology
92 West Dazhi Street, Nan Gang District, Harbin, China
[3] School of Computer Science, Engineering and Mathematics
Flinders University of South Australia
GPO Box 2100, Adelaide, South Australia 5001
[4] School of Computer Science and Technology,
Harbin Institute of Technology Shenzhen Graduate School
518055 Xili, Nanshan, Shenzhen, China
yanlijun@126.com

Abstract. In this paper, two novel image feature extraction algorithms based on directional filter banks and nearest feature line are proposed, which are named Single Directional Feature Line Discriminant Analysis (SD-NFDA) and Multiple Directional Feature Discriminant Line Analysis (MD-NFDA). SD-NFDA and MD-NFDA extract not only the statistic feature of samples, but also the directionality feature. SD-NFDA and MD-NFDA can get higher average recognition rate with less running time than other nearest feature line based feature extraction algorithms. Experimental results confirm the advantages of SD-NFDA and MD-NFDA.

Keywords: Directional Filter Bank, Nearest Feature Line, Feature Extraction.

1 Introduction

As one of a few biometric methods, face recognition and related technology [1-2] have a variety of potential applications in information security, smart card, access control, etc. In the face recognition task, the number of training images per person is smaller than that of the dimensionality of face images. High-dimensional face images lead to high computational complexity and overfitting. Dimensionality reduction is an effective way to alleviate it, and subspace learning algorithms have been widely used.

Principal Component Analysis (PCA) [3], linear discriminant analysis (LDA) [4], and maximum margin criterion (MMC) [5] are among most popular subspace learning algorithms. PCA projects the original data to a low dimensional space, which is

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 119–128, 2012.

spanned by the eigenvectors associated with the largest eigenvalues of the covariance matrix of all samples. PCA is the optimal representation of the input samples in the sense of minimizing the mean squared error. However, PCA is an unsupervised algorithm, which may impair the recognition accuracy. Linear subspace analysis (LDA) finds a transformation matrix U that linearly maps high-dimensional sample $x \in R^m$ to low-dimension data $y = U^T x \in R^n$, where n<m. LDA can calculate an optimal discriminant projection by maximizing the ratio of the trace of the between-class scatter matrix to the trace of the within-class scatter matrix. LDA takes consideration of the labels of the input samples and improves the classification ability. However, LDA suffers from the small sample size (SSS) problem. Many effective approaches have been proposed to solve the problem. Some algorithms using the kernel trick are developed in recent years [6], such as kernel principal component analysis (KPCA) [7-8], kernel discriminant analysis (KDA) [9] and Locality Preserving Projection [10] used in many areas [11]. Researchers have developed a series of KDA algorithms [12-14].

Nearest feature line (NFL) [15] is a new classification tool, proposed by Li in 1998, firstly. In particular, it performs better when only limited samples are available for training. The basic idea underlying the NFL approach is to use all the possible lines consisting of any pair of feature vectors in the training set to encode the feature space in terms of the ensemble characteristics and the geometric relationship. As a simple yet effective algorithm, the NFL has shown its good performance in face recognition, audio classification, image classification, and retrieval. The NFL takes advantage of both the ensemble and the geometric features of samples for pattern classification. In contrast to a nearest neighbor (NN) classifier, the NFL makes better use of the ensemble information for classification [16-18].

While NFL has achieved reasonable performance in data classification, most existing NFL-based algorithms just use the NFL metric for classification and not in the learning phase. While classification can be enhanced by NFL to a certain extent, the learning ability of existing subspace learning methods remains to be poor when the number of training samples is limited. To address this issue, a number of enhanced subspace learning algorithms based on the NFL metric have been proposed, recently. For example, Zheng et al. proposed a nearest neighbour line nonparametric discriminant analysis (NFL-NDA) [19] algorithm, Pang et al. presented a nearest feature line-based space (NFLS) [20] method, and Lu et al. put forward an uncorrelated discriminant nearest feature line analysis (UDNFLA) [21]. However, most of these algorithms share the following shortcomings: 1) The computational complexity of most NFL-based algorithms are all high; 2) In most of these algorithms, the information of between-class is not used efficiently. Neighborhood discriminant nearest feature line analysis (NDNFLA) [22] improves the performance of UDNFLA. However, NDNFLA only extracts the statistic feature of the samples. The face image also is a special two-dimensional signal. Directionality is an important feature of two-dimensional signal. In this paper, two novel image feature extraction algorithms are proposed. Both proposed methods are based on NFL and directional filter banks (DFB) [23-24].

The rest of the paper is organized as follows. In section 2, some preliminaries about DFB, NFL and NDNFLA are given. In section 3, we give an introduction of the

proposed methods. In section 4, several experiments are implemented to evaluate the proposed algorithms. Finally, conclusions are made in section 5.

1.1 Directional Filter Banks

Recently, the new 2-D image representation has become an active research topic. The DFB is a toolkit that can explore the intrinsic geometrical structure of images. It indicates that directionality is an important feature for an image representation. The DFB can split the spectrum into 2^n wedge-shaped slices by using an n-level iterated tree structured filter banks. The complete design details of this filter bank may be found in [25]. These theories include the design of the quincunx filter bank (QFB) by applying dimensionality changing transformations, and complete reconstruction filter bank by applying McClellan transformations. The more general two-dimensional filter bank design algorithms could also be presented in [26-27]. In addition to the two band systems, the first two stages of the filter bank are preceded by a modulator which shifts the input signal spectrum by π in the w_1 direction. The ideal passband characteristics for the analysis/synthesis filters in the first two filter bank stages are diamond shaped and correspond to the anti-aliasing filter for quincunx downsamplers. The remaining stages use four different parallelogram-shaped passband characteristics. The ideal spectral partitioning possible with the directional filter bank family is shown in Fig. 1.

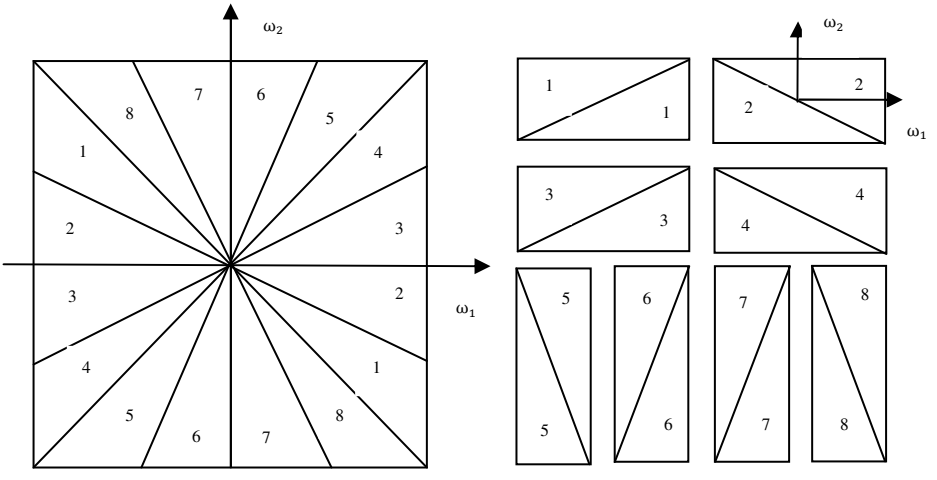


Fig. 1. The ideal frequency partition map for a 3-level DFB. (a) input. (b) outputs

1.2 Nearest Feature Line

Nearest feature line is a classifier. It is first presented by Stan Z. Li and Juwei Lu [15]. Given a training samples set, $X = \{x_n \in R^M : n = 1, 2, \cdots, N\}$, denote the class label of x_i by $l(x_i)$, the training samples sharing the same class label with x_i by

$P(i)$, and the training samples with different label with x_i by $R(i)$. NFL generalizes each pair of prototype feature points belonging to the same class: $\{x_m, x_n\}$ by a linear function $L_{m,n}$, which is called the feature line. The line $L_{m,n}$ is expressed by the span $L_{m,n} = sp(x_m, x_n)$. The query x_i is projected onto $L_{m,n}$ as a point $x^i_{m,n}$. This projection can be computed as

$$x^i_{m,n} = x_m + t(x_n - x_m) \tag{1}$$

where $t = [(x_i - x_n)(x_m - x_n)] / [(x_m - x_n)^T (x_m - x_n)]$. The Euclidean distance of x_i and $x^i_{m,n}$ is termed as FL distance. The less the FL distance is, the more probability that x_i belongs to the same class as x_m and x_n. Fig. 2 shows a sample of FL distance. In Fig. 2, the distance between y_q and the feature line $\overline{y_1 y_2}$ equals to the distance between y_q and y_p, where y_p is the projection point of y_q to the feature line $\overline{y_1 y_2}$.

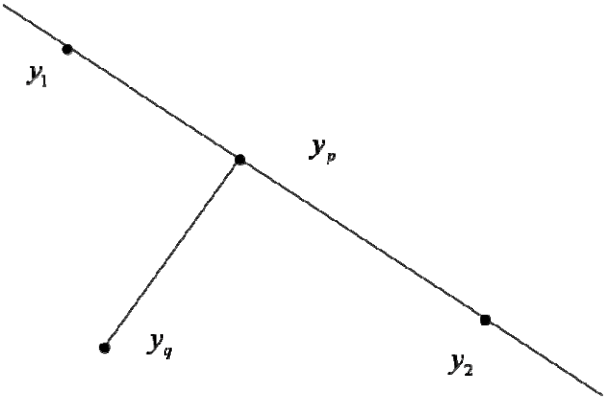


Fig. 2. Feature line distance

1.3 NDNFLA

Neighborhood discriminant nearest feature line analysis (NDNFLA) [22] is a subspace learning algorithm based on nearest feature line. Given a training samples set, $X = \{x_n \in R^M : n = 1, 2, \cdots, N\}$, to introduce NDNFLA, define two types of neighbor-hoods, firstly.

Definition 1. *(Homogeneous neighborhoods). For a sample x_i, its k nearest homogeneous neighborhood N_i^o is the set of k most similar data which are in the same class with x_i.*

Definition 2. *(Heterogeneous neighborhoods). For a sample* x_i *, its k nearest heterogeneous neighborhood* N_i^e *is the set of k most similar data which are not in the same class with* x_i *.*

In NDNFLA approach, the optimization problem is as follows:

$$\max J(W) = (\sum_{i=1}^{N} \frac{1}{NC^2_{|N_i^e|}} \sum_{x_m, x_n \in N_i^e} \left\| W^T x_i - W^T x_{m,n}^i \right\|^2 - \sum_{i=1}^{N} \frac{1}{NC^2_{|N_i^o|}} \sum_{x_m, x_n \in N_i^o} \left\| W^T x_i - W^T x_{m,n}^i \right\|^2) \tag{2}$$

NDNFLA aims to maximize the local feature line between-class scatter and minimize the local feature line within class scatter. The above formula can be simplified as following.

$$\max \ J(W) = \mathrm{tr}[W^T (A - B) W] \tag{3}$$

where

$$A = \sum_{i=1}^{N} \frac{1}{NC^2_{|N_i^e|}} \sum_{x_m, x_n \in N_i^e} [(x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T] \tag{4}$$

$$B = \sum_{i=1}^{N} \frac{1}{NC^2_{|N_i^o|}} \sum_{x_m, x_n \in N_i^o} [(x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T] \tag{5}$$

A length constraint $w^T w = 1$ is imposed on the proposed NDNFLA. Then, the projection W of NDNFLA can be obtained by solving the following eigenvalue problem

$$(A - B) w = \lambda w \tag{6}$$

Let $w_1, w_2, \cdots, w_q$ be the eigenvectors of (6) corresponding to the q largest eigenvalues ordered according to $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$. An $M \times q$ transformation matrix $W = [w_1, w_2, \cdots, w_q]$ can be obtained to project each sample $M \times 1$ x_i into a feature vector $q \times 1$ y_i as follows:

$$y_i = W^T x_i, \qquad i = 1, 2, \cdots, N \tag{7}$$

Then $y=Wx$ is the NDNFLA feature of sample x.

2 Proposed Algorithms

The main idea of this work is to improve the recognition rate of the NDNFLA algorithm by applying a pre-processing based on DFB. First, directional images are generated by applying the DFB to each face image from the database. Then a new face database is composed of the directional images of the face images. The size of the new database is 2^n times that of the original database, where n is the level of DFB. Let $X = \{X_i^j \in R^{m \times n}, i = 1, \cdots, c, j = 1, \cdots, N_i\}$ be a training face database. Its directional face database can be obtained by using n-level DFB. Suppose that $X_i^{j,l}$ denotes the directional image which is generated by the lth directional subband of face image X_i^j.

2.1 SD-FLDA

$X^l = \{X_i^{j,l} \in R^{m \times n}, i = 1, \cdots, c, j = 1, \cdots, N_i\}$ can be treated as a new training database, where l denotes the index of DFB subband, and $1 \le l \le 2^n$. We perform NDNFLA on X^l directly. Let the optimal projection matrix be U, thus the SD-FLDA feature of X_i^j is $Y_i^j = X_i^{j,l} U$ .which is used for classification. The size of U is up to the size of image matrix, the level of DEB and the chosen DFB subband. Given a test face image T, its lth subband T^l of its directional images is needed. Then $FT^l = T^l U$ is the SD-FLDA feature for classification. In SD-FLDA frame, only one subband of face directional images is used, so this face recognition frame is called single directional NDNFLA. The schematic diagram of SD-FLDA is given by Fig. 3.

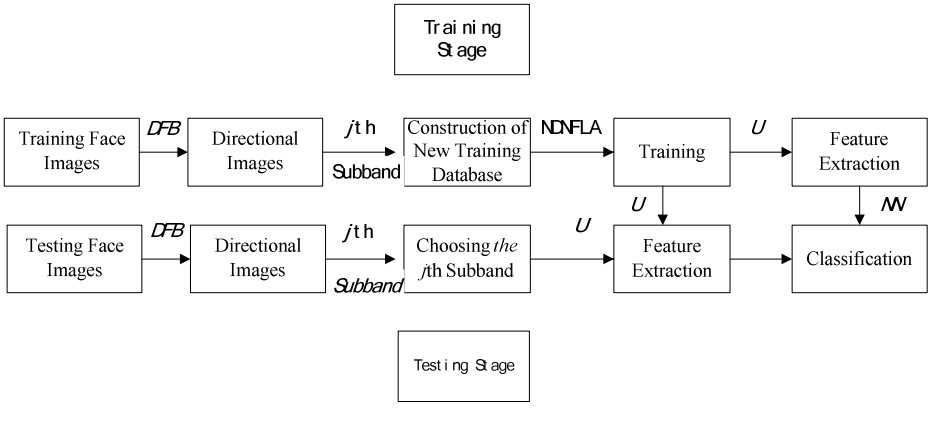


Fig. 3. Schematic diagram of SD-FLDA

2.2 MD-FLDA

For each l, an optimal projection matrix U^l can be calculated by performing NDNFLA on $X^l = \{X_i^{j,l} \in R^{m \times n}, i = 1, \cdots, c, j = 1, \cdots, N_i\}$. That is that there are 2^n training face databases and 2^n optimal projection matrices. Given a testing face image T, its directional images are obtained by applying n-level DFB. For each l, perform SD-FLDA on $X^l = \{X_i^{j,l} \in R^{m \times n}, i = 1, \cdots, c, j = 1, \cdots, N_i\}$, an optimal projection matrix U^l can be gotten, and the MD-FLDA feature of T is $FT = \{F^l = T^l U^l, 1 \le j \le 2^n\}$, then 2^n classification results can be gotten. We apply a voting mechanism to determine which class T will belong to. In MD-FLDA frame, all subbands of face directional images are applied. So it can extract more directional information of face images.

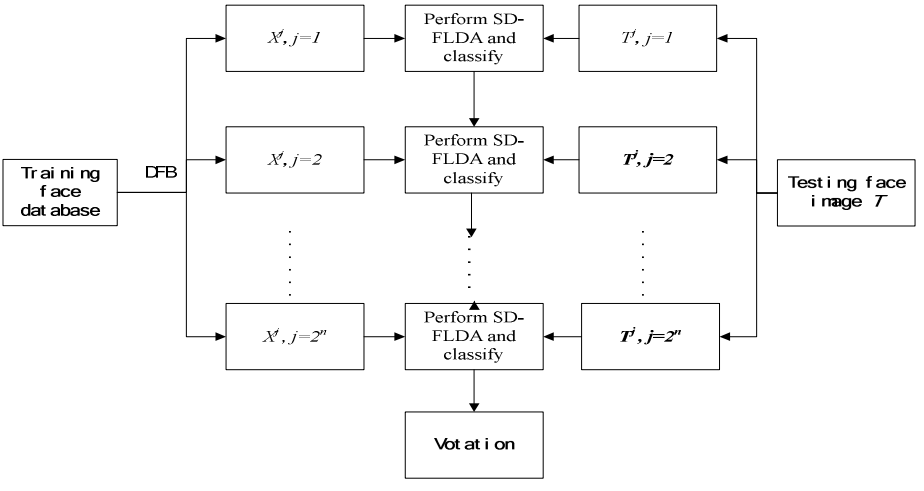


Fig. 4. The main idea of MD-FLDA

3 Experimental Results

In this section, some experiments are implemented to evaluate the effectiveness of the proposed SD-FLDA and MD-FLDA, which are also compared with some conventional subspace learning methods, including PCA, as well as the latest NFL-based subspace learning algorithms, such as NFLS, UDNFLA and NDNFLA. The experiments are implemented on a PC with 2.6-GHz CPU and 3G RAM. NFL classifier is used for classification on the features extracted by the NFL-based learning algorithms. To reduce the computation complexity, PCA is used before the NFL-based learning algorithms. All the energy is retained in the PCA phase.

3.1 Experiments on ORL Face Database

We test our algorithms on ORL face database [27] from Olivetti-Oracle Research Lab firstly. The ORL face database contains 400 face images, 10 different face images per person for 40 individuals. The size of images is 112×92.

In the following experiments, 5 images per person are selected for training and 5 images per person for testing randomly. The system runs 20 times. Table 1 lists the maximal average recognition rates (MARR), average training times (ATT) and corresponding dimensions of feature in different algorithms.

Table 1. Comparison of MARR, ATT and corresponding dimension of feature

Algorithms	MARR (%)	ATT (s)	Dimension of feature
PCA+NN	92.71	0.69	40
PCA+NFL	94.49	0.69	80
Fisherface	91.54	1.39	39
Nflspace	92.84	457.11	190
UDNFLA	88.70	1302.53	180
NDNFLA	95.54	158.52	90
SD-FLDA	96.65	4.4531	70
MD-FLDA	97.15	20.2656	190

3.2 Experiments on AR Face Database

Secondly, the proposed algorithms are evaluated on AR face database [28], which contains more than 4000 face images of 126 persons (56 women and 70 men). We selected a subset of 1652 face images of 118 persons with 14 images of different lighting conditions and expressions per person. The facial part of each image was cropped into 50×40.

In the following experiments, 5 images per person are selected for training and 9 images per person for testing randomly. The system runs 20 times. Fig. 8 displays the variance of ARR versus dimension of feature. Table 2 lists MARR, ATT and corresponding dimensions of feature in different algorithms.

Table 2. Comparison of MARR, ATT and corresponding dimension of feature

Algorithms	MARR (%)	ATT (s)	Dimension of feature
PCA+NN	76.04	1.24	120
PCA+NFL	85.21	1.24	190
Fisherface	94.81	2.68	120
Nflspace	91.26	868.507	190
UDNFLA	93.53	3275.02	120
NDNFLA	96.90	304.62	150
SD-FLDA	97.08	20.7813	140
MD-FLDA	97.82	95.3438	190

From the experiments, we can get the MARR of MD-FLDA is highest among the algorithms. Compared to NDNFLA, SD-FLDA and MD-FLDA have not only higher MARR, but also lower running time than NDNFLA.

3.3 Discussion

SD-FLDA and MD-FLDA extract image statistic feature. At the same time, they use the signal properties of images, too. On one hand, SD-FLDA and MD-FLDA have higher MARR, it illustrates that directionality is effective for classification. On the other hand, when extract the directionality of images, the dimension of the image is lower than the original image, so the ATT is also low.

4 Conclusions

In this paper, SD-FLDA and MD-FLDA are proposed. SD-FLDA and MD-FLDA are two NFL and DFB based image feature extraction algorithms. They extract not only the statistics information, but also the signal feature of the images. SD-FLDA and MD-FLDA have lower computational complexity and higher recognition accuracy than other NFL-based learning algorithms. Experimental results confirm the efficiency of the proposed SD-FLDA and MD-FLDA.

References

1. Lin, C.C., Shiu, P.F.: HighCapacity Data Hiding Scheme for DCT-based Images. Journal of Information Hiding and Multimedia Signal Processing 1, 220–240 (2010)
2. Hu, W.C., Yang, C.Y., Huang, D.Y., Huang, C.H.: Feature-based Face Detection Against Skin-color Like Backgrounds with Varying Illumination. Journal of Information Hiding and Multimedia Signal Processing 2, 123–132 (2011)
3. Belhumeur, V., Hespanha, J., Kriegman, D.: Eigenfaces vs Ffisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. on Pattern Analysis and Machine Intelligence 19, 711–720 (1997)
4. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2002)
5. Hong, Z.Q., Yang, J.Y.: Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane. Pattern Recognition 24, 317–324 (1991)
6. Li, J.B., Pan, J.S., Lu, Z.M.: Kernel Optimization-Based Discriminant Analysis for Face Recognition. Neural Computing and Applications 18, 603–612 (2009)
7. Yang, J., Frangi, A.F., Yang, J.Y., Zhang, D., Jin, Z.: KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 27, 230–244 (2005)
8. Li, J.B., Yu, L.J., Sun, S.H.: Refined Kernel Principal Component Analysis Based Feature Extraction. Chinese Journal of Electronics 20, 467–470 (2011)
9. Pan, J.S., Li, J.B., Lu, Z.M.: Adaptive Quasiconformal Kernel Discriminant Analysis. Neurocomputing 71, 2754–2760 (2008)
10. Li, J.B., Pan, J.S., Chen, S.M.: Kernel Self-Optimized Locality Preserving Discriminant Analysis for Feature Extraction and Recognition. Neurocomputing 74, 3019–3027 (2011)

11. Li, J.B., Pan, J.S., Chu, S.C.: Kernel Class-wise Locality Preserving Projection. Information Sciences 178, 1825–1835 (2008)
12. Li, J.B., Gao, H.J., Pan, J.S.: Common Vector Analysis of Gabor Features with Kernel Space Isomorphic Mapping for Face Recognition. International Journal of Innovative Computing, Information and Control 6, 4055–4064 (2010)
13. Ma, B., Qu, H.Y., Wong, H.S.: Kernel Clustering-based Discriminant Analysis. Pattern Recognition 40, 324–327 (2007)
14. Mika, S., Ratsch, G., Weston, J., Schölkopf, B., Muller, K.R.: Fisher Discriminant Analysis with Kernels. In: Proc. IEEE Int'l Workshop Neural Networks for Signal Processing IX, pp. 41–48 (1999)
15. Li, S.Z., Lu, J.W.: Face Recognition Using the Nearest Feature Line Method. IEEE Trans. Neural Networks 10, 439–443 (1999)
16. Chien, J.T., Wu, C.C.: Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 24, 1644–1649 (2002)
17. Chen, K., Wu, T.Y., Zhang, H.J.: On the Use of Nearest Feature Line for Speaker Identification. Pattern Recognition Letters 23, 1735–1746 (2002)
18. Gao, Q.B., Wang, Z.Z.: Using Nearest Feature Line and Tunable Nearest Neighbor Methods for Prediction of Protein Subcellular Locations. Computational Biology and Chemistry 29, 388–392 (2005)
19. Zheng, Y.J., Yang, J.Y., Yang, J., Wu, X.J., Jin, Z.: Nearest Neighbour Line Nonparametric Discriminant Analysis for Feature Extraction. Electronics Letters 42, 679–680 (2006)
20. Pang, Y., Yuan, Y., Li, X.: Generalised Nearest Feature Line for Subspace Learning. Electronics Letters 43, 1079–1080 (2007)
21. Lu, J.W., Tan, Y.P.: Uncorrelated Discriminant Nearest Feature Line Analysis for Face Recognition. IEEE Signal Processing Letter 17, 185–188 (2010)
22. Yan, L.J., Pan, J.S.: Neighborhood Discriminant Nearest Feature Line Analysis for Face Recognition. In: Proceedings of IBICA 2011, pp. 34–347. IEEE Press, Shenzhen (2011)
23. Bamberger, R.H., Smith, M.J.T.: Narrow Band Analysis of a Filter Bank for the Directional Decomposition of Images. In: IEEE International Conference Trans. on Acoustics, Speech and Signal Processing, pp. 1739–1742. IEEE Press, Albuquerque (1990)
24. Bamberger, R.H., Smith, M.J.T.: A Filter Bank for the Directional Decomposition of Images - Theory and Design. IEEE Trans. on Signal Processing 40, 882–893 (1992)
25. Lien, C.C., Huang, C.L.: The Design of 2-D Nonseparable Directional Perfect Reconstruction Filter Banks. Multidimensional Systems and Signal Processing 5, 289–300 (1994)
26. Park, S.I., Smith, M.J.T., Mersereau, R.M.: Improved Structures of Maximally Decimated Directional Filter Banks for Spatial Image Analysis. IEEE Trans. on Image Processing 13, 1424–1431 (2004)
27. Olivetti, Olivetti & Oracle Research Laboratory Face Database of Faces, `http://www.cam-orl.co.uk/facedatabase.html`
28. Martinez, A.M., Benavente, R.: The AR Face Database, CVC, Tech. 1 Rep. 24 (1998)

A (2, 2) Secret Sharing Scheme Based on Hamming Code and AMBTC

Cheonshik Kim[1], Dongkyoo Shin[1], Dongil Shin[1], and Ching-Nung Yang[2]

[1] Dep. of Computer Engineering, Sejong University, Seoul, Korea
[2] Dept. of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan, ROC
{mipsan,shindk,dshin}@sejong.ac.kr, cnyang@mail.ndhu.edu.tw

Abstract. We present a secret sharing scheme based on absolute moment block truncation coding (AMBTC) and Hamming code theory. AMBTC is composed of bit planes and two representative gray level pixel values, which are the high and low range means. In this secret sharing scheme, a dealer distributes a secret image to the participants as shadow images. It is important in a secret sharing scheme to keep the shadow images confidential. We therefore propose verifying the authenticity of the reconstructed secret image during the revealing and verifying phase. In order to improve confidentiality, our proposed scheme uses meaningful AMBTC compressed images. Experimental results show that the reconstructed secret sharing images are the same as the original secret sharing images and that the proposed scheme exhibits good performance compared to that of previous schemes.

Keywords: Secret sharing, AMBTC, Hamming code, shadow image.

1 Introduction

Secret information such as text, audio, and images is frequently used in medical, commercial, and military fields. Security is a very important issue in these fields, and there has therefore been significant research about security in recent years, including secret hiding and watermarking [1,2,3]. In the case of a single carrier for an application, all secret information such as images, videos, and MP3 files is stored in the carrier. In contrast, in secret sharing that uses two or more carriers, security is stronger than that for data hiding. One multi-carrier secret sharing scheme involves encoding a secret data set into n shares and distributing them to n participants; any k or more of the shares can be collected to recover the secret data, but k-1 or fewer shares will reveal no information about the secret. This technique is known as the (k, n)-threshold secret sharing scheme, and it was first proposed by Shamir [4]. Since then, many researchers have investigated implementations of the (k, n)-threshold scheme [5,6,7,8]. Benaloh and Leichter [9] proposed a more generalized sharing method. Lin and Tsai [10] and Yang et al. [11] proposed a scheme which was shown to have the maximum secret-carrying capacity in the cover image, with a size that is one-quarter the size of

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 129–139, 2012.

shadow images. Other researchers have shown the limited embedding capacity of secret information [12,13,14,15]. To date, most schemes have tried to address the requirements of secret sharing, which are to satisfy the security conditions and to enable reconstruction of the secret image [20,21,22,23]. The traditional schemes [5,6,7,8] require complex computational time. In this paper, we propose a secret image sharing scheme that not only satisfies all of these criteria; in addition, our proposed scheme will be especially important for medical, military, and artistic images, because the method uses meaningful shadow images.

2 Related Works

In Section 2.1, we will describe how to construct an absolute moment block truncation coding (AMBTC) compression image that is used as the shadow image for secret sharing. In Section 2.2, we will explain Hamming code theory, as well as how to encode and decode information.

2.1 AMBTC

AMBTC [16], proposed by Lena and Mitchell in 1984, is a version of Block Truncation Code (BTC), a lossy compression algorithm for grayscale or color images. The processing time for the BTC algorithm [17] has significant computational complexity and therefore is not generally recommended for time-consuming applications. The grayscale image that is to be encoded is divided into non-overlapping blocks of size (4×4) or (8×8), for example, and the average of the blocks is calculated as in Eq. (1):

$$\overline{x} = \frac{1}{W \times H} \sum_{i=1}^{N} x_i \tag{1}$$

where x_i denotes the i^{th} pixel, N is number of pixels in the block, and W and H are the width and height of the block. A pixel of a grayscale image is composed of 8 bits, but that of a binary image is 1 bit. Therefore, we must convert the grayscale image pixels into binary values for AMBTC. Eq. (2) shows how to construct a bit-plane for AMBTC from the grayscale image. If $x \geq \overline{x}$ then 1 is assigned to b_i, and otherwise 0 is assigned to b_i:

$$b_i = \begin{cases} 1 & if \quad x_i, \geq \overline{x} \\ 0 & if \quad x_i, < \overline{x} \end{cases} \tag{2}$$

The means α for the higher range and β for the lower range are calculated with Eq. (3) and Eq. (4):

$$\alpha = \frac{1}{t} \sum_{x_i \geq \overline{x}} x_i \tag{3}$$

$$\beta = \frac{1}{(W \times H) - t} \sum_{x_i < \overline{x}} x_i \tag{4}$$

where t is the number of pixels in a block with gray level greater than $\overline{x}$. A binary block B contains the bit-planes that represent the pixels. The values α and β are used to decode the AMBTC compressed image; every 1 in block B is replaced by α and every 0 is replaced by β:

$$g_i = \begin{cases} \alpha & if \quad b_i = 1 \\ \beta & if \quad b_i = 0 \end{cases} \tag{5}$$

where g_i is a grayscale pixel, and thus the grayscale image is reconstructed. The compression rate for AMBTC is 2 bpp, because the total number of bits required for a block is 32 bits. BTC has a complex computation, while AMBTC has a simple computation and therefore requires less computation time than BTC.

2.2 Hamming Code

In this section, we will explain Hamming code theory [18]. An AMBTC image is a stream of n pixels, which are 0s or 1s, where $n = P \times Q$, the size of the image. Let r be a non-negative integer, the dimension of the parity space, and let $m = 2r - 1$ be the code length and $k = m - r$ be the number of bits that are encoded in each codeword. The codewords will have a minimum Hamming distance of $d = 3$, so that one error can be corrected and two errors detected. We note that in order to correct one error, the position of the erroneous bit must be determined. For an n-bit code, $log_2 n$ bits are therefore required. Eq. (6) shows the parity check matrix for a (7, 4) Hamming code:

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{6}$$

For c to be a codeword, it must be in the null space of this matrix, i.e., $Hc = 0$. Let us assume that there is a sequence of bits which has an error in the first bit position, e.g., 1101010_b. We calculate the syndrome S with:

$$S = H \times (c)^T \tag{7}$$

where H is the parity check matrix, c is a 7-bit binary number and T denote the transpose of a matrix c. That is, the syndrome is $([001])^T$. A syndrome value that is not zero denotes the position of the erroneous bit. If you flip the bit at this position in the codeword, every bit of the codeword will be correct. Binary Hamming codes are $[2^r - 1, 2^r - 1 - r]$ linear codes with a parity check matrix H of dimensions $r \times (2^r - 1)$ and whose columns are binary expansions of the numbers $1, \cdots, 2^r - 1$. For example, Eq. (8) shows the parity check matrix H for $r = 4$. Let us assume that the cover object is an image consisting of $P \times Q$ pixels.

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{8}$$

Example 1. We assume that the codeword c is [1101001]. It is easy to calculate the syndrome using Eq.(7) with the parity check matrix H and the codeword: $S = H \times (c)^t = ([000])^t$. If the computed syndrome vector S is 0, as in this case, there is no error in the codeword. Otherwise, there is an error in the bit at position S in c.

3 Proposed Method

This section proposes our (2, 2) secret sharing scheme for AMBTC images. Our proposed scheme employs Boolean XOR and Hamming operations, and two cover AMBTC images are used as the shadow images. Section 3.1 describes how to construct the shadow images using Hamming code theory, and section 3.2 describes how to reconstruct the shadow images.

3.1 Construction of the Shadow Images

Our scheme is described below in terms of the construction procedure (for computing the shadow images) and the reconstruction procedure (for reconstructing the secret image from the shadows). Two AMBTC images and one secret binary image are required for the secret sharing. We present the shadow image construction step by step:

Input: Original grayscale images AG and BG (with $P \times Q$ pixels), secret image SI, parity check matrix H
Output: Stego images AS and BS, length of secret image $|SI|$

Step 1. First, we construct two AMBTC images AI and BI (Fig. 1(b) and (d)) by applying Eqs. (1) and (2) to the grayscale images AG and BG (Fig.1 (a) and (c)). SI is the binary image (Fig.1(e)) which is used as the secret shared image, and CNT is the length of SI.

Step 2. Divide the two images AI and BI into 4×4 blocks, letting $A = \{a_1, a_2, \cdots, a_{16}\}$, where $a_i \in \{0, 1\}$ for $1 \leq i \leq 16$, be a bit-plane of AI, letting $B = \{b_1, b_2, \cdots, b_{16}\}$, where $b_i \in \{0, 1\}$ for $1 \leq i \leq 16$, be a bit-plane of BI.

Step 3. Read two blocks A and B, one from AI and one from BI, as shown in Figs. 2(c) and (d). Read 7 pixels from each of A and B; call these groups of 7 pixels C and D. Read a block $F = \{f_1, f_2, f_3\}$ from SI, where $f_i \in \{0, 1\}$ for $1 \leq i \leq 3$.

Step 4. Calculate the syndrome S by applying Eq. (7) to the parity check matrix H and $XOR(C, D) : S = H \times (C \oplus D)^T$.

Step 5. Compute the value IX of the exclusive or operation applied to the syndrome value S and the 3 pixels from $SI : IX = F \oplus S$.

Step 6. The syndrome IX denotes the index in the codeword where an error occurs. If IX is 0, this means that there is no error in the codeword. Otherwise, we flip the value at this position in the codeword. Through these steps, it is possible to distribute 3-pixels of the secret data into the block. In this case, it is possible to flip either AI or BI, but we only use the first image AI.

Step 7. Decrease CNT by 3. If CNT is greater than 0, return to step 3 to continue the construction process until there are no more pixels of SI.

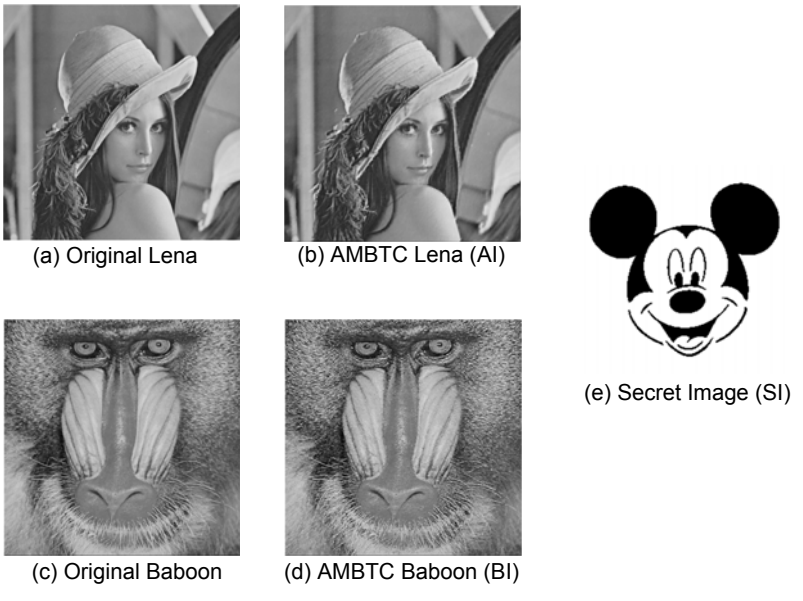


Fig. 1. Four 512×512 images: (a) Original Lena, (b) AMBTC Lena, (c) Original Baboon, (d) AMBTC Baboon, and one 240×240 secret shared image (e) Mickey

Example 2. Codewords $C = [1\ 1\ 1\ 0\ 1\ 1\ 1]$ and $D = [1\ 0\ 0\ 1\ 1\ 0\ 0]$ are pixels from Figs. 3(a) and 3(c), reading from left to right and from top to bottom. The secret stream pixels are $F = [1\ 1\ 1]$. Calculate $S = H \times (C \oplus D)^T = [100]$ and $IX = XOR(F, S) = [011]$. S and IX are syndrome values and S does not include 3-bits of F. Therefore, the XOR operator applies to both F and S. If IX is not zero, this indicates that we should flip a pixel in the codeword C. In this example, the third position of the codeword should be flipped, as shown in Fig. 3(b). As a result, participants can extract secret pixels from the shadow images as they calculate the syndrome of the Hamming code.

(a) Before(Original Lena)

C

162	162	161	160
162	162	161	160
162	162	161	160
162	162	161	160

(b) After (α:162, β:160)

C

1	1	1	0
1	1	1	0
1	1	1	0
1	1	1	0

(e) Mickey (4x4)

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

(c) Before(Original Baboon)

D

145	55	48	88
116	101	39	67
76	114	46	46
70	135	95	48

(d) After (α:87, β:62)

D

1	0	0	1
1	1	0	0
0	1	0	0
0	1	1	0

Fig. 2. Both (a) and (c) are blocks of the grayscale images, (b) and (d) are bit planes corresponding to (a) and (c), and (e) is a block of the binary image

In a 4×4 block, there is only 1pixel flipped for 3-bit secret sharing. In our experiment, we tested a secret sharing scheme with AMBTC using the Hamming Code (15, 11), which makes it possible to conceal 4 bits in a 4×4 block by flipping one pixel in the block.

(a) Before (α:162, β:160)

C

1	1	1	0
1	1	1	0
1	1	1	0
1	1	1	0

(b) After (α:162, β:160)

C

1	1	**0**	0
1	1	1	0
1	1	1	0
1	1	1	0

(e) Mickey (4x4)

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

(c) Before (α:87, β:62)

D

1	0	0	1
1	1	0	0
0	1	0	0
0	1	1	0

(d) After (α:87, β:62)

D

1	0	0	1
1	1	0	0
0	1	0	0
0	1	1	0

Fig. 3. Both (a) and (c) are blocks of bit-planes, and (b) and (d) embed the 3-bits [1 1 1] from (e)

3.2 Reconstruction of the Shadow Images

In section 3.1, we propose how a dealer can distribute a secret sharing image to participants. Participants can then reconstruct the secret sharing image using the Hamming code with the shadow images they receive. In this section, we will explain how participants can reconstruct the secret shared images.

The reconstruction procedure is as follows:

Input: Grayscale shadow images AS and BS, the size of the secret image CNT, parity check matrix H
Output: Secret shared image SI

Step 1. Assume that participants have the shadow images AS and BS. In order for us to decode, we must know the bit-planes for AS and BS. Thus, we convert AS and BS into AI and BI, which are AMBTC images that can be constructed using Eqs. (1) and (2).
Step 2. The shadow images AI and BI are divided into non-overlapping 4×4 pixel blocks.
Step 3. Read two blocks A and B, one from each of AI and BI, as shown in Figs. 3(b) and 3(d). Read 7 pixels from each of A and B, and call these 7 pixels C and D, where $C = \{c_1, c_2, \cdots, c_7\}$ and $D = \{d_1, d_2, \cdots, d_7\}$.
Step 4. Calculate the syndrome using Eq. (7) with parity check matrix H and XOR(C, D), i.e., the syndrome is $S = H \times (C \oplus D)^t$. S is a reconstruction of 3 pixels which are part of the secret shared image.
Step 5. Concatenate SI and S, i.e., $SI = SI||S$.
Step 6. $CNT = CNT - 3$. If $CNT \neq 0$, go to step 3. Otherwise, exit this procedure.

4 Experimental Results

We have proposed a novel (2, 2) secret sharing scheme for AMBTC compressed images. In order to prove that our proposed scheme is correct, we performed an experiment to verify that our scheme ensures that the hidden image can be reconstructed. In addition, our experiment determined the image qualities for the shadow images, which are very important for resisting forgery from attackers. To carry out our experiment, 512×512 AMBTC compressed images were used as shadow images. Fig. 4(a) is a secret shared binary image "left". Its pixels were distributed as the shadow images shown in Figs. 4(b) and 4(c). From Fig. 4, we can see that the shadow image size is the same as the original grayscale images. Moreover, the reconstructed secret sharing image (Fig. 4(d)) is exactly the same quality as the original secret sharing image (4(a)). The qualities of the original AMBTC Lena and Baboon are 33.6835 dB and 26.9827 dB, respectively. The qualities of 4(b) and 4(c) with the concealed secret image are 31.3557 dB and ∞ dB, respectivelythere are no changed pixels in the Baboon image. Using the human visual system, we can see that Lena in Fig. 4(b) has almost the same quality as the original Lena shown in Fig. 1(a).

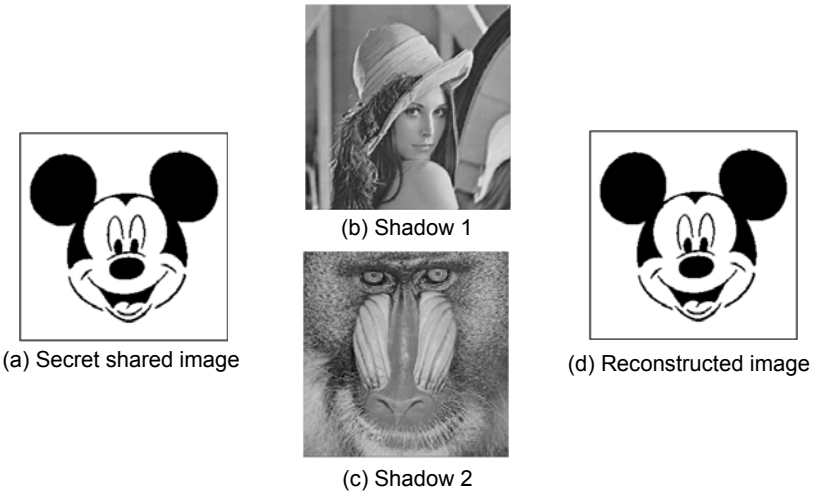


Fig. 4. 4(a) is a 240×240 secret shared image, (b) and (c) are 512×512 shadow images, and (d) is the 240×240 reconstructed image

In our experiments, the qualities of the shadow images are measured by the peak-signal-to-noise ratio (PSNR) [19]. The PSNR is the most popular criterion for measuring distortion between the original image and shadow images. It is defined as follows:

$$PSNR = 10 \times log_{10}(255^2/MSE) \tag{9}$$

where MSE is the mean square error between the original grayscale image and the shadow image:

$$MSE = \frac{1}{m \times n} \sum_{i}^{m} \sum_{j}^{n} [I(i,j) - I'(i,j)]^2 \tag{10}$$

The symbols $I(i,j)$ and $I'(i,j)$ represent the pixel values of the original grayscale image and the shadow image (AMBTC) at position (i,j), respectively, and m and n are the width and height of the original image. Table 1 compares our scheme with other Visual Secret Sharing (VSS) schemes when the secret image is a binary image. From this table, we can see that when our scheme is applied to binary secret images, the reconstructed secret image is exactly the same as the original secret image. The schemes in [20, 21] use a stacking operation to reconstruct the secret images, so the reconstructed images are not the same as the original secret image owing to pixel expansions. While Chang et al. [22] has a fixed shadow size that is the same as that of the original secret image, the schemes [20,21] make a tradeoff between the number of shadows (n) and the shadow size. In general, the shadow size in other VSS schemes is larger than

the original secret image size even when the number of shadows is set to $n = 2$. In other words, our scheme is the only one of these schemes that does not require pixel expansion.

Table 1. Comparison of our scheme and other schemes when the secret image is a binary image

Criteria	Prisco et al. [20]	Tsai et al. [21]	Chang et al. [22]	Cai et al. [23]	Proposed Scheme
Number of shadows (n)	$n \geq 2$	$n \geq 2$	$n \geq 2$	$n \geq 2$	$n \geq 2$
MSE of the reconstructed image	$MSE > 0$	$MSE > 0$	$MSE = 0$	$MSE > 0$	$MSE = 0$
Shadow size	$(2n + n + 1) \times N$	$n \times N$	N	$N \times 5/16$	$N \times N$
Conputational complexity	High	High	Low	Relatively Low	Low
Shadow style	Noise-like shadow	Noise-like shadow	Noise-like shadow	Noise-like shadow	Meaningful shadow
Additional information	No	No	No	No	No
Cheating prevention function	Yes	Yes	Yes	Yes	Yes

In Cai et al. [23], the AMBTC scheme is replaced with a GSBTC scheme (Chang et al.[11]) to preserve features for each block of a color secret image so that the generated shadow can remain almost one-third (5/16) the size of the original color secret image. Our proposed scheme shows MSE = 0, and while most secret sharing schemes use noisy shadow images, our shadow images are meaningful. Thus, there may be fewer suspicious about our shadow images, which is important for secret sharing images.

5 Conclusion

In this paper, we proposed a novel (2, 2) secret sharing scheme based on AMBTC and Hamming code theory. In our scheme the number of shadows is limited to 2, and the scheme can work on binary images. Moreover, our scheme provides authentication for reconstructed secret images by using a concealed binary image without needing to send extra information through the safety channel. Using a Hamming code and XOR operation, the proposed scheme can easily recover the reconstructed image from the collected shadows during the reconstruction phase. The computational cost of the sharing scheme is low, because XOR and Hamming operations are not complex time operations. Experimental results confirm that our proposed scheme produces a reconstructed image that has the same

quality as the original secret sharing image. Moreover, we use meaningful shadow images. Based on the quality of the extracted secret binary image and the MSE between the extracted secret binary image and the original binary image, we conclude that the proposed scheme allows honest participants to correctly reconstruct the secret image.

Acknowledgement. This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2009.

References

1. Yang, C.Y., Lin, C.H., Hu, W.C.: Reversible data hiding by adaptive IWT-coefficient adjustment. Journal of Information Hiding and Multimedia Signal Processing 3(1), 24-34 (2011)
2. Lin, B., He, J., Huang, J., Shi, Y.Q.: A survey on image steganography and steganalysis. Journal of Information Hiding and Multimedia Signal Processing 2(2), 142–172 (2011)
3. Yang, C.Y., Hu, W.C.: High-performance reversible data hiding with overflow/underflow avoidance. ETRI Journal 33(4), 580–588 (2011)
4. Shamir, A.: How to share a secret. Communications of the ACM 22(11), 612–613 (1979)
5. Karnin, E.D., Greene, J.W., Hellman, M.E.: On secret sharing systems. IEEE Transactions on Information Theory 29, 35–41 (1983)
6. Stinson, D.R.: Decomposition constructions for secret-sharing schemes. IEEE Transactions on Information Theory 40(1), 118–124 (1994)
7. Verheul, E.R., van Tilborg, H.C.A.: Constructions and properties of k out of n visual secret sharing schemes. Designs, Codes and Cryptography 11(2), 179–196 (1997)
8. Blundo, C., Santis, A.D.: Lower bounds for robust secret sharing schemes. Information Processing Letters 63(6), 317–321 (1997)
9. Benaloh, J., Leichter, J.: Generalized Secret Sharing and Monotone Functions. In: Goldwasser, S. (ed.) CRYPTO 1988. LNCS, vol. 403, pp. 27–35. Springer, Heidelberg (1990)
10. Lin, C.C., Tsai, W.H.: Secret image sharing with steganography and authentication. Journal of Systems Software 73(3), 405–414 (2004)
11. Yang, C.N., Yu, K.H., Lukac, R.: User-friendly image sharing using polynomials with different primes. International Journal of Imaging Systems Technology 17(1), 40–47 (2007)
12. Tsai, C.S., Chang, C.C., Chen, T.S.: Sharing multiple secrets in digital images. Journal of Systems Software 64(2), 163–170 (2002)
13. Thien, C., Lin, J.C.: Secret image sharing. Computers and Graphics 26(1), 765–770 (2002)
14. Wu, Y.S., Thien, C.C., Lin, J.C.: Sharing and hiding secret images with size constraint. Pattern Recognition 37(7), 1377–1385 (2004)
15. Chang, C.C., Lin, C.C., Lin, C.H., Chen, Y.H.: A novel secret image sharing scheme in color images using small shadow images. Information Sciences 178(11), 2433–2447 (2008)

16. Lema, M., Mitchell, O.: Absolute moment blocks truncation coding and applications to color images. IEEE Transactions on Communications 32, 1148–1157 (1984)
17. Tu, S.F., Hsu, C.S.: A btc-based watermarking scheme for digital images. International Journal of Information Security 15(2), 216–228 (2004)
18. Grossman, J.: Coding theory: introduction to linear codes and applications. Insight: River Academic Journal 4(2), 1–17 (2008)
19. Ichigaya, A., Kurozumi, M., Hara, N., Nishida, Y., Nakasu, E.: A method of estimating coding PSNR using quantized DCT coefficients. IEEE Transactions on Circuits and Systems for Video Technology 16(2), 251–259 (2006)
20. Prisco, R.D., Santis, A.D.: Cheating Immune (2, n)-Threshold Visual Secret Sharing. In: De Prisco, R., Yung, M. (eds.) SCN 2006. LNCS, vol. 4116, pp. 216–228. Springer, Heidelberg (2006)
21. Tsai, S., Chen, T.H., Horng, G.: A cheating prevention scheme for binary visual cryptography with homogeneous secret images. Pattern Recognition 40(8), 2356–2366 (2007)
22. Shang, C.C., Lin, C.C., Le, T.H.N., Le, H.B.: Sharing a verifiable secret image using two shadows. Pattern Recognition 42, 3097–3114 (2009)
23. Cai, K.Y., Wang, S.S., Shiu, P.F., Lin, C.C.: A verifiable secret sharing scheme based on AMBTC. In: Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, ICUIMC 2011, pp. 129–138. ACM, Seoul (2011)

An Automatic Image Inpainting Method for Rigid Moving Object

Jen-Chi Huang[1] and Wen-Shyong Hsieh[2]

[1] Dept. of Computer and Communication, National Pingtung Institute of Commerce, Pingtung City, Taiwan 900
[2] Dept. of Computer Science and Information Engineering, Shu Te University, Kaohsiung, 824, Taiwan
leohuang@cm1.hinet.net

Abstract. Image inpainting is to remove unnecessary objects or reconstruct damaged parts of an image automatically. In order to reduce the capacity of the file, the video film is usually stored after the quad or the octal processing. It is this processing that causes an image of low quality and makes the moving object become indistinct. In this paper, we proposed a new image inpainting method for rigid moving object in the temporal domain, in which the pixels are patched by the neighboring frames. The more neighboring frames are patched, the better a PSNR of the moving object image is obtained. The experiment results have shown that our method has good performance and obtained a better quality of the rigid moving object.

Keywords: We would like to encourage you to list your keywords in this section.

1 Introduction

Over the past few years, the quality of video equipment has improved considerably. As a result the development of video-surveillance systems has attracted the attention of research [1]. The film used in video-surveillance systems is valuable evidence in traffic offences, traffic accidents, a robbery or other criminal offences [2]. However, in order to reduce the capacity of the file, the video film is usually stored after the quad or the octal processing. It is this processing that causes an image of low quality and makes the moving object become indistinct. Therefore image inpainting is an important issue in the development of Video and Image Technology [3-4]. Image inpainting is to remove unnecessary objects or reconstruct damaged parts of an image automatically. Hence, we propose a new method to repair the image of the rigid moving object.

In order to improve the quality of the image, we have some methods available to us in the special domain, such as the dilation method or the mean method [3]. The dilation method is a simple method where one pixel expands into four pixels. The mean method involves averaging a pixel's 8 neighbors which are not null. In this paper, we propose a new method in the temporal domain where the pixels are patched

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 140–147, 2012.

by the neighboring frames. This is possible because an object can move to another location in video sequences, so that the pixels, which are not detected in one frame, may be detected in other frames. In other words, we can combine the pixels, which are from different frames but depict the same object, to become an image of high quality and resolution of the moving object.

Object segmentation [5] is the most import issue for mpeg-4 video coding, and the most popular scheme is the change detection method [6] for inter-frame differences. It is simple to implement, and it enables automatic detection of new appearances. In the change detection method, the Frame Different (FD), which indicates the change in pixels between frame n and frame n-1, is found by the following rule:

for all $(i, j) \in$ the coordinates of the frame n

$FD_n(i, j) = |I_n(i, j) - I_{n-1}(i, j)|$

if $(FD_n(i, j) < V_{thr})$ then $FD_n(i, j) = 0$

end for

$DE_n = \text{Edge\_Detection}(FD_n)$

In which (i,j) is the coordinate of the frame n, I(i,j) is the pixel value of the frame n, Vthr is a threshold determined by the significance test, and the Edge_Detection ( ) is applied to the CDMn to extract the rough object shapes called the edge map of significant difference pixel (DEn). However, the change detection method suffers from a 'fat boundary' and the problem of a 'double-edge'. The new wavelet-based object segmentation [7] was proposed to accurately obtain object shapes without the double-edge problem and with a lower time complexity.

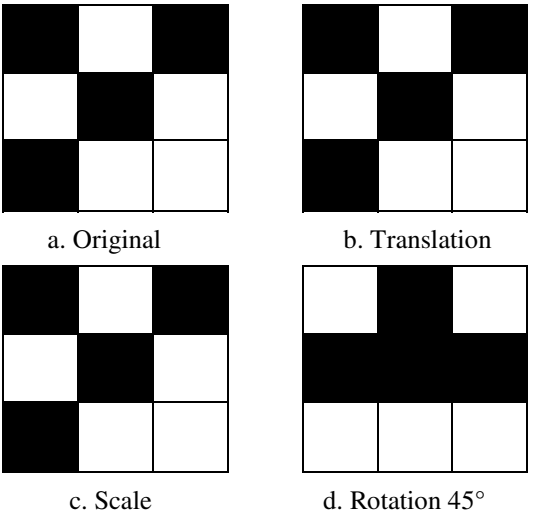


Fig. 1. Three operations of camera motion

However, most moving objects on the road, such as cars or motorcycles move much too fast for the change detection method to work well. It is very easy to extract the newly appearing objects from the pure background without moving objects, and it is easy to find the pure background in a video sequence, because most moving objects are moving in and out in one frame. Hence we propose the pure background method to deal with the problem of the fast moving object.

The object motion estimation is an import issue to estimate the object motion in two successive video frames. The pixel-based or point mapping technique is currently the primary approach, and it usually uses some special feature point to be the matching point. In this paper, we use a robust feature point in a frame called 'cross point', which is located at the corner or the cross-section of the edge lines [8]. Fig. 1 shows the useful characters of the cross point. When the object is translated, rotated or scaled, the cross points in this frame remain the same, because the number of edge points in the 3x3 block are still the same after any motion operation. This allows us to use the relative cross points between two successive frames to find the object motion in the video sequences.

Motion Estimation is the most important issues to quickly find the block motion [9]-[10]. In this paper, we use the automatic feature-based global motion estimation [11] to find the affine model of object motion. The simple and efficient algorithm to detect the cross point is as follows:

Cross Point - Detection Alogrithm

$EM_n = CannyEdgeD\ etection(F_n)$

For all (x, y) in EM_n

if $((EM_n(x, y) = 1)$ and $(\sum_{i=-1}^{1} \sum_{j=-1}^{1} EM_n(x+i, y+j) \geq 4,))$ $CPM_n(x, y) = 1$

else $CPM_n(x, y) = 0$

End

Here Fn is the intensity of frame n, EMn is the edge map after applying the canny edge detection method to frame n, and CPMn is the cross points map in frame n. We define a point as a cross point in the 3x3 block, if this point is on the edge line and more than 3 of its 8-neighbor-points are edge points also. If we detect too many cross points in , then we can define the detection algorithm strictly as the edge points of its 8-neighbor-points equaling 4 or 5 edge-points. The fewer number of cross points detected, the lower the time complexity. In addition, our detection algorithm uses a cluster of edge points to define the cross point. Using a cluster of edge points rather than the shape of the edge can deal with the half point problem and the rotation operation in any angle.

In the cross-point match algorithm [8] in this proposal, a search window of 7x7,is used to define the scope of the match algorithm. For every cross point in frame n, we count the number of the cross points within the search window in frame n-1. And, the cross point which holds only one cross point within the search window in frame n-1 is taken to form a Cross-Point Pair ( ) and then this pair is recorded in the set of Cross Point Pair (SCPP), this is done because more than one cross point within the search

window will create an ambiguous situation. The process of the Cross-Point Match Algorithm is as follows:

Cross - Point Match Algorithm

for all (x, y)

if $(CPM_n(x, y) = 1)$ and $(\sum_{i=-3}^{3} \sum_{j=-3}^{3} CPM_{n-1}(x + i, y + j) = 1)$

Record $(x + i, y + i)$ and (x, y) in SCPP, *here* $CPM_{n-1}(x + i, y + j) = 1$

End if

Applying the Cross-Point Match Algorithm to CPMn and CPMn-1, we will find all Cross-Point Pairs, which will be used to find the affine model of object motion.

In the affine model, we must choice three Cross-Point Pairs to solve the parameters of the affine model, and verify the confirmation of these measured parameters for the rest of the Cross-Point Pairs in the SCPP. The ones with the largest confirmation will be assigned to present the Affine Model of the object motion.

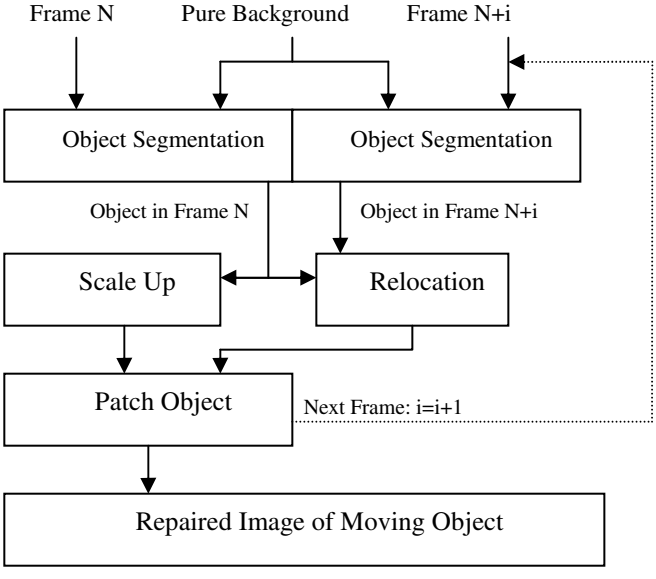


Fig. 2. Overall block diagram of the proposed moving object segmentation algorithm

2 Proposed Method

We propose this method in order to enhance the image resolution of the moving object. Fig. 2 shows the over-all block diagram of the proposed algorithm. First, we use the change detection method with frame N and a pure background to segment the moving objects in frame N. We also segment the moving objects in the successive frames using the same method. We choose the frame N as the base frame, and the

neighboring frames as patching frames. We use the Cross Point feature and the Affine Model to describe the relative location between the same object in the base frame and in the patching frames. Then we up-scale the QIF image of the moving objects in frame N to the base frame image ( ) in CIF format and the algorithm is as follows:

Scaling - up Algorithm

for all (x, y)

$I_{BF}(2x-1, 2y-1) = I_N(x, y);$

End for

We use the affine model to patch the pixels that are not defined in frame N, by means of the moving object image in the successive frames. As a result we obtain an repaired image of the moving object. The algorithm is as follows:

Patch Algorithm

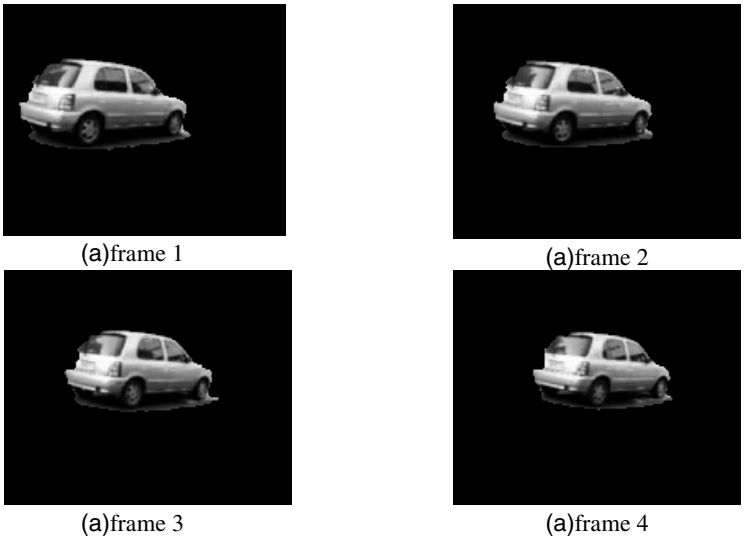
for all (x, y)

$[x' \; y'] = \text{Afine\_Parameter} * [1 \; x \; y];$

$I_{BF}(x', y') = I_N(x, y)$, where $I_{BF}(x', y')$ is null;

End for

In the patch object step, we use the threshold to avoid the luminance effect and the non-rigid object deformation. If the pixel value of the neighboring frames is less than the maximum value of its 8-neighboring pixels and more than the minimum value of its 8-neighboring pixels, then we will patch the pixel value of the neighboring frames in the null value pixel of the base frame.



(a)frame 1 (a)frame 2

(a)frame 3 (a)frame 4

Fig. 3. Segmentation result in the sequence frames

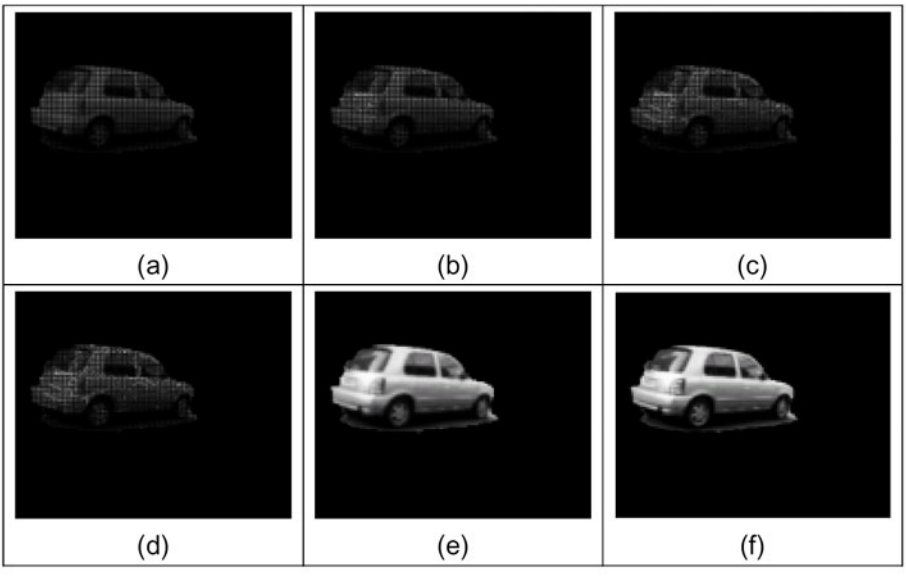


Fig. 4. The sequences of the Image Inpainting results with frames 1 to 4

3 Results

We applied the proposed algorithm to "Car", which is a surveillance type video in 352x288 CIF format. First we simulated the quad processing system to reduce the resolution to 176x144 QIF format by keeping the odd numbered pixels. Then we determined the pure background and applied the change detection method. After that we obtained the object segmentation results in the sequence frames. Fig. 3 shows the segmentation results of the sequence frames. We defined frame 1, see fig. 4 (a), as the base frame, and scaled it up to a 352x288 CIF format from the 176x144 QIF format in the “Small Car” video sequence. We observed that the base frame contained many pixels of null value. Hence, we used the neighboring frames as the patching frames to patch the null value pixels of the base frame. In figs. 4 (b) to (d), we patched the null value pixels from frames 2 to 4. Finally, we used the mean method, which is the average value of its 8-neighboring pixels, to patch the remaining null-value pixels, and the result is shown in fig. 4 (e). Fig. 4 (f) shows the original CIF image in the “small car” video sequence.

Table 1. Comparison with the original CIF frame

PSNR comparison		Small Car 		Large Car 	
		PSNR	Patch Number	PSNR	Patch Number
Quad Processing	(a) Mean with Frame 1	30.83 dB	0	29.07 dB	0
	(b) Only Patch Frame 2	30.96 dB	329	29.16 dB	752
	(c) Only Patch Frame 3	30.88 dB	454	29.15 dB	666
	(d) Only Patch Frame 4	30.91 dB	331	29.09 dB	555
	(e) Patch Frame 2 and 3	30.97 dB	809	29.21 dB	1442
	(f) Patch Frame 2 ,3,4	31.01 dB	1181	29.18 dB	2073
Octal Processing	(g) Mean with Frame 1	22.29 dB	0	21.49 dB	0
	(h) Only Patch Frame 2	22.52 dB	79	21.56dB	70
	(i) Only Patch Frame 3	23.05 dB	195	21.61 dB	105
	(j) Only Patch Frame 4	22.44 dB	62	21.56 dB	96
	(k) Patch Frame 2 and 3	23.23 dB	264	21.68 dB	180
	(l) Patch Frame 2 ,3,4	23.37 dB	341	21.74 dB	287

Table 1 shows the experiment results in two different video sequences, the "small car" and the "large car". Table 1 (a)-(f) used the quad processing. In (a), we only used the mean method, which is the average of its 8-neighboring pixels. In (b)-(d), we used frame 1 as the base frame and patched it by using only one frame, and then we used the mean method to patch the remainder null-value pixels. In (e), we used frame 1 as the base frame and patched it using frames 2 and 3, and then we used the mean method to patch the remainder null-value pixels. In (f), we used frame 1 as the base frame and patched it using frames 2, 3 and 4, and then we used the mean method to patch the remainder null-value pixels. It was evident that the more neighboring frames we patched, the better the PSNR of the moving object image we obtained. However, we found the PSNR of the "Large Car" in Table 1 (d) not satisfactory, because the object of "Large Car" in frame 4 was deformed too much. Hence, the "Large Car" in frame 4 caused a decreased PSNR in Table 1 (f). Table 1 (g)-(l) used the same general procedure as in (a)-(f) but used the Octal Processing.. The result shows that a better performance of enhancing the PSNR is obtained in octal processing compared to quad processing. Table 1 (L) shows the best performance in our experiment; we enhance the 1.08 dB of PSNR in the "small car", after patching it with frames 1 to 4. This also indicates that we obtained a better PSNR by patching more frames.

4 Conclusion

We proposed a new image inpainting method for Rigid Moving Object by patching neighboring frames. The experiment results have shown that our method has good

performance and obtained a better quality of the moving object. The more neighboring frames we patched, the better the PSNR of the moving object image we obtained. The performance of our proposed method depends on the accurate affine parameter estimation of the object motion and the object deformation.

References

1. Hu, W.-C., Yang, C.-Y., Huang, D.-Y.: Robust real-time ship detection and tracking for visual surveillance of cage aquaculture. Journal of Visual Communication and Image Representation 22(6), 543–556 (2011)
2. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts, and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10), 1337–1342 (2003)
3. Guo, H., Ono, N., Sagayama, S.: A structure-synthesis image inpainting algorithm based on morphological erosion operation. In: Congress on Image and Signal Processing, vol. 3, pp. 530–535 (May 2008)
4. Hung, J.C., Hwang, C.H., Liao, Y.C., Tang, N.C., Chen, T.J.: Exemplar-based image inpainting base on structure construction. Journal of Software 3(8), 57–64 (2008)
5. Hu, W.-C.: Real-time on-line video object segmentation based on motion detection without background construction. International Journal of Innovative Computing, Information and Control 7(4), 1845–1860 (2011)
6. Kim, C., Hwang, J.-N.: Object-Based Video Abstraction for Video Surveillance Systems. IEEE Transactions on Circuits and Systems for Video Technology 12(12), 1128–1138 (2002)
7. Huang, J.-C., Hsieh, W.-S.: Wavelet-based moving object Segmentation. IEE Electronics Letters 39(19), 1380–1382 (2003)
8. Brown, L.G.: A survey of image registration techniques. ACM Computing Surveys 24, 325–376 (1992)
9. Liu, P., Jia, K.: Research and Optimization of Low-Complexity Motion Estimation Method Based on Visual Perception. Journal of Information Hiding and Multimedia Signal Processing 2(3), 217–226 (2011)
10. Chen, C.-H., Chen, T.-Y., Wang, D.-J., Li, Y.-F.: Multipath Flatted-Hexagon Search for Block Motion Estimation. Journal of Information Hiding and Multimedia Signal Processing 1(2), 110–131 (2010)
11. Huang, J.-C., Hsieh, W.-S.: Automatic Feature-based Global Motion Estimation in Video Sequencess. IEEE Transactions on Consumer Electronics 50(3), 911–915 (2004)

Automatic Image Matting Using Component-Hue-Difference-Based Spectral Matting

Wu-Chih Hu and Jung-Fu Hsu

Department of Computer Science and Information Engineering,
National Penghu University of Science and Technology, Penghu, Taiwan
{wchu,d98523002}@npu.edu.tw

Abstract. This paper presents automatic image matting using component-hue-difference-based spectral matting to obtain accurate alpha mattes. Spectral matting is the state-of-the-art image matting and it is also a milestone in theoretic matting research. However, the accuracy of alpha matte using spectral matting is usually low without user intervention. In the proposed method, k-means algorithm is used to generate components of a given image. Next, component classification is used based on the hue difference of components to obtain the foreground, background, and unknown components. The corresponding matting components of the foreground, background, and unknown components are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian matrix. Finally, only matting components of the foreground and unknown components are combined to form the complete alpha matte based on minimizing the matte cost. Experimental results show that the proposed method outperforms the state-of-the-art methods based on spectral matting.

Keywords: spectral matting, image matting, alpha matte, hue difference.

1 Introduction

Image matting is a high-value tool in image editing and image composition. Image matting is the process to remove the background from a given image to obtain the foreground object along with opacity estimates (alpha matte). Therefore, image matting takes the problem of estimating the partial opacity of each pixel in a given image. A given image is assumed as a composite of a foreground image and a background image using the compositing equation, where the color of a given image is assumed as a convex combination of corresponding foreground and background colors with the associated alpha value. For each pixel in a given image, compositing equation gives us 3 equations (RGB channels) in 7 unknowns. Therefore, it is a highly under-constrained problem to obtain image matting.

In order to solve the highly under-constrained problem, the existing methods of image matting always require the user to provide additional constraints in the form of a trimap or s a set of scribbles (brush strokes), such as robust matting [1], easy matting [2], and closed-form matting [3]. Therefore, these image matting methods using additional constraints (user intervention) is the semi-automatic image matting. It is a time-consuming and troublesome process for the user. Therefore, it is a challenging task to obtain automatic image matting.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 148–157, 2012.

Spectral matting [4] is an automatic image matting. Spectral matting is the state-of-the-art image matting and it is also a milestone in theoretic matting research. Spectral matting is based on the extended idea of spectral segmentation [5]. In spectral matting, matting components of a given image are automatically obtained based on analyzing the smallest eigenvectors of a suitably defined Laplacian matrix (matting Laplacian [3]). Because the smallest eigenvectors of the matting Laplacian span the individual matting components of the given image, recovering the matting components of the given image is equivalent to finding a linear transformation of the eigenvectors. Finally, these matting components are combined to form the complete alpha matte based on minimizing the matte cost. However, the accuracy of alpha matte using spectral matting is usually low without user intervention. Therefore, it is a challenging task for automatic image matting to obtain high accuracy of alpha matte.

Wang and Li [6] proposed spectral matting based on color information of matting components to increase the accuracy of alpha matte using spectral matting. The intensity/saturation classification [7] is used by analyzing the feature of the HSV color space. The color distinguishing scheme [8] is used to divide the HSV color space into 61 different unequal regions. The measurement of color histograms [9] is used to measure the color similarity between matting components and separate the matting components into the foreground and background groups. Finally, the complete alpha matte is obtained using these matting components of foreground group. However, the similar color between the foreground and background, the bigger distance between color histograms of components, and the area between components all would cause the wrong alpha matte.

Hu et al. [10] proposed spectral matting based on the palette-based component classification to obtain high accuracy of alpha matte. The palette-based component classification is based on the assumption of the angle between the foreground and background palettes is 180^0 in the hue space. Using the palette-based component classification, components are classified as ones of the foreground, background, or unknown regions. Next, the corresponding matting components are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian. Finally, only matting components of foreground and unknown regions are combined to form the complete alpha matte based on minimizing the matte cost. However, the wrong alpha matted would be obtained when the angle between the foreground and background is not 180^0 in the hue space.

In this paper, we propose an automatic image matting using component-hue-difference-based spectral matting to obtain accurate image matting. In the proposed method, k-means algorithm is used to generate components of a given image. Next, component classification is used based on the hue difference of components to obtain the foreground, background, and unknown components. The corresponding matting components of the foreground, background, and unknown components are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian. Finally, only matting components of the foreground and unknown components are combined to form the complete alpha matte based on minimizing the matte cost. Experimental results show that the proposed method can obtain the high-quality alpha matte for natural images without user intervention.

The rest of this paper is organized as follows. A review of spectral matting is presented in Section 2. The proposed spectral matting based on hue difference of components is described in Section 3. Section 4 presents experimental results and evaluations. Finally, the conclusion is given in Section 5.

2 Review of Spectral Matting

Image matting typically assumes that each pixel I_i of a given image is a convex combination of corresponding foreground color F_i and a background color B_i with the associated opacity value (alpha value), as defined in Eq. (1), where α_i is the opacity value.

$$I_i = \alpha_i F_i + (1-\alpha_i) B_i \tag{1}$$

In spectral matting, Eq. (1) is generalized by assuming that each pixel is a convex combination of K image layers $F^1 \sim F^K$, as defined in Eq. (2) [4], where α_i^k must satisfy the condition in Eq. (3). The K vectors α^k are the matting components of the given image, which specify the fractional contribution of each layer to the final color observed at each pixel.

$$I_i = \sum_{k=1}^{K} \alpha_i^k F_i^k \tag{2}$$

$$\sum_{k=1}^{K} \alpha_i^k = 1;\ \alpha_i^k \in [0,1] \tag{3}$$

Suppose that the given image consists of K distinct components $C_1 \sim C_K$ such that $C_i \cap C_j = \phi$ for $i \neq j$. Compute the eigenvectors of $N \times N$ Laplacian matrix L (matting Laplacian [3]) as $E = [e^1, ..., e^M]$, where E is a $N \times M$ matrix (N is the total number of pixels). The distinct components are obtained using spectral segmentation with the k-means algorithm based on the eigenvectors of matting Laplacian. Matting Laplacian is defined as a sum of matrices $L = \sum_q A_q$, each of which contains the affinities among pixels inside a local window w_q as defined in Eq. (4), where δ_{ij} is the Kronecker delta, μ_q is the 3×1 mean color vector in the window w_q around pixel q; $\sum_q$ is a 3×3 covariance matrix in the same window; $|w_q|$ is the number of pixels in the window; and $I_{3\times3}$ is the 3×3 identity matrix.

$$A_q(i,j) = \begin{cases} \delta_{ij} - \dfrac{1}{|w_q|}\left(1 + (I_i - \mu_q)^T \left(\sum_q + \dfrac{\varepsilon}{|w_q|} I_{3\times3}\right)^{-1} (I_j - \mu_q)\right), & (i,j) \in w_q \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

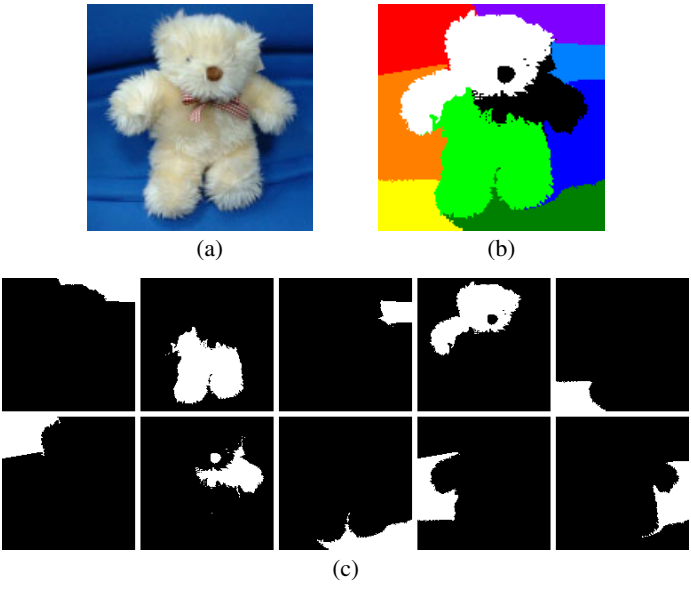


Fig. 1. Distinct component detection: (a) Given image; (b) clustering result; (c) distinct components

Fig. 1 is the result of distinct component detection using spectral segmentation with the k-means algorithm, where Fig. 1(a) is a given image, Fig. 1(b) is the clustering result, and Fig. 1(c) is the result of distinct components.

Next, the corresponding matting components are obtained using a linear transformation of the smallest eigenvectors $\tilde{E} = [e^1,...,e^{\tilde{M}}]$ of the matting Laplacian matrix. Initialize α^k by applying a k-means algorithm on the smallest eigenvectors, and project the indicator vectors of the resulting components $C_1 \sim C_K$ on the span of the eigenvectors $\tilde{E}$ using Eq. (5), where m^C denotes the indicator vector of the component C as defined in Eq. (6).

$$\alpha^k = \tilde{E}\tilde{E}^T m^{C^k} \tag{5}$$

$$m_i^C = \begin{cases} 1 & , i \in C \\ 0 & , i \notin C \end{cases} \tag{6}$$

Compute matting components by minimizing an energy function defined as Eq. (7) subject to $\sum_k \alpha_i^k = 1$ to find a set of K linear combination vectors y^k. The above energy function is optimally minimized using Newton's method, where γ is chosen to be 0.9 for a robust measure. Fig. 2 is the result of matting components.

$$\sum_{i,k} \left|\alpha_i^k\right|^\gamma + \left|1-\alpha_i^k\right|^\gamma, \text{ where } \alpha^k = \tilde{E}y^k \tag{7}$$

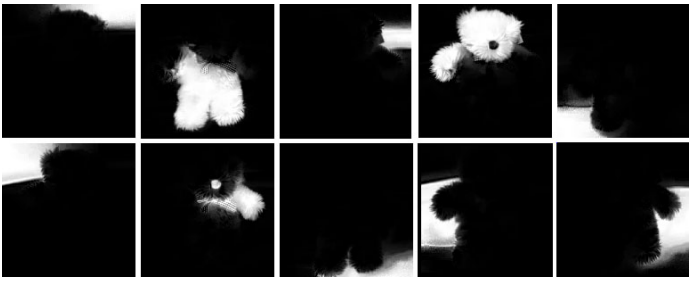


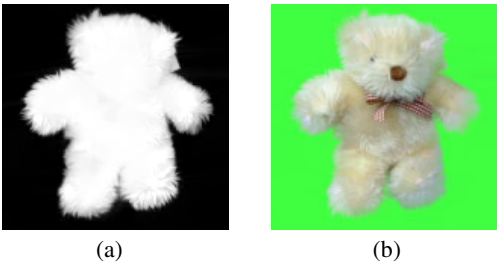
Fig. 2. Matting components

Finally, the matting components are combined to form the complete alpha matte by minimizing the matte cost, as defined in Eq. (8). In order to increase efficiency of minimizing the matte cost, the correlations between these matting components via L are pre-computed and store them in a $K \times K$ matrix Φ, as defined in Eq. (9). Then, the matting cost is computed using Eq. (10), where b is a K-dimensional binary vector indicating the selected matting components. Fig. 3 is the result of image matting, where Fig. 3(a) is the alpha matte and Fig. 3(b) is foreground extraction with a constant-color background.

$$J(\alpha) = \alpha^T L \alpha \tag{8}$$

$$\Phi(k,l) = \alpha^{k^T} L \alpha^l \tag{9}$$

$$J(\alpha) = b^T \Phi b \tag{10}$$



(a) (b)

Fig. 3. Result of spectral matting: (a) Alpha matte; (b) extracted foreground with a constant-color background

3 Spectral Matting Based on Hue Difference of Components

In this paper, spectral matting based on hue difference of components is proposed to obtain automatic and high accuracy alpha matte. The proposed method can overcome the drawbacks of spectral matting [4], spectral matting based on color information of matting components [6], and spectral matting based on the palette-based component classification [10]. A flow diagram of the proposed method is shown in Fig. 4.

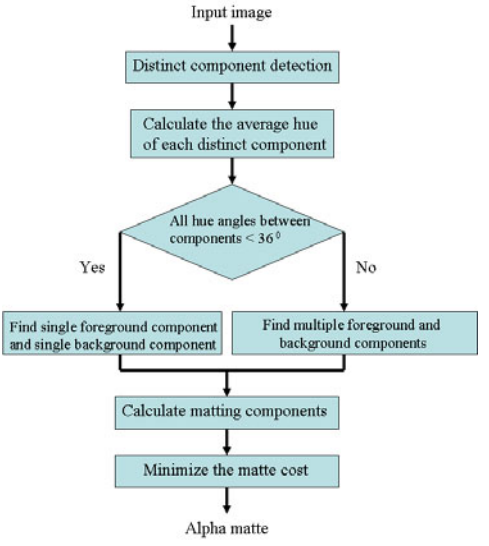


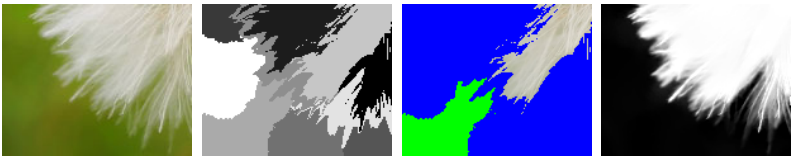
Fig. 4. Flow diagram of the proposed method

The distinct components are firstly obtained using spectral segmentation with the k-means algorithm based on the eigenvectors of matting Laplacian. Next, the average hue of each distinct component is calculated in HSV color space. If all hue angles between components are smaller than 36^0, then single foreground component and single background component are found; otherwise, multiple foreground and background components are found.

For the case of single foreground component and single background component, the background component and foreground component are obtained using Eq. (11) and Eq. (12), respectively, where $C(i)$ is the ith component; I_{boundary} is the image with boundary pixels (with a width of one pixel) of the given image; $C^H(i)$ is the hue angle of the ith component; C_B and C_F are the background component and foreground component, respectively. Next, the corresponding matting components of the foreground, background, and unknown components are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian. Finally, only matting components of the foreground and unknown components are combined to form the complete alpha matte based on minimizing the matte cost. Fig. 5 is the result of image matting, where Fig. 5(a) is a flower image; Fig. 5(b) is clustering result; Fig. 5(c) is the result of component classification; Fig. 5(d) is the alpha matte. In Fig. 5(c), green color is the selected background component; blue color is the unknown components; the other is the selected foreground component.

$$C_B = \arg\max_{i \in k} \{C(i) \cap I_{\text{boundary}}\} \tag{11}$$

$$C_F = \arg\max_{i \in k} \{C_B^H - C^H(i)\} \tag{12}$$



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

Fig. 5. Image matting of a flower image: (a) Original image; (b) clustering result; (c) component classification; (d) alpha matte

For the case of multiple foreground and background components, the adjacent components are firstly merged using the following algorithm, where $C^S(i)$ and $C^V(i)$ are the saturation and intensity of the *i*th component, respectively. Threshold $T_{C(i)}$ is calculated using Eq. (13) [7], where $V_{C(i)}$ is the average intensity of the *i*th component and $\alpha = 4$ is set from experience.

$$T_{C(i)} = \frac{1}{1+\alpha \times V_{C(i)}}, \alpha \in [1,4] \tag{13}$$

Algorithm of component merging

```
If (|C^H(i) − C^H(j)| ≤ 18^0)
    If (C^S(i) ≥ T_C(i)) & (C^S(j) ≥ T_C(j))
        C(i) ∪ C(j)
    else
        If (|C^V(i) − C^V(j)| ≤ 0.1)
            C(i) ∪ C(j)
        else
            Do nothing
else
    Do nothing
```

When the adjacent components have been merged, the background component and foreground component are found using Eq. (11) and Eq. (12), respectively. Next, the foreground component and nonadjacent components are merged using the algorithm of component merging. The background component is processed with the same procedure. Then, the corresponding matting components of the foreground, background, and unknown components are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian. Finally, only matting components of the foreground and unknown components are combined to form the complete alpha matte based on minimizing the matte cost. Fig. 6 is the result of image matting, where Fig. 6(a) is a woman image; Fig. 6(b) is clustering result; Fig. 6(c) is the result of

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

Fig. 6. Image matting of a woman image: (a) Original image; (b) clustering result; (c) component classification; (d) alpha matte

component classification; Fig. 6(d) is the alpha matte. In Fig. 6(c), green color is the selected background component; blue color is the unknown components; the other is the selected foreground component.

4 Experimental Results

The experimental results show that the proposed method performs well. The algorithms were implemented in Matlab R2008b. The number K of clusters using k-means algorithm is set as 10. The number M of the smallest eigenvectors in finding distinct components is set as 20. The number $\tilde{M}$ of the smallest eigenvectors in finding matting components is set as 50.

Five tested images were used to evaluate the performance of the spectral matting [4], spectral matting based on color information of matting components [6], spectral matting based on the palette-based component classification [10] and proposed method. The first column of Fig. 7 are bear image [11], face image [4], flower image [3], pharos image [3], and Kim image [4], respectively. The 2nd- 5th columns of Fig. 7 are the alpha mattes using spectral matting, spectral matting based on color information of matting components, spectral matting based on the palette-based component classification, and proposed method, respectively.

Table 1. Performance evaluation of image matting

Tested image	Spectral matting [4]	Wang and Li's method [6]	Hu et al.'s method [10]	Our method
Bear	0	0	0	0
Face	0	0.1436	0	0
Flower	0	0	0.1108	0
pharos	0	0.2436	0	0
Kim	0.3711	0.2564	0.7616	0

$$MAE = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|\alpha(i,j) - \overline{\alpha}(i,j)\right| \tag{14}$$

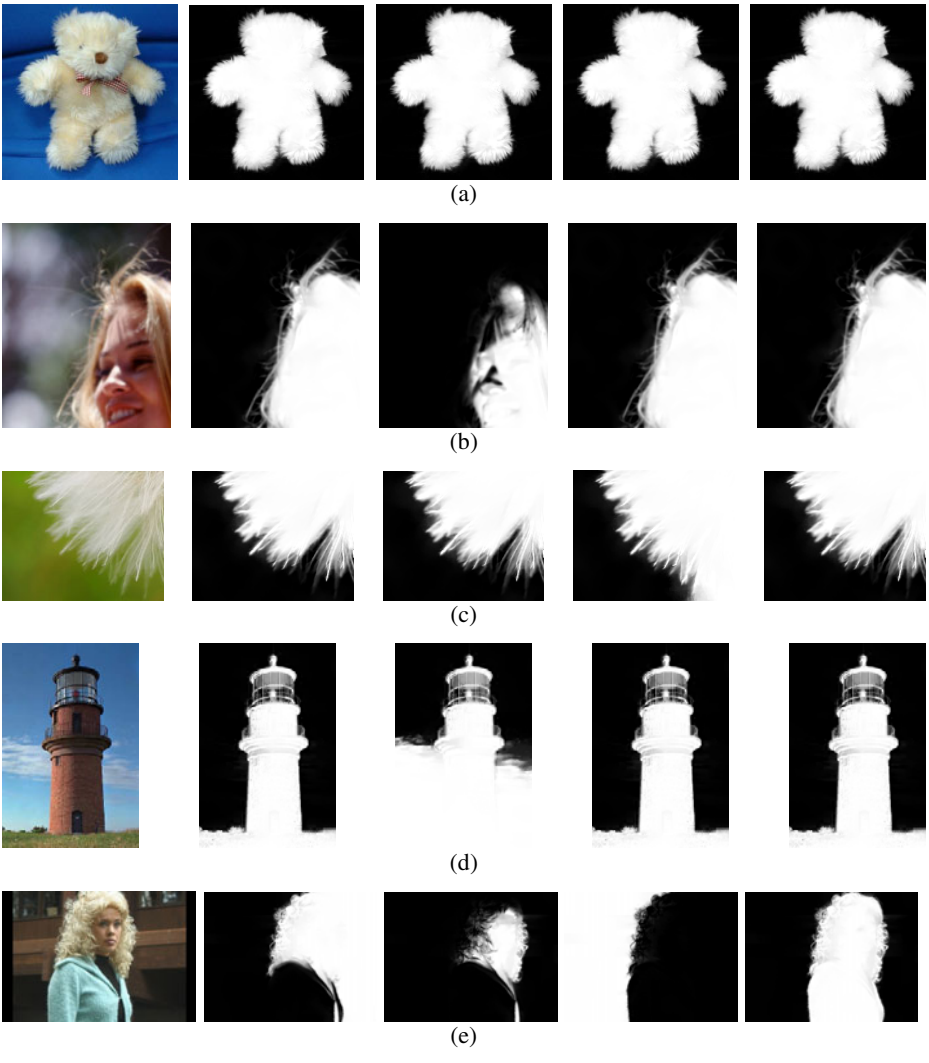


Fig. 7. Image matting of tested images with different methods

In order to show the performance of the proposed method, mean absolute error (MAE) is used to obtain performance evaluation and defined in Eq. (14). Table 1 is the performance evaluation of spectral matting, spectral matting based on color information of matting components, spectral matting based on the palette-based component classification, and proposed method, where the ground truths of the tested images are obtained using spectral matting with user intervention; $\alpha(i, j)$ and $\overline{\alpha}(i, j)$ are the alpha mattes using image matting and ground truth, respectively. Fig. 7 and Table 1 show that the proposed method outperforms the state-of-the-art methods based on spectral matting.

5 Conclusion

The spectral matting based on hue difference of components was proposed to obtain automatic and high accuracy alpha matte. In the proposed method, the distinct components are obtained using spectral segmentation with the k-means algorithm based on the eigenvectors of matting Laplacian. Component classification is then used based on the hue difference of components to obtain the foreground, background, and unknown components. The corresponding matting components of the foreground, background, and unknown components are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian. Finally, only matting components of the foreground and unknown components are combined to form the complete alpha matte based on minimizing the matte cost. Experimental results show that the proposed method can obtain the high-quality alpha matte for natural images without user intervention, and the proposed method has better performance than the state-of-the-art methods based on spectral matting.

Acknowledgments. This paper has been supported by the National Science Council, Taiwan, under grant no. NSC100-2221-E-346-008.

References

1. Wang, J., Cohen, M.F.: Optimized Color Sampling for Robust Matting. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 1–8 (2007)
2. Guan, Y., Chen, W., Liang, X., Ding, Z., Peng, Q.: Easy Matting: A Stroke based Approach for Continuous Image Matting. Computer Graphics Forum 25(3), 567–576 (2008)
3. Levin, A., Lischinski, D., Weiss, Y.: A Closed-Form Solution to Natural Image Matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), 228–242 (2008)
4. Levin, A., Rav-Acha, A., Lischinski, D.: Spectral Matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10), 1699–1712 (2008)
5. Yu, S.X., Shi, J.: Multiclass Spectral Clustering. In: Proceedings of International Conference on Computer Vision, pp. 313–319 (2003)
6. Wang, J.Z., Li, C.H.: Spectral Matting Based on Color Information of Matting Components. In: Luo, Q. (ed.) Advances in Wireless Networks and Information Systems. LNEE, vol. 72, pp. 119–130. Springer, Heidelberg (2010)
7. Sural, S., Qian, G., Pramanik, S.: Segmentation and Histogram Generation using the HSV Color Space for Image Retrieval. In: Proceedings of the 2002 International Conference on Image Processing, pp. 589–592 (2002)
8. Zhengjun, L., Shuwu, Z.: An Improved Image Retrieval Method Based on the Color Histogram. Control and Automation Publication Group 24(2-1), 246–247 (2008)
9. Jou, F.D., Fan, K.C., Chang, Y.L.: Efficient Matching of Large-size Histograms. Pattern Recognition Letters 25(3), 277–286 (2004)
10. Hu, W.C., Huang, D.Y., Yang, C.Y., Jhu, J.J., Lin, C.P.: Automatic and Accurate Image Matting. In: Proceedings of the 2nd International Conference on Computational Collective Intelligence–Technology and Applications, vol. 3, pp. 11–20 (2010)
11. Chang, I.C., Hsieh, C.J.: Image Forgery using Enhanced Bayesian-based Matting Algorithm. Intelligent Automation and Soft Computing 17(2), 269–281 (2011)

Towards Robotics Leadership: An Analysis of Leadership Characteristics and the Roles Robots Will Inherit in Future Human Society

Hooman Aghaebrahimi Samani*, Jeffrey Tzu Kwan Valino Koh**, Elham Saadatian***, and Doros Polydorou†

Keio-NUS CUTE Center
Interactive and Digital Media Institute,
National University of Singapore
hooman@lovotics.com
http://www.lovotics.com

Abstract. This paper aims to present the idea of robotics leadership. By investigating leadership definitions and identifying domains where humans have failed to lead, this paper proposes how robots can step in to fill various leadership positions. This is exemplified by referring to two examples, stock brokering and transportation, and explains how robots could be used instead. Furthermore, this paper aims to provoke discussion by identifying firstly some potential limitations of robots in leadership positions and secondly by proposing that our current technological ecosystem not only is suited for machines to assume leadership positions but rather is inherently headed towards it.

Keywords: Robotics Leadership, Lovotics.

1 Introduction

As the responsibilities of robots within our society takes on increasing importance, consideration of the role of robots is ever evolving. Much like the rights of man has been displaced by the rights of humankind, as well as the elevation of the rights of animals and all living things, robots are increasingly exerting their "right to exist" [17].

Their progress climbing the social ladder can be mirrored by major points in human history. When robots were initially created, the role of master and

* Interactive and Digital Media Institute, National University of Singapore.
** Graduate School for Integrative Sciences and Engineering, National University of Singapore.
*** Department of Electrical and Computer Engineering, National University of Singapore.
† Interaction and Entertainment Research Center, Nanyang Technological University.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 158–165, 2012.

slave was clear. Their invention was intended to assist and even replace the functionality of humans. In some sense, early robots could be perceived as slaves to their human masters.

As technology progresses, robots have evolved from their position of service providers to that of social entities. We see some robots take on the role of companion, even as therapeutic providers in some circumstances. In some cases these robots were no longer perceived as mere machines. In many ways, they have become our companions, much like pets are. Their perceived value to improving human life quality has elevated their own inherent "life worth".

We are currently seeing an emergence of robots inheriting more emotional positions in our lives. Since the advent of Lovotics[1], robots can now participate, albeit in a limited capacity, as emotional counterparts. The ability for robots to love and be loved by humans places in us a responsibility to consider and anticipate the future roles that robots will become responsible for in our society. One aspect of this could be the role of robots as managers of resources, and in this case, the management of human capital.

In this paper we discuss aspects of robot leadership. First we try to understand what it is to be a good human leader by looking to works that discuss ideal leadership qualities. By identifying these, we can attempt to define and outline some of these characteristics within the context of robot leadership. We then postulate what a perfect robot leader could be as well as look to specific domains where robotics leadership could be applied. We conclude with an analysis of possible benefits and drawbacks when using robots as leaders.

2 Definition of Leadership

What characteristics make a good leader continues to be an important question to address. There has been literally thousands of studies regarding the topic of leadership, and the term has been redefined and conceptualized in many different ways [2,19].

These studies share enough points to recommend a standard definition of leadership. According to 221 definitions of leadership by Joseph Rost [14], it could be recognized that leadership is about one person getting other people to do a task. Where the definitions differ is in how leaders motivate their followers and who has a say in the goals of the group or organization [10].

The same concept above is mirrored by other definitions. For instance Ciulla's defines leadership as "the ability to impress the will of the leader on those led and induce obedience, respect, loyalty, and cooperation... leadership is an influence relationship between leaders and followers who intend real changes that reflect their mutual purposes" [6].

Barker reviewed definitions of leadership that are currently used and conclude that leadership is about process and behavior [1].

Kotter defines leadership as "the process of motivating a group (or groups) of people in some direction through (mostly) no-coercive means" [11].

[1] www.lovotics.com

In another study Miller and Sardais [12] tried to propose a baseline conception of leadership by dividing the concepts defined in literature up to that time into two groups of *too broad* and *too narrow* definitions stated in the following.

From the broad definition perspective, "leadership is a process that influences behavior. This conception accounts for 40 percent of the 45 post-1940 leadership definitions that refer to "influence". Leadership is a process that influences the will of followers. Each of these definitions encompasses a wide variety of disparate situations. Influence may include persuasion, setting an example and coercion".

From the narrow definition perspective, "leadership is a process in which a leader changes the convictions of a group of people or an organization or a subordinate. Leadership can certainly have an impact on the mindset of a group, yet it can also concern and influence an individual. Also, it need not occur within a formal organization or between superior and subordinate. It may span organizational boundaries, originating with a subordinate that changes the convictions of a boss".

"Leadership is a process in which a leader exercises his or her power, authority or status . However, leadership also might involve the persuasion of peers or even superiors using argument or emotional, moral or factual appeals".

"Leadership exerts a positive moral force, articulates vision, instils meaning, embodies values, inspires to excellence, or actualizes goal achievement. Leadership is a process in which there is a conscious connection between leader and follower. Leadership may take place even when the follower cannot identify the leader. For example, some role-models change others simply by behaving in a certain way. Leadership is a role, a job, a calling, and is something continuously performed. Leadership can also be an act or an event" [12].

With the above definitions in mind, it can be concluded that they share a common characteristic: "Leadership occurs when someone imparts his or her convictions to another". With this in mind we may conclude for our own purposes that leadership is about managing in order to motivate and convince people to do a specific task.

Definitions of leadership often address the nature of the leader. For example B. Winston et. al. present an integrative definition of leadership that states: "A leader is one or more people who selects, equips, trains, and influences one or more follower(s) who have diverse gifts, abilities, and skills and focuses the follower(s) to the organization s mission and objectives causing the follower(s) to willingly and enthusiastically expend spiritual, emotional, and physical energy in a concerted coordinated effort to achieve the organizational mission and objectives" [18]. Finally, the standard definition of leadership assumes that "leadership is a relation between leaders and followers".

3 Human Failings and How Robots Could Do It Better

There can be no doubt that humans are fallible. More often then most, emotions can greatly influence the performance of a person tasked with a specific responsibility [16]. Through emotions, human beings can either perform for better or for

worse, and it can be assumed that human emotion influences decision-making. The issue with human decision-making is that the human would need to balance logic and emotion in order to make a rational yet empathetic decision. This is easier said then done.

In a robot leader, it can be argued that emotional states can enable a robot to make better decisions, yet it should be noted that some emotional states can be more beneficial then others. As emotional beings, humans feel a torrent and flood of emotions. The filtering of these emotions becomes critical, yet in some cases emotions get the better of people and bad decisions can occur.

Lovotic robots on the other-hand are in the end programmable. With this in mind, a robot leader can be instilled with adequate programming that could include certain emotional states that are beneficial to completing the task at hand, yet exclude emotions that would be detrimental. In this section we look at examples where human leaders have been compromised by their emotions and offer a *what if* perspective, speculating where robotic leaders could possibly perform better in the described situations.

3.1 Stock Brokering

Since the financial crisis in the late 2000s, researchers have discussed the attribution of the financial crisis to many factors. One of the more interesting and compelling ones is that of the contribution of human behavior to the development of the crisis [4]. Rogue trading and other people involved in highly risky activity are adversely effected by stress. Studies such as "How To Be A Rogue Trader" [8] show that stress affects risk aversion in the yellow-eyed junco sparrow. The rogue trader is no different.

What if a robotic trader was deployed in place of a human being to handle high-risk trading? Discipline, research and strategy are some of the key attributes that make a stock broker good, but who is to say that these attributes could not be instilled into a robot? As the previous study mentions, when high stress factors in the junco sparrow occur, the sparrow places itself into higher risk situations in order to get food. In a sense, it takes a gamble. Human traders, also effected by high stress will also make the same gambles [3]. A robot that has no feeling of stress, would not.

Robots as leaders that manage resources in situations such as stock brokering could maintain rational decision-making without the effects of stress. They could base their entire model for stock trading on information pulled form all over the Internet, from stock situations in other markets, to research on how weather effects commodity stocks in wheat and corn, to the effects of Christmas shopping on the chocolate market and beyond, all in real-time, and make an informed decision based on this.

3.2 Transportation

Another interesting area where robots could replace humans as leaders could be the avionics sectors. The Air France flight 447 was shown through data analysis

of the flight recorder (or "black box") that human error was at least partially attributed to the crash. This includes the transcript of the conversation between the pilot and cockpit staff, which paints a picture of confusion and chaos [2].

Would a robot do better in this situation? It is questionable [3], but much of the human error that occurred on Air France flight 447 could have been avoided. Whether human passengers would be willing to place their lives in a robot that is not necessarily concerned with its own self-preservation is a concern, but in the a future where Lovotics robots could be programmed with a level of empathy is possible. Cargo planes could be the first responsibilities for such robots, while the operation of passenger planes could follow.

With so many factors to consider, argument towards and against robots as leaders in these particular fields is up for debate and beyond the scope of this paper. The authors do think it is important to consider these factors when discussing such a subject, and welcome researchers to further explore these topics, as it will be crucial to the development of robots as leaders in the future.

4 Robots That Lead Robots

In order to allocate robots into positions of responsibility, it will be ideal to selectively choose and assign appropriate character traits. This way it can be ensured that the robot will perform the assigned tasks in the optimum way. By referring to to Plutchik's basic emotions and human feelings [13] we can identify that feelings like Aggression and Contempt are unwanted in most situations. These feelings are both associated with the emotion of Anger, an emotion which can be safely left behind when programming a robot aimed to interact in a social environment.

A successful leader needs to develop a relationship with their subordinates and then attempt to influence their will by persuasion, coercion or by setting an example. In order for the robot to be influenced by a leader, they need to have a basic understanding of success and failure, a rating scale of where they rank and the will to rise in those ranks.

By referring to the example mentioned previously, the robot assigned to handle high-risk trading should have not only have character traits which will ensure that it will constantly remain analytical and disciplined, but it should also be instilled with a sense of disappointment and shame towards other robot associates. This will promote a sense of competitiveness which aims to make the robot more resourceful in successfully accomplishing its tasks ahead of the rest of its co-workers. Disappointment and shame can sometimes produce negative consequences in humans like for example, stress and depression. Robots however, since they would not be programmed to understand those emotions would remain largely unaffected.

[2] What Really Happened Aboard Air France 447, `http://www.popularmechanics.com/technology/aviation/crashes/what-really-happened-aboard-air-france-447-6611877-2`

[3] Would You Fly on an Airplane With No Pilot? `http://www.freakonomics.com/2006/12/04/would-you-fly-on-an-airplane-with-no-pilot`

5 Against Robots and Artificial Intelligence

Hubert Dreyfus, an American philosopher and currently a professor of philosophy at the University of Berkeley has constructed a rather harsh critique of artificial intelligence, arguing that computers will never be able to replace humans or live amongst humans as equals [7]. Dreyfus thinks that robots will never be able to understand the world, as it is "organized by embodied beings like us, to be coped with by beings like us". Dreyfus continues and says that in order for the robot to not get completely lost in the space, it needs to be able to gain experiences with each action it performs, like a normal human body. In order for this to happen, AI researchers need to replicate and instill inside the robot a model of the world and a model of the body in order for the associations to be made. Dreyfus says that this so far this has been proved to be unachievable, and without it the world is just utterly un-graspable by computers in the same sense as their human counterparts.

A second claim against the possibility of robot leadership is the limitation of creativity. One necessary ingredient for creativity is the ability to think critically. Goldenburg in his *book* "Creativity in Product Innovation" [9] claims that suspending criticism and thinking that any idea is possible or good may ultimately be destructive to creativity. Humans have the ability to criticize themselves, where as computers cannot. Even though machines can write music and poetry [5] it is eventually up to humans to decide whether the work is of any worth. Will robots be able to think creatively? As creative thinking is considered to be an essential part of leadership it is definitely an interesting topic for further discussion.

6 Critical Thinking, When Does a Robot Stops Being a Tool and Starts Becoming a Leader?

Robots, when taken in the context of programmable machines created to perform specific tasks, have been servicing humans for close to 80 years. Humans have become completely reliant on machines and in more situations than not, they do not understand the complexity of most of their computational requests. By referring to a a simple example, a user searching for the quickest route between two locations on a GPS system, even though the process of actually displaying the information on a screen is not as simple as it appears, the complexity is undertaken by an elaborate artificial intelligence system that runs in the background. This AI system needs to make specific choices and decide on certain actions that are guided solely by the purpose of accomplishing the task it was created to do. The user trusts and blindly accepts what the system has shown, as the computer in this situation is employed simply to perform a given task, making this system a robot laborer.

Going a step forward, if there is more than one route to the required destination, the user could tell the system to choose the quickest one. Even though this gives even more decision making to the system, it still acts as a labour.

Will there be a time when machines are no longer acting as labourers and are therefore actually guiding humans? What if the GPS machine says, "No, I will not show you the fastest way to the grocery store, your fridge is currently full. I will however suggest you visit the barber, your last haircut was 2 months ago". Robot technologies are becoming intelligent enough to become part of a person's every day routine by being instructed to monitor and suggest actions. By programming gadgets that have the ability to perform a number of tasks, we give them the ability to choose for us. Through this exhibition of absolute trust, we are already setting down the ground-work to create robots that can act as our leaders.

Robots have already shown to have a number of advantages over humans [15] making them ideal for assuming leadership positions. Even though imagining now the possibility of replacing our current leaders with robots might sound absurd, by observing the current technological trends, the way technology is penetrating into our daily lives and our open acceptance to the change it affords, we could argue that giving robots positions of responsibility is not only unavoidable but is rather something desired and that we are trying to achieve.

7 Conclusion

In this paper we presented the idea of robots as leaders in the broadest sense. We have presented works by researchers that include theories of human leadership and have identified possible attributes that could be inherited by our robotic counterparts. By presenting instances where human failing could be fixed by possible robotic alternatives, we proposed a future where robots would be elevated in our society to function in roles beyond that of mere service entities, but actual allocators of resources and influencers of people. Finally, we briefly discussed the role of robot leaders and robot subordinates, as well as mentioned the possible failings of robots as leaders, and what makes them leaders as opposed to tools. It is in the hopes of the authors that discussions regarding the above topics takes center stage, as the issues raised will in no doubt become an eventuality. How prepared the human race will be when faced with future challenges regarding robot leadership remains to be seen.

Acknowledgement. This research is carried out under CUTE Project No. WBS R-705-000-100-279 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

References

1. Barker, R.A.: The nature of leadership. Human Relations 54(4), 469–494 (2001)
2. Bass, B.M., Bass, R., Bass, R.R.: The Bass handbook of leadership: Theory, research, and managerial applications. Free Pr. (2008)

3. Bechara, A., Damasio, H., Damasio, A.R.: Emotion, decision making and the orbitofrontal cortex. Cerebral Cortex 10(3), 295–307 (2000)
4. Brooks, D.: The behavioral revolution, 23 (2008)
5. Cheok, A.D., Mustafa, A.R., Fernando, O.N.N., Barthoff, A.K., Wijesena, J.P., Tosa, N.: Blogwall: displaying artistic and poetic messages on public displays via sms. In: Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 483–486. ACM (2007)
6. Ciulla, J.B.: Ethics and leadership effectiveness. In: The Nature of Leadership, pp. 302–327 (2004)
7. Dreyfus, H.L., Dreyfus, S.E., Athanasiou, T.: Mind over machine. Free Press (2000)
8. Gapper, J.: How To Be A Rogue Trader (A Penguin Special). Penguin (2011)
9. Goldenberg, J., Mazursky, D.: Creativity in product innovation. Cambridge Univ. Pr. (2002)
10. Kort, E.D.: What, after all, is leadership? 'leadership' and plural action. The Leadership Quarterly 19(4), 409–425 (2008)
11. Kotter, J.P.: The leadership factor. McKinsey Quarterly (2), 71–78 (1988)
12. Miller, D., Sardais, C.: A concept of leadership for strategic organization. Strategic Organization 9(2), 174–183 (2011)
13. Plutchik, R.: The emotions. Univ. Pr. of Amer. (1991)
14. Rost, J.C.: Leadership for the twenty-first century. Praeger Publishers (1993)
15. Samani, H.A., Cheok, A.D.: From human-robot relationship to robot-based leadership. In: 2011 4th International Conference on Human System Interactions (HSI), pp. 178–181. IEEE (2011)
16. Schacht, A., Dimigen, O., Sommer, W.: Emotions in cognitive conflicts are not aversive but are task specific. Cognitive, Affective, & Behavioral Neuroscience 10(3), 349–356 (2010)
17. Singer, P.: Animal liberation. Vintage (1995)
18. Winston, B.E., Patterson, K.: An integrative definition of leadership. International Journal of Leadership Studies 1(2), 6–66 (2006)
19. Yukl, G.: Leadership in organizations (2002)

Understanding Information Propagation on Online Social Tagging Systems: A Case Study on Flickr

Meinu Quan, Xuan Hau Pham, Jason J. Jung*, and Dosam Hwang

Department of Computer Engineering
Yeungnam University
Dae-Dong, Gyeongsan, Korea 712-749
{jmeenyu,pxhauqbu,j2jung,dosamhwang}@gmail.com

Abstract. Social media have been one of the most popular online communication channels to share information among users. It means the users can give (and take) cognitive influence to (and from) the others. Thus, it is important for many applications to understand how the information can be propagated. In this paper, we focus on social tagging systems where users can easily exchange tags with each other. To conduct experimentation, a tag search system has been implemented to collect a dataset from Flickr.

Keywords: Social tagging, Information propagation.

1 Introduction

Information (or knowledge) can be propagated and disseminated by interactions among people through various communication media. Depending on many internalization processes (e.g., enriching background knowledge and psychological modification), the people can somehow understand the information, and they will exploit it for their tasks in various contexts. Thus, it is important to discover meaningful patterns from this social phenomena of information propagation via certain media [1]. Various patterns of information propagation have been studies in many scientific areas, e.g., information science, cognitive science, neuroscience, and so on.

Especially, online social media have been playing an important role of propagating information to many users [2]. There have been a number of social network services (SNS), e.g., Twitter and FaceBook, to distribute up-to-date information quickly and widely. Various social media analysis schemes have been studied for understanding information flow, identifying experts, and topical authorities [3].

In this paper, we focus on online social tagging systems (also, called folksonomies) which is one of the well known online collective intelligence applications. Online users can represent various resources (e.g., bookmarks, musics,

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 166–174, 2012.

and photos) as a set of keywords, and share them with other users. We can also find the information is propagated among the users through this social tagging system. For example, in Fig. 1, user MQ has added t_3 (Eoljjang[1]) which is a *neologism* (a newly generated term among young generations). User HX will have a chance to recognize the new term and learn the meaning of the term.

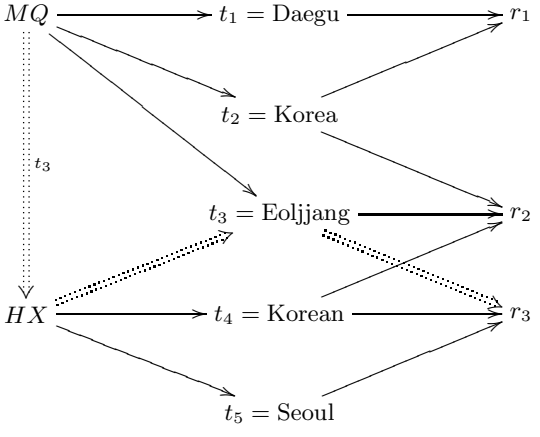


Fig. 1. A simple example on a social tagging system with two online users (i.e., MQ and HX); They have tagged three resources r_1, r_2, and r_3 with five tags t_1, t_2, t_3, t_4 and t_5

Since a certain information, e.g., neologisms, is generated within a social tagging system by some events and someones, various propagation patterns can be revealed. Also, we can regard this propagation process as a contextual synchronization among users [4]. Hence, we propose a novel method to extract such propagation patterns within a social tagging system. Thereby, temporal folksonomy and social pulse are formulated to understand the social propagation patterns.

The outline of this paper is as follows. In Sect. 2, we will describe backgrounds and basic notations. Sect. 3 introduces social pulses for modeling online social tagging systems. Sect. 4 shows an experimental results for evaluating the proposed methods. In Sect. 5, several discussion issues will be addressed. Finally, Sect. 6 draws our conclusions of this work.

2 Backgrounds and Notations

Social tagging process among multiple users can generate a specific information sphere, called a folksonomy, where information can be propagated.

[1] http://en.wikipedia.org/wiki/Eoljjang

Definition 1 (Folksonomy [5]). *Given a set of users U, their own tag sets T can be employed to describe their contexts about resources R. Thus, a folksonomy can be represented as*

$$\mathcal{F} = \langle U \times R \times T \rangle \tag{1}$$

where R indicates a set of resources tagged by the users.

Definition 2 (Temporal folksonomy). *As extended from a folksonomy, a temporal folksonomy is represented as*

$$\mathcal{F}^{\tau} = \langle U \times R \times T \times \tau \rangle \tag{2}$$

where τ is a timestamp of the corresponding tag.

For example, as shown in in Table 1, suppose that a temporal folksonomy has been built. Once we analyze the timestamp of tags, we can find out that t_3 has been propagated from MQ to HX after a certain time delay.

Table 1. An example of a temporal folksonomy

Users	Tags	Resources	Timestamps
MQ	t_2	r_2	2011/02/15 15:12:37 ($\tau^0_{t_2}$)
HX	t_5	r_3	2011/02/16 20:21:53 ($\tau^0_{t_5}$)
MQ	t_2	r_1	2011/02/20 09:12:53 ($\tau^1_{t_2}$)
HX	t_4	r_3	2011/02/22 22:34:41 ($\tau^0_{t_4}$)
HX	t_4	r_2	2011/02/23 13:43:34 ($\tau^1_{t_4}$)
MQ	t_3	r_2	*2011/03/01 21:12:23* ($\tau^0_{t_3}$)
HX	t_3	r_3	*2011/03/01 23:05:31* ($\tau^1_{t_3}$)
MQ	t_1	r_1	2011/03/04 22:54:21 ($\tau^0_{t_1}$)

3 Discovering and Understanding Social Pulses

Once a set of tags of each user is projected from a temporal folksonomy, it can be easily interpreted as a discrete signal which is a sequence of events over time. More importantly, we can set a time window (of which size is δ), since we need to understand how the tags is propagated in the folksonomy. Thus, given a temporal folksonomy $\mathcal{F}_{\tau}$, we can discover a *social pulse*, which is composed of co-occurred tagging patterns within a time window.

Definition 3 (Social pulse). *Given a tag $t \in T$, a social pulse is composed of two parts; a time window $w_{\tau_t} = [\tau_t, \tau_t + \delta]$, and a pitch (a height of the pulse). Especially, the pitch can be represented as either the number of users who have applied the same tag or the number of resources which have been attached by the tag. Thus, it can be formalized by*

$$\mathcal{P}_t = \{\langle w_{\tau_t}, \pi_{\tau_t} \rangle | \tau_t \in \tau\} \tag{3}$$

where the pitch of a time window w_{τ_t} can be computed by

$$\pi_{\tau_t} = |\{u_i | u_i \times t_j \times r_k \times \tau_{t_j} \in \mathcal{F}_\tau, t_j = t, \tau_{t_j} \in [\tau_t + \delta]\}| \quad (4)$$

$$= |\{r_k | u_i \times t_j \times r_k \times \tau_{t_j} \in \mathcal{F}_\tau, t_j = t, \tau_{t_j} \in [\tau_t + \delta]\}| \quad (5)$$

where $u_i \in U$ and $r_k \in R$. These two different measures (Equ. 4 and Equ. 5) are user-based pitch and resource-based pitch, respectively.

From Table 1, the social pulses can be represented like in Table 2. Equ. 4 is used to measuring the pitch. Basically, these social pulses are a sequence of discrete step functions, and they can be easily transformed into a cumulative form.

Table 2. An example on social pulses with Table 1; The size of time window δ is set to a day

Tag	Social pulse
t_2	$\{\langle[\tau^0_{t_2}, \tau^0_{t_2} + \delta], 1\rangle, \langle[\tau^1_{t_2}, \tau^1_{t_2} + \delta], 0\rangle\}$
t_5	$\{\langle[\tau^0_{t_5}, \tau^0_{t_5} + \delta], 1\rangle\}$
t_4	$\{\langle[\tau^0_{t_4}, \tau^0_{t_4} + \delta], 1\rangle, \langle[\tau^1_{t_4}, \tau^1_{t_4} + \delta], 0\rangle\}$
t_3	$\{\langle[\tau^0_{t_3}, \tau^0_{t_3} + \delta], 2\rangle, \langle[\tau^1_{t_3}, \tau^1_{t_3} + \delta], 0\rangle\}$
t_1	$\{\langle[\tau^0_{t_1}, \tau^0_{t_1} + \delta], 1\rangle\}$

Since the social pulses of tags are established from a given temporal folksonomy, we can extract the following two propagation patterns for understudying the temporal folksonomy; *i*) speed, and *ii*) convergence rate.

3.1 Propagation Speed

We can comprehend the *speed* of tag propagation, i.e., how quickly a tag is propagated to other users in a certain time duration. Since a tag t has been firstly used at τ^0_t, we can measure the pitch of the social pulse π_{τ_t}.

Definition 4 (Speed). *Given a tag t and its social pulse $\mathcal{P}_t$, the speed of its propagation can be measured by*

$$\mathcal{S}_t = \max_{\mathcal{P}_t} \left(\frac{\pi_{\tau_t}}{\delta} \right) \quad (6)$$

where δ is a size of window.

The speed simply indicates the maximum propagation power of the corresponding tag. We can measure the speed, when the social pulse of the tag shows the highest pitch. In other words, if we can construct a cumulative curve of the social pulse, we can easily find the speech at the steepest slope.

The speed of tag propagation can be used for measuring spontaneous and prompt response on the the tag. For example, it is easy to understand most online neologisms as well as fun to use them. For example, in Table 2, tag t_3 shows the highest speed, since it can an easy term to understand.

3.2 Propagation Convergence Rate

Also, we can measure the convergence rate of the propagated tag, i.e., how quickly the tag is spread to most of users.

Definition 5 (Convergence rate). *Given a tag t and its social pulse $\mathcal{P}_t$, the convergence rate of its propagation can be measured by a temporal duration*

$$\mathcal{C}_t = |\tau_t^{\Omega} - \tau_t^0| \tag{7}$$

where τ_t^{Ω} is the time ending the social pulse (i.e., $\pi_{\tau_t^{\Omega}} = 0$). It means that there is no more meaningful social pulse after τ_t^{Ω}.

Similar to the propagation speed, convergence rate is also an important indicator for measuring propagation power. Moreover, we assume that convergence rate means how long it take for all users in the folksonomy to realize the tag (or its meaning).

3.3 Relationship between Tags

More interesting task is to discover relationships between tags in a given folksonomy. Most of the previous studies for folksonomy analysis focus mainly on co-occurrence patterns between tags. For example, if two tags are applied together to more resources (or by more users), then we can determine that the relationship between these tags are stronger than others.

Different from such co-occurrence pattern-based schemes, in this paper, we are interested in a new measurement between tags, called *inducibility*. We assume that as the tags in a folksonomy can be propagated to other users, they can encourage the users to apply some relevant tags (especially, to create some new tags, e.g., neologism), depending on various lingual practice of the online users [5].

Definition 6 (Inducibility). *Two given tags t_i and t_j, inducibility $\mathcal{I}_{t_i \to t_j}$ can be measured by the minimized summation of temporal delays between the corresponding social pulses (i.e., $\mathcal{P}_{t_i} = \{\langle w_{r_{t_i}}, \pi_{r_{t_i}} \rangle\}$, $\mathcal{P}_{t_j} = \{\langle w_{r_{t_j}}, \pi_{r_{t_j}} \rangle\}$). It is computed by the following equations*

$$\mathcal{I}_{t_i \to t_j} = \frac{1}{1 + \min \Delta_w(\mathcal{P}_{t_i}, \mathcal{P}_{t_j})} \tag{8}$$

$$= \frac{|\mathcal{P}_{t_i}|}{1 + \min \Delta_w(\mathcal{P}_{t_i}, \mathcal{P}_{t_j})} \tag{9}$$

$$\Delta_w = \sum_{\pi_{r_{t_i}} \geq \zeta, \pi_{r_{t_j}} \geq \zeta} |w_{r_{t_j}} - w_{r_{t_i}}| \tag{10}$$

where ζ is a threshold to remove the trivial social pulses. In Equ. 9, the denominator can normalize the value by choosing the maximum length of the social pulse.

Once we have social pulses from a given temporal folksonomy, we can compare a pair of social pulses for measuring inducibility between tags. Moreover, we take into account the directionality of the inducibility (i.e., $\mathcal{I}_{t_i \to t_j} \neq \mathcal{I}_{t_j \to t_i}$).

4 Experimental Results

In order to evaluate the proposed schemes, we have built a temporal folksonomy $\mathcal{F}^{\tau}$ from Flickr[2]. Particularly, for sampling tags, we have collected 55 Korean internet neologism (i.e., $|T| = 55$) from wikipedia[3]. With $\delta = 28$ days, Fig. 2 and Fig. 3 show temporal distributions of resources (i.e., photos, since Flickr is a social media for sharing photos) and users, respectively. From both cumulative plots, we can find out that there are various temporal distributions.

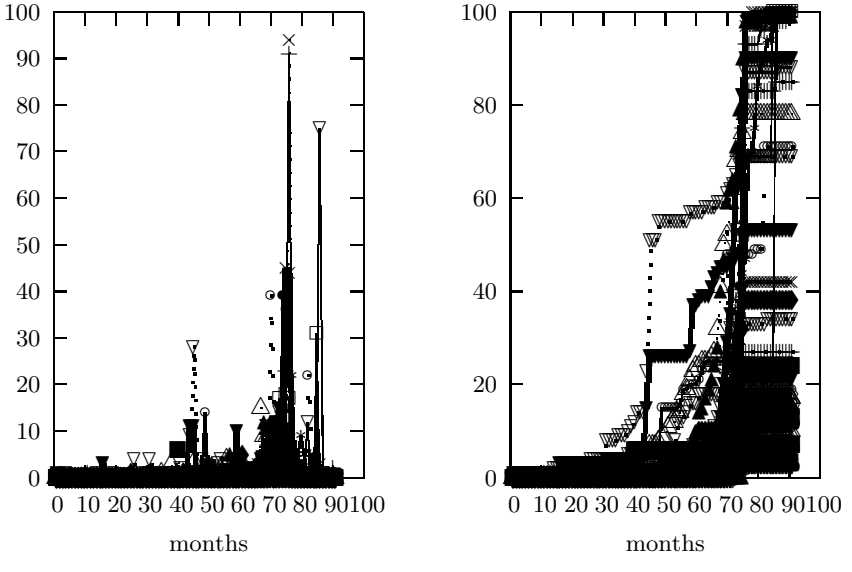


Fig. 2. Temporal distributions of tags by counting (a) a number of resources over time, and (b) a cumulative number of resources over time

Moreover, during analyzing the higher pitches, we have realized that some users have exploited a certain tag to many resources at the same time. It has become even more difficult to consider the propagation effects. Thus, we have decided that the number of users (i.e., Fig. 3) is the more important (precise) factor than the number of resources (i.e., Fig. 2) is.

After removing trivial tags, we have measured inducibility by pairing all possible combinations among the selected tags. Table 3 shows directed inducibility between two arbitrary tags in the form of an asymmetric matrix.

[2] http://flickr.com/
[3] http://ke.yu.ac.kr/wiki/index.php/OnlineNeologism

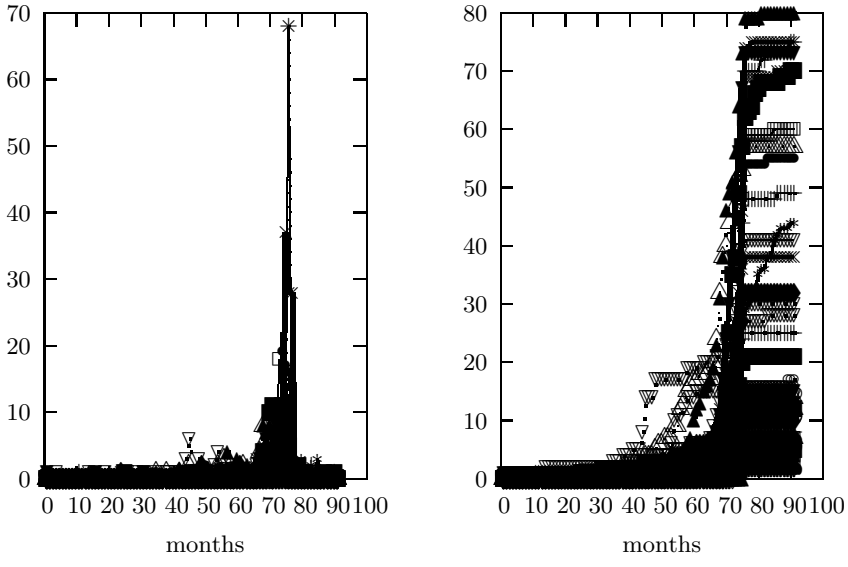


Fig. 3. Temporal distributions of tags by counting (a) a number of users over time, and (b) a cumulative number of users over time

5 Discussion

In this paper, we have focused on understanding how information is propagated via social media (particularly, folksonomy). Two heuristics (i.e., Equ. 4 and Equ. 5) have been designed to build the social pulses by simply counting the number of taggings.

Here, we want to list discuss several important issues on understanding the propagation patterns.

5.1 Pitch as Propagation Speed

Give a social pulse, pitch has been measured in a certain time moment. In this work, we have assumed that the pitch indicates the propagation speed of the corresponding tag. It means that with the higher pitch, there might be a certain event and news that more people have received. Thus, we can understand how an event can be influenced on online users (and social media).

As a research limitation, Flickr API is not providing a tagging timing but a photo uploading timing. We have regarded that both are same. Thus, there might be an error (e.g., temporal difference) if a user has added a new tag to some photos which have been already uploaded, and also modify the tags.

Table 3. Inducibility measurement among the sampled Korean tags by Equ. 8

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}
t_1	-	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0
t_2	0.25	-	0	0.5	0	0.5	0.5	0.5	0	0	0	0.5	0	0	1	1
t_3	0.17	0.33	-	0	0	0	0	0	0	0	0	0	0	0	0	0
t_4	1	0.33	0.5	-	0	0	1	1	0	0	0	0.5	0	0	1	0
t_5	0.5	0.1	0.08	0.5	-	0	0	0	0	0	0	0	0	0	0	0
t_6	1	1	0.33	1	0.5	-	1	1	0	0	0	0	0	0	1	0
t_7	0.5	0.33	0.2	0.5	0.13	0.33	-	1	0	0	0	0.5	0	0	1	1
t_8	1	1	0.33	1	0.5	1	0.33	-	0	0	0	0	0	0.5	1	0
t_9	0.06	0.08	0.09	0.5	0.05	0.33	0.07	0.13	-	0	0	0	0	0	0	0
t_{10}	0.5	0.1	0.08	0.5	1	0.5	0.13	0.5	0.05	-	0	0	0	0	0	0
t_{11}	1	0.11	0.09	1	0.5	1	0.14	1	0.05	0.5	-	0	0	0	0	1
t_{12}	0.5	0.25	0.5	1	1	0.5	0.17	0.5	0.13	1	0.5	-	0	0.08	0	0
t_{13}	0.5	0.2	0.14	0.5	0.17	0.5	0.33	0.5	0.06	0.17	0.2	0.17	-	0	0	0
t_{14}	0.17	0.33	1	0.5	0.08	0.33	0.2	0.5	0.11	0.08	0.09	0.5	0.14	-	0	0
t_{15}	1	0.5	0.5	1	0.5	1	0.25	1	0.5	0.5	1	0.5	0.2	0.5	-	0.2
t_{16}	1	0.5	0.5	1	0.5	1	0.33	1	0.25	0.5	1	0.5	1	0.5	1	-

5.2 Discovering Various Relationships from a Folksonomy

By collecting a set of social pulses, we can measure various relationships (e.g., similarity and distance) [5]. In this work, we have proposed a new measurement, called "inducibility," indicating a temporal closeness. Table 3 has been computed by Equ. 8. More importantly, this measurement is directive, meaning that $\mathcal{I}_{t_i \to t_j} \neq \mathcal{I}_{t_j \to t_i}$.

Again, due to the drawback by Flickr API, we have some problem on obtaining the correct time and computing the inducibility.

5.3 Potential Applications

Since most of collaborative workplaces, e.g., virtual enterprises, have employed online social media, there should be an efficient platform and methodology for monitoring the information propagation within a given workplace. Thus, we are expecting there will be a number of potential domains that the proposed work can be applied, as follows.

- detecting social events (or measuring propagation patterns)
- knowledge management system in various organizations

6 Conclusion and Future Work

Understanding how the information can be propagated is an important task on many applications using online social media. Also, in terms of shared understanding (e.g., contextual synchronization [4]) among multiple users, it is a critical topic how online users are cognitively responding.

In this paper, we have formulated social pulses by taking into account the number of tagging actions on folksonomies. They are similar to a set of discrete

time series dataset where we can extract several main features, e.g., pitches and frequencies. Most importantly, we have proposed an inducibility relationship between two social pulses (in fact, two corresponding tags).

Acknowledgments. This work was supported by the Korea Ministry of Knowledge Economy (MKE) under Grant No.2009-S-034-01.

References

1. Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the flickr social network. In: Quemada, J., León, G., Maarek, Y.S., Nejdl, W. (eds.) Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, pp. 721–730. ACM (2009)
2. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) Proceedings of The 10th IEEE International Conference on Data Mining (ICDM 2010), Sydney, Australia, December 14-17, pp. 599–608. IEEE Computer Society (2010)
3. Pal, A., Counts, S.: Identifying topical authorities in microblogs. In: King, I., Nejdl, W., Li, H. (eds.) Proceedings of the 4th International Conference on Web Search and Web Data Mining (WSDM 2011), Hong Kong, China, February 9-12, pp. 45–54. ACM (2011)
4. Jung, J.J.: Boosting social collaborations based on contextual synchronization: An empirical study. Expert Systems with Applications 38(5), 4809–4815 (2011)
5. Jung, J.J.: Discovering community of lingual practice for matching multilingual tags from folksonomies. Computer Journal (2012), doi:10.1093/comjnl/bxr102

Why People Share Information in Social Network Sites? Integrating with Uses and Gratification and Social Identity Theories

Namho Chung[1], Chulmo Koo[2,*], and Seung-Bae Park[3]

[1] College of Hotel & Tourism Management, Kyung Hee University
Seoul 130-701, Republic of Korea
nhchung@khu.ac.kr
[2] College of Business, Chosun University
501-759 Gwangju, Republic of Korea
helmetgu@gmail.com
[3]School of Business Administration
Sungkyunkwan University
110-745, Republic of Korea
sbpark@skku.edu

Abstract. Social network sites (SNSs) have been drastically increasing in use in recent years and can be cited as a communication tool with diversified forms in comparison with existing media. People are conducting various activities including relaxing entertainments, information sharing, escapism, social interaction, habitual pass time and others through the use of the SNSs. However, people have various motivation when using SNSs, with information sharing drawning a lot of organizational attention. Therefore, this study aims to ascertain motivational factors of SNSs that influence information sharing and conduct empirical analysis based on use and gratification theories and social identity theory. Factors influencing motivation of use of SNSs were divided into self-expression, involvement, interaction and media structure, and analysis was conducted to determine effects on continuance SNSs motivation. Our analysis results show that involvement had the greatest effect on continuance motivation and that remaining factors were also significant. In addition, continuance motivation turned out to have a significant effect on information sharing.

Keywords: Social network sites, continuance motivation, Information sharing, uses and gratification theory, social identity theory.

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 175–184, 2012.

1 Introduction

Social network sites (SNSs) whose users have been drastically increasing in recent years can be cited as a communication tool with diversified forms in comparison with existing media. For example, users of Facebook, a representative SNS tool, can broadcast messages to large audiences through the use of chat, status update or wall posts. Such a variety of diverse methods of communication of SNSs have enable users to have diversified motivational factors (Smock et al., 2011). People use SNSs mostly for the purpose of relaxing entertainment, information sharing, escapism, social interaction and habitual pass time. It is noteworthy that 'information sharing' is cited as one of the reasons for using SNSs. It is due to the fact that many organizations have recently introduced knowledge management (KM) for the purpose of creating high quality knowledge, utilizing knowledge for competitive edge and improving organizational learning and innovation but that the original purpose of KM cannot be fulfilled with an information technology-oriented repository or a system oriented approach (Jeon et al., 2011). Accordingly, if SNSs can be proactively used within an organization, it could make a great contribution to achieving the purposes of KM based on information sharing (Kim et al., 2011). Existing studies on SNSs have focused on examining the types of motivation for using SNSs (Smock et al., 2011), network externalities influencing purchase of products (Lin et al., 2011), factors affecting purchase of products, and risk taking and trust and privacy concerns in social network communities. Studies examining 'information sharing' role of SNSs are a rare to find. In this regard, this study has the following research purposes.

First, what factors influence continuance motivation of SNSs?
Second, how does SNSs' continuance motivation affect information sharing within SNSs?

These research results may make a great contribution to successfully realizing KM by inducing organizations to facilitate information sharing within an organization. In addition, these results are expected to provide marketing suggestions in terms of promotion of performances or exhibitions that require information sharing in addition to KM. In order to fulfill these purposes, the study adopts uses and gratifications (U&G) theory and social identity theory as a theoretical basis.

2 Conceptual Backgrounds

2.1 Uses and Gratification Theory

Uses and gratification (U&G) theory is one of the representative theories on the use of media and is focused on examining what users do through the use of media instead of trying to determine what effects media have on users. According to this theory, users intentionally select and consume specific media that can satisfy their desires based on a specific motivation instead of being preoccupied with passive use (Stafford et al., 2004). In prior studies on uses and gratification, 'use' means

selectively using media expected to satisfy users' needs and 'gratification' refers to the degree of satisfaction acquired in the process of using media. In addition, 'motivation' means stimulation and compensation that induces use of media.

According to existing studies on uses and gratification, users turned out to use media in order to satisfy such psychological desires as pursuit of information, relaxation and escapism. U&G Theory focuses on users and explains use of media in relation to human desires while emphasizing use of media based on psychological motivations.

Furthermore, the communication behavior of each person seems unique due to the different set of social and psychological behavioral factors that affect the gratification extent of a user's needs and interests when using media (Wu et al., 2011). Recently, Papacharissi and Mendelson (2011) used factor analysis to extract nine distinct scales of motives for using SNSs: habitual pass time, relaxing entertainment, expressive information sharing, escapism, cool and new trend, companionship, professional advancement, social interaction and meeting new people (Smock et al., 2011). The U&G Theory has been successfully applied to many new media and communication technologies. Hence, the U&G perspective can be a reasonable theory in explaining when users have been using SNSs rapidly at a time, which would be considered suitable applications for examining already identified factors (Papacharissi and Mendelson, 2011) influencing information sharing with regard to experience-drivenuse of SNSs.

2.2 Social Identity Theory

The U&G Theory explains media users' media selection and uses, but it does not provide a satisfying answer to why people conduct a specific behavior within media. Therefore, Social identity theory (Mael and Ashforth, 1992; Tajfel and Turner, 1986) is adopted to better explain this issue. Social identity theory describes the process of understanding individual identity within a specific group. An individual figures out his or her identity within a specific group by asking the question, "Who am I?", and then, the answer is simply noted like this; "I am a member of a specific group." Such a process of individual identification with a specific group can be applied to the process of using SNSs as it is. According to Social Identity Theory, identity perceived by an individual consists of personal identity embracing personal attributes (psychological characteristics, individual tendency, and ability) and social identity defining his or her own group. According to this theory, people tend to behave in a way that corresponds with their own identity and support an organization that can crystallize their identity.

In addition, the stronger the sense of belonging or membership to a particular group, the more likely for an individual to consider the organizational success as his or her own. Therefore they make efforts to cooperate with one another for the purpose of organizational development and success. If the social identity theory is combined with U&G Theory to explain information sharing within SNS, it can be said that people are trying to keep in touch by sharing information and staying in successful relationships with others conducting activities within SNSs. Then, self-expression to

disclose oneself is believed to represent personal identity. In addition, social identity takes the form of involvement and interactions within SNSs. In particular, this identity needs to be considered along with structure of pertinent SNSs and the breadth and depth of pertinent themes. In this regard, this study has focused on sorting through existing research to conclude that media users' self-expression (Trammell and Keshelashvili, 2005), involvement and interaction (Kim et al., 2011; Park and Chung, 2011), the breadth and depth of a topic (Park, 1996), and the media structure (Wu et al., 2011) of SNSs are all expected to have an impact on motivation. In addition, regarding continuance motivation of SNSs, research models and hypotheses were established based on the presumption that media users share information.

3 Research Model and Hypotheses

3.1 Research Model

Based on the preceding theoretical background and conceptual discussion, the proposed research model for the present study is shown in Figure 1. A research framework was established based on U&G Theory.

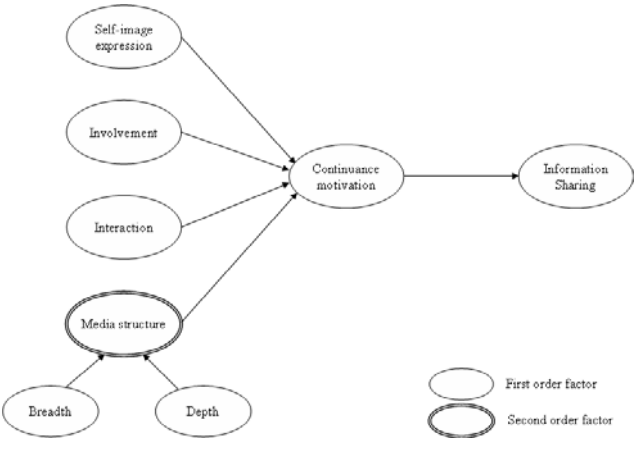


Fig. 1. Research Model

3.2 Hypotheses

Self-image expression means that people play a part in disclosing themselves (Trammell and Keshelashvili, 2005) and it is a constant process of controlling and managing information to continuously deliver one's specific image to others (Leary and Kowalski, 1990). In other words, self-image expression is referred to as expressing the image of one's self, and if possible to effectively express the image of one's self within an organization, media users would define their own identity with continuing efforts. If SNSs satisfy users' self-image expression, users would be able

to further enhance a sense of sharing, standards, tradition and a sense of responsibility, and it is considered that continuance motivation for pertinent SNS occurs. Such online communities as MMORPG (Massive Multiplayer Online Role-Playing Games) induce their members to proactively take part by offering chatting services or messengers that can satisfy members' desire for self-expression (Park and Chung, 2011). Hence, the hypothesis is proposed as follow:

Hypothesis 1: *Self-image expression positively affects media users' continuance motivation to use the SNSs.*

Involvement in the SNSs indicates the level of an individual's attachment and sense of belonging to particular SNSs (Kim et al., 2011). Users' strong involvement in SNSs will lead to enhanced attachment to SNSs. As such, members highly attached to SNSs would have an intention for continued participation while enhancing their affection and loyalty. In particular, if users are strongly involved in specific SNSs, they would develop interest in exchanging with others while getting further absorbed in pertinent SNSs. Accordingly, the stronger involvement in SNSs, the more continuance motivation for pertinent SNS. This is supported by the social identity theory which states that individual self-identity affects the group setting (Tajifel and Turner, 1986). Hence, the hypothesis is proposed as follow:

Hypothesis 2: *Involvement positively affects media users' continuance motivation to use the SNSs.*

In large social networks as SNSs, how well one is getting along and exchanging with others is very important, and it can be defined as interaction.

If members of an SNS proactively interact with one another; they further rely on one another while extending a helping hand, thus leading to enhanced ties among members (Kim et al., 2011). In particular, people with strong interactions tend to conveniently discuss matters with others and share their problems to facilitate social exchange within SNSs and thus increase motivation for continued use of SNS (Park and Chung, 2011). Hence, the hypothesis is proposed as follow:

Hypothesis 3: *Interaction positively affects media users' continuance motivation to use the SNSs.*

It is hard to find a clear definition of media structure, but Wu et al. (2011) defines it as the volume and character that media contents have. It is related to breadth and depth of contents covered by media (Park, 1996). If contents covered by SNSs were simple and shallow, people would not proactively participate in a pertinent SNS organization and thus lose motivation for continuous participation in SNSs. On the other hand, if they are able to have in-depth discussions on various topics, they would be absorbed in pertinent SNSs in order to exchange information with others and continuously participate in pertinent SNSs. Hence, the hypothesis is proposed as follow:

Hypothesis 4: *Media structure positively affects media users' continuance motivation to use the SNSs.*

The continuance motivation mentioned in this study is referred to as a belief and an attitude of sustained participation in particular SNSs based on experiences in the past (Wu et al., 2011). Those who have the determination to continuously take part in SNSs will have stronger affection for and loyalty to SNSs than others. According to social identity theory, the stronger sense of belonging to a particular group or organization (membership), the more likely that he or she regards organizational success as their own success, therefore these people tend to cooperate and endeavor for the purpose of organizational development and success. Accordingly, those with continuance motivation for SNSs are expected to share more information dissemination each other. Consequently, the hypothesis is proposed as follows:

Hypothesis 5: *Continuance motivation positively affects media users' information sharing.*

4 Research Methodology

The main objective of this study is to examine the factors affecting information sharing on SNSs and empirically verify the continuous motivation of user who are actively using SNSs, taking into consideration the effects of self-image expression, involvement, interaction, and media structure (breadth and depth). For this study, data were collected by survey method from users with SNSs usage experience. Respondents were invited from qualified college students who were promised by one of the authors to be given an incentive upon successful completion of the questionnaire survey. A total of 303 questionnaires were collected but only 264 of them were analyzed after excluding those that were incomplete. 137 of particpiants (51.9%) were men, 127 (48.1%) were women, and 248 (93.9%) were high school graduates. As for age of the respondents, 223 of them (84.5%) were under 30 years old. In addition, as for the period of Twitter use, 102 of them (42.5%) were less than 1 year, and 121 of them (45.8%) had used Facebook less than 2 years. The research variables were measured using five-point Likert-type scales. The measurement was based on the respondent's perception of the extent to which they agree or disagree with each item. The measures were adapted from previous research literature.

5 Data Analysis and Results

5.1 Measurement Model

As presented in Figure 2, the measurement model was assessed using the first order model or the second order model. Self-image expression, involvement, interaction, continuance motivation, and information sharing were all measured by the first order model. The remaining one inherent variable, i.e., media structure, was measured by sub constructs using a second order model.

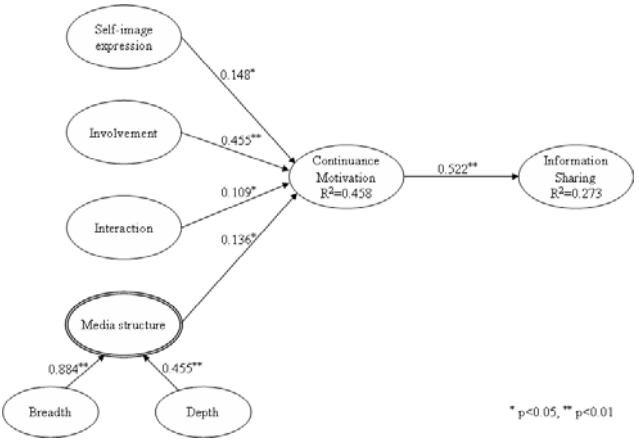


Fig. 2. The Estimated Structural Model

The sub constructs for media structure are breadth and depth. PLS (Partial Least Square) was used to test the measurement model and research hypotheses. PLS is a structural equation modeling approach used to test theoretical and measurement models composed of hierarchically structured variables. PLS is similar to LISREL in testing both theoretical and measurement models at the same time. For this reason, PLS-Graph 3.0 was used for data analysis.

Table 1. Reliability and validity of measures

Latent variables	Items	Loading	t-value	Cronbach' α	Composite Reliability	AVE
Self-image expression	Self1	0.848	37.548	0.913	0.891	0.671
	Self2	0.837	31.494			
	Self3	0.773	24.971			
	Self4	0.818	29.292			
Involvement	Involv1	0.883	55.080	0.932	0.910	0.717
	Involv2	0.845	37.312			
	Involv3	0.848	50.414			
	Involv4	0.810	29.668			
Interaction	Inter1	0.717	14.015	0.909	0.829	0.551
	Inter2	0.678	12.151			
	Inter3	0.848	31.667			
	Inter4	0.714	12.051			
Media Structure	Bread	0.898	12.326	0.731	0.656	0.512
	Depth	0.465	2.878			
Continuance Motivation	Motiv1	0.904	43.336	0.949	0.924	0.802
	Motiv2	0.935	72.002			
	Motiv3	0.918	45.981			
Information Sharing	Share1	0.931	87.192	0.960	0.949	0.862
	Share2	0.935	79.826			
	Share3	0.919	69.888			

Reliability and validity tests were conducted for each latent variable and construct. The coefficient alphas of research variables are indicated in Table 1. All scales exhibit sufficient reliability as they exceed or are close to the acceptable reliability coefficient of 0.7 (Reference). The composite reliability ranged from 0.656 to 0.949, indicating moderate to high reliability. This suggests that a significant portion of variance in the latent variable is explained by the variance of the measured variables. Construct validity is assessed using convergent and discriminate validity. Convergent validity tests if all the items measure a construct cluster, and thereby form a single construct. The average variance extracted (AVE) for all latent variables in this study exceeded or was slightly above 0.5 (Table 1) (Reference). Convergent validity can be investigated from the measurement model by determining whether the estimated parameters of each construct are significant. Table 1 and Table 2 indicate that all the estimated parameters of each construct are significant. The high values of reliability scores and the significant parameter estimates suggest the presence of convergent validity. Discriminate validity refers to the degree to which a latent variable differs from other latent variables. The intercorrelations among the latent variables do not exceed the square root of the average variance extracted. This indicates the discriminant validity of the measures. The factor scores of the sub constructs were used as indicators for the formative constructs. To check the validity of formative constructs, item weights which could be interpreted as β coefficients in a standard regression were examined (Park and Chung, 2011).

Table 2. Descriptive statistics and correlations

Construct	(1)	(2)	(3)	(4)	(5)	(6)
(1) Self-image expression	**0.819**					
(2) Involvement	0.696*	**0.847**				
(3) Interaction	0.627*	0.729*	**0.742**			
(4) Media structure	0.582*	0.664*	0.667*	**0.716**		
(5) Continuance motivation	0.623*	0.760*	0.611*	0.666*	**0.896**	
(6) Information sharing	0.650*	0.744*	0.653*	0.679*	0.767*	**0.928**

* $p<0.01$

Note: Diagonal elements in the "correlation of constructs" matrix are the square root of average variance extracted (AVE). For adequate discriminant validity, diagonal elements should be greater than corresponding off-diagonal elements.

5.2 Test of Structural Model

To evaluate the structural models' predictive power, we calculated the R^2s for continuance motivation and information sharing. Interpreted as multiple regression results, R^2 indicates the amount of variance explained by the exogenous variables. Using a bootstrapping technique, the path estimates and t-statistics were calculated for the hypothesized relationships. The results suggest that distinct antecedents influenced the formation of continuance motivation and information sharing. The result using PLS are shown in Figure 2.

The research hypotheses raised in previous sections are generally proven and the results are statistically significant. H_1, H_2, H_3, and H_4 show that continuance

motivation is significantly influenced by self-image expression (β = 0.148, t-value = 2.454, p < 0.05), involvement (β = 0.455, t-value = 7.963, p < 0.01), interaction (β = 0.109, t-value = 2.046, p < 0.05), and media structure (β = 0.136, t-value = 1.979, p < 0.05). Also, H_5 shows that continuance motivation is significantly influenced by trust transfer (β = 0.522, t-value = 10.232, p < 0.01). In Table 3, we present the resulting standardized parameter estimates and verdicts for hypotheses H_1 to H_5.

Table 3. Results of hypothesis testing

Hypothesis	Path	Coefficient	t- value	Results
H_1	Self-image expression → Continuance motivation	0.140	2.454	Supported
H_2	Involvement → Continuance motivation	0.455	7.963	Supported
H_3	Interaction → Continuance motivation	0.109	2.046	Supported
H_4	Media structure → Continuance motivation	0.136	1.975	Supported
H_5	Continuance motivation → Information sharing	0.522	10.232	Supported

6 Discussion and Conclusion

This study has found some interesting facts in terms of theory and practice. We integrate the uses and gratification theory with social identity theory in the context of SNSs use. First of all, the social identity theory has a significant effect on continuance motivation. This is a similar result as from the study by Kim et al. (2011) and it is also consistent with the previous studies (Kim and Chan, 2007; Park and Chung, 2011). We also found that users' involvement plays a very important role in SNSs. In particular, as it is represented as a more influential factor than self-image expression, interaction and media structure, a lot of thoughts need to be given to how to improve involvement for users' continuance motivation in SNSs. As seen in the study conducted by Park and Chung (2011), self-image expression has a significant effect on continuance motivation, and users more proactively participated and got immersed in SNSs by expressing themselves. In addition, interaction turned out to be a significant factor. As people became more loyal to and more immersed in a pertinent organization through relationships with other members, it was also cited as a significant factor. Last but not least, media structure was also cited as a significant factor, and it means that users' motivation for use is lowered if topics dealt with in SNSs are too simple or shallow. It means that it is necessary to create a venue for diversified topics and in-depth discussions in order to improve continuance motivation on the part of users. Secondly, in terms of uses and gratification theory, in particular, as continuance motivation has a significant effect on information sharing, it turned out to be important for users to stay for a sufficient long time in pertinent SNSs instead of leaving. As mentioned above, this study carries significance in that it has viewed the motivation for using SNSs from the perspectives of information sharing and created and empirically verified our model. For a practical aspect, it is also important for SNS media practitioners to consider how they can be able to engage in

their users or consumers, which interact with a means of expressing themselves with various practical topics in SNSs. Continually, they promote their usage of SNSs by sharing valuable information in each a group or an organization.

Acknowledgments. This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2011 (Grant No. 20111621).

References

1. Kim, H.Y., Zheng, J.R., Gupta, S.: Examining knowledge contribution from the perspective of an online identity in blogging communities. Computers in Human Behavior 27, 1760–1770 (2011)
2. Leary, M.R., Kowalski, R.M.: Impression Management: A Literature Review and Two-Component Model. Psychological Bulletin 107, 34–47 (1990)
3. Lin, K.Y., Lu, H.P.: Why people use social networking sites: An empirical study integrating network externalities and motivation theory. Computers in Human Behavior 27(6), 1152–1161 (2011)
4. Mael, F.B., Ashforth, E.: Alumni and Their Alma Mater: A Partial Test of The Reformulated Model of Organizational Identification. Journal of Organizational Behavior 13, 103–123 (1992)
5. Papacharissi, Z., Mendelson, A.: Toward a new(er) sociability: Uses, gratification and social capital on Facebook. In: Papathanassopoulos, S. (ed.) Media Perspectives for the 21st Century, pp. 212–230. Routledge, New York (2011)
6. Park, M.R.: Making Friends in Cyberspace. Journal of Communications 46, 80–97 (1996)
7. Park, S., Chung, N.: Mediating roles of self-presentation desire in online game community commitment and trust behavior of Massive Multiplayer Online Role-Playing Games. Computers in Human Behavior 27, 2372–2379 (2011)
8. Smock, A.D., Ellison, N.B., Lampe, C., Wohn, D.Y.: Facebook as a toolkit: A uses and gratification approach to unbundling feature use. Computers in Human Behavior 27, 2322–2329 (2011)
9. Stafford, T.F., Stafford, M.R., Schkade, L.I.: Determining Uses and Gratifications for the Internet. Decision Sciences 35, 259–288 (2004)
10. Tajifel, H., Turner, J.C.: The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (eds.) Psychology of Intergroup Relations, 2nd edn., pp. 7–24. Nelson-Hall, Chicago (1986)
11. Trammell, K.D., Keshelaslwli, A.: Examining the New Influence – A Self-Presentation Study of A-List Blogs. Journalism and Mass Communication Quarterly 82, 968–982 (2005)
12. Wu, J.H., Wang, S.C., Tsai, H.H.: Falling in love with online games: The uses and gratifications perspective. Computers in Human Behavior 26, 1862–1871 (2011)

The Impact of Data Environment and Profitability on Business Intelligence Adoption

Chien-wen Shen, Ping-Yu Hsu, and Yen-Ting Peng

Department of Business Administration, National Central University
No.300, Jhongda Rd., Jhongli City 32001, Taiwan
cwshen@ncu.edu.tw, pyhsu@mgt.ncu.edu.tw, amberpeng@hotmail.com

Abstract. The deployment of business intelligence (BI) involves complex processes of data reconfiguration and resource alignment. This study investigated whether the issues of data environment and profitability affect BI implementation for the manufacturers that have already adopted enterprise resource planning systems. We individually considered the factors of data warehousing, online analytical processing (OLAP), and data mining for the data environment, while return on assets, return on sales, and return on investment were transformed into a single component of profitability using principal component analysis. Through logistic regression, we determined that OLAP and data warehousing play important roles in the adoption of BI; however, data mining and profitability indicated no such influence.

Keywords: Business Intelligence, Data Environment, Profitability, Enterprise Resource Planning.

1 Introduction

Business intelligence (BI) is a set of technologies that can improve a firm's decision-making and work-flow through the acquisition and analysis of business data [12]. Because obtaining comprehensive information in a timely manner is critical to the development of new products and the improvement of business operations, BI also plays a central role in producing up-to-date information for operative and strategic decision-making [10]. The implementation of BI enables organizations to acquire, analyze, and disseminate information from internal and external sources in an organized and systematic manner [19]. In addition, BI applications provide tools that can be used throughout the organization to access, analyze and share information from a variety of data sources [3]. Managers involved in manufacturing typically implement BI tools to streamline inventory queries, response orders, and decision making. Hence, BI is considered an important IT investment due to its ability to enhance competitive advantage through the analysis of profitability, product or service usage, and marketing. According to a survey in Network Magazine [22], the total worldwide revenue of the five major BI vendors, which included SAP, Oracle, SAS, IBM and Microsoft, was US$73.63 billion in 2010, representing a 14.3% growth rate from 2009. Although the annual growth rate of the BI market is

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 185–193, 2012.

impressive, the adoption rate of BI for service organizations is only 43%, which is relatively low compared to the 94% adoption rate of enterprise resource planning (ERP) [26]. One of the reasons for the low BI adoption rate is that the benefits of BI are mostly non-financial or intangible [21]. A BI system must often compete with other IT projects for limited capital resources because managers tend to look for IT solutions offering the highest risk-adjusted return on investment. Hence, profitability plays an important role in the investment decisions related to BI projects. In addition, companies must deal with a variety of issues associated with decision process engineering, functional use, information usage, strategic alignment, technical readiness, and IT partnership to ensure the successful implementation of BI. When the readiness of the data environment is critical to the delivery of information and analytical applications, the deployment of BI typically involves complex processes of data reconfiguration and system alignment.

The main objective of this study is to investigate the impact of the data environment and profitability on the implementation of business intelligence, as both factors are important considerations in the purchase of BI. However, there remains a lack of research investigating this connection. Exploring the relationships between the data environment and the adoption of BI enables companies to prioritize the preparation of data infrastructure in capital-constrained circumstances, while assessing the correlation between profitability and BI implementation helps managers to evaluate the financial feasibility of BI projects. Thus, in Section 2, we begin with a discussion of general information environments associated with BI. In Section 3, we describe the methodology of principal component analysis and logistic regression analysis used in this study. Because BI is generally the next IT solution implemented after ERP, this empirical study evaluated how the data environment and profitability influence the adoption of BI systems by manufacturers with existing ERP systems. Results are summarized in Section 4. We conclude our findings and suggest potential future research topics in the final section.

2 Information Environment of BI

An integrated framework of BI elements commonly includes operational data, integrated data, data storage, BI software, and analytical applications [3]. The tier of operational data provides data that are retrieved from CRM applications, ERP applications, and online transaction processing (OLTP). The tier of integrated data consolidates, merges and stores enterprise-wide operational data from various transactional systems [3]. Meanwhile, the tier of data storage contains data warehouse (DW) and data mart to enable rapid, complex, ad hoc queries with drill-down capability. Generally speaking, constructing a BI system progresses through the extraction, transformation, and loading of data dispersed throughout various information systems into a data warehouse system [15]. Data warehousing comprises a shared data warehouse and a subject-oriented data mart for the management of support-oriented data [25]. The construction of data warehouses is commonly designed in a star schema, snowflake schema, or fact constellation. Data warehouses are sometimes called multidimensional databases because they present a

multidimensional, logical view of data. The implementation of data warehousing software is expected to grow rapidly as enterprises are forced to deal with increasing quantities of data, as a means to organize the data and apply it in the decision making process [20]. Effective data warehouse systems enable companies to obtain information required for quality decision making, in a timely manner [8]. Thus, the deployment and effective use of data warehouses, data marts, and extract transform & load tools are necessary for BI systems [6]. However, the evaluation of suitable data warehouse products is a challenging task because such systems are complicated and diverse [18].

In addition, the tier of BI software consists of query and reporting, online analytic processing (OLAP), and data mining (DM) tools [4]. OLAP is a key capability employed in BI that enables the interactive examination and manipulation of large amounts of data from multidimensional aggregates [1]. Relational OLAP, multidimensional OLAP, hybrid OLAP, and desktop OLAP are typical technologies implemented by OLAP systems. The OLAP operations of drill-down, roll-up, pivot, slice, and dice are used to interact with the data cube for multidimensional data analysis [9]. Hence, an OLAP engine is aimed at the generation of interactive reports, which enable users to analyze data from a multi-dimensional database according to pre-defined criteria. While OLAP is an information capability that supports interactive examination and manipulation of large amount of data from multidimensional aggregates [1], data mining is a technique for "selecting, exploring, and modeling large amounts of data to discover previously unknown patterns and correlations, which leads to anticipating future behaviors, events and consequences" [3]. Data mining applies a variety of models, methods, and algorithms to uncover generalizations, regularities, rules, and previously unknown patterns in data that have been cleaned and integrated for consistency [4][7]. Statistical techniques, neural networks, rule induction, econometrics, and clustering are employed by advanced analytics software to make predictions and discover hidden relationships in data [11]. Descriptive data mining discovers associations, relations, exceptions, and deviations for the interpretation of knowledge. By contrast, predictive data mining applies classification, regression, or time series approaches to make forecasts based on historical data. Finally, the tier of BI analytical applications include tools for the evaluation of financial performance, strategy management, the workforce, supply chain management, services operations, and managing customer relations [11]. BI systems must connect analytical engines, OLAP, and/or data mining engines to data warehousing to present analytical reports, which combine tools to submit queries, prepare reports, and perform analysis [2]. The successful implementation of BI depends on the transformation of a dispersed database into decisions; therefore, a reliable data environment is crucial.

3 Methodology

This study applied principal component analysis to transform the performance indexes of profitability into a single indicator; and logistic regression to examine the roles of the data environment and profitability in the implementation of BI. Predictor

variables associated with the data environment and profitability were applied to predict the probability of BI implementation. Suppose that the binary variable Y is equal to 1 if the observed company has implemented a BI system. Based on the research of Kalakota and Robinson [16], our study includes the predictor variables of OLAP, data warehouses, and data mining to represent aspects of the data environment. Each variable was given two response categories: *non-implemented* and *implemented*. By contrast, the financial indicators of return on assets (ROA), return on sales (ROS), and return on investment (ROI) were transformed into a single predictor variable to evaluate the profitability of the companies investigated and these indicators were also used to evaluate the financial impact of ERP systems [14][23]. We employed an orthogonal transformation method [5], and converted observations of potentially correlated variables into a set of principal components to reduce the dimensionality of data [13]. Suppose that we only retrieve one principal component for ROA, ROS, and ROI, we calculate a score of profitability for each subject on this single principal component. According to the above discussions, variables associated with data warehousing (X_1), OLAP (X_2), data mining (X_3), and profitability (X_4) were used to predict the behavior of BI implementation. This was accomplished using logistic regression, in which the probability of a dichotomous outcome is related to a set of potential predictor variables in the model [17]. The framework of the logistic regression model used in this study is shown in Fig. 1.

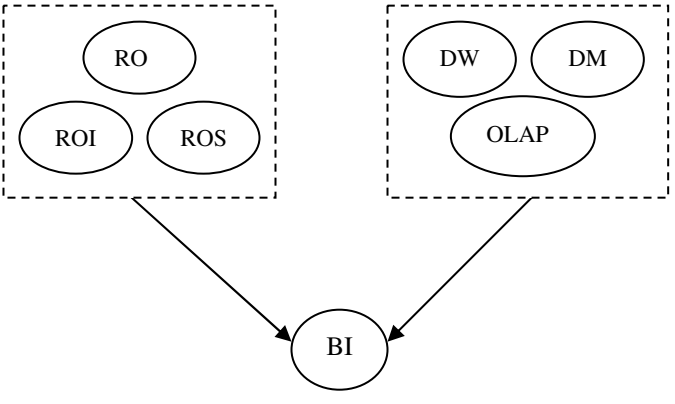


Fig. 1. Logistic regression model

Let Y_i denote the condition of the BI implementation for the i-th company. Under the theory of logistic regression, the probability of $Y_i = 1$ not only equals the expected probability of $Y_i = 1$ given a particular value of X_1, …, X_4 but also follows a binomial distribution. In addition, the logistic regression model presumes that the relationship between Y and the predictor variables X_1, …, X_4 can be accounted for by a logistic function with the following representation:

$$\ln \frac{P(Y)}{1-P(Y)} = \beta_0 + \sum_{j=1}^{4} \beta_j X_j \tag{1}$$

where P is the probability of the outcome, β_1,…., β_4 are the coefficients associated with the predictor variables of data warehouse (X_1), OLAP (X_2), data mining (X_3), and profitability (X_4). The dependent variable is the logarithm of the odds, which is the logarithm of the ratio of $P(Y)$ and 1 – $P(Y)$. Because the probability of BI implementation is the focus of analysis, we can transform equation (**1**) into

$$P(Y) = \frac{\exp(\beta_0 + \sum_{j=1}^{4} \beta_j X_j)}{1+\exp(\beta_0 + \sum_{j=1}^{4} \beta_j X_j)} \tag{2}$$

We applied the maximum likelihood approach to estimate parameter β_i using the iteratively reweighted least squares. This information helps us to understand which predictor variable plays a significant role in explaining the implementation of business intelligent system. An omnibus test was also applied to determine whether the model with predictor variables is significantly better than the model with only the intercept. In addition, the approach of Cox & Snell R^2 and Nagelkerke R^2 were adopted to evaluate the goodness-of-fit of our proposed logistic regression model. Besides, we summarized a classification table to cross-classify the binary response Y with a prediction of whether BI is implemented in order to evaluate the prediction accuracy of logistic regression.

4 Empirical Findings

Our empirical study investigated the companies listed among the top 1000 Taiwanese manufacturers published by CommonWealth magazine in 2007. To obtain information related to the implementation of ERP, BI, OLAP, data warehousing, and data mining, and the year in which ERP went on-line, a questionnaire was designed and distributed to a sample of 350 manufacturers. This study investigated only those companies that had implemented ERP systems because BI solutions are usually the next consideration following the purchase of an ERP system. In addition, many BI applications must be developed on the foundation of an ERP system. A total of 82 companies answered affirmatively that they had implemented an ERP system and these manufacturers were investigated further to determine whether the data environment and profitability factored into the investment in BI. The data source of the variables for the logistic regression model is listed in Table 1. Although the information related to profitability was retrieved from annual financial reports, information related to the implementation of a data environment was obtained from an analysis of the questionnaire. The indexes of ROA, ROS, and ROI were calculated according to the 3-year average performance in the years following the implementation of the ERP system. The variables, BI, OLAP, DW, and DM, were set to 1 to indicate that the corresponding systems were installed by the investigated manufacturers.

Table 1. Data source of investigated variables

Variable		Source	Note
Profitability	ROA	Annual Reports	Based on 3-year average
	ROS		financial performance after
	ROI		the on-line year of ERP
BI Implementation			
Data	OLAP	Questionnaire	'0' Non-Implemented
Environment	DW		'1' Implemented
	DM		

We first needed to evaluate whether the financial performance indicators of ROA, ROS, and ROI were suitable for principal component analysis, using the Kaiser-Meyer-Olkin (KMO) test and Bartlett test of sphericity. As shown in Table 2, principal component analysis is indeed applicable to the sample companies because the KMO statistic exceeds 0.5 and the significance of the Bartlett test of sphericity is zero.

Table 2. Results of KMO statistic and Bartlett test of sphericity

Kaiser-Meyer-Olkin statistic		0.515
Bartlett test of sphericity	Approximate chi-square distribution	231.386
	Degrees of freedom	3
	Significance	0.000

Principal component analysis was also utilized to reduce the profitability dimension of ROI, ROA, and ROS. Table 3 summarizes the results of the eigenvalues and the cumulative percent of variance from the principal component analysis. According to the Kaiser criterion, we only retained 1 component because the eigenvalue of the first component exceeded 1. We named this single component "profitability", which explains approximately 87% of the total variance from the observations of ROA, ROS, and ROI. Meanwhile, all of the factor loadings of ROA, ROS, and ROI exceeded 0.8, which enabled the use the derived component to represent the profitability environment of the investigated manufacturers. The component scores were then calculated using a linear composite of the optimally-weighted observed variables to assign profitability scores to each company.

Table 3. Total variance explained for the component of profitability

Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative%	Total	%of variance	Cumulative%
1	2.533	84.433	87.133	2.533	84.433	87.133
2	0.421	14.033	98.467			
3	0.046	1.533	100.000			

We then examined whether our proposed model depicted in Fig. 1 had adequate explanatory power for the companies in our sample. Instead of the general R^2 index in the ordinary least squares method, we adopted the Cox & Snell R^2 and Nagelkerke R^2 to evaluate the goodness-of-fit of our proposed logistic regression model. As shown in Table 4, both systems indicate that the predictor variables of profitability, data warehouse, data mining and OLAP provide acceptable overall goodness of model fit for the explanation of business intelligence implementation. In addition, we also conducted an Omnibus test to evaluate the suitability of the model for logistic regression. Because the Chi-square values of the Omnibus tests are significant, we can conclude that our logistic regression model with predictors of data warehouse, OLAP, data mining, and profitability is significantly better than a model with only the intercept. Thus, further analysis of logistic regression could be conducted using the proposed model.

Table 4. Results of Cox & Snell R^2 and R^2

Step	-2 Log Likelihood	Cox & Snell R^2	Nagelkerde R^2
1	74.102(a)	0.161	0.242

The classification table based on the prediction results of our logistic regression model is illustrated in Table 5. From the analysis of specificity of prediction, 95% of non-BI adopters were correctly predicted. Although our model only correctly classifies 36.4% of the companies where the predicted event of BI implementation was observed, on the whole our predictions were correct 65 out of 82 times, for an overall success rate of 79.3%. These findings again support the applicability of our logistic regression model for BI implementation.

Table 5. Classification table of logistic regression

	Predicted	BI		Percentage correct
Observed		0	1	
BI	0	57	3	95.0
	1	14	8	36.4
Overall percentage				79.3

To understand which predictor variables play important roles in the implementation of BI, Table 6 summarizes the results of maximum likelihood coefficient estimates and Wald chi-square values from our logistic regression model. Employing a 0.1 criterion of statistical significance, data mining and profitability is not necessarily able to explain the behavior of BI implementation because the corresponding Wald chi-square values are insignificant. Unlike service-oriented companies that generally emphasize the importance of data mining, our findings indicate that manufacturers do not consider data mining tools as important considerations in the implementation of BI. On the other hand, the results of the Wald test imply that the data environments of data warehouse and OLAP have significant partial effects on BI implementation. The positive coefficients also indicate that manufacturers using OLAP and data warehouse are more likely to invest in BI systems.

Table 6. Coefficient estimates of logistic regression

	Coefficient Estimate	Wald	Significance
Data warehouse	1.312	2.941	0.082
OLAP	1.281	2.983	0.081
Data mining	-0.512	0.575	0.433
Profitability	0.256	0.489	0.479
Constant	-2.325	18.713	0.000

5 Conclusions

Although data environment is critical for the performance of BI, its relationship with the BI implementation is seldom addressed. This study attempts to understand how the data environment and profitability influence the implementation of BI systems for firms that have already adopted an ERP system. This issue is particularly important for such firms because managers may implement BI solutions to acquire and analyze up-to-date information for operative and strategic decision-making, following the implementation of an ERP system. From the perspective of the data environment, our empirical findings indicate that data warehousing and OLAP are critical prerequisites for the implementation of BI. Because data warehousing consolidates data from various operational systems and OLAP allows the rapid generation of reports, manufacturers should deal with these issues prior to the implementation of BI. Although the relationship between data mining and BI implementation is not significant, our empirical findings imply that data mining may not be a priority in a BI environment, compared to OLAP and a data warehouse. On the other hand, we determined that profitability generally plays an insignificant role in whether BI is implemented. Because the manufacturers investigated in this study had already installed ERP systems, their financial health was generally adequate to pay for the costly ERP licensing and maintenance expenses.

Future studies could investigate other industries such as banking or retailing to determine whether the implications extend beyond those of this empirical study. Researchers could also apply other methods such as neural networks to evaluate these issues and make comparisons between various approaches. Because the results of Cox & Snell R^2 and Nagelkerke R^2 demonstrate that the goodness-of-fit of the proposed logistic regression model is only acceptable, future studies could examine additional explanatory variables such as decision process engineering, IT partnership, and strategic alignment, and determine which of these influenced the implementation of BI systems.

References

1. Agrawal, R., Srikant, R., Thomas, D.: Privacy Preserving OLAP, Baltimore, Maryland, USA, pp. 14–16 (2005)
2. Cao, L., Zhang, C., Liu, J.: Ontology-based integration of business intelligence. Web Intelligence and Agent Systems: An International Journal 4, 313–325 (2006)

3. Chou, D.C., Tripuramallu, H.B.: BI and ERP Integration. Information Management & Computer Security 13(5), 340–349 (2005)
4. Datamonitor Business Intelligence: From Data to Profit (2001), http://www.researchandmarkets.com/reports/560
5. Duda, R., Hart, P.: Pattern Classification Theory and Systems. Springer, Berlin (1988)
6. Elbashir, M.Z., Collier, P.A., Davern, M.J.: Measuring the Effects of Business Intelligence systems: The Relationship between Business Process and Organizational Performance. International Journal of Accounting Information System, 135–153 (2008)
7. Fayyad, U.M., Piatetsky Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1996)
8. Gorla, N.: Features to Consider in A Data Warehousing System. Communications of the ACM 46(11), 111–115 (2003)
9. Han, J.: OLAP Mining: An Integration of OLAP with Data Mining. IFIP. Chapmen & Hall (1997)
10. Hannula, M., Pirttimaki, V.: Business intelligence: Empirical study on the top 50 Finnish companies. Journal of American Academy of Business 2(2), 593–599 (2003)
11. Heiman, R.V.: IDC's Software Taxonomy 2010. International Data Corporation, Framingham (2010)
12. Herschel, R.T., Jones, N.E.: Knowledge Management and Business Intelligence: The importance of integration. Journal of Knowledge Management 9, 45–55 (2005)
13. Hotelling, H.: Analysis of complex statistical variables into principal components. Journal of Education Psychol. 24, 498–520 (1933)
14. Hunton, J.E., Lippincott, B., Reck, J.L.: Enterprise Resource Planning Systems: Comparing Firm Performance of Adopters and Non-adopters. International Journal of Accounting Information Systems 4(3), 165–184 (2003)
15. Inmon, W.H.: Building the Data Warehouse, 3rd edn. Wiley (2002)
16. Kalakota, R., Robinson, M.: E-business: Roadmap for Success. Addison-Wesley (1999)
17. Leung, P.S., Tran, L.T.: Predicting Shrimp Disease Occurrence: Artificial Neural Networks vs. Logistic Regression. Aquaculture 187, 35–49 (2000)
18. Lin, H.Y., Hsu, P.Y., Sheen, G.J.: A Fuzzy-Based Decision-Making Procedure for Data Warehouse System Selection. Expert system with Applications 32, 939–953 (2007)
19. Lonnqvist, A., Pirttimaki, V.: The Measurement of Business Intelligence. Information Systems Management 23(1), 32–40 (2006)
20. Mukherjee, D., D'souza, D.: Think phased implementation for successful data warehousing. Information Systems Management 20(2), 82–90 (2003)
21. Nelke, M.: Knowledge Management in Swedish corporations. The Value of Information and Information Services, Swedish Association for Information Specialists, Documentation, Stockholm (1998)
22. Network Managzine, http://news.networkmagazine.com.tw/classification/software-application/2011/05/05/24070/
23. Nicolaou, A.I.: Firm performance Effects in Relation to the Implementation and use of Enterprise Resource Planning Systems. Journal of Information Systems 18(2), 79–105 (2004)
24. Saegusa, R., Sakano, H., Hashimoto, S.: Nonlinear principal component analysis to preserve the order of principal components. Neurocomputing 61, 57–70 (2004)
25. Shin, B.: A Case of Data Warehousing Project Management. Information and Management 39(7), 581–592 (2002)
26. SPI Research, 2010 Professional Services Business Application Market Adoption, SPI Research (2010)

The Relationships between Information Technology, E-Commerce, and E-Finance in the Financial Institutions: Evidence from the Insurance Industry

Hong-Jen Lin[1], Min-Ming Wen[2], and Winston T. Lin[3]

[1] Brooklyn College, Cuny
2900 Bedford Ave
Brooklyn, NY 11210
hjlin@brooklyn.cuny.edu
[2] California State University Los Angeles
5151 State University Drive
Los Angeles, CA 90032
mwen2@calstatela.edu
[3] State University of New York at Buffalo,
325A Jacobs Management Center,
Buffalo, New York 14260-4000, USA
mgtfewtl@buffalo.edu

Abstract. This research investigates the impact of IT on the cost efficiencies and cost frontiers of the insurance industry. In addition, we compare the differences of the effect of IT on the insurance industry across countries. Moreover, we test the international market's integration hypothesis and explore the influences of other factors (e.g., scope economy, macroeconomic variables, etc.) on the cost efficiencies of the insurance industry. The evidence from both the two-equation and Tobit models indicates that IT improves cost efficiencies for the developed countries but reduces those for the newly developed economies under study. That is, the international productivity paradox hypothesis is rejected for the insurance firms in the developed countries considered. In addition, results of the cost frontiers for the insurance firms suggest that, IT improves the cost efficiency for the developed countries but not for the emerging economies.

1 Introduction

The whole 20th century has witnessed the invention and innovation of information technology in the insurance industry (Yates [18]). Especially in the past twenty years, information technology (IT)[1] has extensively changed the landscape of international financial institutions. The financial services in the US have become the largest customer of IT in the economy (Frei, et al.[7]). This phenomenon also has taken place in many

[1] IT here refers to the information technology in a broad sense, which includes telecommunications. Some others have called it ICT (informational and communications technology).

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 194–206, 2012.

other countries. The use of IT has improved the quality and speeded up the process of financial services and operations of financial institutions. This dramatic development, known as e-finance, has become a focus of practice in international finance.

E-finance is defined by Hartmann [10] as "transactions in which funding for an economic activity is provided through an electronic communication medium." Therefore, e-finance is a subset of e-commerce, while e-commerce forms a critical part of the management of IT. Allen et al. [1] define it as "the provision of financial services and markets using electronic communication and computation." Figure 1 depicts the relationships between IT, e-commerce, and e-finance.

The impacts of e-finance may be classified into two types. First, the use of electronic communications, such as electronic bill paying, home banking, and internet transaction, has been altering business-to-business (B2B) and business-to-customer (B2C) in the financial markets. The marketing accessibility of financial institutions is extended and increased to remote areas or countries via new telecommunications technology. Second, the applications of electronic computation and databases shorten the processing time of each financial transaction.

In finance literature, many previous studies (e.g., Hunter and Timme [11], among others) have explored the 'technological change' of commercial banks in the U.S. Only a few, such as and Lin et al. [12], have discussed technological changes under the framework of international finance. Furthermore, the concept of the 'technological change' is too broad to capture the contribution of IT. Therefore, it is critical to evaluate the impact of IT on the international financial industry. In addition, it seems that the insurance industry has been lagged behind commercial banks and investment banks in adopting e-finance tools. In terms of literature, more researchers have explored and investigated the effects of e-finance in commercial and investment banking. The impact of e-finance in the insurance industry is important but fewer studies have addressed this issue.

Interestingly, Yates [18] re-discovered the importance of early information technology in the beginning of the 20th century in the life insurance industry, when the invention of the tabulating machinery (an early version of information technology) made complicated insurance transaction easy. Harris and Katz [9] found the evidence that smaller life insurers tend to invest more in IT. Garven [10] and Forman and Gron [6] indicated that IT speeds up distribution between insurers and customers. Generally

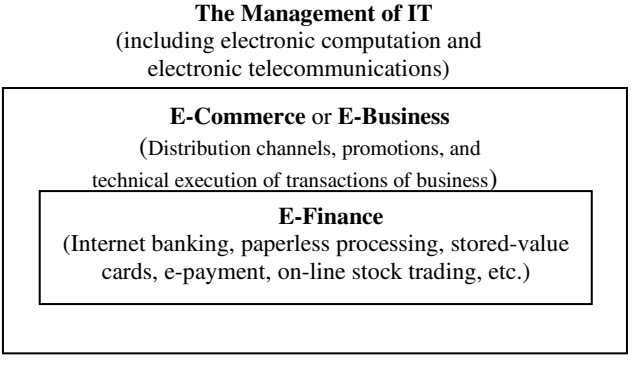


Fig. 1. The Relationships Among IT, E-Commerce, and E-Finance

speaking, the use of IT in the insurance industry has improved the distribution of products and customer services in the prevailing literature and cases studied.

According to Shao and Lin [17], IT investments are costly and they may not be profit-making. This phenomenon is called the IT productivity paradox (Dewan and Kraemer, 2000). It is necessary and important to justify IT investment because financial institutions are the major customers of IT. Alpar and Kim [2] have found that IT saves costs and labors but use more capital. Odoyo and Nyangosi [13] found IT investment enhances productivity of insurance firms; but Prassad and Harker [14] have stated that IT investments have little significance on the productivity of commercial banks. Therefore, there is no definite conclusion on whether or not IT significantly contributes to the profit or cost performance of financial institutions.

Moreover, most previous research on e-finance has been confined to the case of the U.S. An international comparison in this field is rare, except the work of Claessens et al. [4] who have discussed e-finance all over the world by using conceptual reasoning rather than providing strong empirical evidence. Therefore, it is important to investigate the impact of IT on the insurance industry across different countries, given that the world economy is increasingly globalizing.

Although the literature essentially remains silent on the effect of e-finance on international finance, the impact of IT across nations has been discussed extensively. For instance, Dewan and Kraemer [5] have investigated the international productivity paradox of IT. They have found that IT investment is negatively related to the productivity across nations. However, they have found strong evidence to support the contention that IT improves productivity for developed countries, but reduces productivity for newly developed and other developing countries. In other words, the negative relationship between IT and productivity is supported by the observations from the developing countries subsample and there is no "productivity paradox" in the developed countries at all.

The primary objective of this research is threefold. First, this paper investigates the impact of IT on the cost efficiencies and cost frontiers of the insurance industry. Second, we compare differences of the effect of IT on insurance firms across a variety of countries. In addition, as a secondary objective, we explore the influences of other factors (e.g., scope economy, macroeconomic variables, etc.) on the cost efficiencies of financial institutions.

In this study, the insurance industry is chosen as the research sample for the following reasons. First, the choice relates to their importance in the context of a global comparison. Second, insurance firms thrive across developed and developing countries, while other specialized financial institutions exist only in a limited number of developed countries. Second, given our emphasis on a global comparison, data availability becomes a primary consideration, and extensively international data are available for both of these institutions.

The insurance industry consists of property/casualty insurance, health care insurance, and life insurance, and can also be categorized into life and non-life insurances. Although the sources and uses of funds are different across these two types of insurances, they have some basic common features: they receive premiums from customers, pay them for accidents, and invest their reserves in financial markets. According to Rai [16], most international insurance companies (over 60%) are engaged in both life and non-life insurances, but during this deregulation era, the

boundary between these two insurances becomes less important because in many countries an insurance company underwrite both life and non-life insurance policies. This trend justifies our focus on a global comparison rather than comparing different types of insurance companies.

Regarding the methodological issues, efficiency analysis has become dominant in research on financial institutions (Rai, [16]). This paper adopts a more generalized stochastic frontier approach to estimate cost efficiencies, which has been shown empirically and theoretically to be superior to previous approaches.

The remainder of this research is structured as follows: Section 2 discusses methodologies and theories. Section 3 provides data sources and statistical hypotheses. Empirical analyses and results are summarized in Section 4. Section 5 concludes this study.

2 Theoretical Perspectives and Model Specifications

2.1 Theory of Cost Efficiencies

The research on financial institutions finds it difficult to acquire accurate technical (productive) efficiency due to the nature of their multi-product productions. Instead, cost efficiency is used to appraise and evaluate the performance of different firms in the literature. On the other hand, in the management of information technology or information systems (IT/IS), the productivity of IT investments is explained via a production function (Dewan and Kraemer [5]) or technical (productive) efficiency. Since we are interested in the relationship between IT investments and efficiencies of financial institutions, so how to use cost efficiencies to capture productivity is the pertinent question.

According to Shao and Lin [17], productivity (P_{it}) equals the product of technical efficiency change (TE_{it}) and technological change (TN_{it}). That is,

$$P_{it} = TE_{it} \times TN_{it}, \tag{1}$$

where TE_{it} is measured by actual output divided by the ideal (optimal) output and TN_{it} is obtained from the change of the production function. When the technological change remains constant through t, the technical efficiency change is the only factor causing productivity growth.

Technical (productive) efficiency (TE_{it}) captures the extent by which real cost exceeds the cost of the existing technology for a given input mix (or called labor/capital combination), and allocative (cost) efficiency (AE_{it})explains the extra cost of the real input mix over the cost of the ideal input mix. This relationship can be written in Equation (2) as

$$OCE_{it} = TE_{it} \times AE_{it} = \frac{\overline{OB}}{\overline{OA}} \times \frac{\overline{OC}}{\overline{OB}} = \frac{\overline{OC}}{\overline{OA}} = AE_{it} \times P_{it} \div TN_{it}. \tag{2}$$

That is, the overall cost efficiency (OCE_{it}) is composed of technical efficiency (TE_{it}) and allocative efficiency (AE_{it}). Any improvement in TE_{it} and/or AE_{it} will lead to an increase in OCE_{it}. Upon plugging Equation (2) into Equation (1), we can use the overall cost efficiency to depict the productivity progress. Thus, where the relationship between the overall cost efficiency and productivity is particularly of interest. In Equation (2), any improvement in P_{it} and/or AE_{it} leads to an increase in the product of OCE_{it} and TN_{it}. If an estimate of OCE_{it} is developed to capture TN_{it}, it can be used to test the productivity (P_{it}) of IT.

2.2 Stochastic Cost Efficiency Approach

This study adopts the parametric methodology because previous international studies in efficiency of financial institutions (e.g., Rai, [16]) tend to use the stochastic frontier approach. This research follows the two-equation frontier methodology. Theoretically, the two-equation model outperforms the one-equation counterpart because the two-equation approach considers all factors in one step. By doing this, fewer measured errors in the cost/profit efficiency are involved.

The two-equation model consisting of Equations (3) and (4) is to examine the cost efficiency of commercial banks and insurance firms:

$$\ln TC_{it} = f(\mathbf{y_{it}}, \mathbf{p_{it}}; \beta) + u_{it} + v_{ti}, \text{ and (4) } u_{it} = g(IT_{it}, \mathbf{z_{it}}; \alpha) + \varepsilon_{it}, \quad (3)$$

where β : a vector of unknown parameters to be estimated,

u_{it} : a random variable to account for the half-normally distributed cost inefficiency,

v_{it} : a random variable to describe the normally distributed disturbance, and

$f(\cdot)$: the optimized cost function of a given output vector and input prices.

The two-equation (i.e., Equations (3) and (4)) frontier model, featured with the flexibility of the choice of the factors of z_{it}, provides insights into several issues: First, like Lin et al. [12] firm-specific and macroeconomic attributes doomed to importance are considered. Second, IT is considered as an element of the z_{it} vector to explain the relationship between the progress of IT and the efficiencies in two financial industries. Third, z_{it} also can be used to describe the features of insurance firms by assigning different components of z_{it}.

In addition to the variables used in Lin et al. [12], this study emphasizes four country-specific risks in international finance: foreign exchange risk, financial risk, political risk, and economic risk. The financial risk, political risk and economic risk are represented by different risk indices compiled by the PRS (Political Risk Services) Group Inc.[15].

For insurance companies, the translog functional form of Equations (3) and (4) is specified as

$$\ln TC_{it} = \beta^*_0 + \sum_{k=1}^{2} \beta^*_{1k} \ln p_{kit} + \sum_{s=1}^{2} \beta^*_{2s} \ln y'_{sit} + \frac{1}{2} \sum_{k=1}^{2} \sum_{k'=1}^{2} \beta^*_{3kk'} \ln p_{kit} \ln p_{k'it} + \frac{1}{2} \sum_{s=1}^{2} \sum_{s'=1}^{2} \beta^*_{4ss'} \ln y'_{sit} \ln y'_{s'it} + \sum_{k=1}^{2} \sum_{s=1}^{2} \beta^*_{5ks} \ln p_{kit} \ln y'_{sit} + \alpha^*_1 IT_{it} + \ldots + v_{it} + \varepsilon_{it}, \quad (5)$$

Equation (4) will be estimated empirically by employing cost efficiencies and the Tobit model to appraise insurance firms across developed and newly developed countries.

2.3 Choice of the IT Proxies

IT-related variables for a financial institution are not readily available, so the country-level proxies are used. The e-finance variables are measured by using IT and COM proxies. The growth rate of the IT capital stock (ITCS) is employed and ITCS is defined as the revenue paid to vendors for hardware, data communications, software, and services. Due to the rapid development of the internet and telecommunications, the telecommunications investment has increased rapidly in order to foster progress in IT. Thus, the growth rate of telecommunications capital investments (COMCI) is used. The composite effects of IT and COM are important because the development of IT is technically and practically accompanied by that of telecommunications, and vice versa. Based on the specification of IT and COM, we can examine how new technologies influence the performance of financial institutions across nations.

3 Statistical Hypotheses

The major issue in this research is the assessment of the impact of IT: how does IT influence the cost efficiencies of financial institutions?, we propose the following statistical hypotheses.

$H_{1,0}$: IT contributes positively to the cost efficiency of the insurance industry.

$H_{2,0}$: IT contributes negatively or insignificantly to the cost efficiency for developed countries.

In summary, if $H_{1,0}$ for the newly developed and other developing countries are not rejected and the IT effect on the cost or profit efficiency for the developing countries is stronger than that for the developed countries, there exists the leapfrog phenomenon in e-finance. That is, financial institutions in the emerging economies outperform their counterparts in the developed countries in the application of IT or COM.

4 Empirical Results

4.1 Country Average of Firm-Level Variables for Insurance Firms

Table 1 shows the summaries of the variables from the financial statements for insurance firms in different countries. As observed in Table 3, the country average of

the TC/TA variable varies widely from 0.1690 (the Netherlands) to 1.4711 (Brazil). In addition, the high TC/TA could be directly related to a negative or low ROA. For instance, Korea (25) has a negative ROA but a high TC/TA. In contrast, the Netherlands (30) and the US (49), where the financial systems are more reliable and stable, have very low TC/TA's. Some countries such as Brazil (5) and Turkey (47) have earned high ROA's. The PR/TA should reflect the country risk that the insurance industry encounters. Results suggest that a high PR/TA is accompanied by a higher country risk. Several developed countries such as Belgium (4), Denmark (10), the Netherlands (30), Norway (31), Switzerland (44) and the US (49) have recorded low PR/TA ratios while some developing economies (e.g., Brazil (5) and Indonesia (19)) have reported high PR/TA ratios.

4.2 Country-Level Variables

Table 2 summarizes the data on the country-level variables. Several attributes in the z_{it} vectors are collected at the country level. As shown in Table 2, Sri Lanka (42) has the fastest annual growth of capital investments in IT (31.18%) and Japan (23) has the

Table 1. Summary of the Variables for Insurance Firms in Different Countries

Code	Country	TC/TA	PR/TA	FI/TA	ROA	w/TA
3	Austria	0.5233	0.3008	0.8361	0.6671	0.0143
4	Belgium	0.2955	0.1307	0.8307	0.8366	0.0122
5	Brazil	1.4711	0.8747	0.4461	2.3271	0.0003
9	Czech Republic	0.6037	0.3194	0.6522	-1.3673	0.0029
10	Demark	0.3411	0.1948	0.8396	1.4609	0.0131
12	Finland	0.5018	0.2450	0.7745	4.4374	0.0137
13	France	0.3701	0.2681	0.7354	0.9131	0.0069
14	Germany	0.4823	0.3590	0.8189	0.7652	0.0436
15	Greece	0.6597	0.4232	0.5377	0.1187	0.0353
19	Indonesia	0.5635	0.9999	0.5050	7.3136	0.0951
20	Ireland	0.3796	0.2216	0.7184	1.8635	0.0208
22	Italy	0.5710	0.3194	0.7924	0.7889	0.0099
25	Korea	1.3591	0.6793	0.7022	-0.1386	0.0249
27	Malaysia	0.4859	0.4952	0.8363	2.1593	0.0233
30	Netherlands	0.1690	0.0783	0.8960	0.7067	0.0015
31	Norway	0.2298	0.1518	0.8775	0.7616	0.0056
36	Portugal	0.6432	0.5621	0.7816	1.6115	0.0773
40	South Africa	0.4104	0.2218	0.9326	2.9937	0.0048
43	Sweden	0.4140	0.2313	0.9031	0.6795	0.0015
44	Switzerland	0.3118	0.1744	0.7415	0.8601	0.0103
46	Thailand	0.4512	0.4581	0.4782	1.1881	0.0403
47	Turkey	1.4354	0.4019	0.6839	8.8148	0.0274
48	United Kingdom	0.3591	0.2483	0.5720	2.3701	0.0192
49	United States	0.2646	0.1058	0.8368	0.7208	0.0037

Legend:
TC/TA= the ratio of total costs to total assets, PR/TA= the ratio of total premiums to total assets, FI/TA= the ratio of financial investments over total assets, ROA=the return on asset in %, and w/TA=the average wage level over total assets in %. Data Source: the *World Scope* CD and IFS.

Table 2. Summaries of the Country Level Variables

Code	Country	r	IT	COM	TR	IM	EX	FIN	POL	ECO
1	ARGENTINA	7.86	12.78	3.87	16148	27608	24255	34.54	74.62	37.57
2	AUSTRALIA	5.01	6.56	7.54	11675	65245	57521	39.12	85.47	39.36
3	AUSTRIA	2.06	7.07	2.14	15084	65261	57657	45.92	84.33	40.38
4	BELGIUM	3.24	7.50	6.90	11479	156485	170856	44.26	78.56	41.42
5	BRAZIL	34.22	13.67	31.38	32770	55280	48706	32.83	65.33	31.44
6	CANADA	5.01	6.25	7.09	13937	187532	204400	43.58	83.50	39.47
7	CHILE	11.72	8.11	17.42	10751	16891	15240	41.06	76.26	39.35
8	COLOMBIA	25.04	11.85	37.70	6252	13638	10366	35.85	53.43	33.98
9	CZECH	6.02	6.75	18.27	8218	28403	23357	41.88	83.00	37.08
10	DENMARK	2.94	8.46	5.03	11074	43145	47631	45.43	86.71	42.17
11	EGYPT	8.12	23.87	12.52	12587	15321	3463	38.41	64.18	37.46
12	FINLAND	2.18	11.38	13.80	6393	29127	38718	41.52	87.22	41.76
13	FRANCE	3.64	6.13	4.00	25953	274223	283304	42.42	79.99	40.23
14	GERMANY	3.14	6.49	2.19	57292	450078	512943	44.30	82.87	39.81
15	GREECE	12.65	13.08	27.33	11149	22875	10711	36.27	77.49	35.66
16	HONG KONG	4.86	6.01	13.47	54802	189913	175732	44.84	74.33	39.86
17	HUNGARY	11.90	12.73	48.79	6901	22511	19797	38.38	82.00	33.30
18	INDIA	5.00	17.59	34.69	21784	44522	35933	37.63	58.23	33.19
19	INDONESIA	16.42	10.84	6.37	12558	36482	47882	36.59	59.16	32.21
20	IRELAND	0.32	13.85	4.89	5132	38225	53950	42.95	84.95	41.67
21	ISRAEL	10.69	12.59	15.02	12724	30359	22144	39.53	62.53	36.35
22	ITALY	4.63	5.51	8.16	28734	206350	234106	40.52	77.55	39.95
23	JAPAN	0.58	2.21	16.66	149492	315417	413313	47.30	80.84	42.60
24	JORDAN	7.53	25.54	2.71	1617	3855	1822	37.44	71.50	38.51
25	KOREA	8.63	12.76	7.96	32153	125368	132422	41.15	77.21	38.18
26	LUXEMBOURG	3.92	8.26	7.23	60	9653	7361	47.74	90.96	43.23
27	MALAYSIA	5.91	9.13	9.82	18488	70638	77236	40.58	74.28	38.88
28	MEXICO	5.99	11.22	2.76	22851	138993	126065	33.68	68.00	34.27
29	MOROCCO	6.76	21.52	2.71	3079	9908	7093	37.00	69.50	36.25
30	NETHERLANDS	3.53	24.81	7.14	18806	175831	190953	44.08	88.50	41.25
31	NORWAY	4.93	9.58	5.21	15463	33780	43375	46.52	86.88	45.67
32	PAKISTAN	8.50	24.33	3.56	998	10447	8571	30.95	55.37	28.97
33	PERU	14.56	15.11	8.08	6698	9023	5862	35.06	58.45	34.03
34	PHILIPPINES	8.77	15.13	10.17	6426	31772	24175	35.70	66.58	35.35
35	POLAND	16.87	16.71	64.59	14740	39285	24999	39.86	81.61	35.95
36	PORTUGAL	5.54	13.99	27.66	11074	34764	23299	41.71	83.45	40.28
37	RUSSIA	13.27	7.55	10.99	10676	59233	81631	29.75	53.50	27.20
38	SINGAPORE	3.03	4.18	8.96	54223	114647	115833	45.67	86.60	46.09
39	SLOVAKIA	12.13	6.51	46.80	2341	12129	9734	37.50	81.00	34.70
40	SOUTH AFRICA	11.42	8.35	3.63	3436	28985	27953	37.50	70.42	34.91
41	SPAIN	4.82	6.60	5.35	35002	121060	98150	41.35	74.99	38.95
42	SRI LANKA	10.06	31.18	8.01	1404	5513	4179	35.38	59.76	34.18
43	SWEDEN	3.28	7.23	2.74	12576	64391	79803	39.92	84.50	40.12
44	SWITZERLAND	1.49	6.87	1.03	29619	72766	74188	48.43	85.67	42.88
45	TAIWAN	4.70	14.58	12.36	65858	105346	114331	45.77	80.58	42.45
46	THAILAND	8.87	3.99	4.61	22596	59533	54624	40.30	69.32	38.02
47	TURKEY	55.94	14.85	16.72	11594	39610	23746	32.70	54.08	27.81
48	UNITED KINGDOM	3.73	8.72	9.15	27105	286751	255374	42.36	83.88	36.56
49	UNITED STATES	5.20	9.41	4.99	55184	865847	633520	44.03	83.87	38.78
50	VENEZUELA	17.24	22.90	9.74	8402	13015	19588	36.00	64.43	31.42
51	ZIMBABWE	19.58	20.84	7.71	274	2686	2270	26.92	63.33	29.60

Legend:

r : the cost of capital measured by the deposit rate for each country in real term; IT: the growth rate of IT investments; COM: the growth rate of telecommunication investments;

TR: total foreign reserves in million SDR; IM: amount of imports in million US dollars;

EX: amount of exports in million US dollars; FIN: country financial risk; POL: country political risk;and ECO: country economic risk.

Data Sources: *International Financial Statistics*, IDC, and *International Country Risk Guide*.

slowest growth rate (2.21%) among the nations during the time periods considered. Many emerging economies have recorded a two-digit growth rate of the ITCS, due to the rapid expansion in information technology and the low amount of the $ITCS_{t-1}$.Similar to IT, the growth rate of the telecommunications capital investments (COMCI) have increased rapidly in some developing countries. The noted winners of COM are some of the countries, including Poland (35), Hungary (17), and Slovakia (39), in East Europe, due mainly to the de-regulation in the telecommunications industry after the collapse of their communist regimes. TR is measured by the SDR (Special Drawing Rights, defined by the IMF) and IM and EX are denominated in US dollars. The IM and EX measure the amount of international trade and the TR captures a country's capability of international payment. The FIN, POL, and ECO indices are reported in the *International Country Risk Guide*. They are two-digit rating scores based on a survey conducted by the PRS Inc. A larger number represents a lower risk, that is, more desirable. The cost of capital (r) differs substantially from country to country. We have observed that the cost of capital appeared higher for several countries experiencing financial distresses, such as Brazil (5), and Turkey (47). Generally speaking, the financial institutions in developed countries have benefited from low costs of capital.

4.3 Summaries Statistics

Table 3 summarizes the features of these variables for two country groups. Our sample contains 399 observations among which 294 belong to Group 1 (developed countries); and then 105 to Group 2 (newly developed countries or emerging economies). Many variables are measured as proportions to the total assets to mitigate heteroscedasticity problem caused by firm size.

For Group 1, the averages of three international financial risk indicators (FIN, ECO, and POL) are consistently higher than those of Group 2. That is, the developed countries bear lower country risk than the emerging economies. As traditional financial theories state, the higher return accompanies higher risk. Our observation conforms to this statement.

Furthermore, Group 1 countries invest more rapidly in ITCS (i.e., bigger IT) but less quickly in COMCI (i.e., smaller COM) than Group 2 countries. In other words, the newly developed countries emphasize more on the contribution of the communications network than on the renewal of the IT software and hardware, while the developed countries are more interested in gaining competitive strengths in the field of IT. All three international trade indicators, TR, IM and EX, for Group 1 are larger than their counterparts for Group 2. Overall, the amounts of international trades and foreign exchange reserves for the developed countries considered exceed those for the newly developed countries.

It is insightful to compare the descriptive statistics of the variables used for the two country groups. First, Table 3 shows that LNTC for Group 1 is smaller than that for

Table 3. Descriptive Statistics of the Variables for Insurance Firms Based on All Countries, Developed Countries (Group 1) and Developing Countries (Group2)

	All Countries				Group 1				Group 2			
Variable	Mean	Std. Dev.	t-value	Cases	Mean	Std. Dev.	t-value	Cases	Mean	Std. Dev.	t-value	Cases
LNTC	16.12	2.61	6.19	399	15.83	2.34	6.77	294	16.95	3.11	5.45	105
LNC	-3.31	0.77	-4.30	399	-3.57	0.62	-5.78	294	-2.56	0.65	-3.95	105
LNW	-4.93	2.00	-2.46	399	-5.30	2.10	-2.52	294	-3.89	1.20	-3.25	105
LNFI	-0.41	0.45	-0.89	399	-0.36	0.47	-0.76	294	-0.54	0.38	-1.42	105
LNPR	-1.44	1.05	-1.37	399	-1.56	0.74	-2.11	294	-1.11	1.60	-0.69	105
FIN	41.39	4.72	8.78	399	42.48	3.87	10.99	294	38.33	5.49	6.98	105
ECO	38.72	3.91	9.91	399	39.68	2.32	17.08	294	36.03	5.78	6.24	105
POL	81.79	7.55	10.83	399	84.50	4.13	20.47	294	74.21	9.56	7.77	105
ROA	1.29	3.80	0.34	399	1.20	2.10	0.57	294	1.54	6.54	0.24	105
IT	7.60	17.83	0.43	399	7.98	7.58	1.05	294	6.56	32.44	0.20	105
COM	6.19	15.85	0.39	399	5.01	10.69	0.47	294	9.51	24.98	0.38	105
TR	29160	16877	1.73	399	32629	16876	1.93	294	19447	12584	1.55	105
IM	233450	186760	1.25	399	295670	179090	1.65	294	59237	39188	1.51	105
EX	239530	188020	1.27	399	302810	179110	1.69	294	62347	43551	1.43	105

Legend:
LNTC=ln(LC/TA), LNC=ln(r), LNW=ln(w/TA), LNFI=ln(FI/TA), LNPR=ln(PR/TA), FIN=financial risk, ECO=economic risk, POL=political risk, ROA= return on assets in %, IT= the growth rate of IT investments in %, COM=the growth rate of telecommunications investments in %, TR= total foreign reserves in million SDR, IM=imports in million US$, and EX=exports in million US$.t-value= Mean/ Std. Dev.

Table 4. Results of Tobit Regressions for Insurance Firms

Panel I: Cost Efficiency						
	All Countries		Group 1		Group 2	
Variable	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
Constant	0.7904	*** 4.13	-0.1293	-0.39	0.7730	* 1.89
IT	0.0002	0.27	0.0028	* 1.88	-0.0029	*** -2.63
COM	0.0000	0.01	-0.0007	-0.49	0.0018	1.31
IT*COM	0.0000	-1.52	0.0000	-0.22	-0.0001	-1.40
T	0.0045	0.50	-0.0207	* -1.95	0.0235	1.16
TR	-0.0000028	** -2.16	-0.0000020	-1.23	-0.0000016	-0.47
IM	0.0000005	*** 2.61	0.0000004	** 2.36	-0.0000028	-1.45
EX	-0.0000001	-0.62	-0.0000001	-0.25	0.0000020	1.04
FIN	0.0007	0.24	-0.0003	-0.09	0.0043	0.46
ECO	-0.0007	-0.17	-0.0065	-1.37	0.0253	** 2.54
POL	-0.0044	** -2.04	0.0104	*** 3.19	-0.0188	*** -4.79
σ	0.1894	*** 28.25	0.1632	*** 24.25	0.2080	*** 14.49

Legend:
* denotes significance at the 10% level, ** denotes significance at the 5% level and *** denotes significance at the 1% level. IT= the growth rate of IT capital stocks, COM= the growth rate of telecommunications capital investments, ITL= IT lagged one-year, COML= COM lagged one-year, IT*COM= the product of IT and COM, ITL*COML=the product if ITL and COML, T= the trend variable, 1=1993,2=1994,…, 8=2000, TR= total foreign reserves in million SDR, IM=amount of imports in million US dollars, EX= amount of exports in million US dollars, FIN= country financial risk, ECO= country economic risk, POL= country political risk, and σ = the standard error of the Tobit model.

Group 2. As shown, the country level variables (FIN, ECO, POL, IT, COM, TR, IM, and EX) and ROA for Group 2 are more volatile than their counterparts for Group 1. Again, the high volatility of the variables for Group 2 reveals the time-varying characteristics of newly developed countries. Finally, it is noticed that several variables, such as ROA, IT, COM, and LNFI are more volatile than other variables for both groups under study. The highly variant ROA and LNFI describe the fluctuating performance and operations of the insurance firms while the dispersing IT and COM depict the instability of the ICT industry.

4.4 Results of Tobit Regressions for Cost Efficiencies

In order to explore the IT effect on the cost and profit efficiencies of insurance firms, Tobit regressions are conducted and the results are shown in Table 4. Moreover, the statistical efficiency of parameters is weaker than that in the two-equation model. A majority of z_{it} variables are found to be insignificant in relation with cost and profit efficiencies. The insignificance could result from the underestimation of the cost and profit efficiencies from the one-equation model. Nevertheless, the standard deviation (σ) in the sample and subsamples is significant. However, IT for Group 1 improves cost efficiency significantly at the 10% level. In addition, IT*COM tends to be negatively related to cost efficiency for the whole sample and the two subsamples. Moreover, the trend variable T is positively related to cost efficiency for the whole sample but its impact on cost efficiency is significantly negative for Group 1 and insignificantly positive for Group 2.

In sum, according to the results of Tobit regressions, Group 1 employs IT more efficiently than Group 2 and IT is important in improving cost efficiencies for the insurance companies in the developed countries considered. Therefore, we conclude that the international productivity paradox of IT exists in the insurance industry for the emerging economies but not for the developed countries. While other variables do not seem to manifest a significant impact on both the cost and profit efficiencies, the IT variable plays a key role in explaining the cost efficiencies for insurance companies.

5 Conclusions

As we know, the slow adaptation to the e-finance technology in the insurance industry reflects the fact that transaction, interactions between policyholders and insurers are less frequent in insurances (Allen, et al. [1]). On the other hand, the nature of heterogeneity of insurance products may cause the insurance industry to be more “brain intensive” than “capital intensive” in the use of information technology. Consequently, the insurance industry seems to adopt less IT or COM than commercial banks. Our empirical analysis of IT complies with the point of view that insurance firms invest in new ITCS to improve efficiency less actively than commercial banks. The evidence from both the two-equation and Tobit models indicates that IT improves

cost efficiencies for the developed countries but reduces those for the newly developed economies under study.

That is, the international productivity paradox hypothesis is rejected for the insurance firms in the developed countries considered. We should note that the empirical results obtained for cost efficiencies indicate that the tests of the effect of IT via the two-equation models are sound on the cost side. Consequently, results of the cost frontiers for the insurance firms suggest that, IT improves the cost efficiency for the developed countries but not for the emerging economies.

References

[1] Allen, F., McAndrews, J., Stratran, P.: E-Finance: An Introduction. Journal of Financial Services Research 22, 5–28 (2002)

[2] Alpar, P., Kim, M.: A Microeconomic Approach to the Measurement of Information Technology Value. Journal of Management Information Systems 7, 55–69 (1991)

[3] Berger, A.N., Hunter, W.C., Timme, S.G.: The Efficiency of Financial Institutions: A Review and Preview of Research: Past, Present, and Future. Journal of Banking and Finance 17, 221–249 (1993)

[4] Claessens, S., Glaessner, T., Klingebiel, D.: Electronic Finance, Reshaping the Financial Landscape Around the World. Journal of Financial Services Research 22, 29–62 (2002)

[5] Dewan, S., Kraemer, K.L.: Information Technology and Productivity: Evidence from Country-Level Data. Management Science 46, 548–562 (2000)

[6] Forman, C., Gron, A.: Vertical Integration and Information Technology Investment in the Insurance Industry. Journal of Law, Economics, and Organization 27(1), 180–218 (2011)

[7] Frei, F.X., Harker, P.T., Hunter, L.W.: Inside the Black Box: What Makes a Bank Efficient. In: Performance of Financial Institutions: Efficiency, Innovation, Regulation, pp. 259–311. Cambridge University Press (2000)

[8] Garven, J.R.: The Role of Electronic Commerce in Financial Services Integration. North American Actuarial Journal 4(3), 64–70 (2000)

[9] Harris, S.E., Katz, J.L.: Firm Size and the Information Technology Investment Intensity of Life Insurers. MIS Quarterly 15(3), 333–352 (1991)

[10] Hartmann, P.: Comments on Claessens, Glaessner, and Klingebiel. Journal of Financial Services Research 22, 63–71 (2002)

[11] Hunter, W.C., Timme, S.G.: Technological Change in Large U S. Commercial Banks. Journal of Business 64, 339–362 (1991)

[12] Lin, H.J., Lin, W.T., Chen, Y.H.: Technological Progress and Time-Varying Patterns of Cost Efficiencies in Commercial Banks: An International Comparison, in progress (2004)

[13] Odoyo, F.S., Nyangosi, R.: E-Insurance: An Empirical Study of Perceived Benefits. International Journal of Business and Social Science 2(21), 166–171 (2011)

[14] Prassad, B., Harker, P.T.: Examining the Contribution of Information Technology Toward Production and Profitability in U.S. Retail Banking, Report 97-109, The Wharton Financial Institutions Center, University of Pennsylvania, PA (1997)

[15] PRS Group Inc., International Country Risk Guide, East Syracuse, New York (2001)
[16] Rai, A.: Cost Efficiency of International Insurance Firms. Journal of Financial Services Research 10, 213–233 (1996)
[17] Shao, B.B.M., Lin, W.T.: Technical Efficiency Analysis of Information Technology Investments: A Two-Stage Empirical Investigation. Information & Management 39, 391–401 (2002)
[18] Yates, J.: Information Technology and Business Processes in the 20th Century Insurance Industry. Business and Economic History 21, 317–325 (1992)

Applying Data Mining in Money Laundering Detection for the Vietnamese Banking Industry

Dang Khoa Cao and Phuc Do

University of Information Technology, Viet Nam National University-HCMC

Abstract. The applying of data mining techniques in banking is growing significantly. The volume of transaction data in banking is huge and contains a lot of useful information. Detecting money laundering is one of the most valuable information which we can discover from transaction data. This paper will propose the approaches on money laundering detection techniques by using clustering techniques (a technique of data mining) on money transferring data of banking system. Besides, we present an implemented system for detecting money laundering in Viet Nam's banking industry by using CLOPE algorithm.

Keywords: data mining, money laundering, money transferring data.

1 Introduction

In 2006, The World bank warned that Vietnam is becoming an easy target for money laundering activities because of the weakness in the inspection, supervision, auditing and customer relationship management systems. The level of using cash and several kinds of payment method make transactions become out of control [11]. Meanwhile banking plays an important role in cleaning dirty money in the money laundering process. Thus, the requirement for a mechanism to detect money laundering is growing with great importance for all over the world (especially in Vietnam).

This paper will research on anti - money laundering models, we combine the CPU process time of data mining techniques and the human's analyst ability to create an effective method to detect money laundering. Because of the large volume of data, banking creates a convenient environment to hide the original of money laundering. This fact makes money laundering techniques become more sophisticated and hard to trace the crime of money laundering. So the solution for detecting money laundering must be balanced between accuracy and the process time. Finding a suitable algorithm for data mining in banking is the most important step for the overall solution of the problem. This paper will propose a solution for money laundering detection system in Vietnam.

2 Preliminaries

2.1 Data Approaching

Following the research of Linard Moll about the approaches on money laundering models [3], depending entirely on each sort of bank's data, we have 4 approaches as follows:

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 207–216, 2012.

1. **Supervised approaches on labeled data:** This method requires an existing training set (labeled data). The author used one of these techniques: data mining, expert systems, statistic models etc… on the training set [3]. This approach is suitable for experienced banks in detecting money laundering. Therefore data will be in well-form before being mined.
2. **Hybrid approaches with labeled data:** These approaches are similar to "Supervised approaches on labeled data" regarding the aspect of data. The most important difference is that this approach combines multi-techniques to increase the accuracy of overall process. It requires the experience of bank in money laundering detection and the budget invests to the anti money laundering system.
3. **Semi-supervised approached with legal (non-fraud) data:** This approach is different from two approaches before. This approach just requires the valid training set (this means that all members in the training set must be valid data). New transactions will be considered invalid if their behavior doesn't match with the training set. So this approach requires banks which already had a mechanism to distinguish normal and abnormal transactions. With this approach banks can distinguish between valid and invalid data so they have only set up detecting methods on invalid data and let valid data work in normal capacity.
4. **Unsupervised approaches:** This approach is suitable for banks without having any methods for reviewing data (it means that these banks don't have any training sets).

Following the survey at bank X, the approaches of the anti laundering models that require training sets can't be applied for the Vietnamese banking industry (because Vietnam have lacked experience in anti money laundering). Thus, "unsupervised approaches" is the most suitable approach for detecting money laundering of Vietnamese banking industry.

2.2 Money Laundering Process

2.2.1 General of Money Laundering

By the decree 74/2005/NĐ-CP about preventing and anti money laundering of Vietnam's authority, money laundering is defined as: "*Money laundering is the behavior of persons or organizations that tries to validate or cleanse dirty money (money was earned by criminal activities).*

Basically, the money laundering process consists of 3 steps: placement, layering and integration.

Placement: Distributing money from illegal activities in the banks that have weak management mechanisms. Usually, the money will be divided into several parts that fall below the bank alert level. [11]

Layering: In this step, money will be transferred to several banks or several accounts. The real purpose of this step is to hide the illicit money origin by creating a transaction sequence. By creating a sequence of transactions, the origins' money will be hard to detect. [11]

Integration: In this step, money will be invested in legal business, and its profit will be used for criminal activities again to begin new cycle [11].

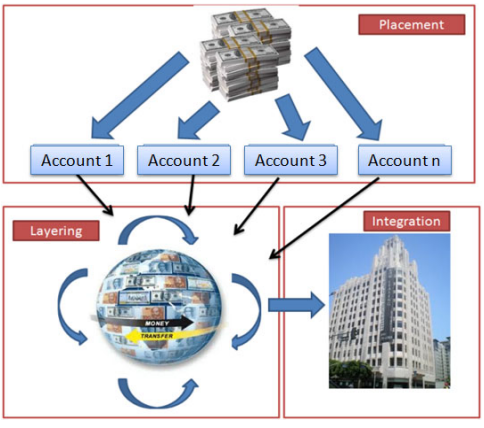


Fig. 1. General money laundering process in real world

2.2.2 Money Laundering Process

The sophistication of money laundering activities depends on the transaction's sequence that is used to hide the relationship between dirty money and its origin. In addition, because of the sophistication of money laundering to be mentioned above, layering becomes the hardest step for applying money laundering mechanism. But by the perspective of the data, this step has the biggest chance for applying automatic money laundering mechanisms.

Following the survey of a bank in Vietnam, the real money laundering processes can be divided into smaller processes that can be considered as sub-processes of the money laundering process. These sub-processes basically have characteristics that make involving transactions become different from normal transactions. So it makes the accounts performing these transactions become suspicious account.

The advantage of this solution is that we can limit the suspicious transactions by checking the suspected account only. This convenience will decrease significantly the operation time and make anti-laundering become realistic. Basically, the processes of money laundering are shown in figure 2:

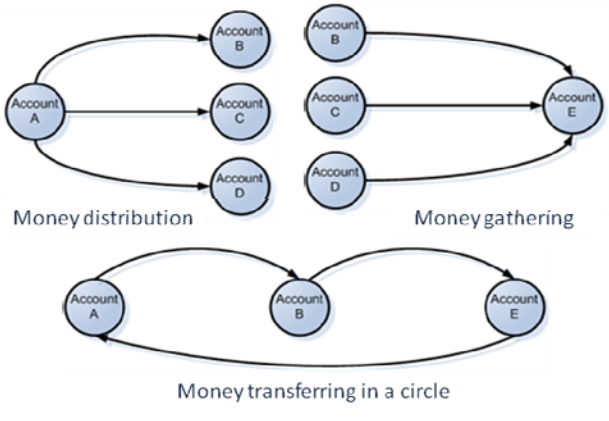


Fig. 2. The money laundering processes

Our paper focuses on the main problem of money laundering. We will identify the accounts that have potential for money laundering instead of finding all money laundering transactions directly.

2.3 The Learning Data Model

Bank transferring transaction consists of five main attributes:

"Sender account ID, receiver account ID, amount of money, money type, transaction date"

All attributes above can express the detail of transferring money data's information. When we take an instance of data, we can draw a graph to present relationship between accounts.

To determine accounts having the potential characteristics of money laundering or not, we must find out the behaviors of these accounts in one period of time. Thus, we will propose a new data set to be created by grouping accounts from transferring information. The new data set can express the special behavior of an account in a determined period of time. The new data has following attributes:

Table 1. Attributes of new data set

Attribute	Explaination
Account	Transaction account
Sum_sending	Sum of sending
Sum_receiving	Sum of receiving
Number_sending	Number of sending
Number_receiving	Number of receiving
Receiving_relationship	Number of account that send money to this account
Sending_relationship	Number of account that receive money from this account
| R_S |	| Sum_receiving – Sum_sending |

Money transferring data is converted to transaction data specifying a particular behavior of accounts. Each transaction data expresses specific behavior for a particular account. Multi-dimensional and large databases are two characteristics of transaction. In the next section we will introduce CLOPE algorithm and prove that CLOPE is the most comprehensive algorithm for processing the transaction data.

2.4 Introducing CLOPE Algorithm

CLOPE algorithm was invented by Yiling Yang, Xudong Guan and Jinyuan You [8]. This algorithm is used for clustering technique. It can works with the nominal variables (string variable) only. The main idea of the algorithm is based on the realistic situation from real life data in string data type. Moreover, the applying of data mining in real life always faces with a multi-dimensional (containing diversified information) data and a large database.

The authors of CLOPE algorithm proved that the distance approach for nominal variables isn't suitable, especially for finance data. Alternatively, CLOPE also defines a global criterion function that is used to optimize clustering process. Section below will introduce briefly about the criterion function and how it affects to the clustering processing. Typically, each clustering algorithm will define a criterion function and use it as a function to optimize the clustering process based on its calculation.

The criterion function will be separate into global criterion function and local criterion function. Global criterion function will determine the optimization for overall clustering instead of each cluster. Otherwise, the local criterion function focuses on optimizing each cluster. Because of the purpose of local criterion function, so it is harder to calculate than the global criterion especially with multi-dimensional data. The use of the global criterion function of CLOPE algorithm proves that it's suitable for multi-dimension data and large database. CLOPE algorithm will model each cluster to histogram that will be displayed in figure 4:

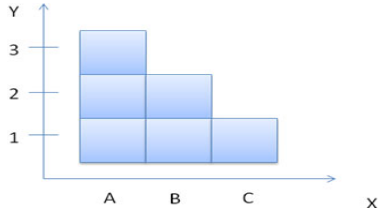


Fig. 3. Modeling clusters to histogram by CLOPE algorithm

- Axis (X): cluster's members ∈ D(C)
- Y-axis (Y): the frequent appearance of cluster members ∈ D(C) C is a cluster.
- Assumption:

$$S(C) = \sum_{i \in D(C)} Occ(i, C) = \sum_{t_i \in C} |t_i| \tag{1}$$

$$W(C) = |D(C)| \tag{2}$$

- S(C): Number of member in cluster C
- W(C): Number of member on x-axis.
- Occ(i,C) :The appearance frequency of member i in cluster C
- H(C) (the high): = S(C)/W(C)

CLOPE's criterion function is as follows:

$$Profit(C) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} |C_i|} \tag{3}$$

r: **Repulsion** is a real number (r > 0). In case arguments $S(C_i)$, $|C_i|$, $W(C_i)$ are given. If *r* increases the more similar data will be grouped or the clustering will bring more profit. Otherwise, if *r* decreases, more similar members will be separated into different groups. The highest purpose of the algorithm is to optimize the clustering to make the highest profit with a given *r*. Because the purpose doesn't focus on managing number of clusters (a cluster will be created if its existence and it makes the overall profit of clustering increasingly). So more cluster was created, it doesn't mean the profit is increased. [8]

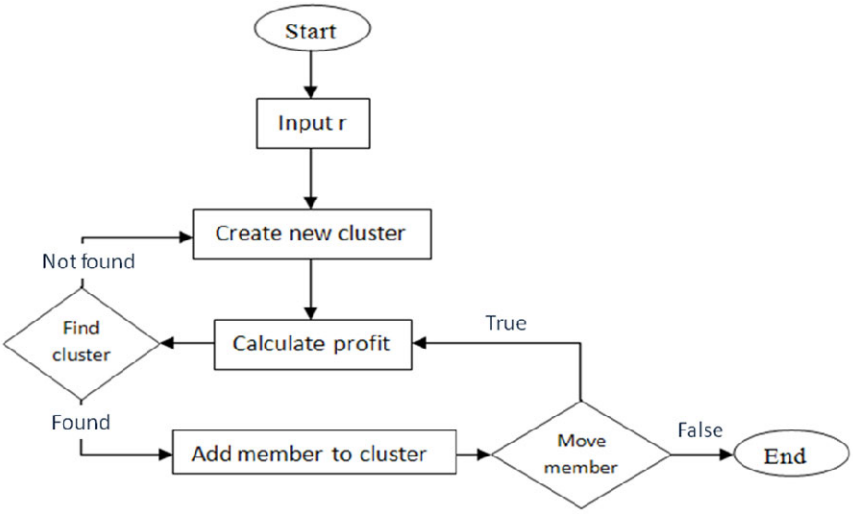


Fig. 4. The process flow of CLOPE algorithm

3 System Implementation

This section will explain the architecture and work flow of the money laundering detection system that utilizes clustering technique. The work flow consists of four stages: data converting, data fragmentation, clustering & analyst, checking the relationship of suspicious accounts:

1. **Data converting:** In this stage, transferring data will be converted to transaction data by separating each pair of accounts (sending and receiving) then make statistics for each account. The final purpose is to create a new data set (transaction data) expressing the behaviors of each account. The new data set will contain *m* records that $n \leq m \leq 2n$ (n is the record of the old data set).
2. **Data fragmentation:** In this stage, the system will convert all number data types into nominal data type by separating number data type to several parts (fragments) that contain specific meaning. Eg: Converted data of attributes "Sum_sending" ∈ [1.000.000.000, 10.000.000.000] into a text like "[1.000.000.000 =>

10.000.000.000]" and assign a meaning for this fragment such as "a big money transaction" (this conversion will make data to become meaningful for data mining before applying CLOPE algorithm).[6]

3. **Clustering & analysis:** In this stage, data will be grouped into clusters by using the CLOPE algorithm. It depends on the optimization of clustering (r argument in criterion function).

To find out which clusters have potential behaviors for money laundering, it must have a set of criteria for each case of money laundering. Based on the survey at bank X in Vietnam, we will examine an example set of criterias to validate cluster .The set of criterias is listed below:

Case 1: Suspicious transferring in a circular pattern: The attribute R_S (|Sum_receiving – Sum_sending|) will be focused on. Because each account in this case will send and receive the same amount of money so R_S attribute of these accounts will near to 0. When focusing on the behavior of account, we can find that all involve accounts belonging to this case will do the same actions: sending and receiving same amount of money. So when the system performs clustering, these accounts will belong to the same cluster as a result. The question is why R_S isn't equal to 0. R_S = 0 is not absolutely right in the real world since the intelligent crime will perform exchange money during the process and the exchange rate of money will affect to the value of original money (just a minute amount). Hence R_S will hardly equal to 0.

Case 2: Suspicious for money distribution: To determine the clusters for this case, we use the following attributes:

- Num_Sending: number of sending times of these account are higher than almost others accounts.
- Sending_Relationship: number of accounts that receive money from each account is higher than almost others account.
- Sum_Sending: Sum of sending is large or very large (higher than alert level of bank).But the sending of money each time usually small or normal when the system query each transferring transaction record of this account.

The complication of this case depends on how large the amount of money in money laundering activities and the state of current finance. Using data mining, the system doesn't care how large the amount of the money laundering activities is, instead of the similar behavior of account that performed these transactions.

Case 3: Suspicion for gathering money from several sources

- Num_Receiving: the frequency of receiving money is high.
- Receiving_Relationship: Number of accounts sent money to this account is higher than almost others account.
- Sum_Receiving: Sum of sending money is very large (larger than alert level of bank). But the receiving of each time is small or normal.

Similarity with the distribution of money case but in this case the current account plays a role to collect money from several accounts (maybe this account is original account in the chain of sequence).

We distinguished 3 basic cases of money laundering activities. Accounts that belong to transferring data in a circle case will have the same behavior so they will be easier to group. But for two remain cases the accounts will be separated into two groups with opposite behavior.

Thus, the result of clustering for money distribution or gathering money will create 2 clusters that have opposite characteristics. The first one contains accounts that send money to several accounts. The second one contains accounts receive money from several accounts. Figure 6.

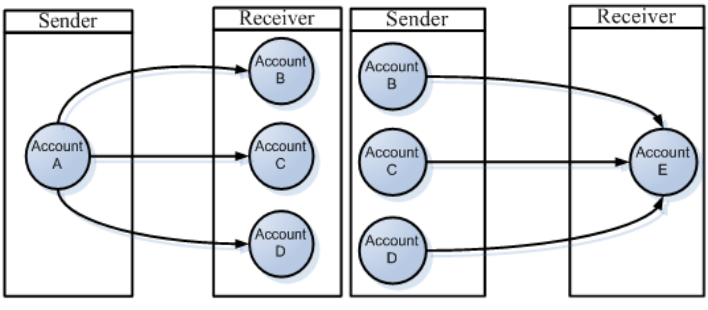


Fig. 5. Distributed & Gathered money difference

4. **Verify the relationships of the suspicious accounts:** After determining the suspicious cluster, the system must check the relationship of each account in each suspicious cluster to indicate which account participating in money laundering activities and in which case the activities belong to. We propose a solution that combines a data management system with n-tree data structure to increase performance for finding the relationship.

4 Experimental Results

Our tested data set contains 8020 records of transferring transaction of bank X. After converting to transaction data, the data set has 12.350 records. This test was performed on software that was implemented by the author based on WEKA open source (learning machine open source of Waikato University). We were simulated 25 records to present each case of money laundering; Structure of money laundering activities was shown in Figure 7:

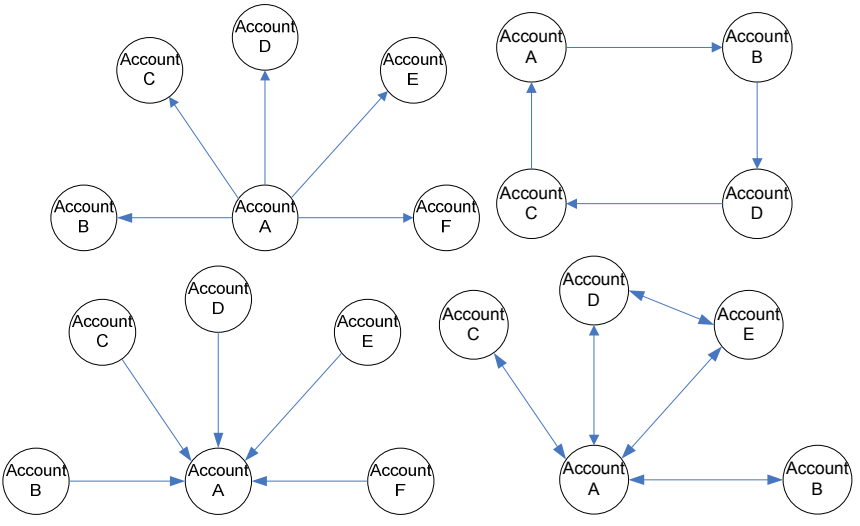


Fig. 6. Simulated money laundering structure cases

Table 2. Experimental result

Case	CLOPE		K-means	
Transferring on a circle	Cluster 1	12 member/12	Cluster 0	4 member /246
	Cluster 19	1 member /1	Cluster 9	9 member /143
	Sum : 13 member / 13		**Sum: 13 member / 389**	
Distributed money	Cluster 6	5 member /1804	Cluster 0	1 member /246
	Cluster 21	1 member /1	Cluster 1	1 member /3178
			Cluster 2	1 member /3039
			Cluster 3	2 member 1091
			Cluster 4	1 member /966
	Sum: 6 member / 1805		**Sum: 6 member / 8520**	
Gathered money	Cluster 2	5 member /1296	Cluster 0	1 member /246
	Cluster 18	1 member /1	Cluster 1	3 member /3178
			Cluster 2	1 member /3039
			Cluster 16	1 member /159
	Sum: 6 member / 1297		**Sum: 6 member / 6622**	

5 Conclusions

According the survey and the requirement specification of bank X (the provider of our data source), we recognized that finding solution for money laundering detection is growing more significantly nowadays for all over the world and especially in Vietnam.

We proposed a new training data set that is converted from banking data to suitable for applying CLOPE algorithms in money laundering detection. The experimented

result proved that CLOPE is a suitable algorithm for money laundering detection. But the system can't run stand alone absolutely, it must base on the ability of analysts in analyzing data and providing a set of rules (criteria set) to validate clusters after clustering. We hope our system can support the money laundering detection and in the near future it can discover automatically the money laundering problems when system receives new transactions.

References

1. Vimal, A., Valluri, S.R., Karlapalem, K.: An Experiment with Distance Measures for Clustering (2008)
2. Rosen, K.H.: Curriculum: Applying discrete mathematics into computer. Translator: Phạm Văn Thiều, Đặng Hữu Thịnh (2002)
3. Linard Moll from Switzerland, Master Thesis: Anti Money Laundering under real world conditions - Findingrelevantpatterns,Universitys of Zurich, 4-15 (2009)
4. Vu Lan, P.: Research and implement some algorithm of data mining. Ha Noi University of Science and Technnology (2006)
5. Le-Khac, N.-A., Markos, S., Kechadi, M.-T.: A Heuristics Approach for Fast Detecting Suspicious Money Laundering Cases in an Investment Bank (2009)
6. Do, P.: Data mining curriculum. National University of HCM City (2008)
7. Wiwattanacharoenchai, S., Srivihok, A.: Data Mining of Electronic Banking in Thailand: Usage Behavior Analysis by Using K-Means Algorithm
8. Yang, Y., Guan, X., You, J.: CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data. Shanghai Jiao Tong University (2002)
9. Webpage : Wikipedia – searching about transaction database, http://en.wikipedia.org/wiki/Database_transaction
10. Webpage :Researching for money laundering forms, http://www.vnecon.vn/showthread.php/3764-R%E1%BB%ADa-ti%E1%BB%81n-l%C3%A0-g%C3%AC-C%C3%A1c-h%C3%ACnh-th%E1%BB%A9c-r%E1%BB%ADa-ti%E1%BB%81n-hi%E1%BB%87n-nay
11. Webpage :anti money laundering in Vietnam (2009), http://www.hids.hochiminhcity.gov.vn/Noisan/32009/mach3.htm

A Web Services Based Solution for Online Loan Management via Smartphone

Théophile K. Dagba[1], Ercias Lohounmè[2], and Ange Nambila[3]

[1] Université d'Abomey-Calavi, 03 BP 1070 Cotonou, Benin Republic
theophile.dagba@eneam.uac.bj
[2] PADME, 08 BP 712, Cotonou, Benin Republic
elohounme@padmebenin.org
[3] Institut Africain d'Informatique, BP 2263 Libreville, Gabon
anambila@yahoo.fr

Abstract. Small microfinance institutions (in underdeveloped countries), despite their determination to expand into rural areas, are limited by geographic isolation and high transaction costs. For this, a solution has been proposed to access some features of the enterprise application via Smartphone. This solution is based on a service-oriented architecture in which the main features are developed as web services. The invocation of the methods is performed from a Smartphone or a PC, while the execution will be on a server that returns an understandable result through WSDL (Web Service Description Language) generated by web services, where the transport is provided by SOAP (Simple Object Access Protocol).

Keywords: Web services, E-finance, Unified Modeling Language, Case studies, Tools and applications.

1 Introduction

Microfinance institutions in developing countries face two major challenges in their efforts to reach rural communities: the transaction costs and risks. The high transaction costs are due to low trading volume, low quality or inadequate telecommunications infrastructure. However, nowadays, the development of mobile computing is helping to overcome these obstacles by encouraging the professional nomadic activities reducing the high transaction costs. Mobile computing deals with all solutions developed on platforms such as mobile phones, Smartphone, notebooks, etc. Mobile computing allows a user, while on the move to keep some of these digital tools, while providing new services [7]. But the costs of acquiring, deploying and maintaining mobility solutions proposed by designers are not easily accessible to small and medium enterprises.

As part of this work, we propose a loan management solution for microfinance institutions whose main objective is to release the mobile officer from geographical constraints. This tool will be used to manage credit operations via Smartphone. This will establish a client-server application with a mobile client and a web client that

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 217–226, 2012.

consume web services in accordance with the architecture to be proposed. The platform's specific objectives are to contribute to the following: improving the efficiency and reliability as well as the speed of loan management, facilitating collaboration between remote offices, accessing information in real time for fast decision-making, improving customer service, reducing the level of risk.

The rest of this paper is organized as follows: section 2 gives an overview of mobile computing with emphasis on service oriented architecture. Sections 3 and 4 present the case study and its implementation respectively. Finally, section 5 discusses the results and section 6 contains conclusions and recommendations for future work.

2 Mobile Computing and Service-Oriented Architecture

The growing development of computer technology and electronics has spurred the miniaturization and proliferation of mobile terminals. Some of these mobile devices are tailored to business needs: PDA (Personal Data Assistant), Smartphone, PC Tablet, etc. Mobile applications are a rapidly developing segment of the global mobile market. They consist of software that runs on a mobile device and performs certain tasks for the user of the mobile phone. Several varieties of applications related to mobility have been developed. Some applications such as SMS/MMS clients, browsers and music players, come pre-installed on the mobile phone whereas others may be provisioned and/or configured post-sales. A classification approach uses the fact that mobility is linked to the user, the data server or data themselves. Thus, there are [5]: mobile client with fixed server, mobile client with mobile server, fixed client with mobile server, and mobile data in a fixed server. The architectures of existing solutions for mobile applications are: client/server model, component model, service-oriented architectures (SOA). A SOA in which the main application functionality will be developed and published as web services better meets our project.

Conceptually, the concept of SOA is based on an architectural style that defines an interaction model between three primary parties: the service provider, who publishes a service description and provides the implementation for the service, a service consumer, who can either use the uniform resource identifier (URI) for the service description directly or can find the service description in a service registry and bind and invoke the service. The service broker provides and maintains the service registry [3]. SOA aims to develop comprehensive software architecture decomposed into services tailored to the business processes of the company. It is based on the reorganization of applications in functional units called services. The service is a key component of SOA. It is a well-defined function or feature. It is also a stand-alone component that is independent of any context or external service, which guarantees reusability. A service is a function encapsulated in a component that can be queried using a query consisting of one or more parameters and providing one or more answers. Many services can be composed to create a wider service. The service can be coded in any language and run on any platform (hardware and software). The service description is to describe the input service parameters, the size and type of data returned. The primary format for describing services is WSDL (Web Services Description Language) [26], standardized by the W3C (World Wide Web

Consortium) [28]. The notion of discovery includes the ability to search a service among those that have been published. The primary standard used is UDDI (Universal Description Discovery and Integration) [24], standardized by OASIS (Organization for the Advancement of Structured Information Standards) [27]. The invocation is the connection and interaction with the customer service. The main protocol used for the invocation of services is SOAP (Simple Object Access Protocol) [22]. By its standard, SOA improves the speed and productivity of software development. A component exposed as a web service can be reused by other applications. SOA can also get all the benefits of client-server architecture including: a modularity allowing easy replacement of a component (service) by another, the reuse of components (as opposed to an all- in-one tailored to an organization), a greater fault tolerance, an easy maintenance, etc.

In a mobile environment, web services have the advantage of significantly reducing the size of mobile device applications. They help facilitate the development of mobile services by offering an easy way to remotely access a data server. Web services provide interoperability between various software programs running on different platforms and they can work through many firewalls without requiring changes to the filtering rules as based on the HTTP protocol. Libraries exist in Java to make use of web services from platforms for mobile applications development such as J2ME (Java 2 Micro Edition).

3 Case Study

The main need for this project is to remotely access credit information relating to a portfolio at anytime and anywhere. To identify needs, the procedures manual of « Association pour la Promotion et l'Appui au Développement des Micro-Entreprises »[1] (PADME) is used. PADME [19] is one of the direct credit institutions in Benin Republic that provide a simple solution to the needs of low-income populations as well as micro enterprises. With its well-developed procedures manual, one can identify the usual operations performed by officers whose function requires certain mobility [14].

For modeling purposes, UML [18] is applied since it has become an essential tool in the modeling process of different types of application [6], [8]. More specifically, it is widely used in the modeling of systems based on web services [1], [2], [4], [10]. In this paper we will present the system through a set of high level UML diagrams (use case diagram, class diagram, and sequence diagram).

The use case diagram is shown in Fig. 1. Based on the requirements of our system, actors are:

- Loan Officer: responsible for implementation, monitoring and collection of funds in his sector
- Office Manager: responsible for the coordination of credit operations in his office area

[1] Association for the Promotion and Development Support of Micro-Enterprises.

- Administrator: responsible for user profiles and passwords management and system administration
- Credit Collector: ensures the recovery of debts
- Internal Auditor: performs controls on granted credits

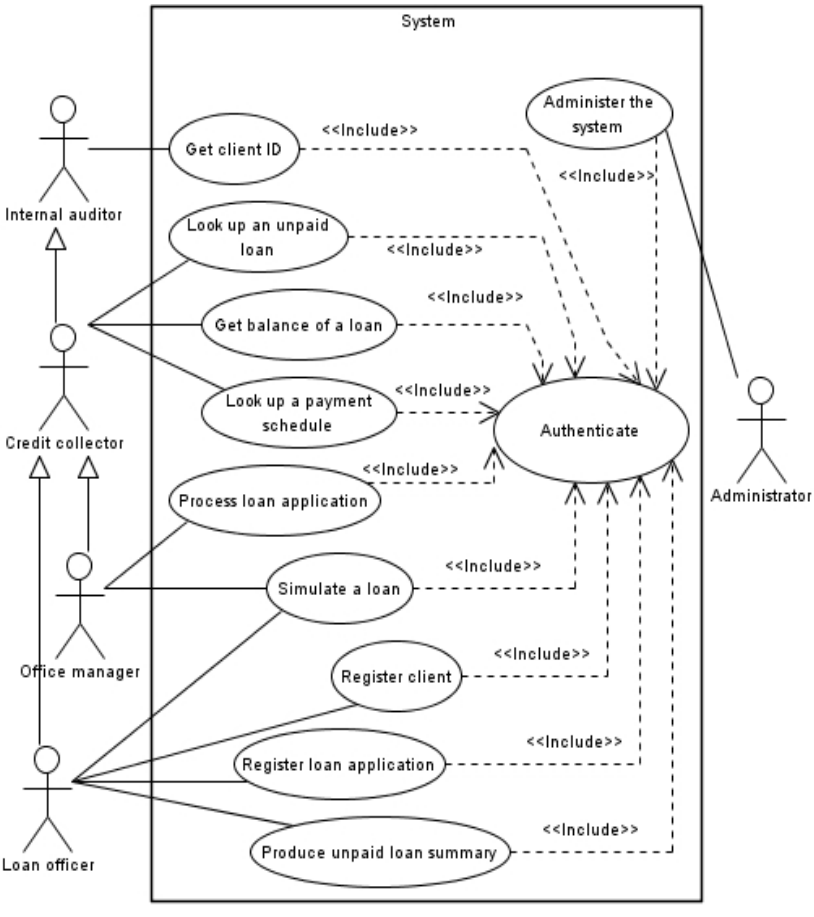


Fig. 1. Use case diagram

The use cases are as follows:

- Get client Id : Get information on the identity of a client
- Look up an unpaid loan : Identify an unpaid loan
- Get balance of a loan : Get the balance of a loan
- Look up a payment schedule : Get information on the payment schedule
- Process loan application : Put in place the credit applied for or reject loan application
- Simulate a loan : Get an overview of information about a loan

- Register client : Register information about a customer
- Register loan application : Register information about a loan application
- Produce unpaid loan summary : Get the list of expected reimbursements
- Administer the system : Administer the system
- Authenticate : Verify the identity of user

The sequence diagram shows the chronological sequence of operations performed by an actor. It shows the objects that the actor will handle and operations that move from one object to another. Although it is a variant of the collaboration diagram, it favors the temporal representation from the spatial representation and it is a good way to model the dynamic aspects of the system. The sequence diagram for "Simulate a loan" use case is in Fig. 2.

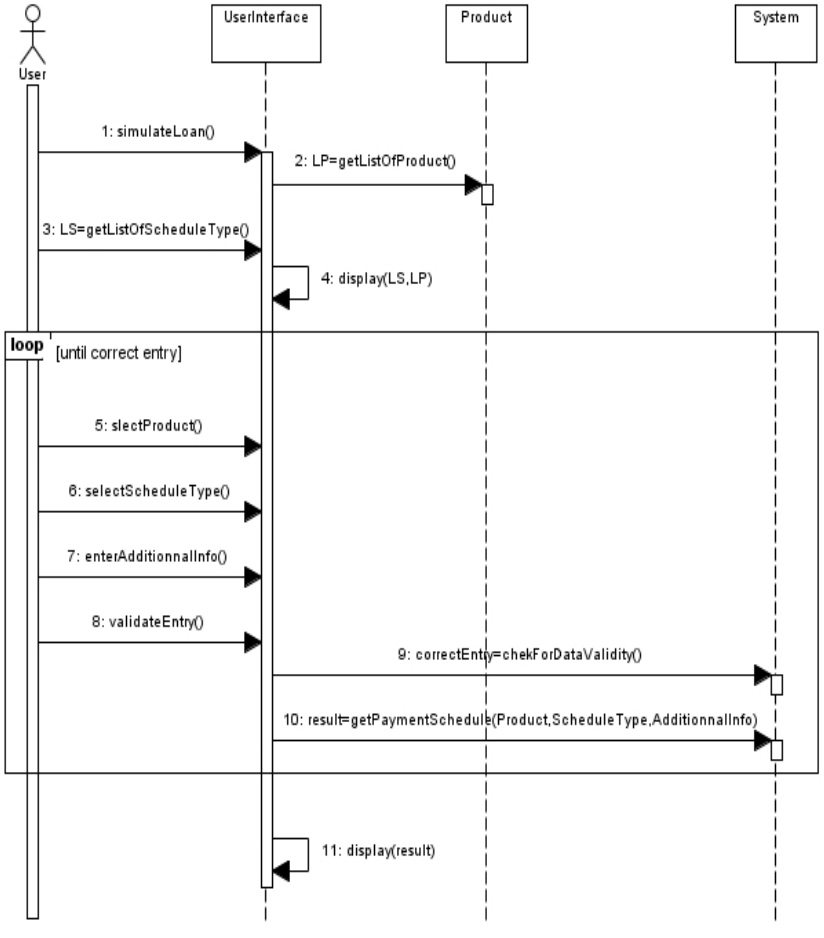


Fig. 2. Sequence diagram of the use case "Simulate a loan"

At the static model, the class diagram is usually chosen as it represents the conceptual architecture of the system. Class diagram is drawn in Fig. 3. Unlike the use case diagram showing the system in terms of actors, it expresses the static structure of the system in terms of classes and relationships between classes. The interest of the class diagram is to model the entities of the information system.

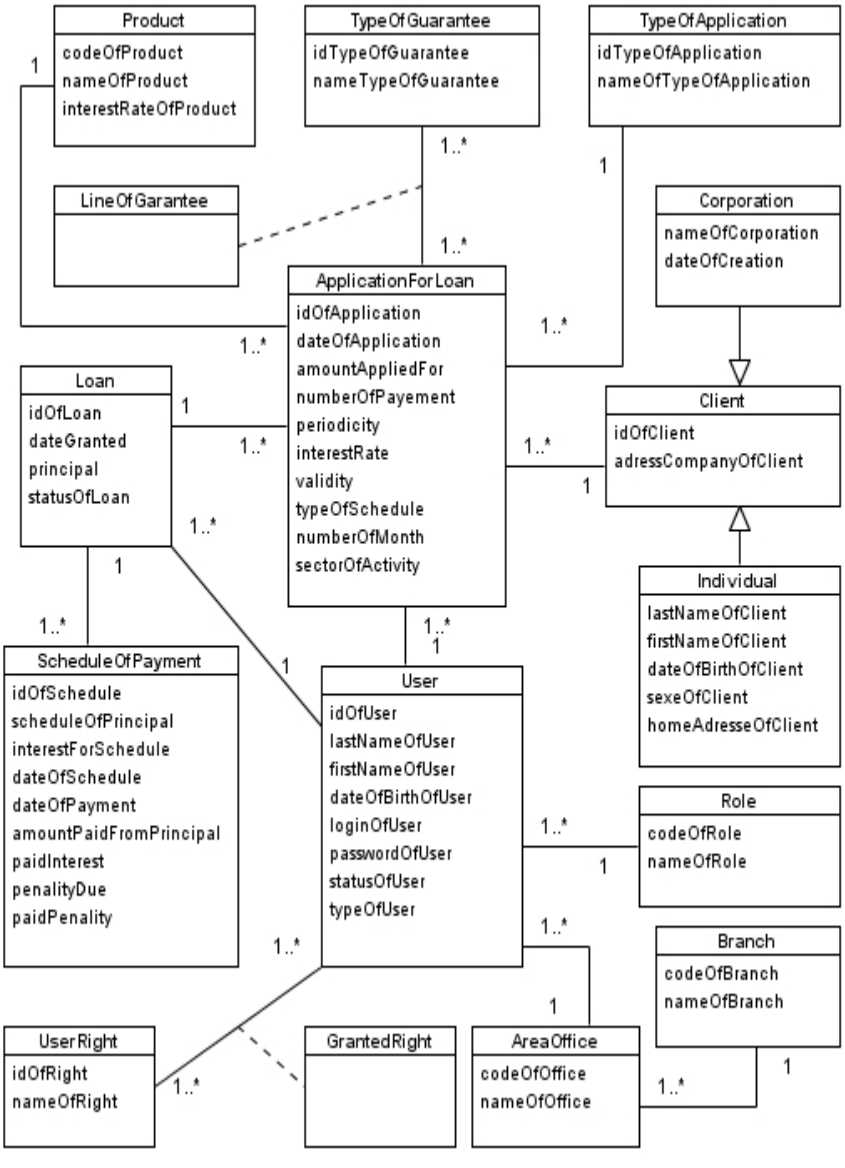


Fig. 3. Domain model **c**lass diagram

In our study, the following classes are identified:
User : Users properties
Product : Products offered to customers
ApplicationForLoan : Information about loan Applications
Loan : Information about Loans granted
ScheduleOfPayment : schedule for reimbursement
Individual : Information on clients that are Individuals
Corporation : Information on clients that are Organization
Client : Beneficiaries of credit
TypeOfApplication : Types of applications
LineOfGuarantee : Guarantee accepted for a loan
TypeOfGarantee : Types of guarantee
RoleOfUser : Function held by a user of the system
AreaOffice : Area office
UserRight : User right
GrantedRight: Rights allocated to certain users of the system
Branch : Information about branch

4 Implementation of the Case Study

An n-tier client/server and service-oriented application is implemented (see Fig.4).

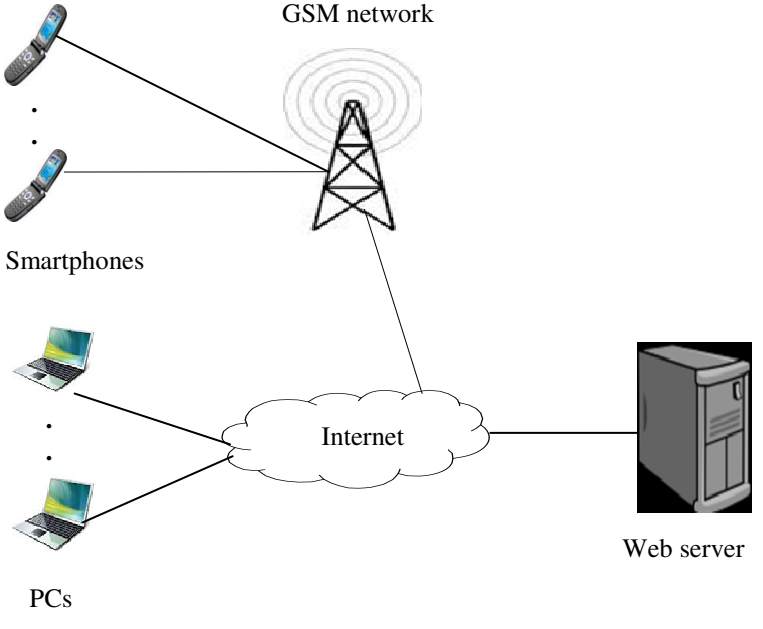


Fig. 4. Block diagram for the proposed solution

The web service methods will be available from the client application. Calling the web service methods will be from Smartphone or PC while the execution will be on the server. The exchange of messages will be provided by SOAP protocol.

From the software point of view, the architecture is based on MVC (Model-View-Controller) [16]. By applying the MVC architecture to our application, we separate core business model functionality from the presentation and control logic that uses this functionality. Such separation allows multiple views to share the same enterprise data model, which makes supporting multiple clients easier to implement, test, and maintain.

The tools used for development are: NetBeans [17] as the Integrated Development Environment (IDE) for the development of J2ME mobile client, PHP5 [20] as the programming language for the web client, PDO (PHP Data Objects) [21] to prevent SQL injections. The database is implemented in MySQL [15]. A SOAP client (for the PC) and a SOAP server are created in PH5. In the case of the mobile client, a J2ME SOAP client is built using KSOAP [12]. This library contains a number of functions for creating or analyzing a SOAP object. It allows creating the instance of the SOAP client. The application for the web client is accessible from a web browser by providing the URL used. Fig.5 shows results on Smartphone for the French version of the solution.

Fig. 5. Results with Smartphone

5 Related Works and Discussion

Software systems of enterprise class usually support business processes and business rules existing in a given domain. Approaches used today to business rules implementation (like in this paper) are very sensitive to changes. Some authors propose to separate business logic layer from business rule layer by introducing an integration layer [8]. A mechanism usually used for separation is Aspect Oriented Programming pioneered by Kiczales et al. [11], and one of the most popular Aspect Oriented Programming languages is AspectJ [2], [13]. As argued by Hnatkowska et al. [8], aspect-oriented languages are rather difficult, so the source code of intermediate layer is complex and hard to understand. They then present a domain specific language (DSL) where models written in the DSL are automatically translated to AspectJ source code. They also pointed out the advantages and disadvantages of the DSL. As advantages, it allows to define connections between rules and business process at higher abstraction level in a declarative way: the syntax is easy and very flexible. The main disadvantage of DSL is the need to know the business classes, relationships among them, the semantics of their methods and the interactions among instances. Therefore, they suggest that the direction of further research should be a formalization of business rules and business processes that allow abstracting from their concrete implementations.

In relation to the security aspect of this work, each business has its own security infrastructure and mechanism, and web services have specialized security requirements. The web services security architecture defines an abstract model for that purpose. The web services security specifications are standardized by OASIS.

In the case of mobile applications, new approaches to security include «reversible data hiding» and «visual cryptography». Huang et al. [9] proposed a new algorithm in reversible data hiding, with application associated with the quick response (QR) codes. The goal of the QR codes aims at convenience oriented applications for mobile users. Visual cryptography also called visual secret sharing is a technique in which a secret is encrypted into several share images and then decrypted later using a human visual system to stack all the share images. Wang et al. [25] proposed a novel approach to visual cryptography for binary images that includes the capabilities of watermarking and verification.

6 Conclusions and Future Work

In this work, technologies related to Web services and J2ME have been experimented. The prototype implementation proposed and developed has been tested live and the results prove encouraging. Of course, there are still improvements to be integrated. In perspective, we can improve our application in terms of security because SOAP messages transmitted over the HTTP protocol are not encrypted, which is a significant weakness. Other aspects to be considered include: disconnected mode and synchronization. The proposed solution is based on the procedures manual of PADME in Benin Republic, but may under minor changes apply to any other business of microfinance.

References

1. Amsden, J.: Modeling SOA, parts I–V. IBM developerWorks (October 2007)
2. AspectJ, http://www.eclipse.org/aspectj
3. Arsanjani, A.: Service-oriented modeling and architecture. IBM developerWorks (November 2004)
4. Belouadha, F.-Z., Roudiès, O.: Un profil UML de spécification de services web composites sémantiques. In: Colloque Africain sur la Recherche en Informatique et en Mathématiques Appliquées, Rabat, pp. 537–544 (2008)
5. Bernard, G., Ben-Othman, J., Bouganim, L., Canals, G., Defude, B., Ferrié, J., Gançarski, S., Guerraoui, R., Molli, P., Pucheral, P., Roncancio, C., Serrano-Alvarado, P., Valduriez, P.: Mobilité et bases de données, Etat de l'art et perspectives, http://www-smis.inria.fr/dataFiles/P03a.pdf
6. Dagba, T.K.: An UML based framework for online and non-traceable E-cash system. In: IEEE 5th International Conference on Broadband and Biomedical Communications, Malaga (2010); doi:10.1109/IB2COM.2010.5723630
7. GFI Informatique: Informatique Mobile, http://www.gfi.fr/gfilabs/common/docs/Offre-Technologique-Informatique-Mobile.pdf
8. Hnatkowska, B., Kasprzyk, K.: Integration of Application Business Logic and Business Rules with DSL and AOP. e-Informatica Software Engineering Journal 4, 59–69 (2010)
9. Huang, H.C., Chang, F.C., Fang, W.C.: Reversible Data Hiding with Histogram-Based Difference Expansion for QR Code Applications. IEEE Trans. Consumer Electronics 57, 779–787 (2011)
10. Johnston, S.: UML 2.0 Profile for Software Services. IBM developerWorks (April 2005)
11. Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C., Loingtier, J.-M., Irwin, J.: Aspect-Oriented Programming. In: Aksit, M., Auletta, V. (eds.) ECOOP 1997. LNCS, vol. 1241, pp. 220–242. Springer, Heidelberg (1997)
12. KSOAP, http://sourceforge.net/projects/ksoap2
13. Laddad, R.: AspectJ in Action: Practical Aspect-Oriented Programming. Manning Publications (2003)
14. Lohoume, E.: Conception et réalisation d'une application de gestion de crédits sur plate-forme mobile pour les institutions de microfinance. MSc thesis, Ecole Nationale d'Economie Appliquée et de Management, Université d'Abomey-Calavi, Cotonou Bénin (2011)
15. MySQL, http://www.mysql.org
16. Model-View-Controller, http://java.sun.com/blueprints/patterns/MVC-detailed.html
17. Netbeans, http://www.netbeans.org
18. Object Management Group, http://www.omg.org
19. PADME, http://www.padmebenin.org
20. PHP 5, http://www.php.net
21. PHP Data Objects, http://php.net/manual/en/book.pdo.php
22. SOAP version 1.2, http://www.w3.org/TR/soap12
23. Rychly, M.: A Case Study on Behavioural Modelling of Service-Oriented Architectures. e-Informatica Software Engineering Journal 4, 71–87 (2010)
24. UDDI version 3, http://uddi.org/pubs/uddi_v3.htm
25. Wang, Z.-H., Chang, C.-C., Tu, H.N., Li, M.-C.: Sharing a Secret Image in Binary Images with Verification. Journal of Information Hiding and Multimedia Signal Processing 2, 78–90 (2011)
26. WSDL 2.0, http://www.w3.org/TR/wsdl20
27. WSDM, http://www.oasis-open.org/commitees/wsdm
28. World Wide Consortium, http://www.w3.org

An Approach to CT Stomach Image Segmentation Using Modified Level Set Method

Hersh J. Parmar and S. Ramakrishnan

Non-Invasive Imaging and Diagnostics Laboratory, Biomedical Engineering Group
Department of Applied Mechanics
Indian Institute of Technology Madras, Chennai, India
hershparmar@gmail.com, sramki@iitm.ac.in

Abstract. Internal organs of a human body have very complex structure owing to their anatomic organization. Several image segmentation techniques fail to segment the various organs from medical images due to simple biases. Here, a modified version of the level set method is employed to segment the stomach from CT images. Level set is a model based segmentation method that incorporates a numerical scheme. For the sake of stability of the evolving zero'th level set contour, instead of periodic reinitialization of the signed distance function, a distance regularization term is included. This term is added to the energy optimization function which when solved with gradient flow algorithms, generates a solution with minimum energy and maximum stability. Evolution of the contour is controlled by the edge indicator function. The results show that the algorithm is able to detect inner boundaries in the considered CT stomach images. It appears that it is also possible to extract outer boundaries as well. The results of this approach are reported in this paper.

Keywords: CT, stomach, segmentation, level set, regularization.

1 Introduction

There are quite a few modalities in medical imaging that depend on the source of the energy that is used to craft the image on the sensor. Some of the modalities that are presently used for medical purposes comprise x-ray, computed tomography, magnetic resonance imaging, ultrasound, optical imaging, positron emission tomography, infrared imaging and many more.

Medical image segmentation involves extraction of the region of interest from medical images for further processing and diagnostic procedures. Process of segmentation in medical images particularly is very tedious, due to which many new approaches have been developed and implemented successfully.

This paper is organized as follows. In Section 2 we take a look at some of the papers in literature based on image segmentation. Section 3 deals with the methodology followed for our work. The preliminary results are presented in Section 4 and we conclude with the conclusions in Section 5.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 227–233, 2012.

2 Literature Review

Classical image segmentation techniques can be narrowly classified into the model driven and data driven types. While the model driven techniques [1-4] extract information concerning objects and build models for the segmentation procedure, the data driven techniques [5-8], on the other hand, incorporate stochastic and/or statistical calculations over image information such as pixel intensities, histogram, spatial correlation etc. Simple methods such as thresholding and detection of edges may not yield an optimal segmentation output, reason behind this being the possiblity of large scale intensity inhomogeneties present in the input image, which is typical in the case of medical images.

Li Xinwu performed volumetric segmentation of medical images in [9] using clustering method. Bianca Lassen, Jan-Martin Kuhnigk, Ola Friman, Stefan Krass and Heinz-Otto Peitgen automatically segmented the lung lobes in CT images through the watershed transformation approach [10]. Ilya Kamenetsky, Rangaraj M. Rangayyan and Hallgrimur Benediktsson explored the split and merge method [11] for segmenting the glomerular basement membrane. Steven Lobregt and Max A. Viergever developed a segmentation technique called discrete dynamic contour model which they tested over X-Ray, CT and MR images in their work [12]. Epidermis tumors in stomach were segmented by application of gradient vector flow snake model in [13] by Zhang Hongbin and Li Guangli. Li Guangli also worked on boundary tracing algorithm [14] to analyze stomach cancer. Study on stomach dynamics were carried out by Mohammed Benjelloun in [15] by applying active contours. Stanley Osher and James Sethian in [16] designed the level sets for tracking deformable isosurfaces, a model driven technique, that relies on partial differential equations to model the evolving contour or surface. Chunming Li, Chenyang Xu, Changfeng Gui and Martin D. Fox [17] introduced a distance regularization term in their work to restrain the abnormalities that arise during the evolution of the level set function. This distance regularization term maintains a signed distance profile near the zero level set contour, thereby eliminating the necessity for reinitialization and avoiding its induced numerical errors and instability.

In this paper, CT stomach images are considered for segmentation using the modified level set method.

3 Methodology

The input CT images used for our purpose were downloaded from an online public domain database.

In two dimensions, the level set method evolves a closed contour by manipulating the level set function [18] $f(X,t)=0$; which is a higher dimension function. Here $X=(x,y)\in R^2$ and t represents time or iterations. The level set function has the following properties [18]:

$$\begin{aligned} f(X,t) &< 0 \quad \text{for } X \in \Omega \\ f(X,t) &> 0 \quad \text{for } X \notin \Omega \\ f(X,t) &= 0 \quad \text{for } X \in \partial\Omega = \Gamma(t) \end{aligned}$$

where Ω is the region bounded by the contour.

The evolving contour can be extracted from the zero level set Γ. Evolution of this contour is governed by a level set equation. By iteratively updating $f(X,t)$ at each time interval, the solution to the partial differential equation can be computed.

The general form of the level set equation is shown below [18].

$$\frac{\partial f}{\partial t} + v.\nabla f = 0 \tag{1}$$

where v is the desired velocity on the zero level set Γ and is arbitrary elsewhere and ∇ is the gradient of the zero level set Γ.

Equation (1) can be written as [18]:

$$\frac{\partial f}{\partial t} + F \mid \nabla f \mid = 0 \tag{2}$$

where $F = v.\frac{\nabla f}{\mid \nabla f \mid}$, is the velocity term that describes the level set evolution.

The level set contour evolves on the basis of minimizing an energy function that is defined as [17]:

$$\varepsilon(f) = \mu D(f) + \lambda L(f) + \alpha A(f) \tag{3}$$

$\mu > 0$ is a constant and $D(f)$ is the distance regularization term given by [17]:

$$D(f) = \int_{\Omega} p(\nabla f) dX \tag{4}$$

where p is the double well potential function given by [17]:

$$p(s) = \begin{cases} \frac{1}{(2\pi)^2}(1 - cos(2\pi s)); & \text{if } s \le 1 \\ \frac{1}{2}(s-1)^2; & \text{if } s \ge 1 \end{cases} \tag{5}$$

and s is a spatial parameter in [0,1]. This potential function has two minimum points at $s = 0$ and $s = 1$.

$\lambda > 0$ is the co-efficient of the (line integral) energy function $L(f)$ which is defined by [17]:

$$L(f) = \int_{\Omega} e\delta(f) \mid \nabla f \mid dX \tag{6}$$

α is the co-efficient of the (area integral) energy function $A(f)$ which is defined by [17]:

$$A(f) = \int_{\Omega} eH(-f)dX \tag{7}$$

δ and H are the Dirac delta function and the Heaviside function respectively.

e is the edge indicator function defined as [19]:

$$e = \frac{1}{1+|\nabla G * I|^2} \tag{8}$$

where G is the Gaussian kernel with a standard deviation σ, I is the image and (*) is the convolution operator.

The Gâteaux derivative of the distance regularized function $D(f)$ is given by [17]:

$$\frac{\partial D}{\partial f} = -\mu div(d_p(|\nabla f|)\nabla f) \tag{9}$$

where $div(.)$ is the divergence operator and d_p is a function given by [17]:

$$d_p(s) = \frac{p'(s)}{s} \tag{10}$$

Usually the Dirac delta function δ and the Heaviside function H are approximated by the following smooth function [20]:

$$\delta_{\varepsilon}(x) = \begin{cases} \frac{1}{2\varepsilon}(1+\cos(\frac{\pi x}{\varepsilon})); & \text{for } x \le \varepsilon \\ 0; & \text{for } x > \varepsilon \end{cases} \tag{11}$$

$$H_{\varepsilon}(x) = \begin{cases} 0; & \text{for } x < -\varepsilon \\ \frac{1}{2}(1+\frac{x}{\varepsilon}+\frac{1}{\pi}\sin(\frac{\pi x}{\varepsilon})); & \text{for } x \le \varepsilon \\ 1; & \text{for } x > \varepsilon \end{cases} \tag{12}$$

Substituting the smooth functions for the Dirac delta function and the Heaviside function into (6) and (7), the energy function is given by [17]:

$$\varepsilon_{\varepsilon}(f) = \mu \int_{\Omega} p(\nabla f)dX + \lambda \int_{\Omega} e\delta(f)|\nabla f|dX + \alpha \int_{\Omega} eH(-f)dX \tag{13}$$

This energy function is minimized to get the optimal result. The solution is obtained by solving the gradient flow given below [17].

$$\frac{\partial f}{\partial t} = \mu div(d_p(|\nabla f|)\nabla f) + \lambda \delta_\varepsilon(f) div(e\frac{\nabla f}{|\nabla f|}) + \alpha e \delta_\varepsilon(f) \tag{14}$$

The initial contour f_0 is user defined and placed approximately either inside or outside the liver. If the initial contour is placed inside the liver, α is initialized with a negative value else it is initialized with a positive value.

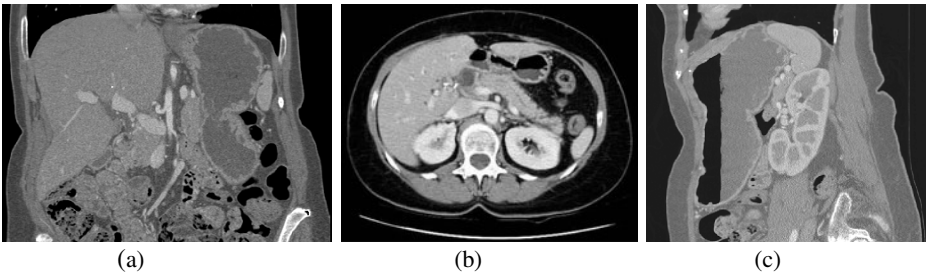
f_0 is a step function defined as [17]:

$$f_0(X) = \begin{cases} -K; & \text{for } X \in Z \\ K; & \text{otherwise} \end{cases} \tag{15}$$

where K is a constant greater than zero and Z is a region in the domain R^2.

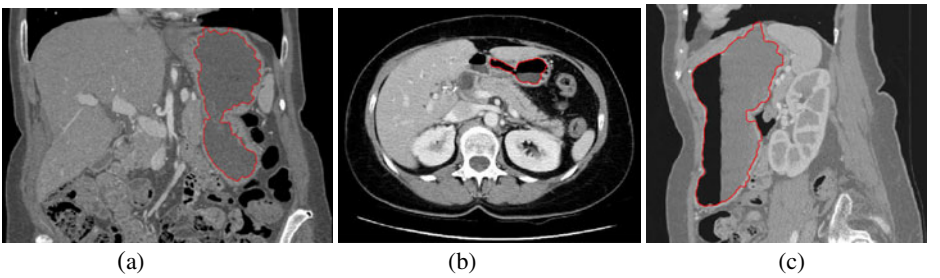
4 Preliminary Results

The original CT stomach images over which the initial contour are input are shown in Fig. 1.a,b and c. Fig. 1.a is of a CT scan viewed in transverse plane whereas Fig 1.b and c are viewed in the axial and sagittal plane respectively. The corresponding segmented output images obtained are shown in Fig. 2.a,b,c respectively. It can clearly be seen that the final zero level set contour demarcates the inner boundary of the stomach without any error and leaks.



(a) (b) (c)

Fig. 1. (a,b,c) are the input images



(a) (b) (c)

Fig. 2. (a,b,c) are the respective segmented output images

5 Conclusion

The proposed algorithm shows appreciable preliminary segmentation results. This method has been assessed successfully in segmenting the liver from CT images in [21]. From the experiments conducted, we can conclude that the evolution of the zero level set contour is influenced by pixel intensity gradient, which is considered in the edge detector function.

Further work on extending this algorithm to extract outer boundaries of the stomach, detect stomach tumors in CT images and the likelihood of performing volumetric estimations are being considered.

References

1. Lei, T., Udupa, J.K.: Performance Evaluation of Finite Normal Mixture Model-Based Image Segmentation Techniques. IEEE Transactions on Image Processing 12(10), 1153–1169 (2003)
2. Sarkar, A., Biswas, M.K., Sharma, K.M.S.: A Simple Unsupervised MRF Model Based Image Segmentation Approach. IEEE Transactions on Image Processing 9(5), 801–812 (2000)
3. Kim, S., Yoo, S., Kim, S., Kim, J., Park, J.: Segmentation of kidney without using contrast medium on abdominal CT image. In: Book 5th International Conference on Signal Processing Proceedings, Beijing, vol. 2, pp. 1147–1152 (2000)
4. Fujimoto, H., Gu, L., Kaneko, T.: Recognition of Abdominal Organs Using 3D Mathematical Morphology. System and Computers in Japan 33, 75–83 (2002)
5. Soler, L., Delingette, H., Malandain, G., Montagnat, J., Ayache, N., Clement, J.M., Koehl, C., Dourthe, O., Mutter, D., Marescaux, J.: Fully automatic anatomical, pathological, and functional segmentation from CT scans for hepatic surgery. Computer Aided Surgery 6, 131–142 (2001)
6. Lamecker, H., Zachow, S., Haberl, H., Stiller, M.: Medical Applications for Statistical 3D Shape Models. In: Weber, S. (ed.) Book Computer Aided Surgery Around the Head, vol. 17, p. 61. VDI, Berlin (2005)
7. Patel, R., Raghuwanshi, M.M., Shrawankar, U.N.: Genetic Algorithm with Histogram Construction Technique. Journal of Information Hiding and Multimedia Signal Processing 2(4), 342–351 (2011)
8. Lin, T.C., Huang, H.C., Liao, B.Y., Pan, J.S.: An Optimized Approach on Applying Genetic Algorithm to Adaptive Cluster Validity Index. International Journal of Computer Sciences and Engineering Systems 1(4), 253–257 (2007)
9. Li, X.: Research on Volume Segmentation Algorithm for Medical Image Based on Clustering. In: 2008 International Symposium on Knowledge Acquisition and Modeling, pp. 624–627 (2008)
10. Lassen, B., Kuhnigk, J.-M., Friman, O., Krass, S., Peitgen, H.-O.: Automatic segmentation of lung lobes in CT images based on fissures, vessels, and bronchi. In: 2010 IEEE International Symposium on Biomedical Imaging, pp. 560–563 (2010)
11. Kamenetsky, I., Rangayyan, R.M., Benediktsson, H.: Segmentation and analysis of the glomerular basement membrane using the split and merge method. In: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2008, pp. 3064–3067 (2008)

12. Lobregt, S., Viergever, M.A.: A Discrete Dvnamic Contour Model. IEEE Transactions on Medical Imaging 14(I), 12–24 (1995)
13. Zhang, H., Li, G.: Application in Stomach Epidermis Tumors Segmentation by GVF Snake Model. In: 2008 International Seminar on Future Bio Medical Information Engineering, pp. 453–456 (2008)
14. Li, G.: Implement for Stomach Epidermis Tumor Diagnosis Based on Watershed Algorithm and Boundary Tracing Algorithm. In: 2009 Third International Symposium on Intelligent Information Technology Application, pp. 704–707 (2009)
15. Benjelloun, M.: Segmentation and feature extraction to evaluate the stomach dynamic. In: Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV 2005), pp. 437–443 (2005)
16. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. 79(1), 12–49 (1988)
17. Li, C., Xu, C., Gui, C., Fox, M.D.: Distance Regularized Level Set Evolution and Its Application to Image Segmentation. IEEE Transactions on Image Processing 19, 3243–3254 (2010)
18. Osher, S., Paragios, N.: Geometric Level Set Methods in Imaging, Vision and Graphics. Springer Science (2006)
19. Malladi, R., Sethian, J.A., Vemuri, B.C.: Shape modeling with front propagation: A level set approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 158–175 (1995)
20. Zhao, H.K., Chan, T., Merriman, B., Osher, S.: A variational level set approach to multiphase motion. J. Comput. Phys. 127(1), 179–195 (1996)
21. Parmar, H.J., Ramakrishnan, S.: Segmentation of Liver in Computed Tomography Images using Distance Regularized Edge Based Level Set Method. In: Hersh, J. (ed.) International Conference on Biomedical Engineering 2011, pp. 247–251 (2011); ISBN: 978-81-8487-195-1

On Fastest Optimal Parity Assignments in Palette Images

Phan Trung Huy[1], Nguyen Hai Thanh[2], Tran Manh Thang[1], and Nguyen Tien Dat[1]

[1] Hanoi University of Science and Technology
phanhuy@hn.vnn.vn, huypt-fami@mail.hut.edu.vn
[2] Ministry of Education and Training
nhthanh@moet.gov.vn

Abstract. By Optimal Parity Assignment (OPA) approach for hiding secret data in palette images, we consider the fastest Optimal Parity Assignment (FOPA) methods which are OPA and do not take any extra reordering procedures on color space of palettes. We show that our rho-method is a FOPA and it is easily implemented for controlling quality of stego-images. As consequences, two algorithms for hiding data in palette images are presented. The first algorithm is taken on the rho-forest obtained by rho-method on a palette and the second is only on the colors which are not isolated (by distance not far from the others) in the palette. To prevent from steganalysis, by controlling high quality of stego palette images, combinations of FOPA method with enhanced CPT schemes for binary images are introduced. Some experimental results are presented.

Keywords: palette image, steganography, rho-method, fastest OPA, enhanced CPT scheme.

1 Introduction

For hiding secret data in palette images, by *Parity Assignment (PA)* approach [4], on the palette P (such as in 8bpp bmp file format in common) of a given palette image G, with some distance d on the colors of P, ones try to split P into two disjoint parts, P_1 and P_2, by a parity function Val: $P \rightarrow E$ from P onto the set E={0,1}, $Val(P_1)$={1} and $Val(P_2)$={0}, Together with the Val, ones need to find a function $Next$: $P \rightarrow P$ satisfying $Next(P_1) \subseteq P_2$ and $Next(P_2) \subseteq P_1$ and minimize the color distances $d(Next(c), c)$ for all $c \in P$. If these claims are satisfied, they immediately imply the minimality of *the cost of the method - which is seen as the sum.*

$Cost(Next) = \sum_{c \in P} d(Next(c),c)$ and we say that this couple of the functions *Next, Val* satisfies *an optimal parity assignment (OPA)* for the given image G and we make use of this to hide secret data in G. Suppose that in G the pixels had been arranged by some linear order. In the step of hiding secret data S as a stream of bits into G, consecutively each bit b of S is "hidden" in a pixel having color x in G as follows: x is kept intact if b=0, otherwise x is changed to $Next(x)$. In the extracting step, to obtain S, each secret bit b obtained easily by taking $b=Val(x)$ from the color x of the corresponding pixel. The optimality of the OPA method leads to the high quality of

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 234–244, 2012.

stego-image G' obtained from G after changing colors. In [4] an algorithm is introduced to obtain an OPA for a given palette image based on the complete graph G_P of the color space P. However, in this algorithm, one step consumes a large amount of time is the step to reorder the set of edges in ascending direction. This step takes time complexity $O(N^2LogN^2)$ with $N=n.(n-1)/2$ edges, where n is the number of colors in P, if for instance the quick sort procedure is applied. The *EzStego method* [5,6] provides examples for PA approach. By virtue of this method, on a selected Hamiltonian cycle H of G_P (P has 256 vertices), on a chosen direction of H, a *Next* function for P is obtained: for each color c in P, $Next(c) = c'$, where c' is the next color of c in this direction. Since the number 256 of colors is even, it is easily to assign a parity function *Val* on P by this *Next* function. However, it is often the case that H is not minimal, this *Next* function is not OPA and the cost of EzStego method is larger than the optimal, as shown in [9]. Considering the whole process to get H, the run-time for the full algorithm is larger. In this paper, we recall our rho-method which is considered as a *near OPA* in [9] and prove that it is actualy an OPA method, the rho-method provides a way to obtain a rho-forest in a fastest way, that is a FOPA method, as shown in our new FOPA algorithms. By their advantages, two algorithms are proposed for hiding secret data in palette image, in any case that the number of colors in palette is even or odd. The first FOPA algorithm is for getting a *Next* function on the whole palette. The second is a modified from the first to obtain a *Next* function on only a part of palette which contains colors not isolated (colors not far from the others by some chosen distance) for controlling high quality of stego-image. To prevent attacks from steganalysis, specially to histogram-based attacks (see for examples some analysis in [7,8], if the alpha ratio of the number of changed pixels to the number of total pixels of a given palette image is lower than 0.1, it is very difficult to guess if the image contains hidden data). Hence, combined with FOPA methods the CPT scheme [1] on binary images or some enhanced CPT scheme (ECPT) can be used to build new schemes on palette images for controlling high quality of stego palette images. Applying these schemes to a concrete palette permits us to gain the small alpha ratio while amount of total hidden bits is large enough. For these purposes, in the part 2 of the paper, the rho- method [9] and OPA method [4] are recalled. In the part 3 our FOPA algorithm is introduced. In the part 4, we propose the second FOPA algorithm with threshold by modifying the first to get a *Next* function on colors which are not isolated. In the part 5, a combination of two algorithms in parts 3 and 4 with ECPT schemes in [10] is introduced for palette images for controlling high quality of stego-images and some experimental results for palette image are presented. The part 6 reserves for conclusion and discusion of future works in relation with other works [11,12].

2 Rho-Forest and Rho-Method

Given a palette image G with its palette $P = \{c_1, c_2, .., c_n\}$ of $n \geq 2$ colors, P is considered as the set of vertices of the weighted complete undirected graph G_P combined with P and some distance function d is chosen on the set of all colors (for instance the Euclidean distance on the vectors of three components Red, Green,

Blue). Each pair (c_i,c_j) is an edge of G_P having the weight $d(c_i,c_j)$. We try to get the optimal *Next* function (for all mentioned methods): $Next$: $P \rightarrow P$ from the condition

$d(c,Next(c)) = \text{Min}_{c \neq x \in P}\, d(c, x)$ together with a function Val: $P \rightarrow E=\{0,1\}$ satisfying $Val(x)=1$ XOR $Val(Next(x))$,$\forall x \in P$.

Let us recall briefly the essence of OPA method in the weight complete undirected graph G_P combined with the palette P of a palette image G [4].

2.1 OPA Method

In terms of the minimal cover forest problem (optimal cover forest) and of the notions *Next* and *Val* function, the algorithm can be expressed as follows.

Algorithm 1. [4] (Getting an optimal cover forest)

Step 1. Sort the set of all edges E by increase in weights.

Step 2. Initiation, put $C=\{c_i, c_j\}$, $c_i \neq c_j \in P$ so that $d(c_i,c_j)$ is the smallest edge; Define *Next* and *Val* on C: $Next(c_i)=c_j$, $Next(c_j)=c_i$, set $Val(c_i)=0$ or 1 and set $Val(c_j)=1$ XOR $Val(c_i)$ randomly (C as the forest F_P with c_i or c_j is chosen as the root of F_P).

Step 3. (loop)

while $C \neq P$ **do**

a. Try to get one smallest edge (c_k,c_l) so that at least one of two c_k or c_l does not belong to C;

b. Append c_k and c_l to C if it is not in C, then set $Next(c_k)=c_l$, $Next(c_l)=c_k$; and set $Val(c_k)=1\text{-}Val(c_l)$, and if c_k , $c_l \notin C$ we assign $Val(c_k)=0$ or 1 randomly;

Endwhile;

Step 4. Return *Val*, *Next*;

Remark 1. By definition of OPA method we easily see that the Next and Val functions provide an OPA since $d(c,Next(c)) = \text{Min}_{c \neq x \in P} d(c,x)$.

2.2 Rho-Method

We call *a rho- path* a sequence $T=(x_1,x_2,..,x_k)$ of vertices of G_P which satisfies: either it contains a unique cycle: $x_i \neq x_j$ $\forall i \neq j < k$ and $x_k=x_t$ for some $1 \leq t < k\text{-}1$, or it does not contain any cycle (i.e. $x_i \neq x_j$ $\forall i \neq j$). A rho- path T *is said to be good* if it does not contain any cycle or the cycle $x_t,x_{t+1},..x_k$ with $k\text{-}t$ is even. Given a palette P of colors, a good rho-path is called *optimal* (for P) if for any edge(x,y) along the rho-path $d(x,y)=\text{Min}_{x \neq c \in P}\, d(x,c)$. A *rho- forest* can be defined inductively as follows:

- The empty set is defined as the smallest rho-forest
- A rho-path is a rho-forest.
- Disjoint union of two rho-forests is also a rho-forest.

- *Hooking* of a rho-path $T = x_1,x_2,..x_k$ into a rho-forest F is a rho-forest whenever x_k belong to F and $x_1,x_2,..,x_{k-1}$ do not. In that case we obtain an expanded rho-forest $F' = F$ union T.
- If F contains all vertices of G_P we say that F is a *cover rho- forest of* G_P, further more if $d(x,y)=\text{Min}_{x \neq c \in P} d(x,c)$ for any edge (x,y) of F, we call F a *minimal cover rho- forest* or *optimal cover rho- forest*. The *cost of a rho-forest* F is the sum of all weights of the edges in F.
- One cover rho- forest or cover forest F *is called near optimal* (*with error* ε) if the ratio of its cost to the cost of a minimal cover forest is not larger than $1+\varepsilon$.

It is easy to see that if an optimal cover rho- forest F exists, the cost of F is exactly the cost of a minimal cover forest in OPA method.

We recall briefly the algorithm in [9] to get a rho-forest F (which will be proved to be minimal in this paper) for the graph G_P without taking any extra reordering procedure on the set of edges of G_P as follows.

Algorithm 2. (finding cover rho- forest for a given palette *P*)

Step 1. (Initiation) Set $F=\varnothing$, $T=\varnothing$ and the set of rest vertices $C = P$.

Step 2. (Define the *Nearest* function)

For i = 1 to n **do**
Begin
+) Compute all $d(x_i,x_j)$, $j \neq i$, $x_i,x_j \in P$;
+) Find an $y = x_t$, $i \neq t$, so that $d(x_i,x_t) = \text{Min}_{1 \leq i \neq j \leq n} d(x_i,x_j)$, set $Nearest(x_i)=y$;
End;

Step 3. (loop, find a new good rho-path T then append it to F or hook T into F)

$imax$=0;
While $C \neq \varnothing$ **do**
Begin /* finding a new rho-path T */
a) choose randomly $x \in C$; Set $imax=imax+1$; $y_{imax}=x$;
mark y_{imax} as the root of a new rho-path T;
set $Val(y_{imax})$=0 or 1 randomly.
b) (loop, suppose y_{imax} is defined and belongs to T)
While T is still updated **do**
Begin
b1) find $x=Nearest(y_{imax})$,
b2) if ($x \notin C$) and (T hooks into F at x)
+ set $Next(y_{imax}) = x$; Mark y_{imax} as
the roof of T; /* $x \notin T$, x *in* F *already* */
+ if $Val(y_{imax}) = Val(x)$ then
invert $Val(y_t) = Val(y_t)$ XOR 1, $\forall y_t \in T$;
/* else keep $Val(y_t)$ intact, $\forall y_t \in T$ */
+ mark T can not update;
b3) if $x \in C$ and $x \notin T$:
+ set $y_{imax+1} = x$ and add x into T;

+ set $Next(y_{imax})= x$; /* = $Nearest(y_{imax})$ */
$Val(x)$=1 XOR $Val(y_{imax})$;
+ update $C=C-\{x\}$; $imax = imax+1$;

b4) if $x = y_t$ for some $y_t \in T$, $t<imax$ *then*
+ mark y_{imax} as the roof of T;
+ two cases happen:

b41) *imax-t* is even: T is good, assign T as a member of F *and* set $Next(y_{imax}) = y_t$;

b42) *imax-t* is odd:
Set $Next(y_{imax})= y_{imax-1}$;
/* Modified T becomes good with Next */.
Mark T can not update;
End; /* While- updating this T */

c) Add T to F *and reset* $T= \varnothing$;
End; /* While $C \neq \varnothing$ */

Step 4. Return *Next, Val*;

As shown in [9], if the case (b42) do not happen, $Next(y) = Nearest(y)$ for all y in the rho-forest F as the result of the algorithm, hence we obtain an OPA. In the following proposition, we prove that, even in case (b42) happens, we always obtain an OPA.

Proposition 1. For any cycle $x_t,x_{t+1},..,x_k$ appears in the case (b42) of the algorithm 2 above, all the weights of the edges along this cycle are equal, that is $d(x_t,x_{t+1})=d(x_{t+1},x_{t+2})= \ldots = d(x_{k-1},x_t)$.

Proof. Indeed, we consider any cycle $x_t,x_{t+1},..,x_k$ which appears on the way updating the rho-path T in the case (b42) of the algorithm 2 above, that is $x_i \neq x_j \ \forall i \neq j < k$ and $x_k=x_t$ for $t < k-1$, and $Next(x_i) = Nearest(x_i)=x_{i+1}$, that means $d(x_t,x_{t+1}) \geq d(x_{t+1},x_{t+2}) \geq \ldots \geq d(x_{k-1},x_k) = d(x_{k-1},x_t)$. These inequalities together with the equation $d(x_t,x_{t+1})= d(x_t,Nearest(x_t))$ imply that $d(x_t, x_{t+1})= d(x_{k-1}, x_t)= d(x_t ,x_{k-1})$, hence $d(x_t,x_{t+1})=d(x_{t+1},x_{t+2})= \ldots =d(x_{k-1},x_k) =d(x_{k-1},x_t)$. The proof is completed.

By this proposition, we deduce that in any case, even if the step (b42) in the algorithm 2 happened, the *Next* and *Val* functions also give us an OPA.

The following corollary is immediate.

Corollary 2. The *Next* and *Val* function obtained in the algorithm 2 provide an OPA**.**

3 Rho-Method as Fastest OPA Method

In this section, we modify the algorithm 2 to obtain a rho-forest F for an OPA, and show that the algorithm is a fastest. That means that it is a FOPA.

3.1 Rho-Forest as an OPA

From Corollary 2, by slightly modifying the algorithm above, we obtain following simple algorithm for rho-method.

Algorithm 3. All steps are the same as in the above algorithm 2, except the substep (b4) in Step 3 is modified as follows

Step 3. **b4)** if $x = y_t$ for some $y_t \in T$, $t<imax$ then
+ mark y_{imax} as the roof of T;
+ set $Next(y_{imax})= y_{imax-1}$;
+ mark T can not update;

3.2 A Fastest OPA Method

We call an OPA method for palette images a *fastest OPA (FOPA) method*, if in this method, after computing the closest distance from each color c of the palette P to the rest colors of P (which is necessary for all OPA methods), the value $c'= Next(c)$ of the *Next* function for this method is immediately defined.

The following algorithm is a modified of the algorithm 3 to get a FOPA.

FOPA Algorithm

Step 1. (Initiation) Given a palette P.
Set $F= \emptyset$, $T=\emptyset$ and the set of the rest vertices $C = P$.

Step 2. Set $imax$=0;
While $C \neq \emptyset$ **do**
Begin (setting new rho-path *T and add T to F*)
a) Choose ramdomly $x \in C$;
+ set $imax = imax+1; y_{imax}=x;$
+ mark y_{imax} as the root of a new rho-path
T; $T=\{y_{imax}\}$;
+ set $Val(y_{imax})$=0 or 1 randomly;
+ mark T can update;
b) (loop,suppose y_{imax} is defined and belongs to T)
While T can update **do**
Begin
b1) find x as the n*earest* of y_{imax} *where*
$y_{imax} \neq x \in P$;
b2) if ($x \notin C$) and ($x \in F$) then (T hooks into F at x)
b21) set $Next(y_{imax}) = x;$
mark y_{imax} as roof of T;
b22) if $Val(y_{imax}) = Val(x)$ then i*nvert*
$Val(y_t) = Val(y_t)$ XOR 1,$\forall y_t \in T$; /* else keep $Val(y_t)$ intact, $\forall y_t \in T$ */
b23) mark T can not update;
b3) if $x \in C$ then /* that is $x \notin T$ */
b31) set $y_{imax+1} =x$;

b32) set $Next(y_{imax})= x$; $Val(y_{imax+1})=$ 1 XOR $Val(y_{imax})$;
b33) update $C=C-\{x\}$; $T= T \cup \{x\}$; $imax = imax+1$;
b34) mark T can update;
b4) if $x = y_t$ for some $y_t \in T, t<imax$ then $\{x \notin C\}$
b41) mark y_{imax} as the roof of T;
b42) set $Next(y_{imax})= y_{imax-1}$;
b43) mark T can not update;
End;/* While- updating this T */
c) update $F = F \cup T$; *Reset* $T=\emptyset$;
End; /* While $C \neq \emptyset$ */

Step 3. Return *Next, Val*;

Let us remark that the main steps for evaluation of time complexity of this algorithm are substeps (2-b1) where they are repeated about n times, where n is the number of colors in P. In each substep, consists in computation of the n distances $d(x,y)$ between x and the remaining colors. Therefore, the time complexity of this algorithm is $O(n^2)$. It is easily seen that the results of this algorithm also give an OPA, and it satisfies the conditions for a FOPA, hence this algorithm provides us a FOPA.

Theorem 4. The algorithm 3 has a time complexity $O(n^2)$, where n is the number of colors in the palette. The *Next* and *Val* result functions constitute a FOPA.

4 FOPA Algorithm with Threshold

In some palette image G, there are colors in the palette whose distances to the others exceed a chosen threshold $\lambda > 0$ that is not safe to hide data in G after changing these colors. We call them *isolated colors* (*iso-color* for short) in the palette. By the flexible property of rho-method in the FOPA algorithm above, it is easily modified to get following *FOPA algorithm with threshold* where $d(x,Next(x))$ does not exceed λ for all x in the rho-forest F.

FOPA algorithm with threshold

Input the palette P, a threshold $\lambda>0$.
Output *Next* and *Val* functions.

Step 1. (initiation) Set $F=\emptyset$, $T=\emptyset$, the list of iso-colors $L=\emptyset$, $C=P$

Step 2. (repeat finding new rho-path T and append it to F, updating *Next* and *Val*)

Set $imax$=0;
While $C \neq \emptyset$ **do**
Begin /* setting new rho-path *T and append it to F* */
a) Repeat
a1) take any $x \in$ C; find the nearest y of x, $x \neq y \in P$;
a2) if $d(x,y) > \lambda$ then
mark x as iso-color and add x to L; update $C=C-\{x\}$;

until $d(x,y) \le \lambda$ *or* $C = \varnothing$;

b) if $C = \varnothing$ then goto step(3) else

c) set /* getting new rho-path T and append it to F after */

c1) $imax = imax+1; y_{imax} = x$; $Val(y_{imax}) = 0$ or 1 randomly;

c2) set $Next(y_{imax}) = y$; $T = \{y_{imax}\}$; $C = C - \{y_{imax}\}$ and mark y_{imax} as the root of T;

c3) If ($y \notin C$) and (T hooks into F at y)

+ mark y_{imax} as the roof of T; /* $y \notin T$, $y \in F$ */
+ if $Val(y_{imax}) = Val(y)$ then *invert* $Val(y_{imax}) = Val(y_{imax})$ XOR 1;
+ mark T can not update;

c4) if $y \in C$ then

+ set $y_{imax+1} = y$; $Val(y_{imax+1}) = 1$ XOR $Val(y_{imax})$;
+ update $C = C - \{y\}$; $T = T \cup \{y\}$; $imax = imax+1$ and mark T can update;

c5) **While** T can update **do** (loop, suppose y_{imax} is defined and belongs to T)

Begin

c51) find z as the *Nearest* of y_{imax} *where* $y_{\text{imax}} \ne z \in P$;

c52) if ($z \notin C$) and ($z \in F$) then

+ set $Next(y_{imax}) = z$ and mark y_{imax} as the roof of T;
+ if $Val(y_{imax}) = Val(z)$ then *invert* $Val(y_t) = Val(y_t)$ XOR 1, $\forall y_t \in T$;
+ mark T can not update;

c53) if $z \in C$ then

+ set $y_{imax+1} = z$; $Next(y_{imax}) = z$; $Val(y_{imax+1}) = 1$ XOR $Val(y_{imax})$;
+ update $C = C - \{z\}$; $T = T \cup \{z\}$; $imax = imax+1$;

c54) if $z = y_t$ for some $y_t \in T$, $t < imax$ *then*

+ mark y_{imax} as the roof of T, set $Next(y_{imax}) = y_{imax-1}$;
+ mark T can not update;

End; /* While can update this T */

d) update $F = F \cup T$; reset $T = \varnothing$;

End; /* While $C \ne \varnothing$ */

Step 3. If $L = P$ then Return "Failure" Else Return F, *Next, Val*;{success}

The following property is obvious from the algorithm.

Proposition 4. In the FOPA with threshold, for any x in P, x belongs to the rho-forest F if and only if x is not an iso-color. If x in F, then $Next(x)$ is also in F and $d(x, Next(x)) = Nearest(x) \le \lambda$.

Application to hiding and extracting data algorithms

For application of the FOPA algorithm with threshold, in practice, using proposition 4, we can "hide a secret bit *b"* in a pixel having color x if and only if $x \in F$. If this case happens, b can be hidden in this pixel as follows: x is kept intact if $b = Val(x)$ or x is changed to $x' = Next(x) \in F$ otherwise. Therefore, the algorithms use F to hide and extract secret data can be constructed easily. By restriction of the paper's framework, they are not written in details.

5 Experimental Results

5.1 CPTE Schemes for Palette Images

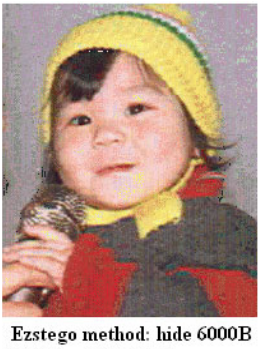
As an application of FOPA algorithms, to prevent steganalysis, we use two *enhanced CPT schemes* which are modifications of the *CPTE schemes* for binary images [10] to palette images as presented in [9]. These schemes can give stego-images with high quality, since only a small number of colors need to be changed in the images, and they are chosen randomly, while the number of hidden bits is large enough for real problems.

1. The results of FOPA algorithms – the *Next* and *Val* functions are used together with the CPTE1 scheme in [10] to build a new scheme for hiding and extracting data in palette images. In this scheme, we split a stego palette image G into blocks $F_1,F_2,..,F_N$, each such a block, say F, can be considered as a matrix $F=(F_{ij})$ of size $m \times n$ and containing $q=m.n$ pixels F_{ij}. For CPTE1 scheme, we consider the binary matrix $Val(F)=(Val(F_{ij}))$ of the same size $m \times n$. In each block F we can hide approximately $2r$-1 bits where $r = \lfloor \log_2(q+1) \rfloor$, by changing at most two entries of F, as in CPT scheme, while in CPT scheme [1] ones can hide only r bits. To change a pixel F_{ij} in F, we take $F'_{ij}=Next(F_{ij})$ so that $Val(F'_{ij})=1-Val(F_{ij})$ and this affects $Val(F)$ as in the scheme CPTE1 (see [10] for more details). Hence we can hide approximately $2r$-1 bits in each block F after changing at most two pixels, approximately $(2r-1)N$ bits in the whole G.
2. The results of FOPA algorithms with threshold λ– the *Next*, *Val* functions and the rho-forest F are used together with the modified scheme CPTE2 in [10] to build a new scheme, for which stego-images has very high quality if $\lambda \leq 25$.

5.2 Experimental Results for FOPA and CPTE Schemes

Results of our program are presented in the following BMP palette images.

Little singer: original palette image, BMP 8bpp, size 200x240



Ezstego method: hide 6000B PSNR = 27.596 dB

FOPA by Rho method: hide 6000B, PSNR = 33.54 dB

FOPA with threshold λ=25
hide 4855B, PSNR 35.2 dB

CPTE1 using FOPA, block size
5x6, hide 1584B, PSNR 42.57 dB

CPTE2 using FOPA with threshold λ=25
block size 5x6, hide 1107B, PSNR 44.52 dB

They are in 8bpp format. The original image "Little singer" has the size 200×240. In the two next images, we use the Next and Val functions obtained by Ezstego and FOPA with Rho method to hide 1 secret bit per pixel. In the fourth by the FOPA algorithm with threshold we hide at most 1 bit per pixel and in the last two images we apply CPTE1, CPTE2 schemes with block size 5×6, using *Next* and *Val* functions defined by FOPA algorithm, and FOPA algorithm with threshold $\lambda = 25$. The *PSNR* is defined by $PSNR = 10.\log_{10}((255)^2 / MSE)$ where for palette images 8bpp, *MSE* is the sum over all squared value differences of colors red, green, blue components of colors of pixels divided by image size and by three. The experiment results show that the third image by FOPA is better than the second by EzStego method for which one can see some distortion. The fourth, fifth have higher qualities. and the last image has the best quality and look smooth, according as their *PSNR*.

6 Conclusion

In the paper, we present two new FOPA algorithms, without threshold in the first and with threshold in the second to control high quality of stego-images, to prevent some attacks from revealing wherether the images are used to embed data or not. Two FOPA algorithms can also be applied flexibly together with any CPTE scheme to build new schemes, providing a very high quality of stego-images in case ones interest in quality of stego-images, or apply to other media format such as audio or video by its optimal characteristics. Applying FOPA algorithms together with the *Pseudo-Random Pixel Rearrangement Algorithm Based on Gaussian Integers* proposed by Aleksey Koval, Frank Y. Shih, and Boris S. Verkhovsky [11] (2011), one can improve even more the quality of stego-images together with high embedded data ratio. In the area of copyright protection, some authors as in [12] interest in techniques to embed metadata of images in. FOPA algorithms can provided new techniques with high quality. These directions will be deverloped in our future works.

References

1. Chen, Y., Pan, H., Tseng, Y.: A secret of data hiding scheme for two-color images. In: IEEE Symposium on Computers and Communications (2000)
2. Colour metric, http://www.compuphase.com/index_en.htm
3. Alman, D.H.: Industrial color difference evaluation. Color Research and Application (18), 137–139 (1993)
4. Fridrich, J., Du, R.: Secure Steganographic Methods for Palette Images. In: Pfitzmann, A. (ed.) IH 1999. LNCS, vol. 1768, pp. 47–60. Springer, Heidelberg (2000)
5. Romana Machado. EZStego[EB/OL], http://www.stego.com
6. Por, L.Y., Lai, W.K., Alireza, Z., Ang, T.F., Su, M.T., Delina, B.: StegCure: A omprehensive Steganographic Tool using Enhanced LSB Scheme. WSEAS Transactions on Computers 7(8) (August 2008); ISSN: 1109-2750 1309
7. Zhang, X., Wang, S.: Vulnerability of pixel-value differencing steganography to histogram analysis and modification for enhanced security. Pattern Recognition Letters 25, 331–339 (2004)
8. Zhang, X., Wang, S.: Analysis of Parity Assignment Steganography in Palette Images. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 1025–1031. Springer, Heidelberg (2005)
9. Phan, T.H., Nguyen, H.T.: On the maximality of secret data ratio in CPTE schemes. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part I. LNCS (LNAI), vol. 6591, pp. 88–99. Springer, Heidelberg (2011)
10. Thanh, N.H., Huy, P.T.: Fast and Near Optimal Parity Assignment in Palette Images with Enhanced CPT Scheme. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010, Part II. LNCS, vol. 5991, pp. 450–459. Springer, Heidelberg (2010)
11. Koval, A., Shih, F.Y., Verkhovsky, B.S.: A Pseudo-Random Pixel Rearrangement Algorithm Based on Gaussian Integers for Image Watermarking. Journal of Information Hiding and Multimedia Signal Processing 2(1), 60–70 (2011)
12. Huang, H.C., Fang, W.C.: Metadata-Based Image Watermarking for Copyright Protection. Simulation Modelling Practice and Theory 18(4), 436–445 (2010)

Setting Shape Rules for Handprinted Character Recognition

Daw-Ran Liou, Chia-Ching Lin, and Cheng-Yuan Liou

Department of Computer Science and Information Engineering,
National Taiwan University, Republic of China
cyliou@csie.ntu.edu.tw

Abstract. This work shows how to set shape rules and convert them into logical rules to skip incorrect templates and reduce the number of candidate templates in the spatial topology distortion method [1]. The recognition rate is also improved by including shape constraints in the self-organizing process. This will drastically reduce the number of computations with improved recognition.

Keywords: pattern recognition, self-organizing map, morphing process, shape rule, shape constraint.

1 Introduction

In the spatial topology distortion method [1], named STD, the distortion between a candidate template and an unknown pattern can be computed by using the self-organizing algorithm [7]. This distortion is used to rank its candidate. It is cost to obtain such fine distortions for all templates. It is expected that certain incorrect templates can be skipped by imposing the rules among template features. The STD will be operated only for those templates that meet the rules for fine discrimination.

The STD is a parallel morphing process. It starts from a set of features that represent the skeleton of a template and gradually fits them to an unknown skeleton. The process comprises two phases in its training iterations, the match phase and the evolution phase [2,3,4]. These two phases operate iteratively until a final match is reached. All templates simultaneously compete to match the unknown pattern. The distortions can be obtained after the final matches.

To skip incorrect templates, we manually construct several reliable shape rules among the features for each template, such as the relative positions among features. For example, the first feature and the second feature are on the opposite sides of the third feature. These kind rules can be set with inequalities and can be converted into logical rules to speed the computation. They are used to disable incorrect templates. Some rules are general for all templates and the rest rules are only suitable for individual templates.

We also refine the evolution phase of STD by imposing shape constraints to increase distortions for incorrect templates. This will benefit the discrimination. During the STD evolution phase, the shape topology of an incorrect

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 245–252, 2012.

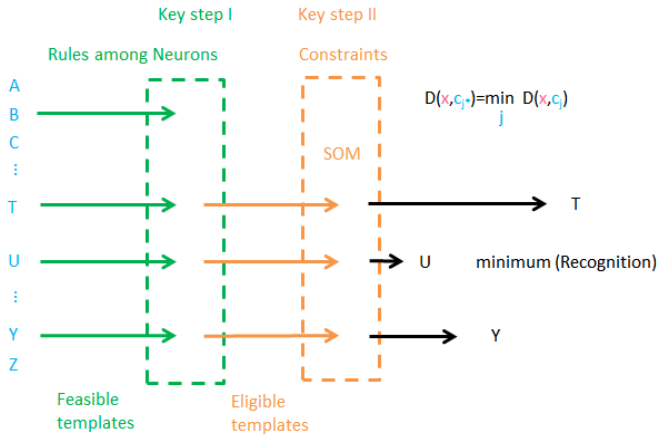


Fig. 1. Illustration of the two steps augmented in STD

template could be severely deformed. These constraints will keep the integrity of a template shape, such that the distortions will increase only for those incorrect templates.

The overall process is illustrated in Fig. 1, where $D(x, c_j)$ is the similarity between the feature set x of an unknown pattern and the feature set c_j of the jth template. A difficult character may have many templates such as **B** and B for the character B. Every template has its own rule set. There are few eligible templates remained to be trained and recognized by the STD method. The unknown pattern is classified to the template with the lowest distortion value D.

In Fig. 1, the key step I is a rule-checking step where one manually constructs a set of rules to skip incorrect templates before applying the STD method. The key step II is an integrity step that applies shape constraints in the evolution phase of STD to increase the distortion values for incorrect templates. These two steps are presented in Section 2. Section 3 shows experimental results and includes the performance under both constraints and rules. Some improvements and applications are discussed in the section.

Review the STD method. A thick stroke is represented by ellipses after feature extraction, Fig. 2(a). Each ellipse is a four-dimensional (4D) feature [5,6] containing the (x, y) coordinate of the center and the orientation (u, v) of the major axis, Fig. 2(b). Each neuron represents one ellipse feature. This work uses the vector $w_{jk}(t) = [x_{jk}(t), y_{jk}(t), u_{jk}(t), v_{jk}(t)]^T$ to denote the current four weights of the kth neuron in the jth template at training time t. This work calls the features of an unknown character as unknown features.

After the feature extraction, we operate the two phases of STD. In the match phase, sample an unknown feature e_i and find the winner neuron n^* from each template whose weight vector is most similar to the feature e_i. In the evolution

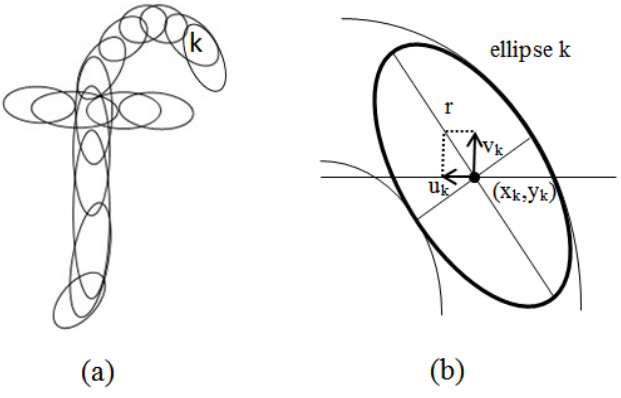


Fig. 2. (a) The ellipses fitted in the stroke 'f'. (b) A well fitted ellipse k and its 4 weights (x, y, u, v).

phase, update all neurons' weights inside a specific neighborhood of the neuron n^* by a factor, where the neighborhood is defined as the ellipse range of the feature e_i centered at the neuron n^* to allow more neighborhood neurons on similar strokes being modified.

These two phases are operated repeatedly until a satisfied convergence. After convergence, the STD computes the distortion between the unknown features x and those of the jth template c_j by using the distortion measure

$$D(x, c_j) = \frac{1}{|c_j|} \sum_{1 \le k \le |c_j|} |w'_{jk}(0) - w'_{jk}(T)|,$$

where $w'_{jk}(t)$ is a $2D$ vector which contains the $x_{jk}(t)$ and $y_{jk}(t)$ components of $w_{jk}(t)$, the center position of the kth neuron, at training time t. $|c_j|$ is the total number of features in the jth template. The training starts from $t = 0$. The end time of training is at $t = T$. The unknown pattern is classified as the template j^* if

$$D(x, c_{j^*}) = \min_j D(x, c_j), 1 \le j \le M,$$

where M is the total number of templates.

2 Shape Rules and Constraints

2.1 Rules among Features

We show how to set shape rules for each individual template. Several representative neurons are selected to represent the geometry structure of a template, Fig. 3(a). The template pattern 'W' is represented by four neurons.

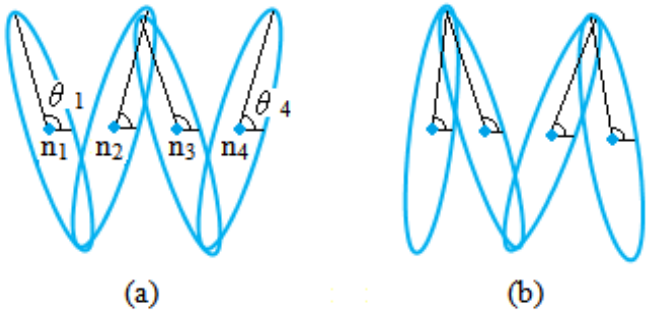


Fig. 3. (a) The representative neurons for the template pattern 'W'. (b) The representative neurons for the template pattern 'M'.

We manually set relative angle and position rules for these neurons. As an example, for the two ellipses 'n_1' and 'n_2' in Fig. 3(a), the angle of 'n_1' must be larger than that of 'n_2', and the center position of 'n_1' must be on the left of 'n_2'. For the template 'M' in Fig. 3(b), the angle of 'n_1' must be smaller than that of 'n_2'. Such rules can be set with inequalities and converted into logical rules. Based on these rules, one can discriminate a huge number of templates. The rules for pair features can be included in a matrix, Fig. 4. One can check these rules by logical relations and find the violations. Some skipped templates for the unknown pattern 'W' are plotted in Fig. 5(a). Those eligible templates are in Fig. 5(b).

The rule-checking step is processed as follows. In the very first match phase of STD, each representative neuron finds several similar features in the unknown features. For each template, we check all combinations among those similar features of all representative neurons by applying the template rules on relative angles and positions. The number of candidate combinations can be reduced after applying each rule. There is no need to check all rules for each combination. This is a tree-like divide and conquer reduction. The reduced number of templates highly relies on the resolution quality of each rule. For those combinations satisfying all rules, their templates are eligible for the STD method. When there are several templates that satisfy the rules, the top ranked template will be picked by the distortion measure D.

2.2 Setting Shape Constraints

In the evolution phase, we manually construct several shape constraints for each template to keep the integrity during the self-organizing process. This will increase the distortion values for incorrect templates. The convergence time t will be reduced also.

During the match phase, the topology of a template may be severely deformed, Fig. 6(a). To maintain the integrity of the template topology, connected strokes

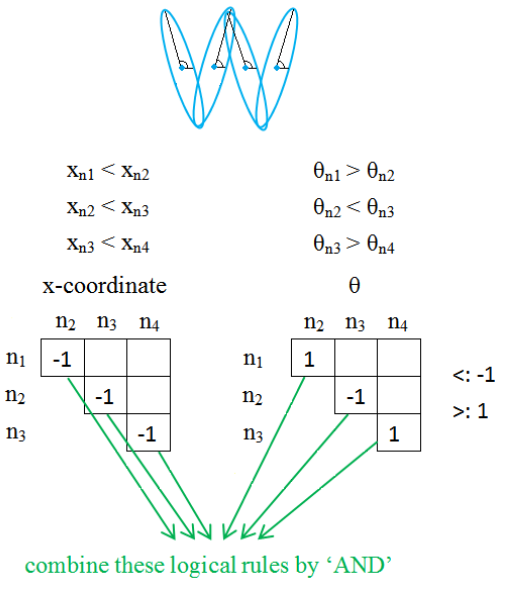


Fig. 4. Logical rules for pair representative features of the template 'W' are obtained from their inequalities and can be included in a matrix

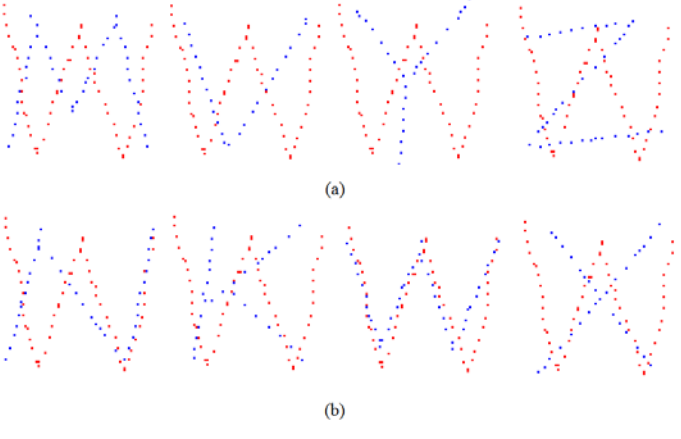


Fig. 5. (a) Skipped templates. (b) Eligible templates.

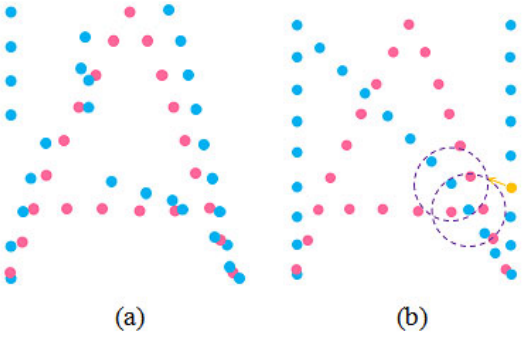


Fig. 6. (a) The updated blue neuron is too close to the neurons on different strokes. (b) The updated neuron is not allowed to enter the purple circular region of a different neuron.

can not be split, and unconnected strokes are not allowed to join together. Also, the shape rules in Section 2.1 should be maintained.

As a trial, we set a circle for each neuron as its motion border. When updating a neuron, its new position can not enter other's border except its neighborhood neurons. On the other hand, each neuron can not leave it own border, Fig. 6(b). The border of each neuron is examined in each iteration to check if the updated position violated it. If a neuron entered other's border, it will move back along the opposite direction of the updated vector to the outside of other's border, Fig. 6(b). The radius of the circle is manually designed.

3 Simulations

All characters are properly normalized. Based on the proposed rules and constraints in Section 2, we compare the simulation results with those in [1]. The template characters are the English alphabets in uppercase.

In Fig. 7, the template 'A' and all other templates are trained to resemble the unknown character 'A'. The solid orange line records the average moving distance of all neurons of the correct template 'A' at each training epoch. The dotted orange line shows the average moving distance of all neurons of incorrect templates.

The distoration D values between the correct template 'A' and the unknown during the training process are plotted with the solid black line. The D values between incorrect templates and the unknown are plotted with the dotted black line. As the training proceeds, the correct template 'A' fits in the unknown pattern gradually. The simulations without any constraints are plotted in Fig. 7(a). Fig. 7(b) shows the results with all constraints. With constraints, this figure shows that incorrect templates are harder to move and thus less similar to the unknown. The correct template is not restricted by these constraints so that it can fit very well.

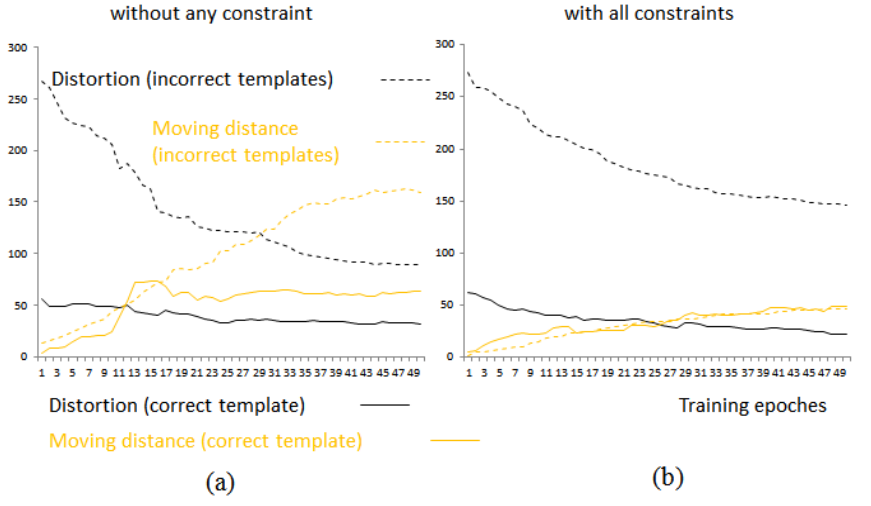


Fig. 7. The average moving distances of all neurons of the templates (the solid orange line for the correct template and the dotted orange line for all incorrect templates) versus the distoration D values between the templates and the unknown pattern (the solid black line for the correct template and the dotted black line for incorrect templates). (a) Without any constraint, (b) With all constraints.

Table 1. Comparison of simulations (Recognition rate / Time)

Constraints / Rules	Without any Rule	All Rules
Without any Constraint	72 % / 6.55 sec	82 % / 1.71 sec
All Constraints	87 % / 9.25 sec	87 % / 4.39 sec

The recognition rate and running time with rules and constraints are recorded in Table 1. With constraints, the recognition is more accurate but the running time is increasing because of checking the border constraints for all neurons. The rules eliminate many incorrect templates so that fewer templates are needed to be trained. Overall, applying rules reduces the training time and adding constraints increases the recognition rate.

Finally, we summarize the proposed techniques. A set of reliable key rules are constructed to filter out incorrect templates before training. We construct shape constraints for the self-organizing process. The rules can be well developed among neurons. Unrecognized patterns can be included as new templates. The proposed two steps can be applied to license plate recognition, zip code recognition, face identification and many topics in medicine cosmetology.

References

1. Liou, C.-Y., Yang, H.-C.: Handprinted Character Recognition Based on Spatial Topology Distance Measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 941–945 (1996)
2. Burr, D.J.: Elastic Matching of Line Drawings. IEEE Transactions on Pattern Analysis and Machine Intelligence 3, 708–713 (1981)
3. Burr, D.J.: Designing a Handwriting Reader. IEEE Transactions on Pattern Analysis and Machine Intelligence 5, 554–559 (1983)
4. Hinton, G.E., Williams, C.K.I., Revow, M.D.: Adaptive Elastic Models for Hand-Printed Character Recognition. In: Moody, J.E., Hanson, S.J., Lippmann, R.P. (eds.) Advances in Neural Information Processing Systems, pp. 512–519. Morgan Kaufmann (1992)
5. Daugman, J.G.: Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression. IEEE Transactions on Acoustics, Speech and Signal Processing 36, 1169–1179 (1988)
6. Gabor, D.: Theory of Communication. Journal of the Institution of Electrical Engineers 93, 429–457 (1946)
7. Kohonen, T.: Self-organization and Associative Memory, 3rd edn. Springer, New York (1989)

A Block-Based Orthogonal Locality Preserving Projection Method for Face Super-Resolution

Shwu-Huey Yen, Che-Ming Wu, and Hung-Zhi Wang

Department of Computer Science and Information Engineering,
Tamkang University, 151 Yingzhuan Road, Tamsui Dist., New Taipei City, Taiwan 25137
Republic of China (ROC)
105390@mail.tku.edu.tw, 294197073@s94.tku.edu.tw,
600410640@s00.tku.edu.tw

Abstract. Due to cost consideration, the quality of images captured from surveillance systems usually is poor. To restore the super-resolution of face images, this paper proposes to use Orthogonal Locality Preserving Projections (OLPP) to preserve the local structure of the face manifold and General Regression Neural Network (GRNN) to bridge the low-resolution and high-resolution faces. In the system, a face is divided into four blocks (forehead, eyes, nose, and mouth). The super-resolution process is applied on each block then combines them into a complete face. Comparing to existing methods, the proposed method has shown an improved and promising result.

Keywords: Orthogonal Locality Preserving Projections (OLPP), manifold, super-resolution, General Regression Neural Network (GRNN).

1 Introduction

Based on security considerations for modern society, there is a growing interest in visual surveillance system for security in public areas. However, the vision supervisory system usually confronts with the problem that the CMOS camera chip resolution is not sufficient, which causes people in the image are difficult to identify. This issue has been studied by many researchers. But there still exist a number of difficult problems such as estimating facial pose, facial expression variations, resolving object occlusion, changes of lighting conditions, and in particular, the low-resolution images captured in visual surveillance system at public areas.

To improve the capability in facial identification, many image processing methods are developed using one or more low-resolution face images to synthesis a high resolution one. Baker et al. coined the term "face hallucination" [1], to refer to the process of inferring a high-resolution face image from a low-resolution one. One common approach is learning-based technique, such as Eigenfaces (PCA), Fisherfaces (LDA) and Laplacianfaces (LPP) [2-5]. While the Eigenfaces method aims to preserve the global structure of the image space, and the Fisherfaces method aims to preserve the discriminating information; the Laplacianfaces method aims to preserve the local structure of the image space [4].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 253–262, 2012.

The Eigenfaces method applies Principal Component Analysis (PCA) [2] to project the data points along the directions of maximal variances. Its goal is to find a set of mutually orthogonal basis functions (i.e., Eigenfaces) that capture the directions of maximum variance in the data and for which the coefficients are pairwise decorrelated. The Eigenfaces method guarantees to discover the intrinsic geometry of the face manifold when it is linear. The Fisherfaces method applies Linear Discriminant Analysis (LDA) [3] to search for the project axes on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other. Both Eigenfaces and Fisherfaces reflect the global Euclidean structure. Thus, when the face images lie on a nonlinear submanifold hidden in the image space, these two methods can not discover the underlying structure. To solve this problem, some non-linear methods are proposed. Locally Linear Embedding (LLE) is one of the most common approaches [6]. Modeled by a nearest-neighbor graph, LLE successfully projects data lying on a nonlinear submanifold with the local structures preserved. However, when the given data are in large quantity and high dimension, the computation becomes complicated. Besides, LLE is defined only on those trained data. Locality Preserving Projections (LPP), a linear one, has been proposed for face images dimensionality reduction that preserves local relationships within the data set and uncovers its essential manifold structure.

When a high dimensional data lies on a low dimensional manifold embedded in the ambient space, the LPP is obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold. As a result, LPP shares many of the data representation properties of nonlinear techniques such as LLE. Yet LPP is linear and more crucially is defined everywhere in ambient space rather than just on the training data points. However, the basis functions obtained by the Laplacianfaces method are non-orthogonal. This makes it difficult to reconstruct the data. In [7], they proposed Orthogonal Locality Preserving Projections (OLPP) technology which shares the same locality preserving character, but at the same time the basis functions to be orthogonal. There are researches indicating that OLPP can have more locality preserving power than LPP does [4][7].

As in LPP or OLPP, they accomplish a dimension reduction that mapping a high dimension data point to a low dimension space. This contradicts with the goal of super-resolution. Thus, there are different methods proposed to resolve this problem. In [8], they divided images into multiple overlapping patches and perform super-resolution for each patch by employing Maximum a posterior (MAP) estimator to infer LPP coefficients for each patch. In [9], authors also divided images into multiple overlapping patches, and super-resolution for each patch was completed by employing Kernel Ridge Regression to infer OLPP coefficients for each patch. In this paper, we employ the OLPP and Generalized Regression Neural Network (GRNN) to reconstruct a given low-resolution face image.

2 Overview of LPP, OLPP and GRNN

A brief introduction on algorithms of LPP, OLPP, and GRNN, the main theme for the proposed algorithm, is given in the following.

2.1 Locality Preserving Projections (LPP)

Since PCA and LDA, the two popular methods used for face recognition, see only the global Euclidean structure, they fail to discover the underlying structure if the face images lie on a non-linear submanifold. Unlike PCA and LDA, LPP [4] is a manifold learning technique which preserves the local structure. The LPP algorithm is described below.

(1) Input: A set of image data space $X = [x_1, x_2, ..., x_m]$ (m training images) where x_i is a column vector representing a point in n-dimensional space.

(2) Compute distances between any two training images
The distance between two arbitrary x_i and x_j is calculated as Euclidean distance:

$$dist_{ij} = \left\| x_i - x_j \right\|. \tag{1}$$

(3) Construct the adjacency graph
Let G denote a graph with m nodes for m training images. Each node (or training image) has K nearest neighbors according to the distance metric in (1), and the neighbors are connected by edges.

(4) Setting the weights between neighbors
Construct the (symmetric) weight matrix S (size $m \times m$) such that if x_i and x_j are not connected by edge, set $S_{ij} = 0$, otherwise set

$$S_{ij} = e^{-\left(\left\| x_i - x_j \right\|^2 / t \right)}. \tag{2}$$

The justification for this choice of weights can be traced back to [10].

(5) Eigenmaps: To preserve the local structure, find the projection matrix A such that

$$\arg\min_A \sum_{ij} (A^T x_i - A^T x_j)^2 S_{ij} \cdot \tag{3}$$

To solve Eq. (3) is equivalent to compute the eigenvectors a and eigenvalues λ for the generalized eigenvector problem:

$$XLX^T a = \lambda XDX^T a\,, \tag{4}$$

where D is a diagonal matrix whose entries are column sums of S, $D_{ii} = \Sigma_j S_{ji}$, and $L = D - S$ is the Laplacian matrix. Let the column vectors $a_1, \ldots, a_l$ be the solutions of equation (4), ordered according to their eigenvalues, $\lambda_1 < \ldots < \lambda_l$. Thus, the projection matrix is $A = (a_1, a_2, \ldots, a_l)$, and the embedding is as follows:

$$x_i \rightarrow y_i = A^T x_i , \tag{5}$$

where x_i is an n-dimensional vector, y_i is an l-dimensional vector, and A is an $n \times l$ matrix.

2.2 Orthogonal Locality Preserving Projections (OLPP)

OLPP [7] is based on the LPP method. It shares the same locality preserving character as LPP with the orthogonal basis functions. The major difference of LPP and OLPP is in step (5) described below.

(5)' Computing orthogonal basis functions: (***D*** and ***L*** are defined just as in LPP.) Let $a_1, a_2, \ldots, a_l$ ($l \geq 2$) be orthogonal basis vectors, we define

$$A_{l-1} = [a_1, \ldots, a_{l-1}], \text{ and } B_{l-1} = A^{\mathrm{T}}_{l-1} (XDX^{\mathrm{T}})^{-1} A_{l-1} . \tag{6}$$

The orthogonal basis vectors $\{a_1, a_2, ..., a_l\}$ can be computed as follows.

a) Compute a_1 as the eigenvector of $(XDX^T)^{-1} XLX^T$ associated with the smallest eigenvalue.

b) Compute a_l as the eigenvector of associated with the smallest eigenvalue of M_l where M_l is shown in (7). A detailed derivation of LPP and OLPP is referred to [4] and [7].

$$M_l = \{I - (XDX^T)^{-1} A_{l-1} B_{l-1}^{-1} A_{l-1}^T\} (XDX^T)^{-1} XLX^T . \tag{7}$$

2.3 Generalized Regression Neural Network (GRNN)

Both LPP and OLPP are used for dimensional reduction that project images from high-dimension (n) to low-dimension (l). But our purpose is to reconstruct a low-resolution image to a high-resolution image. Methods such as Maximum a posterior (MAP) [8], Generalized Regression Neural Network (GRNN) [11], and Kernel Ridge Regression [9] are proposed to deal with the reconstruction problem. In particular, a nonlinear GRNN is often used for function approximation. The architecture for a GRNN is shown in Fig. 1. It has two layers: radial basis layer and special linear layer. In our application, radial basis layer, R is the input OLPP coefficients, Q_1 is the results of Gaussian distribution. A parameter spread should be defined here. The value of spread determines the scope of Gaussian distribution. In the special linear layer, Q_3 is used to output the results of regression calculation. In our algorithm, the corresponding high-resolution image is the output Q_3.

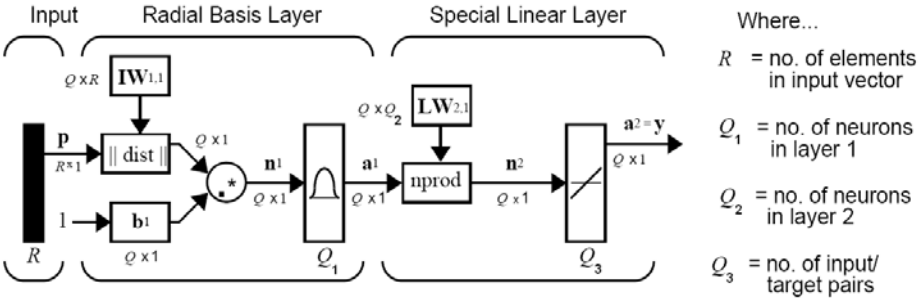


Fig. 1. The architecture for a GRNN

3 The Proposed Algorithm

Patch-based approaches for super resolution using small image patches are less dependent on person-identity. However, there is a symmetric relation in most of facial components such as left eye and right eye, left cheek and right cheek, etc. So in the proposed algorithm we adopt block-based approach according to facial components. For each high-resolution training image (192 × 144), from top to bottom in the order of forehead, eyes, nose, and mouth, it is divided into 4 blocks (32 × 144, 32 × 144, 56 × 144, 72 × 144) then down-sampled into 1/8 in both directions (i.e., 4 × 18, 4 × 18, 7 × 18, 9 × 18). With respect to each block, the proposed algorithm will train one transformation matrix W and one GRNN. To reconstruct a high-resolution face image from a given low-resolution one, the algorithm employs the constructed transformation matrix W and GRNN to super-resolution each block respectively, then combine these 4 parts into a high-resolution face image. The following algorithm is described in two-phase, one is training phase and the other is super-resolution phase.

3.1 Training Phase

The training purpose is to obtain the OLPP transformation matrices and the GRNNs for blocks, respectively. Flow chart of the training process is outlined in Fig. 2.

(1) Construct the transformation matrix W

To avoid singularity problem, the input data needs to perform PCA first to remove those components corresponding to zero eigenvalues [11][12]. Let $X = [x_1, x_2, ..., x_m]$ be the training data consisting of m low-resolution blocks corresponding to one of forehead, eyes, nose, or mouth. Denote the transformation matrix of PCA by W_{PCA}. Next, follow the steps described in 2.1 and 2.2 to construct the OLPP transformation matrix $W_{OLPP} = [a_1, a_2, \ldots, a_l]$. Then the transformation matrix W is $W = W_{PCA}W_{OLPP}$.

(2) Train the GRNN

After W is constructed, the mapping of $X = [x_1, x_2, ..., x_m]$ is

$$Y = W^T X . \tag{8}$$

where $Y = [\, y_1, y_2, ..., y_m \,]$, y_i is the coefficients from the mapping of x_i. Note that x_i is n-dimension and y_i is l-dimension. To train the GRNN, Y is the input and X_h is the target output where $X_h = [x_{h1}, x_{h2}, ..., x_{hm}]$ be the vector of high resolution blocks of X.

3.2 Super-Resolution Phase

For a given low-resolution face image of size 24 × 18, we divide the image into 4 blocks as mentioned above. For each low-resolution facial-block, we used the corresponding constructed transformation matrix to obtain the coefficients. Then use the coefficients as the input of the trained GRNN and the output is the super-resolution result. Repeat the steps for each facial block. The reconstruction flow chart is given in Fig. 3. The super-resolution face image is obtained by combining these 4 reconstructed high-resolution blocks.

To have seamless effect on combining reconstructed blocks, we take 2 pixels overlapping for two neighboring blocks then combine them by linear interpolation. To have a better visual quality, we fuse the above reconstructed results with the bi-cubic images. From the given low-resolution image of 24 × 18, bi-cubic method is used to enlarge it to the size of 192 × 144. Comparing to the original high resolution image, the reconstructed face image (I_{Re}) provides good similarity in facial components and the bi-cubic enlarged face image ($I_{bi\text{-}cu}$) provides a good similarity in facial complexion. Based on this, a facial weighting mask is used for fusing I_{Re} and $I_{bi\text{-}cu}$. As shown in Fig. 4 (a), the facial mask is indicated by three colors (3 different weights) such that blue, yellow, green represent the degrees of importance in I_{Re} by descending order. If f is the fused face image, then

$$f(x, y) = \alpha \cdot I_{Re}(x, y) + (1 - \alpha) \cdot I_{bi-cu}(x, y)\,, \tag{9}$$

where α is the weight on (x, y) position from the facial mask. As shown in Fig. 4, (b) is the reconstructed one I_{Re} and (c) is the bi-cubic enlarged one $I_{bi\text{-}cu}$, and (d) is the final image f. The facial mask has the size of 192 × 144 (same with the face image) with points on blue having weights $\alpha = 0.8$, points on yellow having weights $\alpha = 0.6$, and points on green having weights $\alpha = 0.4$.

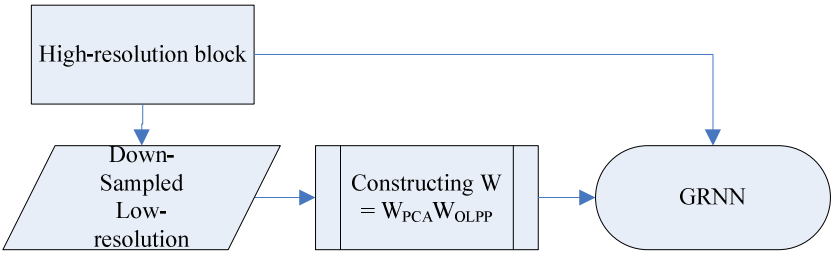


Fig. 2. The training flowchart

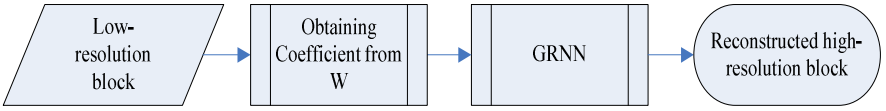


Fig. 3. Facial super-resolution flowchart

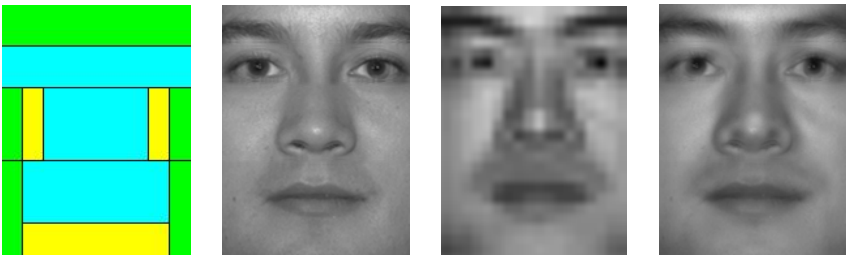


Fig. 4. (a) is the facial mask (weight by *Blue* = 0.8, *Yellow* = 0.6, *Green* = 0.4), (b) is the reconstructed face I_{Re}, (c) is the bi-cubic enlarged image $I_{\text{bi-cu}}$, and (d) is the fusion result f.

4 Experimental Results

For experiments, subset of Yale B [13] and color FERET [14] database are used. We selected 200 images with neutral expression and frontal pose, and used 190 images for training and 10 images for testing. Before experiments, the face images were aligned with given eye coordinates, cropped to 192 × 144 images, and normalized by intensity. The high-resolution images were down-sampled to low-resolution 24 × 18 images. In the proposed algorithm, there are three parameters (in Section 2): (1) the value K in constructing the adjacent graph for K nearest neighbors, (2) the value of l in constructing the OLPP transformation matrix to map an n-dim into l-dim, and (3) the value of spread in GRNN. After many experiments, we use the parameters as shown in Table 1. Figure 5 illustrates some experimental results of the proposed algorithm where input low-resolution images (24 × 18) are on the top row, the super-resolution results after fusion (192 × 144) are on the middle row, and the original high-resolution images are on the bottom row.

Table 1. Parameters for reconstructing facial blocks

Blocks	K	l	spread
Forehead	30	40	100
Eyes	30	40	100
Nose	15	100	100
Mouth	15	100	100

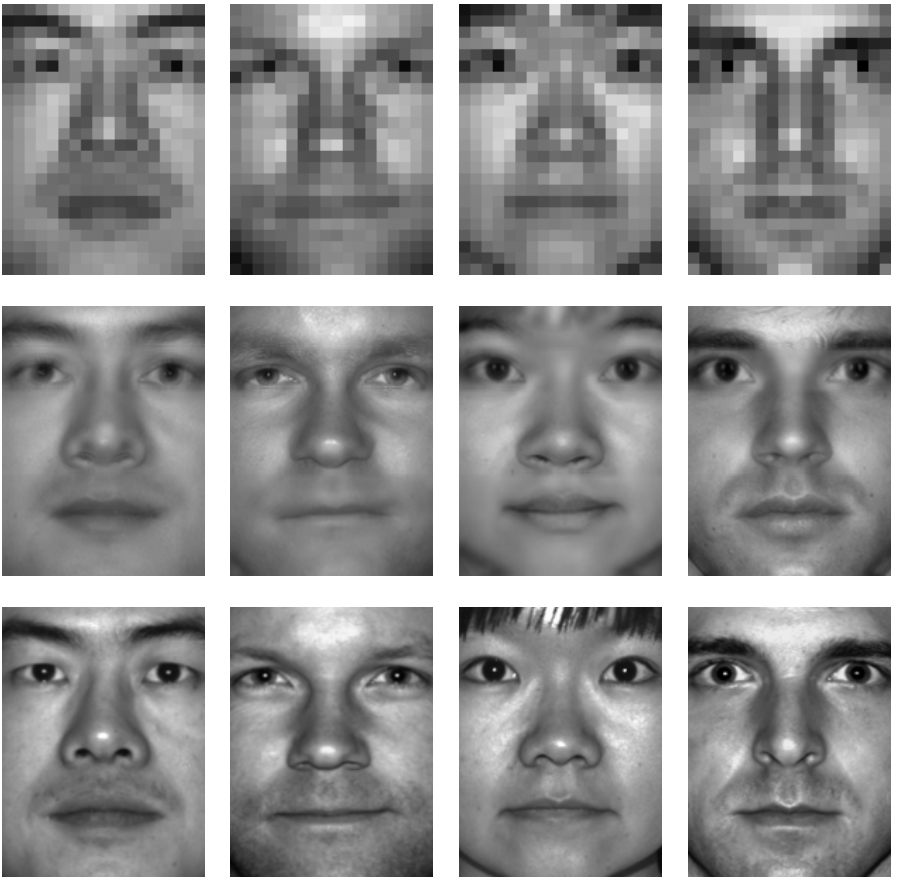


Fig. 5. Four examples where images on the first row are the low-resolution face images (24×18), second row show the 192×144 hallucinated result of our approach, and the third row are the original high-resolution face.

We also compared our method with patch-based methods in [8] and [9]. To reconstruct face images, [8] uses LPP and MAP and [9] uses OLPP and Kernel Ridge Regression. In Fig. 6, (c) and (d) are the results cited from their papers, and (e) is ours method. Note that (c) and (d) are reconstructed from 32×24 to be 128×96 (4 times in both width and height) and our method is reconstructed from 24×18 to be 192×144 (8 times in both width and height). Comparing to (c) and (d), although the reconstructed image (e) is not very clear visually, but (e) shows a better resemblance in important facial components such as in the eyebrows, eyes, nose, mouth.

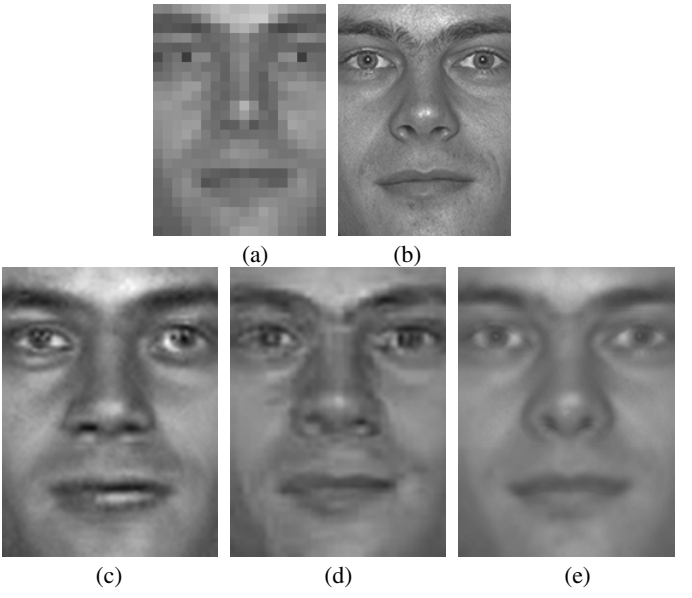


Fig. 6. Reconstruction comparison. (a) the low-resolution face 24×18, (b) the original high-resolution face, (c) Park et al. [8], (d) Vijay Kumar et al. [9], (e) ours.

5 Conclusion

In this paper, we presented an algorithm that it reconstructs a face image up to 8 times in both width and height of a low-resolution face image. To be less dependent on person-identity, the training and reconstructing are performed on 4 blocks according to facial components, respectively. To have the final result, 4 reconstructed blocks are combined by linear interpolation and fused with bicubic enlarged image. By this way, the overall resemblance is improved. However, some artifacts are observed too. The future work includes solving the artifact problem as well as extending this study to faces with varying pose and illumination.

References

1. Baker, S., Kanade, T.: Hallucinating Faces. In: 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 83–88 (2000)
2. Turk, M., Pentland, A.: Eigenfaces for Recognition. Cogn. Neurosci. 3(1), 71–86 (1991)
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (2001)
4. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face Recognition Using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)

5. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Extensions: A General Framework for Dimensionality Reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(1), 40–51 (2007)
6. Roweis, S., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290, 2323–2326 (2000)
7. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Orthogonal Laplacianfaces for Face Recognition. IEEE Transactions on Image Processing 15(11), 3608–3614 (2006)
8. Park, S.W., Savvides, M.: Breaking Limitation of Manifold Analysis for Super-resolution of Facial Images. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 573–576 (2007)
9. Vijay Kumar, B.G., Aravind, R.: Face Hallucination Using OLPP and Kernel Ridge Regression. In: 2008 IEEE International Conference on Image Processing (ICIP), pp. 353–356 (2008)
10. Belkin, M., Niyogi, P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: Advances in Neural Information Processing Systems, vol. 14, pp. 585–591. MIT Press, MA (2002)
11. Ahmed, S., Rao, N.I., Ghafoor, A., Sheri, A.M.: Direct Hallucination: Direct Locality Preserving Projections (DLPP) for Face Super-Resolution. In: 1st International Conference on Advanced Computer Theory and Engineering (ICACTE), pp. 105–110 (2008)
12. Hu, D., Feng, G., Zhou, Z.: Two-dimensional Locality Preserving Projections (2DLPP) with It's Application to Palmprint Recognition. Pattern Recognition 40(1), 339–342 (2007)
13. Lee, K.C., Ho, J., Kriegman, D.: Acquiring Linear Subspaces for Face Recognition under Variable Lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5), 684–698 (2005)
14. The FERET Database, `http://www.nist.gov/humanid/colorferet`

Note Symbol Recognition for Music Scores

Xiaoxiang Liu

College of Electrical Engineering and Information,
Jinan University, Zhuhai, China
tlxx@jnu.edu.cn

Abstract. Note symbol recognition plays a fundamental role in the process of an OMR system. In this paper, we propose new approaches for recognizing notes by extracting primitives and assembling them into constructed symbols. Firstly, we propose robust algorithms for extracting primitives (stems, noteheads and beams) based on Run-Length Encoding. Secondly, introduce the concept of interaction field to describe the relationship between primitives, and define six hierarchical categories for the structure of notes. Thirdly, propose an effective sequence to assemble the primitives into notes, guided by the mechanism of giving priority to the key structures. To evaluate the performance of those approaches, we present experimental results on real-life scores and comparisons with commercial systems. The results show our approaches can recognize notes with high-accuracy and powerful adaptability, especially for the complicated scores with high density of symbols.

Keywords: optical music recognition, music notes recognition, primitive extraction, structure analysis.

1 Introduction

Optics music recognition (OMR) addresses the problem of musical data acquisition, with the aim of converting optically scanned music scores into a versatile machine-readable format. Note symbol recognition plays a fundamental role in the process of an OMR system, since the notes are most ubiquitous in a score and they carry the most important semantic information (the pitch and duration). The recognition of notes is almost the hardest part in an OMR system. The shape of notes is ever-changing. A note feature is more intricate than a common character, consisting of variable numbers of primitives that dramatically vary in size, orientation and position. Regarding this case, Note recognition can be divided into two steps: extracting the primitives which make up notes; assembling them into constructed symbols, guided by musical knowledge [1], [2].

The common methods for extracting primitives include projection [3], skeleton [4], Line Adjacency Graph [5], template matching [6], mathematics morphology [7], [8], feature matching [9] and so on. The methods for assembling primitives often use high-level grammars such as string grammar [10][11], graph grammar [12] and rule-based model [6]. However, these existing methods show the following common limitations: first, the capability to separate overlapping and touching symbols is weak in case of

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 263–273, 2012.

complex scores; second, the relationship between primitives is described too rigidly, and the ambiguous relationship cannot be handled exactly; third, structural features of notes are not elaborately classified into hierarchical categories, so there is no point during the recognition process, and the high-level knowledge rules cannot be made fully use. For overcoming these limitations, this paper proposes new algorithms for extracting primitives based on Run-Length Encoding (RLE), and a new method for assembling primitives by analyzing the structural features of notes.

2 Staff Lines Detection and Removal

In most OMR systems, location and removal of staff lines is the initial phase. Staff lines bring two important parameters:

- ***Staff lines height***, the thickness of staff lines, gives a fundamental unit used to normalize line width.
- ***Staff space***, the space between the staff lines, conveys the scale of scores and gives a fundamental unit used to normalize distances and sizes.

The values of them can be estimated with good accuracy as the most frequent black (*Staff lines height*) and white (*Staff space*) vertical run-length respectively [13].

The paper applies an efficient and robust algorithm for staff lines detection proposed by Fujinaga [13]. The algorithm first detects staff lines by horizontal projections and then deskews each staff line by correlating the horizontal project profiles of adjacent vertical strips; each strip is sheared to the position with maximal correlation. The adopted staff line removal method is a skeleton-based algorithm proposed by Dalitz [14], and the code of this algorithm is freely available in [15].

3 Primitive Extraction

Notes (e.g. simple stemed notes, beam-group notes), are composed of geometric primitives plus a set of graphical and syntactic rules on those primitives. In this section, three RLE based algorithms are presented respectively for extracting stems, note heads and beams.

3.1 Stem Extraction

The existing methods for extracting lines are prone to break a stem into several pieces due to touching or overlapping parts (such as noteheads, flags and beams) on the stem. To solve this problem, the paper proposes a RLE based algorithm, described as follows (referring to Fig. 1):

Step1: Scan the image vertically and generate vertical runs of black pixels by RLE. If the length of a vertical run is longer than a certain threshold value (2 times of *Staff space*), and then this run will be marked as a Seed Run (SR).

Step2: Scan the image horizontally and generate horizontal runs of black pixels by RLE. For each pixel in a SR, find the horizontal run which contains this pixel. By locating the two end pixel of the horizontal run, gain the left and right edge pixels of the SR.

Step3: For each SR, if a set of its edge pixels satisfies the following three conditions simultaneously, then this SR will be accepted as a stem.

***Condition1*:** The distance between the left edge pixel and the right edge pixel must be shorter than a certain threshold value (2 times of *Staff lines height*).

***Condition2*:** The left edge pixel and the right edge must be bilateral symmetrical relative to the center of the SR.

***Condition3*:** The edge pixels satisfying the condition 1 and 2 must be arranged in succession and add up to a certain number (1.2 times of *Staff space*).

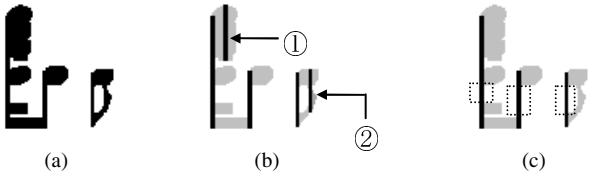


Fig. 1. Stem extraction. (a) Samples of notes, (b) SR in notes, ① will be ruled out by condition1, ② will be ruled out by condition2,3, (c) Last results, in the rectangles all conditions are satisfied.

3.2 Notehead Extraction

Noteheads often touch each other, and also are possible to touch other symbols such as sharps, flats and natures. These segmentation problems are impossible to solve without any contextual information. Hence, the prior domain knowledge of scores is used here, and it includes:

- Staff lines either go through the center of noteheads, or appear at the top or bottom edge.
- On the left or right edge of each notehead, there is one stem linking with it.

The knowledge implies that the location of staff lines and extracted stems can provide a basis for segmentation. Based on this, the detail algorithm of noteheads extraction is described as follows:

Step1: Scan the image horizontally and generate horizontal runs of black pixels by RLE. Remove the runs whose length is shorter than a certain threshold value (3 times of *Staff lines height*). This step separates noteheads from stems.

Step2: Label connected components (ccs) in the image. Find the ccs whose width is longer than a certain threshold value (1.5 times of *Staff space*), and if at least one of their top and bottom edges is flat they will be removed. This step removes the ccs that have characteristic feature of beams, and ensures they will not be identified as noteheads by mistake in the next process.

Step3: Split ccs along each extracted stem. Find the ccs whose height is longer than a certain threshold value (2 times of *Staff space*). If their top or bottom edges overlap with staff lines, split them along staff lines, otherwise split them along the middle lines between two adjacent staff lines. (Referring to Fig. 2)

Step4: Re-label connected components. If a cc satisfies the following two conditions simultaneously, then this cc will be accepted as a notehead.

Condition1: The width is approximately a certain threshold value (1.5 times of *Staff space*); the height is approximately a certain threshold value (*Staff space*).

Condition2: The ratio of pixel area (the number of black pixels in the cc) and width×height is approximately 0.8.

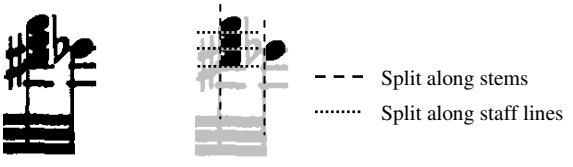


Fig. 2. Notehead extraction: (a) Sample of a note. (b) Noteheads segmentation.

3.3 Beam Extraction

The conventional method for extracting beams is line detection, but it is really hard to deal with the following two difficulties: 1. a beam is possible to be broken into several pieces due to overlapping with stems; 2. beams are possible to touch each other in a beam-group note. This paper gives up the method of line detection, but sees the beams in a beam-group note as a block region. The block region is identified by relying on at least one flat top or bottom edge. The number of beams is counted with the method of doing calculation with the thickness of a single beam. The detail algorithm runs as follows:

Step1: Scan the image horizontally and generate horizontal runs of black pixels by RLE; remove the runs whose length is shorter than a certain threshold value (3 times of *Staff lines height*). This step separates beams from stems.

Step2: Label connected components in the image. If a cc satisfies the following two conditions simultaneously, then this cc will be accepted as a beam region.

Condition1: The width is longer than a certain threshold value (2 times of *Staff space*); the height is longer than a certain threshold value (*Staff space*).

Condition2: Either the top or bottom edge is flat.

Step3: Calculate the thickness of a single beam. Generate an image that only contains beam regions. Scan this image vertically and generate vertical runs of black pixels by RLE. The most common black run-length represents the thickness of a single beam.

Step 4: Count the number of beams. Given a beam region and a stem, divide the thickness of the beam region at the stem position by the thickness of a single beam, and the result is the number of corresponding beams.

4 Structure Analysis

The essential task of structure analysis is that of expressing a valid taxonomy of note structures. It explains how the extracted primitives are assembled into musical notes.

In this section, the paper first introduces ***interaction field*** to describe the relationship of primitives, and then defines six structures of notes elaborately; next proposes an effective sequence with priority mechanism for assembling primitives by analyzing the features of the six structures.

4.1 Description of the Relationship between Primitives

Definition 1 [Interaction Field of Note Primitives (IFNP)]: F represents an IFNP, F is four tuples, $F=<V_1,V_2,F_h,F_v>$, V_1 and V_2 are interaction objects, namely two different primitives. The object has several attributes including coordinates of bounding box, corresponding IFNP array, and a check tag. F_h is the horizontal interaction potential and F_v is the vertical interaction potential. Here the potential is a scalar value returned by a specific function which takes the distance between objects as an independent variable.

In terms of different types of objects, IFNP is divided into two kinds: stem-notehead IFNP and stem-beam IFNP. Table 1 shows the function of vertical interaction potential for stem-notehead IFNP. As space is limited, the other potential functions are no longer presented here.

Table 1. Vertical potential of the stem-notehead IFNP (B returns the y-coordinate of the bottom of an object's bounding box; T returns the y-coordinate of the top of an object's bounding box; D=*Staff space*/2; SP=*Staff space*).

Interaction Space	f -Interaction Potential
① $B(nh)\leq T(stem)-D$	Φ
② $B(nh)>T(stem)-D \wedge B(nh)<T(stem)$	$[T(stem)-B(nh)-D]/D$
③ $B(nh)\geq T(stem) \wedge T(nh)\leq T(stem)-D$	-1
④ $T(nh)>T(stem)-D \wedge \mathrm{T}(nh)<\mathrm{T}(stem)+D$	$[T(nh)-T(stem)-D)]/SP$
⑤ $T(nh)\geq T(stem)+D \wedge B\ (nh)\leq B(stem)-D$	0
⑥ $B(nh)>B(stem)-D \wedge B(nh)<B(stem)+D$	$[\mathrm{B}(nh)-\mathrm{B}(stem)+D]/SP$
⑦ $B(nh)\geq B(stem)+D \wedge T(nh)\leq B(stem)$	1
⑧ $T(nh)>B(stem) \wedge \mathrm{T}(nh)<\mathrm{B}(stem)+D$	$[D-T(nh)+B(stem)]/D$
⑨ $T(nh)\geq B(stem)+D$	Φ

The IFNP based method for describing the relationship between primitives has the following advantages:

- The potential function is a function of distance, which can express the relationship directly and precisely. It has good qualities with both robustness and precision, and is flexible for complex data environment.
- The IFNP carries the knowledge of note structures. The primitives that may assemble a suitable note produce an IFNP with strong interaction potential, otherwise they produce an IFNP with weak interaction potential. Therefore, IFNP discloses the characteristics of note structures, and can provide great convenience for the next step of structure classification.
- Under the effect of an IFNP, some primitives will be gathered into a single collection. As a collection is just a potential note, search and judgment

operators are performed only within such small collection. Therefore, it can significantly reduce searching space and computational complexity compared with the full search method.

4.2 Definition of Note Structures

Definition 2 [Key Note Structure (KNS), Key Notehead (KN), Key Note IFNP (KN-IFNP)]: *fsn* is a stem-notehead IFNP produced by a *stem* and a *notehead* (*nh*), $fsn=<stem,nh,f_h,f_v>$. If $f_h \times f_v > 0$ then the *stem* and the *nh* can be assembled into a KNS. The *nh* is called KN, and the *fsn* is called KN-IFNP (referring to Fig. 3(a)).

KNS is the indispensable structure for any notes. The stem and the notehead in KNS have the most stable relationship (maintain a relatively fixed position).

Definition 3 [Chord Note Structure (CNS), Chord Note IFNP (CN-IFNP)]: *fsn* is a KN-IFNP, $fsn=<stem,nh,f_h,f_v>$. *fsn'* is an another IFNP produced by the *stem* and another *notehead* (*nh'*), $fsn'=<stem,nh',f_h',f_v'>$. If $f_v'=0 \wedge f_h \times f_h' > 0$ then the *stem* and the *nh'* can be assembled into a CNS. The *nh'* is called CN, and the *fsn'* is called CN-IFNP (referring to Fig. 3(b)). If the *stem* is connected to a beam, the vertical space between the CN and the beam must be longer than *Staff space*.

CNS depends on KNS. The position of CN is unstable (non-fixed).

Definition 4 [Alternative Chord Note Structure (ACNS), Alternative Chord Notehead (ACN), Alternative Chord Note IFNP (ACN-IFNP)]: *fsn* is a KN-IFNP, $fsn=<stem,nh,f_h,f_v>$. *fsn'* is an another IFNP produced by the *stem* and another *notehead* (*nh'*), $fsn'=<stem,nh',f_h',f_v'>$. If $f_v'=0 \wedge f_h \times f_h' < 0$ then the *stem* and the *nh'* can be assembled into an ACNS. The *nh'* is called ACN, and the *fsn'* is called ACN-IFNP (referring to Fig. 3(c)).

ACNS depends on KNS or CNS. The position of ACN must be near to KN or CN.

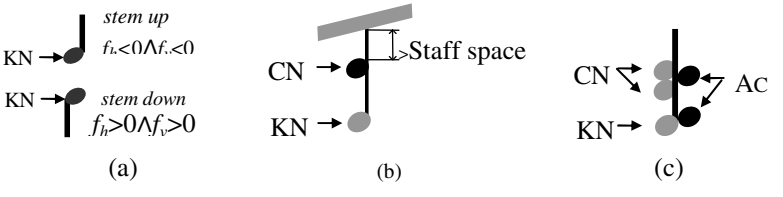


Fig. 3. Examples of the structures of stemed notes: (a) KNS, (b) CNS, (c) ACNS

Definition 5 [Frame Beam Structure (FBS), Frame Beam (FB), Frame Beam IFNP (FB-IFNP)]: *fsb* is a stem-beam IFNP produced by a *stem* and a *beam*, $fsb=<stem,beam,f_h,f_v>$, and the *stem* belongs to a KN-IFNP. If $f_h \times f_v \neq 0$ then the *stem* and the *beam* can be assembled into a FBS. The *beam* is called FB, and the *fsb* is called FB-IFNP (referring to Fig. 4(a)).

FBS is the indispensable structure for beam-group notes. A beam-group note must have only one FB, and there must be one FBS at each end of a FB. The stem and the beam in FBS have a stable relationship.

Definition 6 [Inner Beam Note Structure (IBNS), Inner Beam Note IFNP (IBN-IFNP)]: There is a *stem* belongs to a KN-IFNP, and a *beam* belongs to a FB-IFNP. *fsb* is a stem-beam IFNP produced by the *stem* and the *beam*, *fsb*=<*stem*,*beam*,f_h,f_v>. If f_h=0∧f_v≠0 then the *stem* and the *beam* can be assembled into an IBNS. The *fsb* is called IBN-IFNP (referring to Fig. 4(b)).

IBNS depends on KNS and FBS. The position of the stem in IBNS is unstable.

Definition 7 [Subordinate Beam Structure (SBS), Subordinate Beam (SB), Subordinate Beam IFNP (SB-IFNP)]: *fsb* is a stem-beam IFNP produced by a *stem* and a *beam*, *fsb*=<*stem*,*beam*,f_h,f_v>, and the *stem* belongs to a KN-IFNP. If f_v=0 and the vertical space between two adjacent beams is shorter than *Staff space* then the *stem* and the *beam* can be assembled into a SBS. The *beam* is called SB, and the *fsb* is called SB-IFNP. (referring to Fig. 4(c)).

SBS depends on KNS and FBS. The position of SB is unstable.

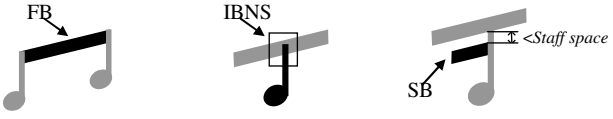


Fig. 4. Examples of the structures of beam-group notes: (a) FBS, (b) IBNS, (c) SBS

4.3 Structure Analysis

The most existing methods for analyzing note structures can be concluded as a two-phase sequence (2PS): phase one assembles stems and noteheads into stemed notes; phase two assembles stems and beams into beam-group notes. However, when a person observes notes, the case is not always like 2PS. In the human sense, the eyes always give priority to the global, important, and stable structures, and then the local, normal and unstable ones. With the above elaborate definition of note structures, this behavior pattern of humans can be imitated. Table 2 shows up the differences among the six note structures concerning their importance, stability and dependence. Based on it, the paper proposes a new sequence (PMS) for assembling primitives, guided by priority mechanism of giving priority to the key structures.

Table 2. Differences among six structures of notes

Categories	Stability	Dependence	Importance
KNS	*Strongest*	*None*	*Highest*
FBS	*Strong*	KNS	*High*
IBNS	*Medium.*	KNS *and* BNS	*Medium.*
CNS	*Medium.*	KNS	*Low*
SBS	*Weak*	KNS *and* BNS	*Low*
ACNS	*Weakest*	KNS *and* CNS	*Lowest*

PMS runs as follows (referring to Fig. 5):

Step1: Initialization. For each interaction object (primitive), generate the corresponding IFNP array and mark its check tag as "*Unchecked*".

Step2: For each stem, search the KN-IFNP (Definition 2) within its IFNP array. If a KN-IFNP exists, the check tags of the stem and the KN will be marked as "*Checked*".

Step3: For each beam, if there is a FB-IFNP (Definition 5) at each end of it, its check tag will be marked as "*Checked*".

Step4: For each beam whose check tag is "*Checked*", find the IBN-IFNP (Definition 6) within its IFNP array.

Step5: For each stem whose check tag is "*Checked*", do the followings:

Step5.1: Search the CN-IFNP (Definition 3) within its IFNP array. If a CN-IFNP exists, the check tag of the CN will be marked as "*Checked*".

Step5.2: Search the SB-IFNP (Definition 7) within its IFNP array. If a SB-IFNP exists, the check tag of the SB will be marked as "*Checked*".

Step5.3: Search the ACN-IFNP (Definition 4) within its IFNP array. If an ACN-IFNP exists, the check tag of the ACN will be marked as "*Checked*".

Step6: Transform the "*Checked*" stems, noteheads and beams into musical events to be performed.

The main contributions of PMS are as follows:

- It can reduce the additional cost which pays because of the nonessential judgment, and improve the recognition efficiency. For example, if KN is not found within the IFNP array of a stem, the stem must be not acceptable, so all other primitives related to the stem will be omitted immediately.
- It can reduce the recognition difficulty for the complex notes, and improve the recognition rate. In PMS the key structures are given priority, and usually they are recognized with high accuracy, so they can instruct next recognition of the other structures. In this way, the high-level knowledge rules can be used more deeply, and it helps to get rid of incorrect redundancy successfully.

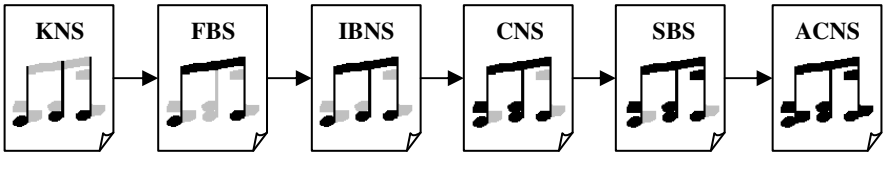


Fig. 5. PMS for assembling primitives

5 Performance Evaluation

All the algorithms described in this paper have been implemented, and based on it we have developed a prototype system called IOMRS. It is written by Visual C++, and runs on Windows XP platforms. In this section, we present the performance evaluation in terms of experimental results with real-life score images, and compare IOMRS with two pieces of commercial software. Here we choose SmartScore X Pro [16] and SharpEye 2.68 [17], which are the most popular and well-known OMR systems for Windows, and their last version are freely available for demonstration on the Web.

The method of evaluation is: for each note, transform it into music events (stemed note corresponds to one event and beam-group note corresponds to multiple events); for each event, only if the pitch (position of the noteheads) and the duration (number of the beams/flags) are both correct, this note is deemed correct. We select eight typical images as a test set (scanned with a resolution of 150 dpi, size: 1200×1600 pixel^2), including polyphonic piano scores with various degrees of difficulty due to symbol density and rhythms complexity. Table 3 shows global evaluations on the test set respectively of IOMRS, SmartScore and SharpEye. From image1 to image8, the degree of difficulty increases gradually.

Table 3. Test Results with IMORS, SmartScore and SharpEye (C: Correct count, F: False count, M: Misses, T=C+F+M, R=C/T)

Image	*T*	*IOMRS*		*SmartScore*		*SharpEye*	
		C,F,M	*R(%)*	*C,F,M*	*R(%)*	*C,F,M*	*R(%)*
Image1	107	107,0,0	100	107,0,0	100	107,0,0	100
Image2	109	108,0,1	99.0	109,0,0	100	109,0,0	100
Image3	119	117,1,1	98.3	118,1,0	99.1	119,0,0	100
Image4	120	119,01	99.1	120,0,0	100	120,0,0	100
Image5	127	125,1,1	98.4	125,2,0	98.4	126,1,0	99.2
Image6	146	144,2,0	98.6	142,2,2	97.2	143,2,3	97.9
Image7	155	152,1,2	98.0	150,3,2	96.7	148,5,2	95.4
Image8	157	154,2,1	98.0	145,8,4	92.3	147,2,8	93.6
Average Rate		98.65%		98.00%		98.28%	
Time (*s*)		0.5-0.7		2-3		3-4	

The results show that:

- For the first four images with low density and complexity, either SmartScore or SharpEye provides excellent performance (close to 100%), and IMORS is little weaker than them (close to 99%). The limitation of IOMRS is that it can not recognize grace notes and tuplets.
- For the last four images with relatively high density and complexity, SharpEye appears to a noticeable declining trend in performance (from 99.21% to 93.63%), and SmartScore is even worse than SharpEye (from 98.43% to 92.36%). However, IOMRS gives good performance at sustained high rate (more than 98%). It is verified that IOMRS is robust against the impacts of complex data environment.
- For all eight images, the average recognition rate of IOMRS is 98.65%, and is slightly higher than that of SmartScore (98.00%) and SharpEye (98.23%).
- The run time of IOMRS is 0.5-0.7s for one image. It is really a satisfactory result. But this does not reflect that IOMRS is certainly superior to SmartScore (2-3s) and SharpEye (3-4s). The major reason is that the range of symbols to be recognized in IOMRS is just limited to the notes rather than the full musical symbols in SmartScore and SharpEye.

Fig.6 illustrates the results of a real-life score (partial section of the image8 in Table 3) respectively recognized by IOMRS, SmartScore and SharpEye.

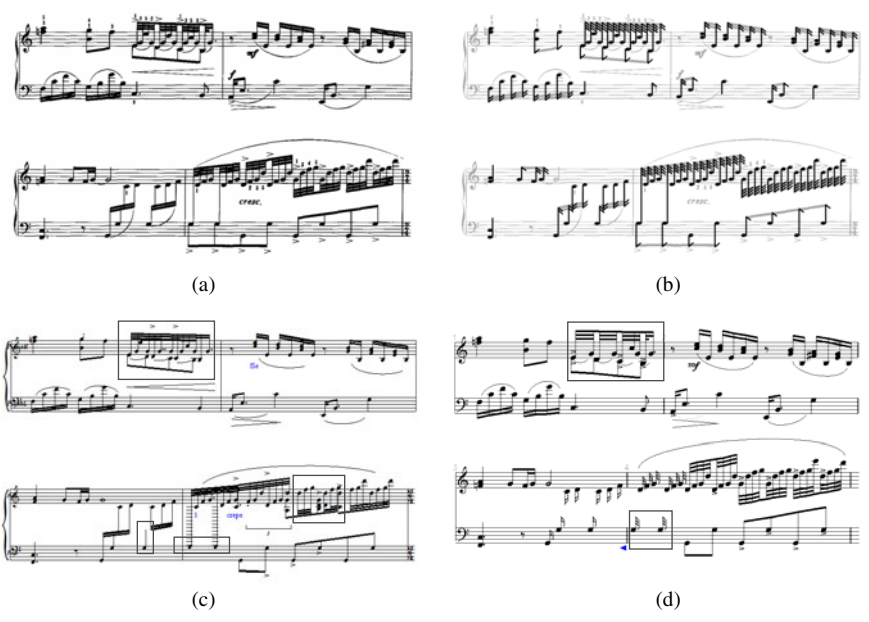


Fig. 6. Comparison among IOMRS, SmartScore and SharpEye: (a) An excerpt of polyphonic score image; (b) Results with IOMRS: no error; (c) Results with SmartScore: several significant errors in the rectangles; (d) Results with SharpEye: several significant errors in the rectangles.

6 Conclusion

Considering the variety and complexity of the music notes, a structure-based recognition strategy is chosen. In the procedure of primitive extraction, three algorithms based on RLE are proposed respectively for extracting stems, note heads and beams. In the procedure of structure analysis, interaction field is introduced to describe the relationship of primitives, and six structures of notes are defined hierarchically, next a sequence with priority mechanism for assembling primitives is proposed. Experimental results have shown that these approaches can resolve the troublesome issues of overlapping and touching symbols, reduce the calculation complexity, and improve the recognition rate, especially for the situation of high density of symbols.

Further research of our system will focus on improving current limitations and the global semantic analysis. The symbols should be expanded from the notes to the full musical symbols.

Acknowledgments. The author thanks all reviewers for their valuable comments and suggestions that helped me for improving our work. This work is funded by the Natural Science Foundation of Guangdong, China under Grant 7300090, and the Talent Introduction Foundation of Jinan University under Grant 510061.

References

1. Ng, K.C., Boyle, R.D.: Recognition and reconstruction of primitives in music scores. Image and Vision Computing 14(1), 39–46 (1996)
2. Blostein, D., Baird, H.S.: A Critical Survey of Music Image Analysis. In: Baird, H.S., Bunke, H., Yamamoto, K. (eds.) Structured Document Image Analysis, pp. 405–434. Springer, Heidelberg (1992)
3. Bellini, P., Bruno, I., Nesi, P.: Optical Music Sheet Segmentation. In: Proceedings of the First International Conference on WEB Delivering of Music (WEDELMUSIC 2001), Florence, Italy, pp. 183–190 (2001)
4. Martin, P., Bellissant, C.: Low-level analysis of music drawings imags. In: Proceedings of the first International Conference on Document Analysis and Recognition (ICDAR 1991), Saint-Malo, France, pp. 417–425 (1991)
5. Carter, N., Bacon, R.: Automatic recognition of printed music. In: Baird, H.S., Bunke, H., Yamamoto, K. (eds.) Structured Document Image Analysis, pp. 456–465. Springer, Heidelberg (1992)
6. Rossant, F., Bloch, I.: Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection. EURASIP Journal on Applied Signal Processing, 815–841 (2007)
7. Modayur, B.: Music score recognition – a selective attention approach using mathematical morphology. Technical Report, Electrical Engineering Department, University of Washington, Seattle (1996)
8. Szwoch, M.: Guido: a Musical Score Recognition System. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brasil, pp. 809–813 (2007)
9. Rebeto, A., Caplea, G., Cardoso, J.S.: Optical recognition of music symbols. International Journal on Document Analysis and Recognition 13, 19–31 (2010)
10. Bainbridge, D., Bell, T.C.: A music notation construction engine for optical music recognition. Software-Practice & Experience 33(2), 173–200 (2003)
11. Coüasnon, B., Brisset, P., Stephan, I.: Using logic programming languages for optical music recognition. In: International Conference on the Practical Application of Prolog (PAP 1995), Paris, France, pp. 115–134 (1995)
12. Fahmy, H.: A Graph-rewriting Papadigm for discrete relaxation: application to sheet-music recognition. International Journal of Pattern Recognition and Artificial Intelligence 12(6), 763–799 (1999)
13. Fujinaga, I.: Staff Detection and Removal. In: George, S.E. (ed.) Visual Perception of Music Notation: On-Line and Off-Line Recognition, pp. 1–39. IRM Press, Idea Group Inc., Hershey (2004)
14. Dalitz, C., Droettboom, M., Czerwinski, B., Fujigana, I.: A comparative study of staff removal algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(5), 753–766 (2008)
15. Dalitz, C., Droettboom, M., Czerwinski, B., Fujigana, I.: Staff removal tookit for gamera, `http://music-staves.sourceforge.net`
16. SmartScore X Pro Demo, `http://www.musitek.com/smartscre.html`
17. SharpEye 2.68 Demo, `http://www.visiv.co.uk/`

Network Vulnerability Analysis Using Text Mining

Chungang Liu[1], Jianhua Li[1,2], and Xiuzhen Chen[2,*]

[1] School of Electronic Information and Electric Engineering, Shanghai Jiaotong University, Shanghai 200240, China
[2] School of Information Security Engineering, Shanghai Jiaotong University, Shanghai, 200240, China
{gangsir,lijh888,chenxz}@sjtu.edu.cn

Abstract. The research on network vulnerability analysis and management has gained increased attention during last decade since many studies have proved that combination of exploits is typical means to compromise a network system. This paper presents an intelligent method for analyzing and classifying vulnerabilities based on text mining technology. The proposed mechanism can automatically classify vulnerabilities into different predefined categories and obtain valuable information from abundant vulnerability texts. A series of experiments on 1060 new reported vulnerabilities in last three years by CERT are performed to demonstrate the efficiency of this mechanism. The results generated by this study can be applied to detecting multistage attack, correlating intrusion alerts, and generating attack graph.

Keywords: Vulnerability analysis, Vulnerability classification, Text mining, k-NN, Naïve Bayes, SVM.

1 Introduction

According to reports published by CERT in 2008, more than 44,000 vulnerabilities have been identified since 1995 [1]. What is more, it is expected that the number of new discovered vulnerabilities will keep its rapid-growing tendency in the future since the computing systems and networks have been keeping growing larger and more complex. The huge amount of vulnerabilities hidden on internet and intranet has led to a great threat to all network systems.

Fortunately, most companies and organizations have been aware of urgency of protecting their information systems, which have become important assets for them with fast development of e-business. In fact, a lot of computer administrators and cyber security researchers have focused on automatic network vulnerability analysis and management.

In this work, we proposed an intelligent approach based on text mining technology to extract useful information automatically from abundant texts provided by software vendors, security experts and institutes. Our study offers a distinctive feature compared to the related research on vulnerability analysis and management. That is

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 274–283, 2012.

we collect vulnerability texts directly from cyber websites and security bulletins which usually contain the most updated information. By classifying them into any predefined category, we can extract information about new discovered vulnerability without the assumption of any other existing repository.

The remainder of this paper is organized as follows. In section 2, we briefly review some related techniques of text mining and vulnerability analysis and management. Section 3 gives the structure of vulnerability analysis using text mining. The research details of text mining including text classification and information extraction are presented in section 4. Section 5 presents the empirical test results using more 1000 new reported vulnerabilities in last three years. Finally, section 6 outlines conclusion and future network.

2 Related Work

The goal of this paper is to achieve automatic classification of vulnerability documents by text mining algorithm. The following section presents related work about vulnerability analysis, management, and text mining.

Vulnerabilities in information technology systems and software can be caused by various factors, but commonly faulty system configuration, bad system design and poor quality of implementation. The work of identifying and handling vulnerabilities has evolved into a crucial cyber security branch called vulnerability analysis and management which includes a cyclical practice of identifying, classifying, accessing and remediating [2].

A lot of research has focused on automatically network vulnerability analysis since Baldwin for the first time proposed a system called Kuang to detect all attack paths in a network [3]. After that, Ou proposed a logic programming approach to network security analysis in his thesis [4]. Geraldine Vache presents the quantitative characterization of vulnerability life cycle and of exploit creation by probability distribution [5]. However, current solutions mainly focus on vulnerability analysis and management at system level. All these systematic solutions require extensive information about vulnerabilities which is usually obtained from existing repository such as NVD.

Text mining refers to the application of methods and algorithms which can be used to obtain non-trivial information from large collection of unstructured texts [6]. Text classification, a subject involved in this paper, is to label the natural language texts with thematic categories from a predefined set [7]. In a mathematical view, the progress of classification can be formally described as a mapping function which assigns a predefined category to a new document.

In this paper, text mining technology is firstly introduced for analysis of vulnerability by classifying vulnerability documents into a set of pre-defined categories and extracting useful information from abundant vulnerability text resources. The result of this study lays the foundation of attack graph generation, intrusion alert correlation and multi-step attack generation.

3 System Architecture of Vulnerability Discovery and Analysis

The architecture of text mining based vulnerability analysis is mainly composed of two parts: vulnerability discovery tool (VULDis) and text mining server, illustrated in Figure 1. The VULDis part is responsible for discovering vulnerabilities and collecting related documents. And the core component: text mining server implements the analysis of vulnerability text, including document labeling, preprocess, feature selection, text classification, evaluation and so on. A prototype demo system is developed using C++ and Python as programming languages as well as Mysql as a database engine. The functional details and related information flow are depicted in Figure 1.

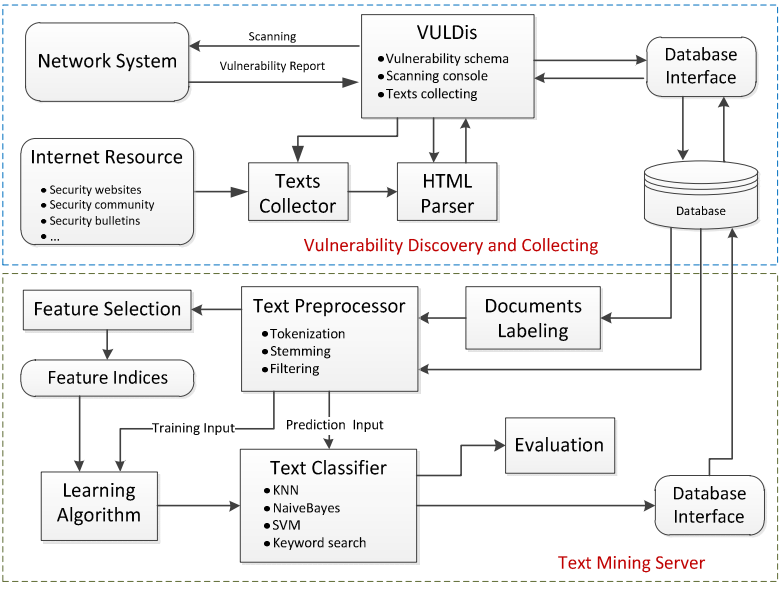


Fig. 1. Architecture of vulnerability analysis using text mining

- VULDis

VULDis is a vulnerability scanning and management tool that we developed using C++. It is based on open vulnerability assessment language (OVAL), adopts the client/server system architecture and has two components: agent and server.

Once a user has installed the agent component to computers in target system, the user can start scanning process by configuring the console. A scanning report in format of XML will be generated and sent to console. After that, the VULDis will search and collect text resources about the identified vulnerabilities from internet as many as possible. These text resources are then processed and stored into database for further mining.

- Text Mining Server

When enough text data is prepared, the task of text mining can be invoked by user after configuring some parameters. This task includes handcrafted labeling, text

preprocessing, feature selection, training and text prediction. The user can change the dimension of feature indices and specify algorithm to obtain different performance of classification. More technique details of this module are elaborated in section 4.

4 Text Mining Based Vulnerability Analysis

The text mining task in this paper mainly consists of three steps: text preprocessing, text classification and evaluation. Their concrete design schemes are shown as follows:

4.1 Text Preprocessing and Vector Representation

The documents we need to handle are gathered by VULDis from security community's bulletins, software vendor's websites and so on. It is necessary to preprocess these documents and store the text information in a relative structured format, which should be more suitable for further mathematical and statistical analysis. The main preprocessing steps consist of tokenization, stop words filtering, stemming, term frequency and document frequency count. We will not detail any of these subtasks, because there have been considerable research theses about them [8]. In this section, we mainly discuss how a text document is mathematically represented using vector representation model.

In order to formally describe our work, the following variables are given firstly. $D=\{d_1, d_2,..., d_s\}$ represents the collection of all documents we need to cope with. $TF(t_i, d_j)$ refers to the absolute frequency of t_i occurring in the document of $d_j (d_j \in D)$ and $DF(t_i)$ indicates the frequency of documents containing t_i occurring in D. $C=\{c_1,c_2...,c_m\}$ is the set of categories we predefined. $P(c_i)$ indicates the probability of a random document belongs to category $c_i (c_i \in C)$ while $P(c_i|t_j)$ indicates the conditional probability of a document containing term t_j belongs to category $c_i (c_i \in C)$.

Then we can represent any document by a corresponding vector, of which each weight is calculated based on *TFIDF* (Term-Frequency Inverse-Document-Frequency) [9]. Each element v_j ($1<j<|T|$), usually called term weight, reflects the importance of the corresponding term to categorization of all documents.

$$V_{d_i} = \{v_1, v_2, \cdots, v_M\} \quad with \quad v_j = TF(t_j, d_i) * \log\left(|D| / DF(t_j)\right) \quad . \tag{1}$$

4.2 Vulnerability Classification

The task of classification starts with a set of training text documents D_{train} which are already labeled with a correct class c_i. The set of C can be predefined according to different criteria. For example, according to the way that vulnerability is exploited, C can be defined as {remotely, locally} or {with authentication, without authentication} while from the view of privilege escalation after exploit, C can be defined as {none, user, root}. In this study, the classifier is not overlapped.

For the application of this study, we use three prevalent distinct algorithms for classification. In the following part of this section, we will briefly introduce them.

4.2.1 k-Nearest Neighbors Classifier

k-Nearest Neighbors (k-NN) is a lazy learning and example-based algorithm which was first proposed by Yang and Chute [11]. In this algorithm, a random document is classified by determining the class of its *N* nearest neighbors (i.e. *N* most similar training documents $D_{neighbor}$). The similarity of two document vectors in our work is defined as their cosine angle $S(d_i,d_k)$.

An accumulated score for each category can be calculated by adding up all the similarity score with respect to the neighbor documents belong to this category. Then, the class c_j with the highest score $W(d_i,c_j)$ is assigned to document d_i.

$$W(d_i,c_j)=\sum_{d_k\in D_{Neighbor}} S(d_i,d_k)\tau(d_k,c_j),\ \ \tau(d_k,c_j)=\begin{cases}1, & \varphi(d_k)=c_j\\ 0, & \varphi(d_k)\neq c_j\end{cases}. \tag{2}$$

However, in our application of classifying vulnerability texts from different websites, the traditional k-NN algorithm does not perform so satisfactory as a result of sparse term problem. The sparse term refers to the term which occurs in only one or a few texts but has high entropy or information gain. These sparse terms which appears frequently in the web texts about network vulnerability give rise to large amount of zeros in the indexing vector of documents. To solve this problem, we set up a threshold for the *DF* of each term to filter sparse terms.

4.2.2 Naïve Bayes Classifier

Naïve Bayes is probabilistic classifier based on the assumption that the words of a document d_i have been generated by a probabilistic mechanism and are independent of each other [12]. This Naïve assumption can be described as:

$$P(t_1,...,t_M\,|c_i)=\prod_{j=1}^{M}P(t_j\,|c_i)\ . \tag{3}$$

For the sake of classification, it is supposed that the class of document d_i has some relation to the words which appear in the document. This relation can be described by a conditional probability. The task of classification turned out to be a problem of finding out the label with the highest estimated conditional probability.

$$\max.P(c_i\,|t_1,...,t_M)=\max.P(t_1,...,t_M\,|c_i)P(c_i)=\max.P(c_i)\prod_{j=1}^{M}P(t_j\,|c_i)\ . \tag{4}$$

4.2.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised classification algorithm which was first presented by Cortes and Vapnik [13]. One single SVM classifier can only separate two classes: a positive class (*L1:* $y = 1$) and a negative class (*L2:* $y = -1$). After receiving training examples in the form: $\{\boldsymbol{x_i}, y_i\}$, $\boldsymbol{x_i}\in D_{train}$, $y_i\in C$, a hyperplane

which is located between positive and negative examples is determined by SVM algorithm. The hyperplane may be defined in a linear or non-linear equation by setting y = 0. In this work, we used linear SVM.

$$y = f(\vec{x_i}) = \vec{w}^T \vec{x_i} + b \quad where \quad \vec{w} \in R^d, b \in R \,. \tag{5}$$

Then, the parameter vector $\boldsymbol{w}$ is adapted in such a way that the distance between the separation hyperplane and closest positive and negative training cases is maximized. This amounts to a constrained quadratic optimization problem which can be described as follows:

$$min \ \frac{1}{2}\left\|\vec{w}\right\|^2 \quad subject\ to \quad y_i(\vec{w}^T \vec{x_i} - b \geq 1), \forall i \ . \tag{6}$$

These closest example documents are called support vectors and the maximum distance is called margin. Then a new document with term vector $\boldsymbol{x_j}$ is labeled with *L1* if $f(\boldsymbol{x_j}) > 0$, otherwise with *L2*. In this work, we combined several two-class SVM classifiers to perform a multi-class prediction [14].

4.3 Measures of Classification Performance

True Positives (TP_i), True Negatives (TN_i), False Positives (FP_i) and False Negatives (FN_i) are four common measures for evaluating how successful a classifier is. TP_i are the relevant cases that belong to category C_i and are correctly identified as yes. TN_i are the irrelevant cases that do not belong to category C_i and are correctly identified as no. FP_i are the irrelevant cases that do not belong to category C_i but are incorrectly identified as yes. FN_i are the relevant cases that belong to category C_i but are incorrectly identified as no.

Given above four measures, another two common measures can be defined. The precision ρi estimates the probability that a document really belongs to the category ci which the classifier assigns to it. However, the recall γi indicates the probability that a document is correctly assigned to the category ci which it actually belongs to. These two measures are separately defined as:

$$\rho_i = \frac{TP_i}{TP_i + FP_i} \quad and \quad \gamma_i = \frac{TP_i}{TP_i + FN_i} \ . \tag{7}$$

Moreover, precision and recall rate are generally combined into one single metric called F. F is defined to be a harmonic mean of precision and recall. The parameter α (0<α<1) serves to balance the importance of precision and recall rate. By setting α to be 0.5, we can get a simpler expression:

$$F = \frac{\rho\gamma}{\alpha\gamma + (1-\alpha)\rho} \quad and \quad F^1 = \frac{2\rho\gamma}{\rho + \gamma} \quad when \quad \alpha = 0.5 \,. \tag{8}$$

In order to measure the global performance of the classifier taking into account of all predefined categories, we introduced two averaged measure called macro *F* value and micro *F* value. Macro *F* value, identified as *MF* focus more on frequent categories while micro *F* value, identified as μF gives more emphasis on rare ones. They are defined by:

$$MF^1 = \frac{\sum_{i=1}^{|C|} F_i^1}{|C|} \quad and \quad \mu F^1 = \frac{\sum_{i=1}^{|C|} 2TP_i}{\sum_{i=1}^{|C|} (2TP_i + FP_i + FN_i)} \quad . \tag{9}$$

5 Experiment and Discussion

5.1 Data Description

In order to evaluate the performance of our research, we performed some off-line tests on basis of 1060 new reported vulnerabilities in last three years by CERT. These vulnerabilities are selected at random and assigned with a thematic label by hand. In the empirical test, we predefined the category set to be {SQL Injection, Code Injection, Buff Errors, Access Issue} according to the way vulnerability is exploited. The definition of these category labels refer to Common Weakness Enumeration (CWE) which is a community developed dictionary of software weakness types. The category set is incomplete and just includes part of vulnerabilities.

Then, thousands of web documents are collected by VULDis from different security websites, such as IBM Internet Security Systems Ahead of the threat, securityTracker.com, securityfocus.com, and so on. The testing vulnerabilities are depicted in Table 1.

Table 1. Vulnerabilities used in empirical test

Category Label	For Training			For Testing		
Year	2009	2010	2011	2009	2010	2011
SOL Injection	5	61	11	40	193	24
Code Injection	18	34	1	68	117	4
Buff Errors	0	13	21	2	121	108
Access Issue	4	25	7	6	118	59
Total	200			860		

5.2 Vulnerability Classification Result

In order to build classification model of three categories, we randomly selected two hundred vulnerabilities as a training set and collected 546 web documents about them. VULDis includes a HTML parser to clean the document content by removing meta tags, image maps, JS codes, forms and tables. After that several documents about the same vulnerability are combined to one and labeled with a proper category tag.

After training, 860 vulnerabilities are prepared and sent to text classification model to do prediction. In this study, we used three prevalent classification algorithms as well as the simple keyword search method.

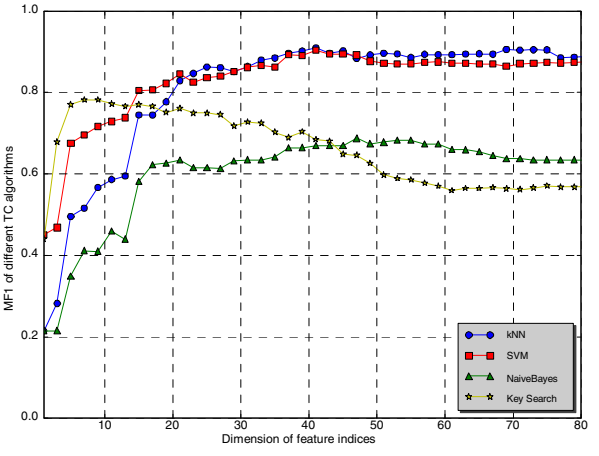


Fig. 2. MF1 of four classifying algorithms for different dimensions of feature indices

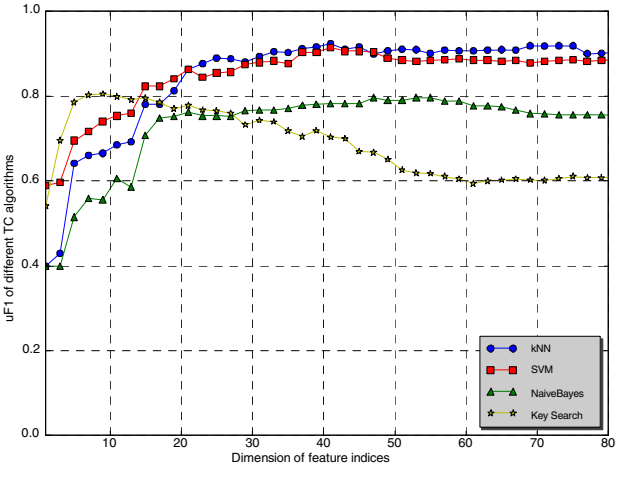


Fig. 3. μF1 of four classifying algorithms for different dimensions of feature indices

Figure 2 and 3 display two averaged measures of the classification performance. It can be noted that SVM and k-NN classifier achieved better results than Naïve Bayes classifier. As the dimension of feature indices gets bigger, the performance of SVM and k-NN classifier gets similar with a MF1 of 0.87 to 0.92. We can also note that all three text classification algorithms produced better performance than the simple keyword searching method.

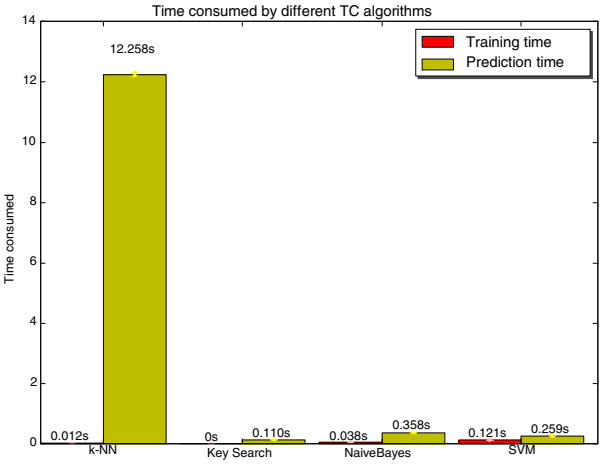


Fig. 4. Time consumed by different classifying methods

We also tested system cost of each classifying algorithm. The results are presented in Figure 4. It can be observed that k-NN algorithm requires the least training time and its example-based nature makes it slow when predicting new documents. SVM classifier takes much less time while Naïve Bayes classifier offers the fastest performance. Although the keyword searching method requires no previous training and offers fast predicting speed, the accuracy performance is much worse than the other three text classification methods.

6 Conclusions

How to efficiently manage and analyze vulnerabilities has become a great challenge for network administrators with the exponentially increase of network vulnerabilities during last decades. In this paper, we propose a new method using text classifying technology to obtain useful information about vulnerabilities from abundant web texts automatically. This study introduced a completely new artificial intelligent mechanism: text mining for managing and analyzing vulnerability. Feature selection and three text classification algorithms are applied to select text feature terms and train vulnerability classifier, respectively. A series of experiment result shows that the correctness of SVM based classification exceeds 90%. This mechanism proposed in this paper greatly reduced the need for human effort to manage and analyze the ever-identified vulnerabilities. The classification results provide us with much information about vulnerability type, fundamental generation reason, and so on. Moreover, the output of text mining based vulnerability analysis system can be used to detect multistage attack, correlate intrusion alerts and generate attack graph.

There is much more to be explored in the future. Our future work will focus on automatically mining the prerequisites and consequences of vulnerability exploitation from vulnerability texts so as to enlarge the knowledge database used in attack graph generation.

Acknowledgments. This paper was supported by State Key Development Program of Basic Research of China (No.2010CB731403/2010CB731406) and National Natural Science Foundation of China (No. 61071152).

References

1. CERT Statistics (1995-2208), http://www.cert.org/stats/
2. Foreman, P.: Vulnerability Management. Taylor & Francis Group (2010)
3. Baldwin, R.: Rule based analysis of computer security, Technical Report TR-401, MIT LCS Lab (1988)
4. Ou, X., Govindavajhala, S., Appel, A.W.: MulVAL: A logic-based network security analyzer. In: 14th USENIX Security Symposium, Society for Industrial and Applied Mathematics (2005)
5. Vache, G.: Vulnerability analysis for a quantitative security evaluation. In: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement (2009)
6. Ben-Dov, M., Feldman, R.: Text Mining and Information Extraction. Part 6, 809–835 (2010)
7. Hearst, M.A.: Untangling text data mining. In: Proceedings of the 37th Conference on Association for Computational Linguistics. Association for Computational Linguistics, College Park, Maryland (1999)
8. Porter, M.: An algorithm for suffix stripping. Program, 130–137 (1980)
9. Metzler, D.: Generalized inverse document frequency. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management (2008)
10. Pudil, P., Somol, P.: Current Feature Selection Techniques in Statistical Pattern Recognition. In: Computer Recognition Systems. Advances in Soft Computing, vol. 30 (2005)
11. Yang, Y., Chute, C.G.: An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems 12(3), 252–277 (1994)
12. Lewis, D.D.: Naive (Bayes) at forty: the independence assumption in information retrieval. In: Nedellec, C., Rouveirol, C. (eds.) Proceedings of ECML1998, 10th European Conference on Machine Learning. Springer, Heidelberg (1998)
13. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
14. Hsu, C.W., Lin, C.J.: A comparison on methods for multi-class support vector machines. IEEE Transactions on Neutral Networks (2002)

Intelligent Information System for Interpretation of Dermatoglyphic Patterns of Down's Syndrome in Infants

Hubert Wojtowicz[1] and Wieslaw Wajs[2]

[1] University of Rzeszow, Faculty of Mathematics and Nature, Institute of Computer Science, Rzeszow, Poland
[2] AGH University of Science and Technology, Faculty of Electrical Engineering, Institute of Automatics, Cracow, Poland

Abstract. The paper describes design of an intelligent information system for assessment of dermatoglyphic indices of Down's syndrome in infants. The system supports medical diagnosis by automatic processing of dermatoglyphic prints and detecting features indicating presence of genetic disorders. Application of image processing and pattern recognition algorithms in pattern classification of fingerprints and prints of hallucal area of the sole is described. Application of an algorithm based on multi-scale pyramid decomposition of an image is proposed for ridge orientation calculation. A method of singular points detection and calculation of ATD angle of the palm print is presented. Currently achieved results in dermatoglyphic prints enhancement, classification and analysis are discussed. Scheme used in classification of dermatoglyphic prints is described. RBF and triangular kernel types are used in the training of SVM multi-class systems generated with one-vs-one scheme. Results of experiments conducted on the database of Collegium Medicum of the Jagiellonian University in Cracow are presented.

1 Introduction

Early detection of genetic disorders in infants, allowing for making therapeutic decision and starting a treatment, is largely dependent on the ability of carrying out fast and reliable observations and obtaining results of these observations. One of the methods of detecting genetic disorders is dermatoglyphic analysis carried out by an anthropologist. Dermal ridge patterns on fingers, palms and soles used in this method become visible at about three months and are completed by the sixth month of prenatal development. Factors disturbing normal development of fetus may also influence formation of dermal ridges structures. Down's syndrome (trisomy 21), one of the most common chromosome disorders and Turner's syndrome can be detected using dermatoglyphic patterns analysis [6] [8]. Dermatoglyphic patterns of infants with genetic disorders differ from normal patterns found in healthy population. Determining the presence of genetic disorder requires a simultaneous analysis of dermatoglyphic prints of fingers,

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 284–293, 2012.

palms and soles. The presence of a single pattern typical for the particular genetic syndrome in any of the considered areas is not indicative of Down's or Turner's syndrome. Many of these patterns can be found in healthy infants. However, when several, or all, of the patterns characteristic for a genetic disorder are present together, they are indicative of its presence. For the detection of Down's syndrome a diagnostic index was developed called dermatoglyphic nomogram [7]. Diagnostic index used in screening tests for Turner's syndrome's presence was also developed [5]. Both of these indexes rely on a correct recognition of dermatoglyphic patterns by the anthropologist.

2 The Aim of the Work

Problem of pattern recognition and pattern understanding of genetic traits of infants with Down's syndrome is a difficult and complex issue. Available data estimate that around 60000 people living in Poland were diagnosed with Down's syndrome. Development of modern telecommunication networks and Internet in particular, enables creation of the telemedical system localized in Collegium Medicum of the Jagiellonian University in Cracow. Clinique of Jagiellonian University has many years of experience as a centre of genetic disorders diagnostics in neonates. Scientific faculty of the Clinique of Jagiellonian University provides substantial support in realization of the project. Conception of the project assumes development, in the Clinique of Jagiellonian University, of a prototype of the telemedical system, in which data in the form of dermatoglyphic images is transferred through the telecommunication networks from the distant hospital database centers and processed. The aim of the telemedical system is screening analysis of incoming data. Another goal is a support of the diagnosis process conducted by medical personnel in cases of ambiguous classification of the telemedical system.

The aim of the research conducted is the creation, based on gathered data and domain knowledge described in medical literature, of an automatic system supporting the diagnosis process and detecting infants' genetic disorders, as follows:

1. The system recognizes characteristic combinations of particular patterns of soles, palms and fingers and on that basis, infers the occurrence of genetic disorders. It is expected that the application of this system improves treatment's effectiveness, i.e. the number of complications caused by the treatment in the later years of infants' life is going to be lower.
2. The system supports doctor's work by the analysis of large amounts of patients' data and decreases the probability of a mistake in strenuous biometric analysis such as counting the number of ridges, determining ridge width or calculating the ATD angle.

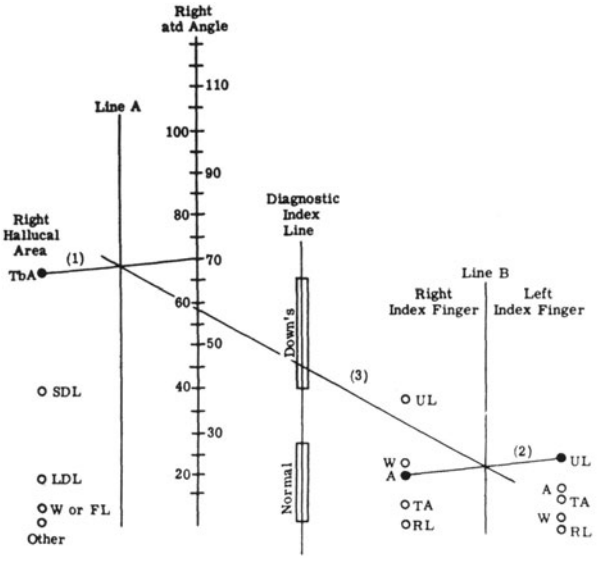


Fig. 1. Example of the dermatoglyphic nomogram. Abbreviations: TbA - tibial arch; SDL - small distal loop; LDL - large distal loop; FL - fibular loop; W - whorl; UL - ulnar loop; A - arch; TA - tented arch; RL - radial loop.

3 Description of Dermatoglyphic Nomogram Used in the Analysis of Down's Syndrome

Dermatoglyphic nomogram is based upon four pattern areas and uses four dermatoglyphic traits chosen for their high discriminant value. These traits include: pattern types of the right and left index fingers, pattern type of the hallucal area of the right sole and the ATD angle of the right hand.

Figure 1 presents an example of the nomogram [7]. To determine the patient's index score three lines are constructed:

1. Line 1 is drawn connecting the appropriate point on the ATD axis to the circle corresponding to the right hallucal pattern type.
2. Line 2 connects the corresponding circles for the pattern types of the right and left index fingers.
3. Line 3 connects the points of intersections of lines 1 and 2 with lines A and B, respectively. The point at which line 3 crosses the diagnostic index line determines whether the individual has dermatoglyphics within the Down's syndrome or normal distribution (bars) or whether it falls into the overlap area. In the example of the patient scored, line 3 crosses the diagnostic index line within the distribution of the Down's syndrome patients.

Designations referring to the anatomy of the hand are used in describing palmar dermatoglyphics and in presenting methods of interpreting them. Terms of

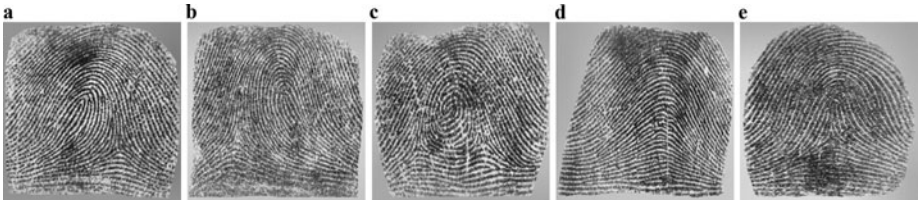


Fig. 2. Example fingerprints: (a) left loop; (b) right loop; (c) whorl; (d) plain arch; (e) tented arch

anatomical direction (proximal, distal, radial, ulnar) are employed in describing the locations of features and in indicating directions toward the respective palmar margins. Fingerprint pattern left loop is called ulnar loop when found on any of the fingers of the left hand and it is called radial loop when found on any of the fingers of the right hand. Pattern right loop in the left hand is called radial loop and in the right hand is called ulnar loop [2].

4 Classification Method of Fingerprint Images

Fingerprint classification is one of the tasks of dermatoglyphic analysis. Many classification methods were developed and described in the literature. Classification method used in dermatoglyphic analysis is called the Henry method. It classifies fingerprints into five distinct classes called: left loop (LL), right loop (RL), whorl (W), arch (A) or plain arch (PA) and tented arch (TA) (Fig. 2).

Classification scheme based on the Henry method is a difficult pattern recognition problem due to the existence of small interclass variability of patterns belonging to the different classes and large intraclass variability of patterns belonging to the same class. In the paper we present a classification scheme based on the extraction of fingerprint ridge orientation maps from the enhanced images. Vectors constructed from the directional images are used for training of SVM multi-class algorithms. For the induction process of SVMs we use RBF and triangular type kernels.

5 Classification of the Hallucal Area of Sole Patterns

Classification of the patterns in the hallucal area of the soles is another task of dermatoglyphic analysis. The patterns in the hallucal area are classified into five distinct classes called: large distal loop (LDL), small distal loop (SDL), whorl (W), tibial arch (TA) and tibial loop (TL) (Fig. 3).

Classification of the patterns in the hallucal area of the soles is also a difficult pattern recognition problem due to the existence of small interclass variability of patterns belonging to the different classes and large intraclass variability of patterns belonging to the same class. The upper row in Fig. 4 shows three impressions of different topology all belonging to the whorl class, the lower row in Fig. 4 shows

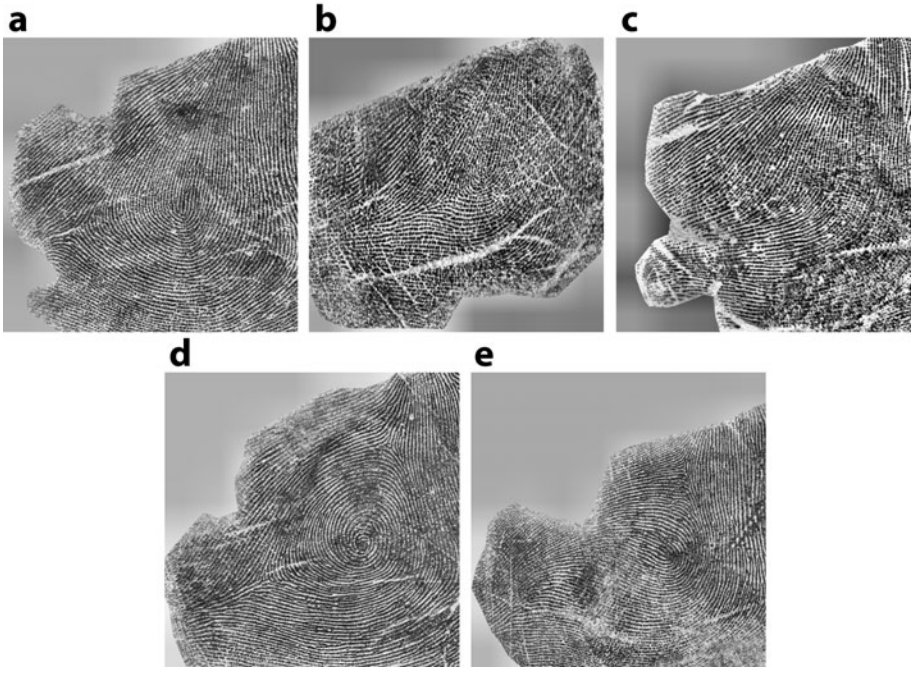


Fig. 3. Example patterns in the hallucal area of the soles: (a) large distal loop , (b) small distal loop, (c) tibial arch, (d) whorl, (e) tibial loop

on the left impression belonging to the class tibial arch, and on the right impression belonging to the class small distal loop.

6 Method of the ATD Angle Calculation

Value of the ATD angle of the right palm is calculated by finding locations of digital triradii labeled A and D and location of axial triradius labeled t. Mean value of the ATD angle for normal infants is $\sim 48°$. Mean value of the ATD angle for infants with Down's syndrome is $\sim 81°$.

Locations of palmprint's singular points are calculated from the orientation map. Orientation map is computed using algorithm based on the principal component analysis and multi-scale pyramid decomposition of the image. Principal Component Analysis is applied to the image to calculate maximal likelihood estimations of the local ridge orientations and the multi-scale pyramid decomposition of the image helps to improve the accuracy of the calculated orientations. Algorithm calculating ridge orientations is robust to noise and allows for reliable estimation of local ridge orientations in the low quality areas of images. After the process of estimation of ridge orientations a Poincare index is calculated, which

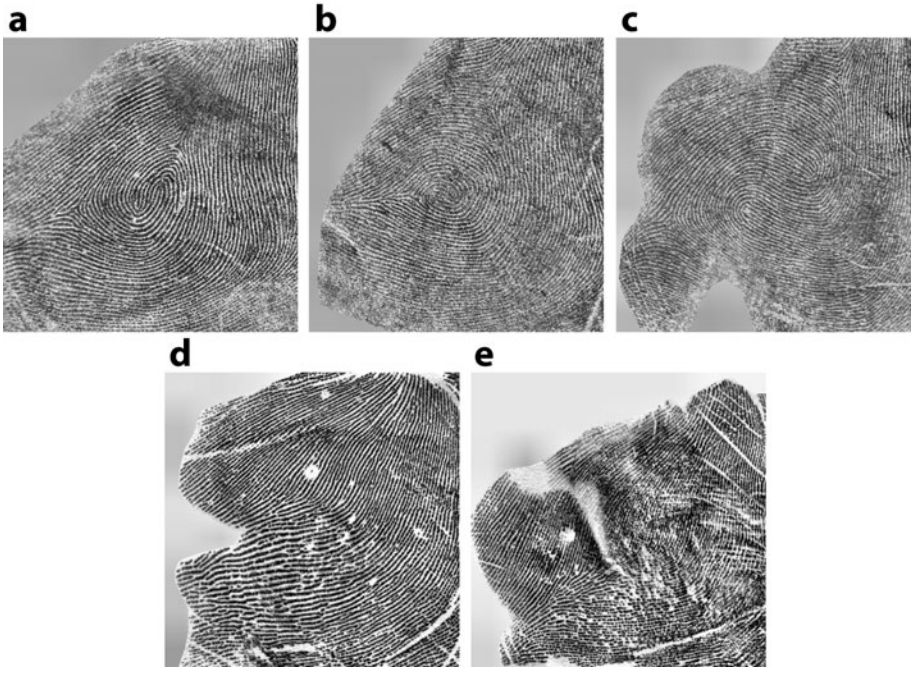


Fig. 4. The upper row contains patterns of the hallucal area of the soles belonging to the same class but of different topology (large intraclass variability), the lower row contains patterns belonging to the different classes but of similar topology (small interclass variability).

in turn allows for extraction of singular points from the palmprint image. Basing on the locations of singular points the ATD angle of the palmprint is calculated.

A two stage algorithm is applied in order to reliably estimate positions of singular points. In the first stage improved formula of Poincare index is used to calculate candidate locations of singular points [9]. In the second stage a coherence map of the palmprint image is calculated. The coherence map is calculated for each pixel in the image on the basis of its eigenvalues. Pixel eigenvalues are calculated from the confusion matrix containing values calculated for each pixel (i,j) by applying the combination of two dimensional orthogonal Gaussian-Hermite moments to the segment of the image centered in pixel (i,j). The proposed approach combines information about singular points obtained from the improved Poincare index calculated from the orientation map and the information obtained from the coherence map calculated from the pixel map of the image. The two stage approach allows for reliable detection of only true singular points even in low quality areas of the palmprint image [9]. Example of calculation of palmprint's singular points using proposed algorithm is presented in Fig. 5.

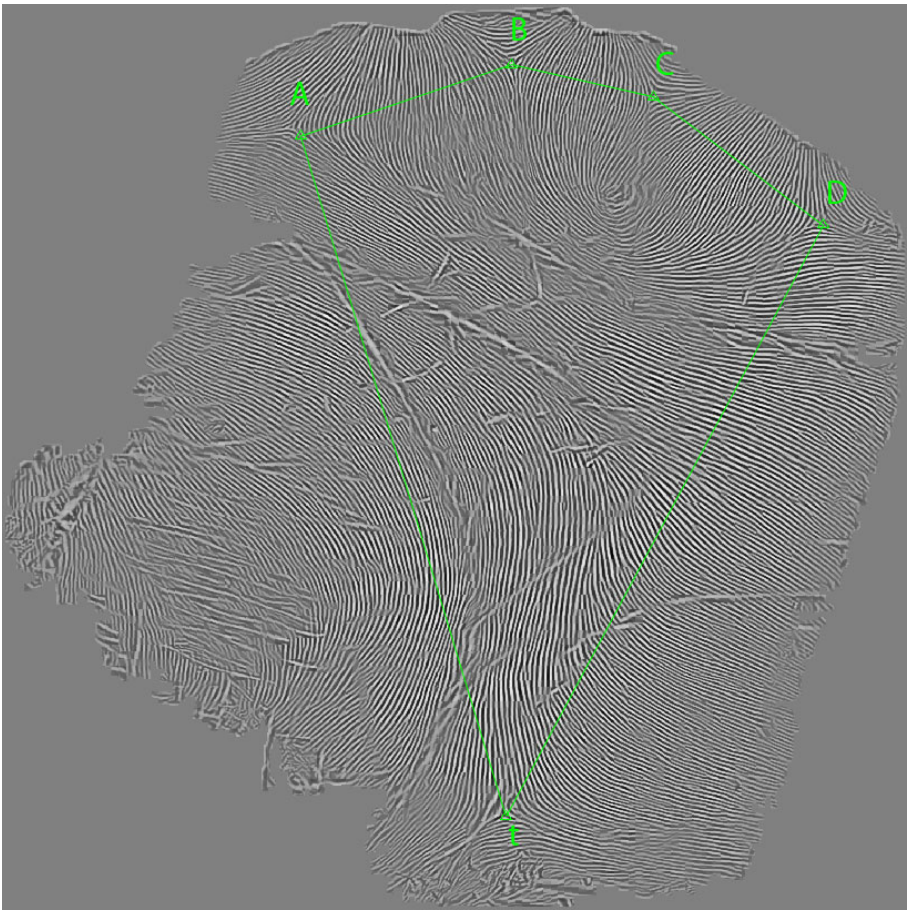


Fig. 5. Singular points of the palmprint located using two stage algorithm taking advantage of improved Poincare index and Gauss-Hermite moments

7 Feature Extraction

Accurate extraction of features and classification of fingerprints depends on the quality of the images containing the impressions. Determination of the quality of the impressions partly depends on subjective criteria. It corresponds to the clarity of the dermal ridges structure. Impressions of good quality are characterized by high contrast and clearly discernible structure of ridges and valleys, poor quality impression are of low contrast and ill-discernible structure of ridges and valleys [1].

Factors that can adversely affect the quality of impressions are:

(a) Presence of furrows interrupting ridge continuity. Although high number of furrows on the fingerprint makes the recognition task harder in case of genetic disorders it may be an indication of Downs syndrome.
(b) Dry fingers yield impressions of fragmentary ridge structure and low contrast.
(c) Sweat on fingerprints leads to smudges and false connections of parallel ridges.

Image pre-processing of fingerprint impressions consists of several stages. In the first stage image segmentation takes place, which separates background from the parts of the image where ridges are present. After the region mask representing the foreground area is determined, the background area is removed from the image. Coordinates of the boundary points of the region mask are found. The image is truncated according to the boundary points coordinates and then using bicubic interpolation resized to the frame of 512 x 512 pixels. In the next stage image contrast enhancement is performed using locally adaptive contrast enhancement algorithm. In the last stage of pre-processing image quality enhancement is carried out using a contextual image filtration STFT (Short Time Fourier Transform) algorithm [1], which generates information about ridges flow directions, frequency of ridges and local image quality estimation. Ridge orientation maps are computed using algorithm based on Principal Component Analysis and multi-scale pyramid decomposition of the images [3]. Principal Component Analysis is applied to fingerprint image to calculate maximal likelihood estimations of the local ridge orientations. PCA based estimation method uses Singular Value Decomposition of the ensemble of gradient vectors in a local neighborhood to find the dominant orientation of the ridges. Multi-scale method provides robustness to noise present in the image.

8 Fingerprint Classification with SVM Algorithm

The data set used in the experiment consists of 600 fingerprint 8-bit grey scale images. The size of the images is 512 x 512 pixels. The classes of left loop, right loop, whorl arch and tented arch are uniformly distributed in the data set and consist of 120 images each. A number of images containing partial or distorted information about the class of the pattern for which two or more image copies of the same impression were available were registered using non rigid registration algorithm and then mosaicked. Fingerprint classification was accomplished using a SVM algorithm. For multi-class problem an ensemble of SVM classifiers was created trained with one vs one voting method. Classifiers were using RBF type kernel functions and triangular kernel [4]. Training dataset consists of 300 ridge orientation maps calculated from the fingerprint images. It contains 60 maps for each of the five classes. The dataset used for testing of the SVM is comprised of 300 ridge orientation maps. There are 60 maps of LL, RL, W, A and TA classes in the testing set. Images were in the first stage pre-processed using CLAHE algorithm and then filtered using STFT. Ridge orientation maps were computed from the filter enhanced images using PCA and multi-scale pyramid decomposition algorithm [3]. Cross-validation and grid search were used to obtain kernel

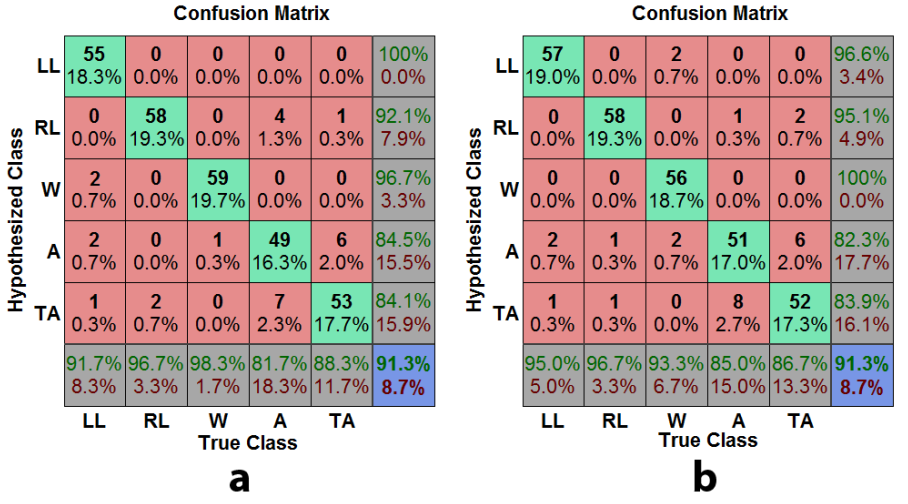


Fig. 6. Test results for the SVM algorithm trained with: (a) RBF kernel function, (b) triangular kernel function

parameters for the training of the SVM algorithms. Test results of the SVM algorithm with RBF and triangular kernel functions are presented as confusion matrices in Fig. 6a and Fig. 6b, respectively. Both the SVM trained with the RBF kernel and the SVM trained with the triangular kernel achieved classification accuracy of 91,3 % on the test data set. The experiments described in this paper were performed on the images of varying quality. Most of these images were of low contrast due to the gradual fading of the chemical compound used to create impressions of the infants' fingerprints. The training dataset contained partially blurred fingerprint impressions or impressions of incomplete patterns. An accurate recognition of ridge directions in the completely blurred areas of the fingerprint is a difficult task. Principal component analysis and multi-scale pyramid decomposition of the image allows a reliable estimation of local ridge directions in the blurred areas, but with the increase of the blurred area size certainty of accurate estimation decreases. Accurate estimation of ridges direction may not be possible if the blurred area contains characteristic points such as cores or triradii. Inclusion of low quality impressions negatively influences selection of the parameters for the training of SVM classifiers and also has a negative impact on the testing accuracy. Achieved classification accuracy is made possible thanks to the pre-processing of the images and application of the RBF and triangular kernel functions in the SVM training scheme.

9 Summary

In this paper we presented a design of the intelligent information system for the detection of Down' syndrome basing on the dermatoglyphic nomogram. Three

pattern recognition tasks required for the determination of patient's diagnostic score were described. These tasks are fingerprint pattern classification, hallucal area of the sole pattern classification and calculation of palm print's ATD angle. Currently achieved methods for fingerprint classification were presented. Our fingerprint classifier allows a reliable determination of the left and right index finger patterns which are two of the traits used in the nomogram diagnostic index. The most common pattern types in Down's syndrome which are left and right loops are classified with a 95.0% accuracy for the left loop type and a 96.7% accuracy for the right loop type when using the triangular kernel function in the SVM algorithm. In our future work the same approach consisting of image enhancement, feature extraction and classification stages will be applied to the recognition of the pattern types of the hallucal area of the soles. Filter methods and algorithm calculating ridge orientation information described in this paper is also used in the detection of singular points of the hand allowing for automatic measurement of the ATD angle.

References

1. Chikkerur, S., Cartwright, A.N., Govindaraju, V.: Fingerprint image enhancement using STFT analysis. Pattern Recognition 40, 198–211 (2007)
2. Cummins, H., Midlo, C.: Fingerprints, palms and soles - Introduction to Dermatoglyphics. Dover Publications Inc., New York (1961)
3. Feng, X.G., Milanfar, P.: Multiscale principal components analysis for image local orientation estimation. In: Proc. of the 36th Asilomar Conf. on Signals, Systems and Computers, vol. 1, pp. 478–482 (2002)
4. Fleuret, F., Sahbi, H.: Scale-invariance of support vector machines based on the triangular kernel. In: Proc. of the IEEE Int. Conf. on Computer Vision, ICCV (2003)
5. Preuss, M.: A screening test for patients suspected of having Turner syndrome. Clinical Genetics (10), 145–155 (1976)
6. Reed, T.: Dermatoglyphics in Down's syndrome. Clinical Genetics (6), 236 (1974)
7. Reed, T.E., et al.: Dermatoglyphic nomogram for the diagnosis of Down's syndrome. J. Pediat. (77), 1024–1032 (1970)
8. Tornjova-Randelova, S.G.: Dermatoglyphic characteristics of patients with Turners syndrome. Medicine Anthropologie 43(4), 96–100 (1990)
9. Yin, Y., Weng, D.: A New Robust Method of Singular Point Detection from Fingerprint. In: Proc. of Int. Symposium on Information Science and Engieering. IEEE (2008)

Using a Neural Network to Generate a FIR Filter to Improves Digital Images Using a Discrete Convolution Operation

Jakub Pęksiński and Grzegorz Mikołajczak

West Pomeranian University of Technology,
Faculty of Electrical Engineering
26 Kwietnia 10, 71-126 Szczecin, Poland
{jakub.peksinski,grzegorz.mikolajczak}@zut.edu.pl

Abstract. The aim of the article is to show correction possibilities of digital images, achieved by image acquisition tools built on low class CCD matrices, such as their quality were close to images achieved by high class tools. For this purpose the authors used the linear filter with 3x3 mask, which were generated with neural network. Digital images were compared using the quality metrics such as MSE, NMSE and Q.

Keywords: digital image, neural network, FIR filter.

1 Introduction

Dynamic development of image acquisition tools based on matrices CCD (*Charge Coupled Device*) became a must have almost in every home. Few years ago high class scanners were very expensive and were used as equipment in professional companies, handling with digital image processing. It was related to production expanses of CCD matrices, which satisfy high requirements of sensivity and quantum efficiency. Professional equipment that has high quality parameters and really maps image is still expensive and is owned primarily by professional companies that deal with digital image processing. The purpose of this article is to present opportunities to improve the quality of images obtained with low-class devices, so their quality is comparable to images obtained by high-end devices.

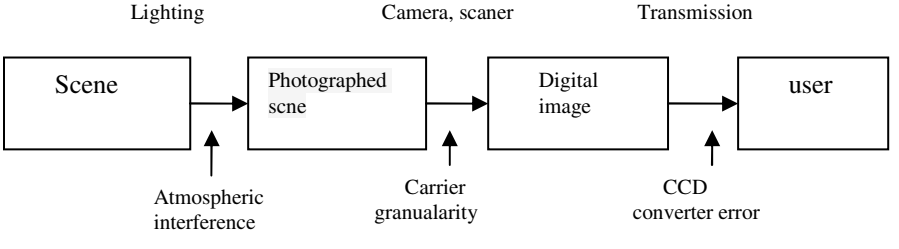


Fig. 1. Block diagram of digital image acquisition process and points of interfere

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 294–300, 2012.

A digital image is a more or less accurate mapping of an analog image [1]. A typical image processing path / track, indicating points where interferences occur, is shown on Fig. 1. Typical errors arising in image processing track, which significantly affect the quality of a digital image, have been shown in Table 1. Due to all interferences listed in Table 1, the digital image is a more or less accurate representation of the analogue image. One can safely say that using better image acquisition equipment we shall acquire much higher quality of images than when using low class equipment.

Table 1. Examples of errors occurring in image processing track

Primary errors	Exemplary errors of image converter	Errors resulting from image converter lighting
Of image converter	Resulting from resolution	Thermal noise
From interferences occurring in transmission track	Resulting from sensitivity	Fixed pattern noise
Connected with data processing	Resulting from noise level	Transfer noise
Connected with data interpretation	Resulting from performance quality	Charge interference noise

The better the quality of the equipment located in the image acquisition path, the more accurate the mapping. Using a device to convert an analog image to a digital form has a considerable impact on the quality of the mapping. Nowadays, the devices featuring construction based on the CCD structure are commonly used for conversion. Various interferences and errors affecting the quality of the obtained digital image occur during acquisition of data by means of such devices. Following factors are responsible for the quality of images acquired by those devices:

- Number of pixels (resolution),
- Quantum efficiency (sensitivity),
- Noises in matrice,
- Quality of the A/C converter.

The occurrence of any error always causes a decline in quality of the obtained images. The basic method of improving the quality of an image is the linear filtration of an image [2]. Alternatively, any device of very favorable parameters can be used but it usually implies a high device cost. It can be easily found that the quality of the obtained digital images depends strongly on the quality of a device applied for image acquisition. The purpose of the paper is proving that it is possible to obtain an image generated by means of a low class scanner of the quality similar to the quality of a digital image generated by means of a high class scanner. Firstly, an image is subject to the linear filtration. The paper features application of a neural network for generation of the impulse response as a FIR filter. Design and optimization of FIR filters using a neural network [3] features a huge potential in the field of digital image processing. The results

of the tests proved that the mask of a filter used for correction of the quality of an image can be obtained as the result of the neural network learning process.

2 Method Comparison of Digital Images

A practical experiment was performed in order to show a difference in the quality of digital images generated by means of a low class scanner and images generated by means of a high class scanner. For the comparison of images, the following criteria:

Mean Square Error (MSE) according to formula no. 1 [4]:

$$MSE = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left[f_{in}(x,y) - f_{out}(x,y)\right]^2 \tag{1}$$

Normalized Mean Square Error (NMSE) according to formula no. 2 [4]:

$$NMSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}\left[f_{in}(x,y) - f_{out}(x,y)\right]^2}{\sum_{x=1}^{M}\sum_{y=1}^{N}\left[f_{in}(x,y)\right]^2} \tag{2}$$

Universal Image Quality Index (Q) according to formula no 3.

$$Q = \frac{4\cdot\mu_{f_{in},f_{out}}\cdot\overline{f_{in}}\cdot\overline{\overline{f_{out}}}}{\left(\mu_{f_{in}}^2+\mu_{f_{out}}^2\right)\cdot\left[\left(\overline{f_{in}}\right)^2+\left(\overline{f_{out}}\right)^2\right]} = \frac{\sigma_{f,f`}}{\sigma_{f,f`}\cdot\sigma_{f,f`}}\cdot\frac{2\cdot\overline{f}\cdot\overline{f`}}{\left(\overline{f}\right)^2+\left(\overline{f`}\right)^2}\cdot\frac{2\cdot\sigma_f\cdot\sigma_{f`}}{\sigma^2{}_{f`}+\sigma^2{}_{f`}} \tag{3}$$

Where: f_{in} - the image generated by means of a high class scanner;

f_{out} - the image generated by means of a low class scanner.

For the experiment used ten pairs of images obtained from two different scanners. One scanner was considered high class and second as low class. Photos from the reference scan indicated by the letter "a". The results of comparison of the images are presented in table no. 2.

Table 2. The values of criteria MSE , NMSE, Q

Image pair	MSE	NMSE	Q
Image 1a and 1b	52,381	0,771	0.324
Image 2a and 2b	57,293	0,765	0.436
Image 3a and 3b	57,511	0,791	0.478
Image 4a and 4b	50,022	0,712	0.398
Image 5a and 5b	54,135	0,783	0.278
Image 6a and 6b	56,772	0,737	0.411
Image 7a and 7b	58,635	0,754	0.509
Image 8a and 8b	54,494	0,777	0.491
Image 9a and 9b	55,765	0,798	0.297
Image 10a and 10b	49,987	0,734	0.563

The analysis of the results shown in table 2 proves that slight differences between the images can be noticed. This fact is caused by the errors specified in section 1 of the paper. It means that two digital images of different quality were obtained. The images generated by means of a low class scanner can be fitted to the quality of the images generated by means of a high class scanner by applying the digital image processing. The digital image filtration is one of the most often used methods of improving the quality of the digital images. The operation guarantees that some undesirable objects such as interferences or noises can be eliminated from an image. However, it is necessary to possess the knowledge related to the image interferences in order to perform the operation efficiently.

3 Generate a Filter Mask

The authors applied a neural network for generation of the impulse response as a FIR filter according to the following assumptions:

- A filter shall be a 3x3 mask
- A filter shall execute the discrete convolution according to formula no. 4

$$f'_{out}(x,y) = \sum_{i,j \in k} f_{out}(x-i, y-j) \cdot w(i,j) \tag{4}$$

Where: $f_{out}(x,y)$ - low class image before filtering; $f'_{out}(x,y)$ - image after filtering; $w(i,j)$ - filter mask.

Thanks to the above specified assumptions, a filter can be efficiently applied for improving the quality of the images using common graphical software such as *PhotoShop* or *Paint Shop Pro*.

A neural network used for generation of a FIR filter and applied by the authors is a unidirectional network operating in the mode of supervised learning. This configuration implies that the input images in such networks are processed by neural network from the input layer to the output. It means that the network outputs at a given moment t are input-dependent only at the same moment. Therefore, the connection weights serve as parameters.

To obtain a digital filter with *3x3* mask the authors of the article have used one direction neural network working in mode with teacher – figure no 2. This network has 9 inputs $(x_0,x_1,..x_8)$, into which in the course of learning, we give image acquired by low class scanner (y_1). In the course of learning, in which the actualization of wages was based on recurrent presentation of the images acquired by low class scanner, the change in wage rate of the neural network occurred $(w_0,w_1,..,w_8)$ according to formula no 5 [4].

$$w_i(j+1) = w_i(j) + n \cdot [x_i - w_i \cdot (j)] \tag{5}$$

Where: j – step number, n – learning parameter, $w_i(j)$ - wage, n - error

The whole process of generating the filter shows a diagram in Figure number 2.

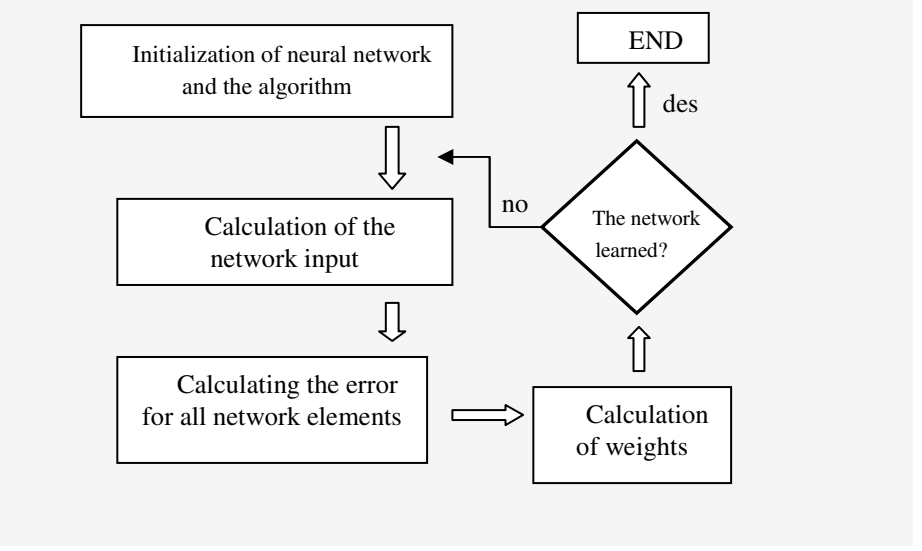


Fig. 2. Diagram of learning process

The result of the operation of the neural network is the filter of 3x3 mask (9 wages), which in the course of realization of discrete tangle operation described by formula no 5, improve the quality of digital images acquired by low class scanner in such a way, that they will correspond with images from the model scanner.

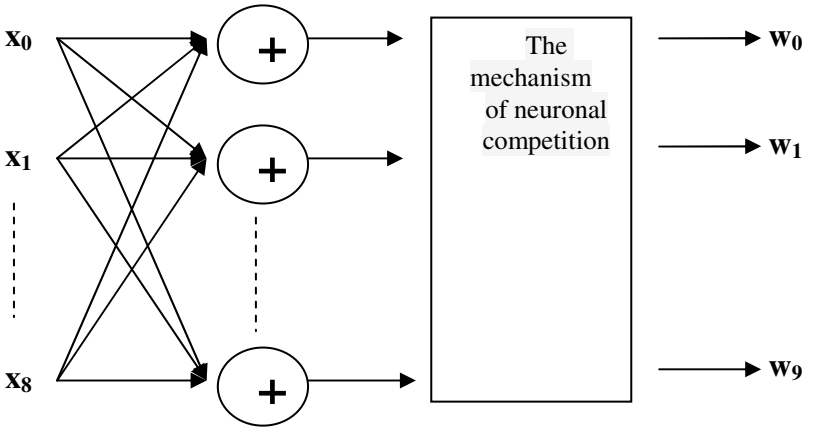


Fig. 3. Neural network diagram

During the network learning process were obtained for 10 image pairs ten filter masks, some of which were presented in formula:

$$w = \begin{bmatrix} 0.0006 & 0.0902 & 0.0130 \\ 0.1116 & 0.5893 & 0.1135 \\ 0.0027 & 0.0958 & 0.0008 \end{bmatrix} \tag{6}$$

The amplitude characteristics of the generated filters are presented in figure no 4 , it can be found that the filters generated during the neural network learning process are low-pass. Therefore, most interferences of an image generated from a low class scanner can be classified as noises. In order to check the efficiency of operation of the obtained filters, each of ten images generated from a low class scanner was subject to the filtration process according to formula no. 3. A new image so generated was compared with a reference image by means of the criteria according to formulas 1, 2, 3. The results are shown in table 3.

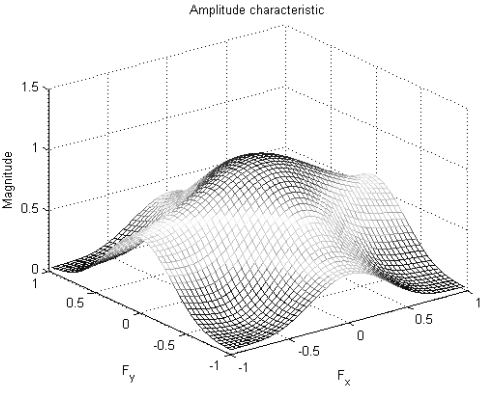


Fig. 4. The amplitude characteristics of the generated filters

Table 3. The values of criteria for the images after filtering by filter *w (6)*

	MSE before	MSE after filtering	NMSE before	NMSE after	Q before filtration	Q after filtration
Image 1a	52,381	33,311	0,771	0,898	0.324	0.632
Image 2a	57,293	40,764	0,765	0,882	0.436	0.699
Image 3a	57,511	42,349	0,791	0,824	0.478	0.754
Image 4a	50,022	31,576	0,712	0,843	0.398	0.658
Image 5a	54,135	36,023	0,783	0,897	0.278	0.521
Image 6a	56,772	39,351	0,737	0,864	0.411	0.767
Image 7a	58,635	44,976	0,754	0,872	0.509	0.809
Image 8a	54,494	27,009	0,777	0,898	0.491	0.621
Image 9a	55,765	29,731	0,798	0,851	0.297	0.478
Image 10a	49,987	20,097	0,734	0,878	0.563	0.732

The results from table 3 prove explicitly the efficiency of the filters generated by means of a neural network. The efficiency is related to the objective evaluations performed by means of MSE , NMSE and Q criterion.

4 Final Conclusions

Analyzing the experimental results shown in Table No. 3 the following conclusions can be drawn:

- Indication of the quality metrics improved significantly;
- Generated filter can be used for digital convolution operation;
- Interference in the images are noise;
- The generated filter is the low-pass filter;
- Generated filter can be used in popular graphics programs;
- Filtration efficiency can be incrased by generating a large mask of filter.

References

1. Pratt, W.K.: Digital Image Processing. PIKS Inside. Willey (2001)
2. Bovik, A.C.: Handbook of Image and Video Processing, Department of Electrical and Computer Engineering, 2nd edn. The University of Texas, Elsevier Academic Press, Elsevier Inc., Austin (2005)
3. Gupta, P.K., Kanhirodan, R.: Design of a FIR Filter for Image Restoration using Principal Component Neural Networks. In: IEEE International Conference on Industrial Technology, ICIT 2006, December 15-17, pp. 1177–1182 (2006)
4. Wang, Z., Bovik, A.C.: Mean Squared Error: Love it or Leave it? IEEE Signal Processing Magazine (2009)
5. Wang, W., Bovik, A.C.: A Universal Image Quality Index. IEEE Signal Processing Letters 9(3) (2002)
6. Korbicz, J., Obuchowicz, A., Uciński, D.: Sztuczne sieci neuronowe. Akademicka Oficyna Wydawnicza, Warszawa (1994)
7. Grainger, E.M., Cuprey, K.N.: An Optical Merit Function (SQF) Which Correlates With Subjective Image Judgments. Fotografic Science and Engineering 16(3) (1992)
8. Wyszecki, G.: Color appearance. In: Boff, K.R., Kaufman, L., Thomas, J.P. (eds.) Handbook of Perception an Human Performance (1986)
9. Ackenhusen, J.G.: Real-Time Signal Processing: Desing and Implementation of Signal Processing Systems. PTR Prentice Hall, Upper Saddle River (1999)
10. Bellanger, M.: Digital Processing of Signal. Theory and Practice. Wiley, Chichester (1989)
11. Thimm, G., Fiesler, M.: Neural network initialization. In: Mira, J. (ed.) Form Neural to Artificial Neural Computation, Malaga, pp. 533–542 (1995)
12. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Processing 12(11), 1338–1351 (2003)
13. Papas, T.N., Safranek, R.J., Chen, J.: Perceptual criteria for image quality evolution. In: Handbook of Image and Video Processing, 2nd edn. (May 2005)
14. Cammbell, C.: Neural Network Theory. Bristol University Press (1992)

Hybrid Genetic Simulated Annealing Algorithm (HGSAA) to Solve Storage Container Problem in Port

Riadh Moussi, Ndèye Fatma Ndiaye, and Adnan Yassine

Laboratory of Applied Mathematics of Le Havre (LMAH), Le Havre University. 25 rue Philippe Lebon B.P. 540 - 76058 Le Havre Cedex, France
Superior Institute of Logistics Studies (ISEL). Quai Frissard B.P. 1137 - 76063 Le Havre Cedex, France

Abstract. Container terminals play an important role in marine transportation; they constitute transfer stations to multimodal transport. In this paper, we study the storage of containers. We model the seaport system as a container location model, with an objective function designed to minimize the distance between the vessel berthing locations and the storage zone. Due to the inherent complexity of the problem, we propose a hybrid algorithm based on genetic (GA) and simulated annealing (SA) algorithm. In this paper, three different forms of integration between GA and SA are developed. In order to prove the efficiency of the HGSAAs proposed are compared to the optimal solutions for small-scale problems of an exact method which is Branch and Bound using the commercial software ILOG CPLEX. Computational results on real dimensions taken from the terminal of Normandy, Le Havre port, France, show the good quality of the solutions obtained by the HGSAAs.

Keywords: Container terminal, storage container, hybrid genetic simulated annealing algorithm (HGSAA).

1 Introduction

Container terminals are essential intermodal interfaces in the global transportation network. The container storage is one of the most important services in a container terminal. To increase the efficiency of a container terminal, containers are optimally stacked in the storage areas in the form of stacks. The maximum height of stacks is fixed by the port authorities, based on the equipments used. The container stacks are arranged in rows aligned, called sides. A set of sides form a block. Each storage area consists of many blocks (see Fig. 1.).

This paper examines the method employed in the storage of containers at a port container terminal. The focus is to utilize the storage area in a more optimal manner; thus reducing the time required for the transfer of containers. The objective of the model is to determine the optimal storage strategy for containers. A new container location model is designed to address the objective of the research (see Moussi et al [13]). This model is based on three major constraints: (1) consider the state of the storage area before the arrival of containers, (2) for

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 301–310, 2012.

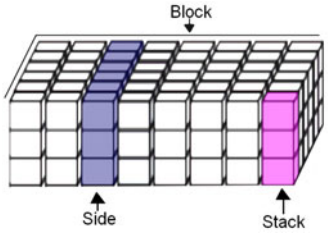


Fig. 1. Blocks in port

each stack, containers are stored in the decreasing order of their departure time from the yard, (3) containers are stored by respecting the constraint of dimension compatibility. The goal of this work is to minimize the unloading time of a number of containers and to determine an optimal storage strategy.

The problem treated is known to be NP-hard category. Thus, Moussi et al [13] proposed a GA to solve this problem. We have applied two variants of a GA, the first uses the crossover operator 'one-cut-point' and the second uses the 'two-cut-point'. By comparing the results obtained by these two variants of GA with the exact results provided by ILOG CPLEX on problems of small dimensions, we found that, the results of the variant 'one-cut-point' are more efficient than the variant "two-cut-point". But, the major inconvenient of this algorithm that is far to the optimum results. For these reasons, hybrid genetic simulated annealing algorithms (HGSAA) are proposed to solve the problem treated. In this paper, we propose three methods combining a GA based on 'one-cut-point' and a SA algorithm. The rest of this paper is organized as follows: a brief overview and related literature is presented in Section 2. We provide a description of the problem treated in Section 3. The mathematical model representing the concrete problem is formulated in Section 4. We propose the three variant of HGSAA which are developed in Section 5. The results of numerical simulations of all the implemented algorithms and a comparative study are presented in Section 6, and finally Section 7 is devoted to conclusion.

2 Literature Review

Various aspects of the operational problems in the container terminal are developed in the literature. Stahlbock et al [16] have presented the current state of the art in container terminal operations and operation research. Steenken et al [17] have described and classified the main logistic processes and operations in container terminals.

Storage containers are a critical resource in container terminals. In the terminals, the loading sequence of containers affects significantly the productivity of port operations. However, the optimal allocation of containers allows a sequence

of optimal loading. It affects the efficiency of delivery and loading operations. Peter et al [14] developed a model to determine optimal storage strategies and container handling schedules. They proposed a heuristic method with a GA. In [7], they treated the same problem with an improved algorithm. Indeed, they developed a GA, a Tabu Search algorithm (TS) and a hybrid algorithm between TS and GA. Mohammad et al [12] solved an extended Storage Space Allocation Problem (SSAP) in a container terminal by an efficient GA. The SSAP is defined as the temporary storage of the inbound and outbound containers of the storage blocks. The objective of the SSAP developed is to minimize the time of storage and retrieval time of containers. Changkyu et al [2] focused on the planar storage location assignment problem (PSLAP). The PSLAP can be defined as the assignment of the inbound and outbound containers of the storage area in order to minimize the number of obstructive moves. The PSLAP is resolved by a GA. Chuqian et al [3] treated the SSAP by using a rolling-horizon approach. For each planning horizon, the problem is decomposed into two levels: At the first level, they defined for each period the number of containers to be placed in each storage block. At the second level, they found out the number of containers stored in each block at each period associated with each vessel. The objective of the work of Chuqian et al. is to minimize the total distance to transport containers between their storage blocks and the vessel berthing locations.

There is another problem developed by some academic researchers to treat the storage container problem which is the storage of inbound containers and outbound containers. The storage location assignment problem for outbound containers is treated by, Lu et al [11], the objective of the problem is to minimize the rehandling operations by cranes in order to maintain the stability of the ship. The problem is decomposed into two stages. In the first stage, the numbers of locations in each yard bay are determined by a mixed integer programming model. In the second stage, the exact storage location for each container is determined by a hybrid sequence stacking algorithm. The same problem is discussed by Kap et al [9]. They formulated a basic model as a mixed-integer linear program and they suggested two heuristics algorithms to solve the problem. The storage location assignment problem for inbound containers is one of the problems developed to solve the storage container problem. Kap et al [10] discussed a method of determining the optimal price schedule for storing inbound containers. Kap et al [8] proposed a mathematical model to allocate storage space for import containers using the segregation strategy in order to minimize the number of rehandles.

3 Context

One of the major problems of a terminal is to store containers in an optimal way. The goal of this work is to minimize the unloading time of containers and to determine an optimal storage strategy. In our case, the rehandling operations are not accepted, i.e., when a container is stored, it is not moved from its position until the departure time. We model the problem with a new mathematical model

that reflects reality and takes into account most of the constraints imposed by port authorities. This model treats the following hypotheses: (1) We don't mix on the same block and in the same period the loading and the unloading containers. (2) Before the beginning of each period, we know the state of the storage area. For each stack, we know: the number of container stored, the departure time of each container and the type of the stack (dimension of containers in the stack). (3) For each stack, containers are stored in the decreasing order of their departure time. (4) Containers are stored by respecting the constraint of dimension compatibility. All containers stored in the same stack have the same dimension. (5) The maximum number of container stored in each stack is 3.

4 Mathematical Model

The container location problem is formulated as a new and original mathematical programming model. This model is applied on each period in order to determine an optimal storage strategy based on the following assumptions. Note that p represents stacks, i is the index of the empty position in a stack and k represents containers.
N: Represents the number of containers.
N_p: Represents the number of stacks.
c_p: Represents the number of empty position for stack p.
r_p: Represents the type of stack p.
t_p: Represents the date of stack p. The date of a stack is equal to the departure time of the container stored on the top else if the stack is empty $t_p = M$, M is a big number.
R_k: Represents the type of container k.
T_k: Represents the departure time of container k.
d_{pk}: Represents the shortest way between the position of the ship of container k and the stack p.

$$\lambda_{pik} = \begin{cases} 1 \; if \; container \; k \; is \; stored \; in \; position \; i \; of \; stack \; p \\ 0 \; otherwise \end{cases}$$

The problem treated can be formulated as the following model:

$$Minimiser \; = \sum_{p=1}^{N_p} \sum_{i=1}^{c_p} \sum_{k=1}^{N} \lambda_{pik}.d_{pk} \tag{1}$$

$$\sum_{p=1}^{N_p} \sum_{i=1}^{c_p} \lambda_{pik} = 1, \; k = 1, ..., N \tag{2}$$

$$\lambda_{pik} + \lambda_{pik'} \leq 1 \tag{3}$$

$$k = 1, ..., N, \; k' = 1, ..., N, \; k \neq k', \; p = 1, ..., N_p, \; i = 1, ..., c_p$$

$$\sum_{k=1}^{N} \lambda_{pik} \geq \sum_{k'=1}^{N} \lambda_{pi+1k'}, \quad p = 1, ..., N_p, \; i = 1, ..., c_p \tag{4}$$

$$(r_p - R_k)\lambda_{pik} = 0, \quad k = 1, ..., N, \ p = 1, ..., N_p, \ i = 1, ..., c_p \tag{5}$$

$$T_k \leq M(1 - \lambda_{pik}) + \lambda_{pik} t_p, \quad k = 1, ..., N, \ p = 1, ..., N_p, \ i = 1, ..., c_p \tag{6}$$

$$M(1 - \lambda_{pik}) + T_k \lambda_{pik} \geq T_{k'} \lambda_{pi+1k'} \tag{7}$$

$$k = 1, ..., N, \ k' = 1, ..., N, \ p = 1, ..., N_p, \ i = 1, ..., c_p$$

The objective of this model is to minimize the distance between the quay and the storage position of each container, see constraint (1). The second objective of the model is to determine an optimal storage strategy which is developed by the following assumptions. Constraints (2) ensure that, each container is stored in one storage position. Each position in each stack receives only one container. This idea is developed by constraint (3). Constraint (4) ensures that, an empty intermediate positions between containers stored in the same stack are not accepted. Each container is stored in a stack with the same type; this idea is developed in Constraint (5). Constraint (6) and (7) ensure that containers must be stored in a stack in the decreasing order of their departure time from the yard. Constraint (6) takes into account containers stored before the beginning of the new period. In constraint (7), Containers will be stored in the same stack, they must be stored in the decreasing order of their departure time.

5 Hybrid Genetic and Simulated Annealing Algorithm (HGSAA)

The container storage problem is a very important problem in port logistics. It is formulated as a linear integer programming problem. The problem is known to be NP-hard and the computation complexity increases exponentially. This makes it difficult his resolution in reasonable time with an exact method (Branch and Bound Method, for example).

A HGSAA is an algorithm based on the combination between a GA algorithm and SA algorithm. A GA has been developed by J. Holland in the 1970s to understand the adaptive processes of natural systems. Their operation is as follows: We start with an initial population composed of a finite set of potential solutions, called chromosomes, randomly selected. We estimate their relative performance. We create a new population of potential solutions using simple evolutionary operators: selection, crossover and mutation.

On the other hand, SA has been developed by Kirkpatrick et al., 1980. SA is a stochastic algorithm that enables under some conditions the degradation of a solution. It begins with a random solution, and then proceeds in several iterations. At each step, we choose a random neighbor of the current solution. This neighbor is the solution of the next iteration if that neighbor is better than the current soultion, or it is selected with a given probability that depends on a control parameter, called temperature (T), and the amount of degradation of the objective function (ΔE). ΔE represents the difference in the objective value between the current solution and the generated solution. With each iteration, we

decrease the value of the temperature T. The algorithm stops when the temperature becomes low. This probability flows, in general, the Boltzman distribution: $P(\Delta E, T) = e^{-\frac{f(s)'-f(s)}{T}}$.

The combination of SA with GA provides a strong possibility to GA to avoid local minima and have a marked improvement of its performance. In recent years, HGSAAs have received significant interest and are being increasingly used to solve real-world problems. A GA is able to incorporate other techniques within its framework to produce an hybrid algorithm that reaps the best from the combination. Elmihoub et al [5], treated the effects of learning strategy and probability of local search on the performance of HGSAA. Elmihoub et al [4], reviewed diverse forms of incorporation between GA and other search and optimization techniques which are considered as a local search tool. GA and SA can be seen as complementary tools that can be brought together to achieve an optimization goal. In this paper, three different forms of integration between genetic algorithms and SA are developed. To see more variants of GA, SA and HGSAA, you can review the thesis of Bourazza [1], the work of Rahila Patel et al [15] and the book of El-Ghazali [6].

5.1 Chromosomes

Each chromosome is represented as a table with two lines. The first line contains the empty positions of each stack and the second line contains containers assigned to each stack. The number of columns is equal to the number of the empty positions of all stacks.

5.2 Selection Method

The most common method for the selection mechanism is the roulette wheel, in which each chromosome is assigned a slice of a circular roulette wheel and the size of the slice is proportional to the chromosomes fitness.

5.3 Crossover Operators: One-Cut-Point

We choose two chromosomes and we fix a cutting position. The cutting position divides the two parents into two segments. In order to obtain the new children, we permute the two segment situated on the right of the cutting position between the two parents. For one-cut-point after correction, we obtain two valid children and we introduce to the next generation the best one.

5.4 Mutation and Neighborhood Principle

Mutations with GA and neighborhood principle with SA introduce random changes to the solutions by altering the value to a gene. In our case, we choose randomly a container and we affect it to another stack respecting all constraints. The GA used to solve the problem treated in this paper is presented by

Moussi et al [13]. All GA runs used the following standard characteristics: Crossover rate: 0.75, mutation rate (P_m): 0.025, number of generation: 1000 and population size: 30.

5.5 Hybridizations Methods Proposed

In HGSAA, a GA incorporates SA to improve the performance of the genetic search. There are several ways in which a SA can complement the genetic search. In this paper, we propose three ways to combine GA and SA algorithm:

HGSAA1: The initial solution of the SA algorithm is the chromosome chosen with the mutation probability (P_m). Each application of mutation is followed by SA algorithm (see (a) in Fig. 2.). All HGSAA1 runs have the following standard characteristics: number of generation: 700, temperature: 140 and K: 0.72.

HGSAA2: The initial solution of the SA algorithm is the best of each generation (see (b) in Fig. 2.). The SA algorithm is applied with a probabilityP_{Best}. All HGSAA2 runs have the following standard characteristics: number of generation: 700, P_{Best}: 0.01, temperature: 120 and K: 0.71.

HGSAA3: A simple combination method of GA and SA algorithm is to produce new individuals with GA, then these individuals are processed with SA with a probability of application (P_{SA}) and the results are used as the initial individuals of the next generation. This method can be illustrated in (c), Fig. 2. All HGSAA3 runs have the following standard characteristics: number of generation: 600, P_{SA}: 0.9, temperature: 120 and K: 0.71.

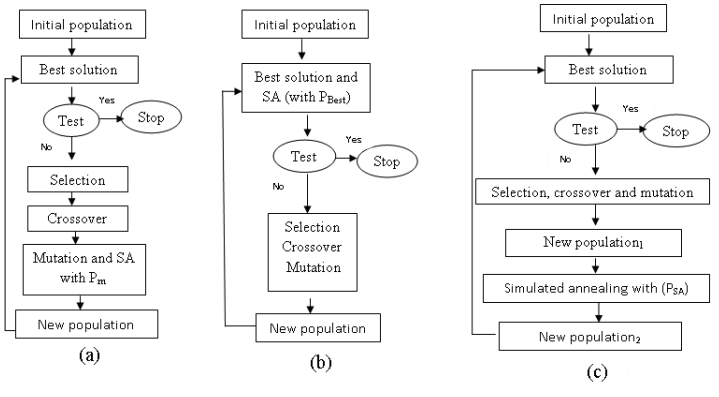


Fig. 2. The HGSAAs proposed

6 Numerical Results

In this section, firstly, we compare the results provided by the proposed HGSAAs with those of the Branch and Bound method applied to small dimension problems to prove the robustness and performance of HGSAAs. Second, we generate large dimension instances that will be solved by HGSAAs (the exact methods, including the Branch and Bound method, are unable to solve these instances).

6.1 Small-Scale Problems

In this section, the GA and the three hybrid algorithm proposed are compared to an exact method which is Branch and Bound developed by ILOG CPLEX with small scale problems. This comparison is carried out in order to verify the quality of the HGSAAs proposed. Fig. 3. represents the percentage deviation ($\frac{Algorithm\ proposed\ result - BB\ result}{BB\ result} * 100$) of the GA and the three HGSAAs proposed from the optimum results given by ILOG CPLEX for 31 instances generated. We note that HGSAA3 is the closest algorithm to the optimum results.

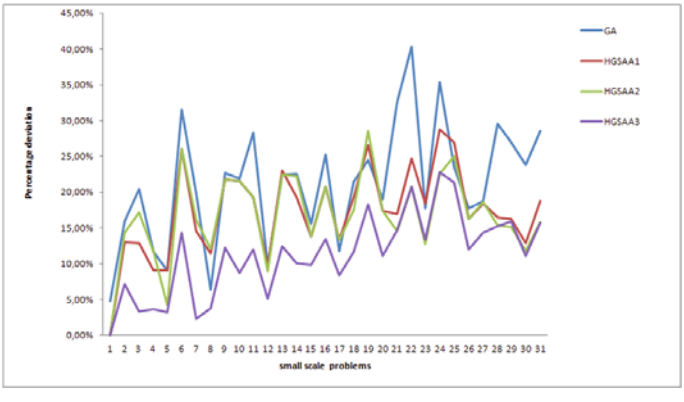


Fig. 3. Comparison between GA and HGSAAs with Branch and Bound

We calculate, also, the average of percentage deviation which is equal to 21.26% for the GA, 17.29% for HGSAA1, 16.68% for HGSAA2 and 10.22% for HGSAA3. The HGSAA3 is the closest algorithm to the optimum.

6.2 Large-Scale Problems

Based on real-life terminal operators taken from the terminal of Normandy Le Havre port, we generate some large-scale problems. The problems generated are resolved by the GA and the HGSAA proposed. Fig. 4, shows the variation of each hybrid algorithm from the results given by the GA (for each instance, we calculate the difference between the results given by the GA and each hybrid algorithm). We note that HGSAA3 improve the GA results for all the instances generated and it is more efficient then the other hybrid algorithm.

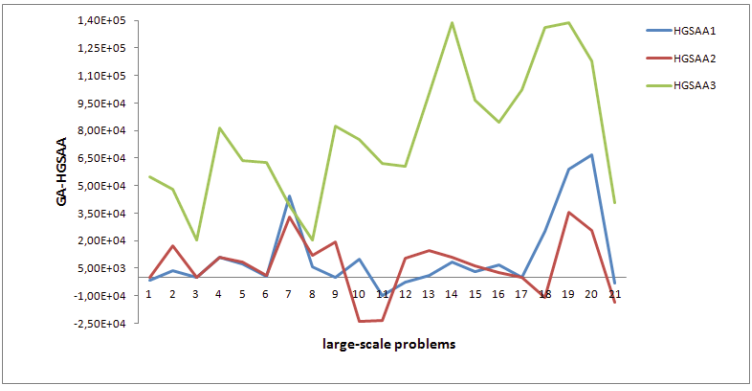


Fig. 4. Comparison between GA and HGSAAs

7 Conclusion

In this paper, a new container location model is designed to minimize the unloading time of containers and to determine an optimal storage strategy. This problem, very important in port logistics, is NP-Hard. This requires the use of meta-heuristics methods to find an approximated optimal solution for the large instances where it is impossible to determine the optimal solution by exact methods. The model treated is solved in previous work by a GA based on the one-cut-point as a crossover method. But, the inconvenient of this algorithm that is far to the optimum results with an average of percentage deviation equal to 21.26%. For these reasons, our perspective was to find another strategy to find solutions very close to the optimum.

The idea is to improve the results provided by the GA by combining it with another local search method. For this motivation, we propose a hybrid algorithm HGSSA based on the combination of GA and SA algorithm. The SA had been chosen for its simplicity and efficiency to be incorporated with the GA. By comparing the results obtained by the three variants of HGSAA with the exact results provided by ILOG CPLEX on problems of small dimensions, we found that, the results of HGSAA3 are more efficient than the other hybrid algorithms proposed. In addition, applied to large dimension problems, we get the same result: HGSAA3 is more effective than the other two hybrid algorithms proposed.

The novelty of this work lies in the treatment of a new model to solve the Container Location Problem in ports. The model developed is based on real-life terminal operators taken from the terminal of Normandy Le Havre port, France. In order to solve the model, we propose three hybrid algorithms HGSAA (combination of GA algorithm and SA algorithm). The HGSAA3 is the best one and its efficiency is proved by the quality of solution obtained in large dimension problems (real problems). But, the major inconvenient of this algorithm that its

solution is not very close to the optimum (The average of percentage deviation is equal to 10.22%). For these reasons, our perspective is to find another strategy to improve outcomes obtained in this work and find better solutions.

References

1. Bourazza, S.: Variants of genetic algorithms applied to scheduling problems. PhD Thesis, University of Le Havre, France (2006)
2. Changkyu, P., Junyong, S.: Mathematical modeling and solving procedure of the planar storage location assignment problem. Computers and Industrial Engineering 57, 1062–1071 (2009)
3. Chuqian, Z., Jiyin, L., Yat-Wah, W., Katta, M.: Storage space allocation in container terminals. Transportation Research 37, 883–903 (2003)
4. El-Mihoub, T.A., Hopgood, A.A., Nolle, L., Battersby, A.: Hybrid Genetic Algorithms: A Review. Engineering Letters 13, 2–16 (2006)
5. El-Mihoub, T.A., Hopgood, A.A., Nolle, L., Battersby, A.: Performance of Hybrid Genetic Algorithms Incorporating Local Search. In: Horton, G. (ed.) 18th European Simulation Multiconference (ESM 2004), Magdeburg, Germany, pp. 154–160 (2004)
6. El-Ghazali, T.: Metaheuristics from design to implementation. John Wiley and Sons (2009)
7. Erhan, K., Peter, P.: Mathematical modeling of container transfers and storage locations at seaport terminals. OR Spectrum 28, 519–537 (2006)
8. Kap, H.K., Hong, B.K.: Segregating space allocation models for container inventories inport container terminals. Int. J. Production Economics 59, 415–423 (1999)
9. Kap, H.K., Kang, T.P.: A note on a dynamic space-allocation method for outbound containers. European Journal of Operational Research 148, 92–101 (2003)
10. Kap, H.K., Ki, Y.K.: Optimal price schedules for storage of inbound containers. Transportation Research 41, 892–905 (2007)
11. Lu, C., Zhiqiang, L.: The storage location assignment problem for outbound containers in a maritime terminal. Int. J. Production Economics 48, 991–1011 (2010)
12. Mohammad, B., Nima, S., Nikbakhsh, J.: A genetic algorithm to solve the storage space allocation problem in a container terminal. Computers and Industrial Engineering 56, 44–52 (2009)
13. Moussi, R., Ndiaye, F., Yassine, A.: A genetic algorithm and new modeling to solve container location problem in port. In: The International Maritime Transport and Logistics Conference, A Vision For Future Integration, Alexandria, Egypt (December 2011)
14. Peter, P., Erhan, K.: An approach to determine storage locations of containers at seaport terminals. Computers and Operations Research 28, 983–995 (2001)
15. Patel, R., Raghuwanshi, M.M., Shrawankar, U.N.: Genetic Algorithm with Histogram Construction Technique. Journal of Information Hiding and Multimedia Signal Processing 2(4), 342–351 (2011)
16. Stahlbock, R., Vob, S.: Operations research at container terminals: a literature update. OR Spectrum 30, 1–52 (2008)
17. Steenken, D., Vob, S., Stahlbock, R.: Container terminal operation and operations research-a classification and literature review. OR Spectrum 26, 3–49 (2004)

Satellite Payload Reconfiguration Optimisation: An ILP Model

Apostolos Stathakis[1], Grégoire Danoy[2], Pascal Bouvry[2], and Gianluigi Morelli[3]

[1] Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg
apostolos.stathakis@uni.lu
[2] CSC Research Unit, University of Luxembourg
{gregoire.danoy,pascal.bouvry}@uni.lu
[3] SES Engineering, Betzdorf, Luxembourg
gianluigi.morelli@ses.com

Abstract. The increasing size and complexity of communication satellites has made the manual management of their payloads by engineers through computerised schematics difficult and error prone. This article proposes to optimise payload reconfigurations for current and next generation satellites using a novel Integer Linear Programming model (ILP), which is a variant of network flow models. Experimental results using CPLEX demonstrate the efficiency and scalability of the approach up to realistic satellite payloads sizes and configurations.

Keywords: Integer-Linear programming, network optimisation, satellite payload.

1 Introduction

With an expected lifetime of 15 years or more, telecommunication satellites require flexibility and efficiency to answer potential changing service requirements or hardware failures during that period. To meet this demand, the complexity of satellites' payloads, which play the main role in signal transmissions, has increased significantly. A payload is composed of a number of hardware components, such as amplifiers and channel filters. In addition, a large number of switches, interconnected to compose switch matrices, are used to ensure signal routings among payload components (selectivity) and functionality in case of failures (redundancy). Currently, payload engineers use computerised schematics to graphically represent the payload structures and find workable configurations. However, finding optimal payload reconfigurations and signal re-routing solutions has become a hard task due to the increasing complexity of the switch matrices, making this manual management of the payload difficult and time consuming. Commercial software solutions exist that have built-in optimisers [4,7], however, as they are closed packages, they do not provide the advantage of the flexibility achieved with computerised schematics.

The work proposed in this paper will help payload engineers to configure current and future satellites payloads via a novel optimisation tool that works

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 311–320, 2012.

in conjunction with the existing computerised schematics. More precisely, we here propose to solve the payload configuration process using an integer-linear programming (ILP) model, i.e. a variant of network flow problems, which to the best of our knowledge has never been tackled in the literature. This mathematical model should be flexible and scalable respectively to different types and sizes of real satellite payloads.

The remainder of this paper is organized as follows. The next section provides a state of the art analysis on payload reconfiguration softwares and network flow problems. Section 3 presents a detailed description of the payload reconfiguration problem and a description of the proposed mathematical model. Experimental results using the ILOG CPLEX solver are presented in Section 4. Finally, section 5 concludes our work and provides some perspectives.

2 State of the Art

The main function of the satellite payload system is to receive the uplink signals from the ground stations, to route those signals to apply frequency conversion and amplification, and to finally retransmit those modified signals on the downlink [8]. The signals' routing flexibility is currently achieved through reconfigurable payload components, i.e. different switch types organized in matrices. These permit to modify the signals routings in order to meet new operational needs. As an example, in case of failure of an amplifier, the signals are rerouted to a spare one. This reconfiguration process is initiated through telecommands, which are submitted from the ground stations, by the payload engineers. However this flexibility comes to the expense of a higher payload size and complexity. Indeed, finding an optimal reconfiguration that satisfies operational objectives has become a time consuming and error prone task for engineers.

Two commercial softwares dedicated to payload reconfiguration optimisation exist. Smart Rings [4] uses a recursive search combined with automated deduction to compute all possible payload reconfiguration solutions and provides information on each solution quality. TRECS [7] is a second software package that allows the user to define the payload model per satellite and proposes an AutoSolve feature that can plan a reconfiguration to limit or minimize the number of necessary changes. However, payload engineers would benefit from an open interaction between the computerised schematics they use to describe the payload models and the optimisation tools, rather than adapting to new interface environments. Besides, these tools act as 'black boxes', neither allowing flexibility nor knowledge extraction concerning the algorithms used. Finally, they lack of openness since their adaptation, e.g. in case of newer payload components, might be difficult or time consuming.

In this paper, an original Integer Linear Programming (ILP) model, based on network models, is proposed for solving the satellite payload reconfiguration problem. Network optimisation problems is a well studied area [1], which consist of supply and demand points, together with several routes that connect these points and are used to transfer the supply to the demand [3]. Network problems

such as shortest path, assignment, max-flow, transshipment, vehicle routing, and multicommodity flow problems, constitute the most common classes of practical optimisation problems [3]. Such models are widely used in telecommunications and satellite networks for optimal routing [2] and network design [6,5].

In our approach, we consider electromagnetic signals (flows) crossing the payload network structure to reach the suitable outputs. However, a direct application of network flow models is not possible due to some payload properties. Indeed each solution to the problem, that consists in a set of switches positions, will define a different network topology. In order to handle these problem properties, e.g. types of switches or priority on switches to move, and considering the requirements for full control by the payload engineers, we propose an ILP model of the satellite payload reconfiguration problem that to the best of our knowledge has not been tackled in the literature.

3 Satellite Payload Reconfiguration: An ILP Model

3.1 Problem Description

Switch matrices are used in satellite payloads to provide system redundancy and to enhance the satellite capacity by providing full and flexible interconnectivity between received and transmitted signals. Through telecommands, the state of switches can be changed and each state defines different connectivities, as shown in Figure 2. Through such connectivities, the input channel signals cross the input switch matrix and are propagated to the appropriate amplifiers. Signals are then routed through the output switch matrix to the corresponding outputs. The objective of this optimisation problem is therefore to minimize the required number of switch changes to reconfigure the payload, which will intrinsically limit the risk of switch failures.

A simple problem instance with 32 switches of C-type (2 states) and R-type (4 states), is shown in Figure 1. In this example, the input channels 1 and 3 are activated and need to be connected to the corresponding output channels 1 and 3. The proposed solution requires 11 switch changes (shown in darker color) and activates amplifiers 2 and 3. In this work we tackled small to medium switch matrices sizes which correspond to realistic satellite payloads and investigated the problem of finding optimal configurations for establishing an initial set of connectivities.

3.2 The Mathematical Model

In our proposed model the signals are flowing through the network and each signal is assigned a distinct integer flow value. These values are propagated through the network according to flow propagation constraints. The capacity on each link is equal to one, i.e., a physical link can not be shared by different signals. Initially and for simplicity we consider only amplifiers, channels and switch matrices as payload components.

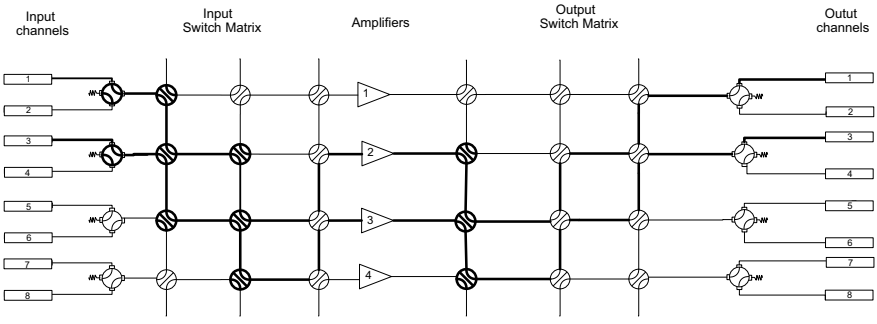


Fig. 1. A small problem instance with C and R switch types with a solution for connecting channels 1 and 3 with 11 switch changes, using amplifiers 2 and 3

Constants and Sets

Let define:

- q as the number of channels to be connected.
- n as the maximum number of states among all switches of the network (e.g. if R switches with 4 states and C switches with 2 states are used, n=4).
- P as a set of size n, with integer values from 0 to $n-1$, representing the maximum number of positions a switch can have. The position of a switch refers to the number of steps the switch should take from its current state to reach a new state.
- S as the set of all switches.
- T as the set of all amplifiers.
- C as the set of all channels.
- L as the set of all links. L contains the following subsets:
  - set $L = CL \cup TL \cup SL$ with:
  - CL as the set of all links connected to the channel filters. $CL = CL_{in} \cup CL_{out}$, where CL_{in} of size $|C|$, the set of all links neighboring with the channel filters on the input and CL_{out} of size $|C|$, the set of all links neighboring with the channel filters on the output. Let cl_{in_c} be the link connected with channel c on the input and cl_{out_c} be the link connected with channel c on the output.
  - TL as the set of all links connected to amplifiers. $TL = TL_{in} \cup TL_{out}$, with TL_{in} of size $|T|$, the set of input links of all amplifiers and TL_{out} of size $|T|$, the set of output links of all amplifiers. Let tl_{in_t} be the input link of amplifier t and tl_{out_t} be the output link of amplifier t.
  - SL as the set of all links between any two switches.
- l_o as a special link, or link 0, when signal can not be propagated. For instance in Figure 2, with a switch in state 2, link a is connected with l_0.
- $CL_{conn} \subseteq CL_{in}$ of size q, as the set of links neighboring with the input channels that will be connected.

– We define M a matrix of size $|S| * |P| * |L| * |L \cup \{l_0\}|$ with:

$$m_{s,p,l_i,l_j} = \begin{cases} 1 & \text{if } l_i \text{ is connected with } l_j \text{ via switch s at position p,} \\ 0 & \text{otherwise.} \end{cases}$$

Variables

– Let define *Pos* of size $|S|$, as the solution vector, providing the position for each switch. For instance, if an R-type switch (with 4 states) is initially at state 3, and needs to be in $pos_s = 1$, it will have to move to state 4.
– Binary vector *Change* of size $|S|$, indicating whether a switch needs to change its initial state or not.

$$change_s = \begin{cases} 1 & \text{if } pos_s > 0, \\ 0 & \text{otherwise.} \end{cases}$$

– Integer vector *Flow* of size $|L|$, showing the flow value (from 0 to q) that is carried by each link. It follows:

$$flow_l = \begin{cases} x & \text{with } 0 < x \leq q, x \in \mathbb{Z} \quad \text{if link l is used,} \\ 0 & \text{otherwise.} \end{cases}$$

– Binary vector B of size $|S| * |P|$, is used to activate or de-active the flow propagation constraints, such that:

$$b_{s,p} = \begin{cases} 1 & \text{if } pos_s = p, \\ 0 & \text{otherwise.} \end{cases}$$

– boolean vector *Ampused* of size $|T|$, indicating whether an amplifier is active (used) or not.

Constraints

– To start the flow distribution,we assign positive distinct flow values to each link of set CL_{conn}:

$$flow_{l_i} = k_i, 0 < k_i \leq q, k_i \in \mathbb{Z}, \forall l_i \in CL_{conn}.$$

– To avoid flow paths starting from an input channel and returning in another input channel, flow values must be set to 0 for all unused input channels:

$$flow_{l_i} = 0, \quad \forall l_i \in \{\{CL_{in}\} - \{CL_{conn}\}\}.$$

– We ensure that each input channel signal reaches the corresponding output:

$$flow_{cl_{inc}} = flow_{cl_{outc}}, \quad \forall c \in C.$$

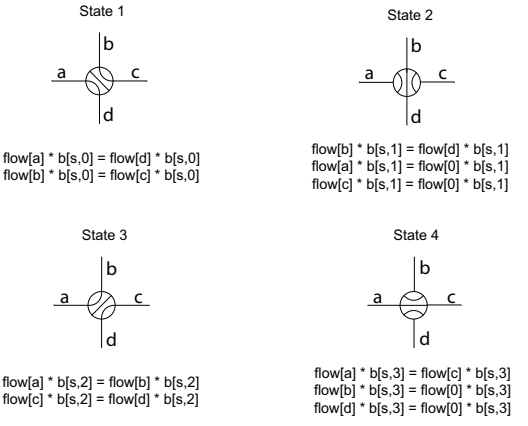


Fig. 2. Flow propagation constraints of R-type switch with state 1 as initial state

– The flow propagation constraints are expressed for each switch and describe the established connections at each switch state. We ensure that connected links have the same flow value propagated. Figure 2 shows the non-linear expressions for flow propagation constraints for all states of an R-type switch. The linear equivalent constraints are expressed as follows:

$$flow_{l1} + b_{s,p} * q - q \leq flow_{l2} \leq flow_{l1} - b_{s,p} * q + q :$$
$$\forall s \in S, \forall p \in P, \forall l_1, l_2 \in (L \cup \{l_0\})^2, \text{ such as } m_{s,p,l_1,l_2} = 1.$$

If $b_{s,p} = 0$ the constraint is deactivated. If $b_{s,p} = 1$ the flow value will be propagated. For the C-type switches that have only 2 available states we set $b_{s,2} = b_{s,3} = 0$. The flow propagation constraints for the initial state of each switch are described for $b_{s,0} = 1$.

– We also ensure that only one switch position is selected each time:

$$\sum_{p \in P} b_{s,p} = 1, \quad \forall s \in S.$$

$$pos_s = \sum_{p \in P} p * b_{s,p}, \quad \forall s \in S.$$

– We ensure that $change_s = 1$, if and only if $pos_s > 0$:

$$change_s * n \geq pos_s, \quad \forall s \in S.$$

– We ensure the flow propagation through an amplifier and that the number of active amplifiers equals the number of connected channels:

$$flow_{tl_{in_t}} = flow_{tl_{out_t}}, \quad \forall t \in T.$$

$$ampused_t * q \geq flow_{tl_{in_t}}, \quad \forall t \in T.$$

$$\sum_{t \in T} ampused_t = q.$$

– Link $\{l_0\}$ never carries any signal.

$$flow_{l_0} = 0.$$

Objectives

$$Min \sum_{s \in S} change_s.$$

Complete ILP Model

$Variables$:	
$pos_s \in \mathbb{Z}$	$\forall s \in S$
$change_s \in \{0; 1\}$	$\forall s \in S$
$flow_l \in \mathbb{Z}$	$\forall l \in L \cup \{l_o\}$
$b_{s,p} \in \{0; 1\}$	$\forall s \in S, \forall p \in P$
$ampused_t \in \{0; 1\}$	$\forall t \in T$
$ObjectiveFunction$:	
$Min \sum_{s \in S} change_s$	
$Constraints$:	
$flow_{l_i} = k_i$	$0 < k_i \leq q, k_i \in \mathbb{Z}, \forall l_i \in CL_{conn}$
$flow_{l_i} = 0$	$\forall l_i \in \{\{CL_{in}\} - \{CL_{conn}\}\}$
$flow_{cl_{in_c}} = flow_{cl_{out_c}}$	$\forall c \in C$
$flow_{l1} + b_{s,p} * q - q \leq flow_{l2} \leq flow_{l1} - b_{s,p} * q + q$	$\forall s \in S, \forall p \in P, \forall (l_1, l_2) \in (L \cup \{l_0\})^2$, s.t. $m_{s,p,l_1,l_2} = 1$.
$\sum_{p \in P} b_{s,p} = 1$	$\forall s \in S$
$pos_s = \sum_{p \in P} p * b_{s,p}$	$\forall s \in S$
$change_s * n \geq pos_s$	$\forall s \in S$
$flow_{tl_{in_t}} = flow_{tl_{out_t}}$	$\forall t \in T$
$ampused_t * q \geq flow_{tl_{in_t}}$	$\forall t \in T$
$\sum_{t \in T} ampused_t = q$	
$flow_{l_0} = 0$	

3.3 Flexibility of the Model

The proposed model permits to set backup (or failed) amplifier(s) by setting their flow value to 0. A switch may also fail in a usable or non-usable state, which respectively implies fixing its position or removing it from the model. Due to frequency constraints, not all amplifiers are appropriate for all channels. It is possible to define which amplifier to connect on which channel by assigning the same flow value (or a range of values for subset of channels) to the amplifier links.

Table 1. Experimental results on different payloads configurations

Number of switches	Switches type	Number of amplifiers	Number of required connections	Average number of switch changes	Average Time(sec)
40	R	10	5	$4.2_{\pm 2.17}$	$0.206_{\pm 0.32}$
			6	$6.6_{\pm 1.52}$	$0.447_{\pm 0.47}$
			7	$7.6_{\pm 2.61}$	$0.692_{\pm 0.71}$
			8	$9.4_{\pm 2.30}$	$1.913_{\pm 1.26}$
			9	$11.2_{\pm 3.70}$	$2.478_{\pm 1.45}$
			10	$13.6_{\pm 3.51}$	$2.212_{\pm 2.40}$
50	R	10	5	$9_{\pm 3.67}$	$1.975_{\pm 1.52}$
			6	$9.8_{\pm 3.27}$	$4.332_{\pm 3.30}$
			7	$10.2_{\pm 3.19}$	$4.212_{\pm 3.01}$
			8	$14.8_{\pm 1.92}$	$16.231_{\pm 7.29}$
			9	$17_{\pm 3.94}$	$106.611_{\pm 83.57}$
			10	$20.2_{\pm 2.17}$	$316.153_{\pm 183.86}$
90	R	30	15	$11.2_{\pm 1.10}$	$43.322_{\pm 61.96}$
			20	$10.8_{\pm 4.55}$	$217.003_{\pm 460.53}$
			25	$15.8_{\pm 4.27}$	$818.926_{\pm 1498.70}$
			30	$19.4_{\pm 2.51}$	$3329.019_{\pm 3443.70}$
100	T,C,R	30	15	$9_{\pm 3.16}$	$1197.784_{\pm 2523.85}$
			20	$9.8_{\pm 2.86}$	$43.661_{\pm 47.24}$
			25	$14.4_{\pm 3.78}$	$480.446_{\pm 667.91}$
			30	$14.8_{\pm 2.77}$	$31.303_{\pm 57.35}$

4 Experimental Results

The proposed model was implemented in AMPL and experiments were conducted on different payload sizes using an ILP solver with a time limit: CPLEX 11 on a machine with two Intel Xeon X5355 2,66 GHz processors with 4 cores each, 32GB of RAM and running Ubuntu 8.04.

Experiments have been conducted on 4 groups of random payload instances, each one described in Table 1 by the number and type of switches used and the number of usable amplifiers. Each instance consists in connecting 5 sets of randomly chosen channels with no pre-existing connection (see column 'Number of required connections'). This will permit to analyze the influence of the subset of channels to connect on the solution quality and computational time. No pre-existing connection represents 'worst case' problems, since reconfiguring payloads with some pre-established connections would imply smaller search spaces. The last two columns respectively provide the average value and standard deviation of the objective function (number of switch changes) and the average required computational time for solving our problem instances.

Table 1 provides a scalability study of the proposed approach. All solutions have been validated using the computerised schematics. In terms of solutions quality, the number of switch changes increases as the number of required connections increases, in all but one case (90 switches, 30 amplifiers and 20 channels to connect). It can be noticed that the number of switch changes is lower for all the 100 switch instances that the 90 switch ones. The solution quality thus

depends on the topology of the switch network and on the type of switch used. Indeed in the 100 switch instance most of the switches are of type T and C, which ensure all the necessary connectivities with respectively only 3 and 2 possible states. The high standard deviation in the number of switch changes also indicates that the chosen set of channels to be connected has an important impact on the solution.

In terms of computational time taken by CPLEX, it generally also increases with the number of switches that compose the switch matrix, and the number of required connectivities. The exceptions are with 40 switches and 10 channels; 50 switches and 7 channels; and with 100 switches. As for the solution quality, this can be explained by the topology and type of switches of the payload. The standard deviation of the computational time is very high, especially for our large size payload systems (90 and 100 switches with 30 amplifiers). The chosen set of channels to connect has thus a key influence on the result. One possible explanation is that some channels may require less switch crossings to reach a set of amplifiers. To investigate this impact, it is necessary to compute for any channel the number of switch crossings that is required to connect to any amplifier. This will be conducted in our future work.

5 Conclusions and Future Work

In this paper, we proposed to tackle for the first time the optimisation of satellites' payload reconfiguration with an ILP model, based on network flow models. Experimental results using the CPLEX solver have been validated using computerised schematics and the scalability of the model has been studied on payloads of different sizes, up to realistic ones (from 40 to 100 switches).

As future work, we plan to further enhance the model in order to meet new operational objectives. Among them, the minimization of the path losses and number of signals interruptions is a direct next step. The robustness of the solution will also be considered, in order to ensure the functionality of the system in cases of potential failures. Finally, after determining new objectives, we will consider the definition and optimisation of multi-objective versions of the payload reconfiguration problem.

References

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B., Reddy, M.: Applications of network optimization. In: Network Models. Handbooks in Operations Research and Management Science, vol. 7, pp. 1–83. North-Holland (1995)
2. Antonio Capone, F.M.: A multi-commodity flow model for optimal routing in wireless mesh networks. Journal of Networks 2(3), 31–54 (2007)
3. Bertsekas, D.: Network optimization: continuous and discrete methods. Optimization and neural computation series. Athena Scientific (1998), http://books.google.lu/books?id=afYYAQAAIAAJ

4. Chaumon, J., Gil, J., Beech, T., Garcia, G.: Smartrings: advanced tool for communications satellite payload reconfiguration. In: 2006 IEEE Aerospace Conference, p. 11 (2006)
5. Gamvros, I.: Satellite network design, optimization and management. Ph.D. thesis, University of Maryland (2006)
6. Minoux, M.: Multicommodity network flow models and algorithms in telecommunications. In: Resende, M.G.C., Pardalos, P.M. (eds.) Handbook of Optimization in Telecommunications, pp. 163–184. Springer, US (2006)
7. TRECS: Transponder reconfiguration system, http://www.integ.com/TRECS.html
8. Wyatt-Millington, R.A., Sheriff, R.E., Hu, Y.F.: Performance analysis of satellite payload architectures for mobile services. IEEE Transactions on Aerospace and Electronic Systems 43(1), 197–213 (2007)

DC Programming and DCA for Large-Scale Two-Dimensional Packing Problems

Babacar Mbaye Ndiaye[1], Le Thi Hoai An[2],
Pham Dinh Tao[3], and Yi Shuai Niu[3]

[1] Laboratory of Mathematics of Decision and Numerical Analysis
Cheikh Anta Diop University - Dakar. BP 45087, Dakar-Fann Senegal
babacarm.ndiaye@ucad.edu.sn
[2] Laboratory of Theoretical and Applied Computer Science
UFR MIM, University of Paul Verlaine, Metz, Ile du Saulcy, 57045 Metz, France
lethi@univ-metz.fr
[3] Laboratory of Mathematics,
National Institute for Applied Sciences - Rouen,
University Avenue, BP 08, 76801 Saint-Etienne-du-Rouvray, France
{pham,niuys}@insa-rouen.fr

Abstract. In this paper, we propose a global optimization method based on DC (Difference of Convex functions) programming and DCA (DC Algorithm) involving cutting plane techniques for solving two-dimensional Bin packing and Strip packing problems. Given a set of rectangular items, we consider problems of allocating each item to larger rectangular standardized units. In two-dimensional bin packing problem, these units are finite rectangles, and the objective is to pack all the items into the minimum number of units. In two-dimensional strip packing problem, there is a single standardized unit of given width, and the objective is to pack all the items within the minimum height. These problems are characterized as BLP (Binary Linear Programming) problems. Thanks to exact penalty technique in DC Programming, the BLP can be reformulated as polyhedral DC program which can be efficiently solved via the proposed DC programming approach. Computational experiments on large-scale dataset involving up to 200 items show the good performance of our algorithm.

Keywords: Two-dimensional packing, Bin packing, Strip packing, BLP, DC Programming, DCA, Cutting plane techniques.

1 Introduction

Binary Linear Programming (BLP) techniques, dealing with optimization problems formulated in linear form, have demonstrated their advantages that help the industry to solve very complex planning problems. Several industrial problems typically include problems where the number of decision variables is so large that some (BLP) techniques are not able to deal with the exponential growth of the solution space. Such problems include packing problems (e.g. allocating a set of rectangular items to larger rectangular standardized units with minimum waste).

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 321–330, 2012.

In several industrial applications (for example wood and glass industries, transportation, warehousing etc.), the standardized stock units are rectangles, and a common objective function is to pack all the requested items into the minimum number of units: the resulting optimization problems are known in the literature as two-dimensional bin packing problems. In other contexts, such as paper or cloth industries, we have instead a single standardized unit (a roll of material), and the objective is to obtain the items by using the minimum roll length: the problems are then referred to as two-dimensional strip packing problems. Several contributions in the literature are devoted to the case where the rectangles to be packed cannot rotated 90 degrees. In this article, it is assumed that the items have fixed orientation, i.e., they cannot be rotated. Gilmore and Gomory [8] proposed the first model for two-dimensional packing problems, by extending their column generation approach for one-dimensional packing problem (see [7]). For a comprehensive overview of models and algorithms for two-dimensional packing problems see Martello and Vigo [15] and Fekete and Schepers [5], which present several lower bounds on the solution value using, respectively, partitioning of rectangles in various classes and dual feasible functions. Boschetti and Mingozzi [1] presented a new lower bound that dominates the bounds of Martello and Vigo [15] and Fekete and Schepers [5]. Considering the same variant of the problem, Dell'Amico et al. [3] presented a lower bound and an exact branch-and-bound algorithm. General surveys on packing problems can be found in Martello and Vigo [15] and Dyckhoff et al. [4]. Let us introduce the problems (see [13] for more details). We are given a set of N rectangular items $j \in J = \{1, ..., N\}$, each having *width* w_j and *height* h_j.

- in the Two-Dimensional Bin Packing Problem (2BP), we are further given an unlimited number of identical rectangular *bins* of width W and height H, and the objective is to allocate all items to the minimum number of bins;
- in the Two-Dimensional Strip Packing Problem (2SP), we are further given a bin of width W and infinite height (hereafter called *strip*), and the objective is to allocate all the items to the strip by minimizing the height to which the strip is used.

In both cases, the items have to be packed with their w-edges parallel to the W-edge of the bins (or strip). We will assume, without loss of generality, that all input data are positive integers, and that $w_j \leq W$ and $h_j \leq H$ $(j \in J = \{1, ..., N\})$. Both problems are strongly NP-hard, as is easily seen by transformation from the strongly NP-hard (one-dimensional) Bin Packing Problem (1BP), in which N items, each having an associated size h_j, have to be partitioned into the minimum number of subsets so that the sum of the sizes in each subset does not exceed a given capacity H. For both 2BP and 2SP, we consider the special case where the items have to be packed into rows forming levels. The first attempt to model two-dimensional packing problems was made by Gilmore and Gomory [8], through an extension of their approach to 1BP (see [7]).

Let A_j be a binary column vector of N elements a_{ij} $(i = \{1, ..., N\})$ taking the value 1 if item i belongs to the jth subset of items (pattern), and the value 0 otherwise. The set of all feasible patterns is then represented by the matrix A,

composed by all possible A_j columns $j \in J = \{1, ..., M\}$ and the corresponding mathematical model is (2BP):

$$\min \left\{ \sum_{j=1}^{M} x_j \ : \ \sum_{j=1}^{M} a_{ij} x_j = 1 \ \ (i = 1, ..., N), \ x_j \in \{0, 1\} \ \ (j = 1, ..., M) \right\}. \quad (1)$$

where x_j takes the value 1 if pattern j belongs to the solution, and the value 0 otherwise. Observe that (2BP) is a valid model for 1BP as well, the only difference being that the A_j's are all columns satisfying $\sum_{i=1}^{N} a_{ij} h_i \leq H$.
Due to the immense number of columns that can appear in A, the only way for handling the model is to dynamically generate columns when needed (for an example see [8]). While for 1BP Gilmore and Gomory [7] had given a dynamic programming approach to generate columns by solving, as a slave problem, an associated 0-1 knapsack problem, for 2BP they observed the inherent difficulty of the two-dimensional associated problem. Hence they switched to the more tractable case where the items have to be packed in rows forming levels for which the slave problem was solved through a two-stage dynamic programming algorithm.

Most of the approximation algorithms in the literature for 2BP and 2SP pack the items in rows forming levels. The first level is the bottom of the bin (or strip), and items are packed with their base on it. The next level is determined by the horizontal line drawn on the top of the tallest item packed on the level below, and so on (see Figure 1 (a)). Let us denote by 2BLBP (resp. 2BLSP) problem 2BP (resp. 2SP) restricted to this kind of packing.
We assume in the following, without loss of generality, that (see Figure 1 (b))

(i) in each level, the leftmost item is the tallest one;
(ii) in each bin/strip, the bottom level is the tallest one;
(iii) the items are sorted and re-numbered by non-increasing h_j values.

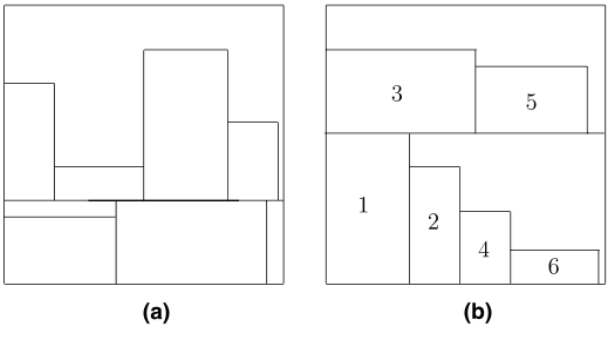


Fig. 1. (a) level packing; (b) normalized level packing

We will say that the leftmost item in a level (resp. the bottom level in a bin/strip) *initializes* the level (resp. the bin/strip). Problem 2BLBP can be efficiently modeled by assuming that there are N potential levels (the ith one associated with

item i initializing it), and N potential bins (the kth one associated with potential level k initializing it). Hence let y_i, $i \in J$ (resp. q_k, $k \in J$) be a binary variable taking the value 1 if item i initializes level i (resp. level k initializes bin k), and the value 0 otherwise. The problem can thus be modeled as:

$$\min \quad \sum_{k=1}^{N} q_k \tag{2}$$

subject to

$$\sum_{i=1}^{j-1} x_{ij} + y_j = 1 \qquad (j = 1, ..., N) \ . \tag{3}$$

$$\sum_{j=i+1}^{N} w_j x_{ij} \leq (W - w_i) y_i \qquad (i = 1, ..., N-1) \ . \tag{4}$$

$$\sum_{k=1}^{i-1} z_{ki} + q_i = y_i \qquad (i = 1, ..., N) \ . \tag{5}$$

$$\sum_{i=k+1}^{N} h_i z_{ki} \leq (H - h_k) q_k \qquad (k = 1, ..., N-1) \ . \tag{6}$$

$$q_k, y_i, x_{ij}, z_{ki} \in \{0, 1\} \qquad \forall i, j, k \ . \tag{7}$$

where x_{ij}, $i \in J \setminus \{N\}$ and $j > i$ (resp. z_{ki}, $k \in J \setminus \{N\}$ and $i > k$) takes the value 1 if item j is packed in level i (resp. level i is allocated to bin k), and the value 0 otherwise. Restrictions $j > i$ and $i > k$ easily follow from assumptions (i)-(iii) above. Eqs. (3) and (5) impose, respectively, that each item is packed exactly once, and that each used level is allocated to exactly one bin. Eqs. (4) and (6) impose, respectively the width constraint to each used level and the height constraint to each used bin. By modifying the objective function and eliminating all constraints (and variables) related to the packing of the levels into the bins, we immediately get the model for 2BLSP:

$$\min \ \Big\{ \sum_{i=1}^{N} h_i y_i \ : \ (3), (4), \ y_i, x_{ij} \in \{0, 1\} \quad \forall i, j \Big\} \ . \tag{8}$$

DC programming and DCA were introduced by Pham Dinh Tao in 1985 and extensively developed by Le Thi Hoai An and Pham Dinh Tao since 1994 to become classic and increasingly popular (see http://lita.sciences.univ-metz.fr/ lethi/) DCA is based on local optimality conditions and the duality in DC programming. In the paper, we present a cutting plane method based on these theoretical and algorithmic tools for solving binary linear programs (8). Unlike usual outer approximation algorithms, our algorithm introduces cutting planes from local solution computed by DCA applied to an equivalent penalized DC program of

(8). DCA converges to a local solution after finitely many iterations and it consists of solving a linear program at each iteration. Moreover, although our DCA is a continuous approach, it provides an integer solution. To ensure globality of solutions DCA is combined with the classical cutting plane techniques. This algorithm is known as DCA-Cut algorithm.

The following section describes an algorithm for solving the two problems. In Section 3 computational experiments are presented. Finally, summary and conclusions are provided in Section 4.

2 DCA-Cut Algorithm for Solving 2BLBP and 2BLSP

In this section, we propose a global method based on new cutting planes. The most important step of cutting plane algorithms is the separation problem. The strategy is to construct a cutting plane that cut off the non-binary optimal solution. Cutting plane algorithms can depend on the problem structure or can be generals (Mixed Integer Gomory cuts, MIR cut algorithm etc). For DCA-Cut algorithm, in contrast to the classical approaches [6,14], an equivalent problem with continuous variables is solved instead, derived from exact penalty techniques in DC Programming. The cutting plane is obtained from a local minimum of the penalty function in the relaxed domain of (2BLBP) (resp. (2BLSP)). This new method has been first introduced in [16,17]; where comparison results therein, in nonconvex real problems have shown the efficiency of this algorithm.

The problems (2BLBP) and (2BLSP) formulations are both of the form:

$$\alpha := \min\{C^T u : Au \leq b, u \in \{0,1\}^n\} \qquad (\mathcal{P})$$

where n is the number of binary variables, $C \in \mathbb{R}^n$, b is a m-vector, A is a $m \times n$ matrix and m is the number of constraints.

2.1 DC Reformulation and DCA for Solving the DC Program (11)

In this section, we illustrate how DCA-Cut algorithm is applied to 2BLBP problem (since 2BLSP is obtained by modifying 2BLBP). Without loss of generality we can denote 2BLBP by $(\mathcal{P})$. Using the well known results concerning exact penalty (see [12], Theorem 1), the problem $(\mathcal{P})$ is formulated as a concave minimization programming. The simplified DCA applied to the problem can be formulated as follows: Denote by:

- $q = (q_k)$, $y = (y_i)$, $x = (x_{ij})$, $z = (z_{ki})$ and $u = (q, y, x, z) \in \{0,1\}^n$.
- H the set of feasible points $u = (q, y, x, z)$ determined by the system of the constraints $\{(3), \ldots, (6)\}$; and $S = \{u = (q, y, x, z) \in H \ : \ u \in \{0,1\}^n\}$.
- $K = \{u = (q, y, x, z) \in H \ : \ u \in [0,1]^n\}$, is nonempty, bounded polyhedral convex set in $\mathbb{R}^n$.
- $p(u) = \sum_{i=1}^{n} \min(u_i, 1 - u_i)$. It is clear that p is concave and finite on K, and $p(u) \geq 0$ for all $u = (q, y, x, z) \in K$.

DC Reformulation by Exact Penalty Technique: The problem $(\mathcal{P})$ can be expressed in the form

$$\alpha = \min\{C^T u : u \in S\} \ . \tag{9}$$

Next, problem (9) can be rewritten as follows: $\alpha = \min\{C^T u : u \in K, p(u) \leq 0\}$ From the exact penalty theorem (see [12], Theorem 1), there exists $t_0 \geq 0$, such that we get for a sufficiently large number t $(t \geq t_0)$, the equivalent concave minimization problem to (9):

$$\alpha(t) = \min\{C^T u + tp(u) : u \in K\} \ . \tag{10}$$

$$= \min\{g(u) - h(u) : u \in \mathbb{R}^n\} \ . \tag{11}$$

with

$$\begin{cases} g(u) = \chi_K(u) \\ h(u) = \langle -\mathrm{C}, u\rangle - t \sum_{i=1}^{n} \min(u_i, 1 - u_i) \ . \end{cases} \tag{12}$$

It is easy to see that g and h are convex functions, thus (10) is a DC program with the form (11). Recall that if either g or h is polyhedral convex, then (11) is called a polyhedral DC program for which DCA has a finite convergence [9,10]. Consequently DCA applied to (10) has also a finite convergence.
According to (12) (definition of h), a subgradient $(a, b, c, d) \in \partial h(q, y, x, z)$ can be chosen as follows: $a = -C + (a_i)$ and $b = c = d = a_i$ with:

$$a_i = \begin{cases} t & \text{if } \ u_i \geq \dfrac{1}{2} \\ -t & \text{otherwise} \ . \end{cases} \tag{13}$$

The DCA to solve the problem $(\mathcal{P})$ is expressed through the following algorithm.

Algorithm 1. [DCA for (11)]
Let $\epsilon > 0$ be small enough and (q^0, y^0, x^0, z^0) a given initial point. $k \leftarrow 0$; $er \leftarrow 1$
While $(er > \epsilon)$ **do**
Compute $(a^k, b^k, c^k, d^k) \in \partial h(q^k, y^k, x^k, z^k)$ via (13)
Solve the linear problem:
$\min\{-\langle (a^k, b^k, c^k, d^k), (q, y, x, z)\rangle : u = (q, y, x, z) \in K\}$
to obtain $\{q^{k+1}, y^{k+1}, x^{k+1}, z^{k+1}\}$
$er \leftarrow \|(q^{k+1}, y^{k+1}, x^{k+1}, z^{k+1}) - (q^k, y^k, x^k, z^k)\|$
$k \leftarrow k + 1$
End While

Convergence properties of DCA and its theoretical basis can be found in [9,10,11]. Our algorithm converges to a critical point after a finite number of iterations and consists of solving a linear program at each iteration.

2.2 Construction of New Cutting Planes

A Valid Inequality: In this section, it is shown that, from a local minimum of the penalty function p, a valid inequality for all solutions in S can be constructed.

For all $u^* \in K$, let's denote $I := \{1, \ldots, n\}$; $J_0(u^*) := \{j \in I : u_j^* \leq \frac{1}{2}\}$; $J_1(u^*) := \{j \in I : u_j^* > \frac{1}{2}\}$; $p(u) := \sum_{i=1}^{n} \min(u_i, 1-u_i)$ and $l_{u^*}(u) := \sum_{j \in J_0(u^*)} u_j + \sum_{j \in J_1(u^*)} (1 - u_j)$. We have the following properties:

- $l_{u^*}(u) \geq p(u), \forall u \in \mathbb{R}^n$
- $l_{u^*}(u) = p(u)$, if and only if $u \in R(u^*) := \{u \in \mathbb{R}^n : x_j \leq \frac{1}{2} \ \forall j \in J_0(u^*); \ x_j > \frac{1}{2} \ \forall j \in J_1(u^*)\}$.

Lemma 1. *[17] Let u^* be a local minimum of the penalty function p on K, then the inequality*

$$l_{u^*}(u) \geq l_{u^*}(u^*) \tag{14}$$

is valid $\forall u \in K$.

Let u^* be a KKT (Karush-Kuhn-Tucker) point of the problem

$$\alpha = \min\{p(u) \ : \ u \in K\} \ . \tag{15}$$

If $u^* \neq \frac{1}{2}$ (i.e. u^* is a local minimum of p), from Lemma 1 we obtain the valid inequality (14). Otherwise, we solve the following linear program

$$\xi = l_{u^*}(\tilde{u}) = \min\{l_{u^*}(u) \ : \ u \in K\} \tag{16}$$

and test if $l_{u^*}(u^*) > l_{u^*}(\tilde{u})$. If the inequality is valid then $p(u^*) = l_{u^*}(u^*) > l_{u^*}(\tilde{u}) \geq p(\tilde{u})$, starting from $\tilde{u}$ we could find a local minimum of p where a valid inequality is constructed according to Lemma 1. Otherwise, a local minimum of p could be also found from $\tilde{u}$ if $\tilde{u} \notin R(u^*)$ since $p(u^*) = l_{u^*}(u^*) = l_{u^*}(\tilde{u}) > p(\tilde{u})$.

Construction of a Cutting Plane from an Infeasible Local Solution: Let u^* be a local minimum of the problem (15), and let u^* not being a feasible solution of the problem $(\mathcal{P})$. Then, at least there exists an index j_0 such that $u_{j_0}^* \notin \{0,1\}^n$. In addition, we have two cases: $p(u^*)$ is not integer or $p(u^*)$ is integer.

Case 1: If u^* is a local minimum such that $p(u^*)$ is not integer, a cutting plane which cuts off u^* from S is straightforwardly obtained by the following theorem.

Theorem 1. *[17] If u^* is an infeasible minimum and $p(u^*)$ is not integer then the following inequality is a cutting plane which cuts off u^* from S.*

$$l_{u^*}(u) := \sum_{j \in J_0(u^*)} u_j + \sum_{j \in J_1(u^*)} (1 - u_j) \geq \lfloor l_{u^*}(u^*) \rfloor + 1 \ . \tag{17}$$

Case 2: In the case where u^* is an infeasible local minimum with the integer value $p(u^*)$, it is possible to obtain feasible solutions $\acute{u}$ such that $l_{u^*}(\acute{u}) = l_{u^*}(u^*)$. If there exists $\acute{u}$, the best feasible solution and the upper bound of $(\mathcal{P})$ could be updated. Otherwise, for all $u \in S$, we have $l_{u^*}(u) > l_{u^*}(u^*)$, i.e.

$$l_{u^*}(u) \geq l_{u^*}(u^*) + 1 \ . \tag{18}$$

is a cutting plane which cuts off u^* from S.

By applying a procedure (Procedure P) (see [16,17] for more details) , either a feasible solution or a potential point or a cutting plane can be obtained.

Construction of a Cutting Plane from a Feasible Local Solution: If u^* is a feasible minimum of p, then we have the same cutting plane as (17) which cuts off the point u^* from S. In this case, $l_{u^*}(u^*) = p(u^*) = 0$, the cutting plan is turned into:

$$l_{u^*}(u) := \sum_{\{j:u_j^*=0\}} u_j + \sum_{\{j:u_j^*=1\}} (1 - u_j) \geq 1 \ . \tag{19}$$

The Idea of DCCUT Algorithm: DCA applied to (11) gives both local solution of $(\mathcal{P})$ and for constructing cutting planes (instead of searching local minimum of p, we can find local minimum of (11) by DCA).

Let α be the optimal value of the problem $(\mathcal{P})$. We can build two sequences $\{\gamma_k\}$ (upper bounds, the feasible solution of $(\mathcal{P})$ found by DCA) and $\{\beta_k\}$ (lower bounds, solution of the linear relaxation of $(\mathcal{P})$) decreasing and increasing respectively, such that: $\beta_k \leq \alpha \leq \gamma_k \ \forall k = 1, 2, \ldots$. At each iteration, either a feasible solution and/or a cutting plane is found. If a feasible solution is found, the upper bound γ_k and the best known feasible solution are updated and a cutting plane can be also constructed; Otherwise, we obtain a cutting plane. The feasible set of the problem is reduced in each iteration due to the introduced cutting planes. The algorithm is terminated when the gap between the upper and the lower bounds is smaller enough or the final reduced set does not contain feasible solution of S. For more details about the theoretical basis of DCA-Cut Algorithm and its convergence properties, the reader is referred to [16,17].

3 Numerical Experiments

In the basic model (2BP) presented in Section 1, due to the immense number of columns that can appear in the matrix A (representing the set of all feasible subsets of items), the only way for handling the model is to dynamically generate columns when needed. Taking into account the three assumptions in Section 1, the basic bin packing model (2BP) could be dynamically presented as the formulations (2BLBP) and (2BLSP) which are quite useful in practice.

To test the performance of the algorithm DCA-Cut proposed in Section 2, we consider eight classes of randomly generated problems. If we consider instances (involving less than 150 items) proposed by Martelo and Vigo [15] (they use it for solving the base model (2BP)), the solutions of the two models are obtained. For each value of N (number of items), two instances are generated where a realistic situation is considered. For 2BLBP classes, w_j and h_j uniformly random integers in [0, 100], W=H=100. Small size instances with $N = 200$ and $N = 400$ were solved to optimality. For 2BLSP classes, w_j and h_j uniformly random integers in [0, 12], W=H=12.

The algorithm is implemented in C++, and tested on Dell computer: 2×Intel(R) Core(TM)2 Duo CPU 2.00GHz, 4.0Gb of RAM, under UNIX system. The number of items represents the numbers of rectangular items, each rectangular is defined by a width w_j, and a height h_j. We denote: DCCUT = DCA-Cut Algorithm; CPLEX 11.0 = CPLEX *Optimization* ILOG, version 11.0 [2]; time(s) = total execution

time in seconds; iter = number of iterations; lb = lower bound; ub = upper bound; obj = objective value of Cplex; nv = number of variables; nbcut = number of cuts of DCCUT. Computational experiments on large dataset involve up to 200 items. The proposed DCCUT algorithm can efficiently produce very good solutions (optimal solutions are found).

To verify the obtained solutions of DCCUT, we compare with solutions given by CPLEX 11.0 (we give to CPLEX the original problem), see Tables 1 and 2.

For all scenarios of 2BP, the value $lb = ub = 1.0$ is obtained. DCCUT gives the optimal value, the same value found by CPLEX (obj=1.0). Also, DCA is run only once, see Table 1. Thus, the first level is the bottom of the bin, and items are packed with their base on it. We find the same optimal value as CPLEX. The CPLEX is fast on these examples and made? few iterations.

What's interesting here is that we always need adding one cut to obtain the same optimal value as CPLEX, and DCA is run only once. More specifically, the optimal value is obtained from a single cut added via a feasible solution of DCA. Our method usually provides a best upper bound after two iterations, for packing all the items into the minimum number of units or within the minimum height. DCCUT is inexpensive for handling large-scale problems as DCA often requires few number of iterations (2 in average) and the introduced DCCUT can intensely improve the upper bound.

Table 1. DCCUT and Cplex for 2BLBP

Number	nbre of	DCCUT		
of items	variables	time(s)	iter	nbcut
600	360600	16.03	2	1
600	360600	16.48	2	1
800	640800	22.58	2	1
800	640800	22.56	2	1

Table 2. DCCUT and Cplex for 2BLSP

Number	nbre of	DCCUT					CPLEX 11.0
of items	variables	time(s)	iter	lb	ub	nbcut	obj
200	20100	0.60	2	6.0	6.0	1	6.0
200	20100	0.59	2	8.0	8.0	1	8.0
600	180300	6.08	2	4.0	4.0	1	4.0
600	180300	6.11	2	12.0	12.0	1	12.0
800	320400	10.98	2	7.0	7.0	1	7.0
800	320400	11.24	2	10.0	10.0	1	10.0
1000	500500	17.53	2	12.0	12.0	1	12.0
1000	500500	17.51	2	7.0	7.0	1	7.0

4 Conclusion

In this paper, we address the two-dimensional bin packing and strip packing problems. The combined DCCUT is interesting, the results show that it gives

optimal solution within very short CPU running times. We notice that, the DC cutting plane technique combined with DCA can satisfactorily solve the problem. Finally, we point out that the presented algorithm can be applied to solve the general mixed-binary linear programs to produce lower bounds for bin-packing and strip-packing models not restricted to this kind of packing.

References

1. Boschetti, M.A., Mingozzi, A.: The two-dimensional finite bin packing problem. Part I: New lower bounds for the oriented case. 4OR 1, 27–42 (2003)
2. CPLEX Optimization ILOG, IncUsing the CPLEX^{R} Callable Library and CPLEX Barrier and Mixed Integer Solver Options Version 11.0 (2007)
3. Dell'Amico, M., Martello, S., Vigo, D.: A lower bound for the non-oriented two-dimensional bin packing problem. Discrete Appl. Math. 118, 13–24 (2002)
4. Dyckhoff, H., Scheithauer, G., Terno, J.: Cutting and packing (C&P). In: Dell'Amico, M., Maffioli, F., Martello, S. (eds.) Annotated Bibliographies in Combinatorial Optimization, pp. 393–413. Wiley, Chichester (1997)
5. Fekete, S.P., Schepers, J.: New classes of lower bounds for the bin packing problem. Math. Programming 91, 11–31 (2001)
6. Gomory, R.E.: Solving Linear Programming Problems in Integer. In: Bellman, R.E., Hall Jr., M. (eds.) Combinatorial Analysis, pp. 211–216. American Mathematical Society (1960)
7. Gilmore, P.C., Gomory, R.E.: A linear programming approach to the cutting stock problem - part II. Operations Research 11, 863–888 (1963)
8. Gilmore, P.C., Gomory, R.E.: Multistage cutting problems of two and more dimensions. Oper. Res. 13, 94–119 (1965)
9. Le Thi, H.A., Pham Dinh, T.: Convex analysis approach to d.c. programming: Theory, Algorithms and Applications. Acta Mathematica Vietnamica 22(1), 289–355 (1997)
10. Le Thi, H.A., Pham Dinh, T.: The DC Programming and DCA revisited with DC models of real world nonconvex optimization problems. Annals of Operations Research 133, 23–46 (2005)
11. Le Thi, H.A., Pham Dinh, T.: DC optimization algorithms for solving the trust region subproblem. SIAM J. Optimisation 8, 476–505 (1998)
12. Le Thi, H.A., Pham Dinh, T., Le Dung, M.: Exact penalty in dc programming. Vietnam Journal of Mathematics 27(2), 169–178 (1999)
13. Lodi, A., Martello, S., Monaci, M.: Two-dimensional packing problems: A survey. European Journal of Operational Research 141, 241–252 (2002)
14. Marchand, H., Martin, A., Weismantel, R., Wolsey, L.: Cutting Planes in integer and mixed integer programming. Core Discussion Paper 9953, 1–50 (1999)
15. Martello, S., Vigo, D.: Exact solution of the two-dimensional finite bin packing problem. Management Science 44, 388–399 (1998)
16. Ndiaye, B.M., Pham Dinh, T., Le Thi, H.A.: DC programming and DCA for SS-CRP. In: Modelling, Computation and Optimization in Information Systems and Management Sciences. CCIS, vol. 14, pp. 21–30. Springer, Heidelberg (2008)
17. Nguyen, V.V.: Exact methods for polyhedral DC program with mixed 0-1 variables, based on DCA and New cutting planes. PhD Thesis, National Institute for Applied Sciences, Rouen (July 2006)

Gaussian Kernel Minimum Sum-of-Squares Clustering and Solution Method Based on DCA

Le Hoai Minh[1], Le Thi Hoai An[1], and Pham Dinh Tao[2]

[1] Laboratory of Theoretical and Applied Computer Science - LITA EA 3097
UFR MIM, University of Paul Verlaine - Metz, Ile de Saulcy, 57045 Metz, France
{lehoai,lethi}@univ-metz.fr
http://lita.sciences.univ-metz.fr/lethi/
[2] Laboratory of Modelling, Optimization & Operations Research
National Institute for Applied Sciences - Rouen,
Avenue de l'Universit?- 76801 Saint-Etienne-du-Rouvray Cedex, France
pham@insa-rouen.fr

Abstract. In this paper, a Gaussian Kernel version of the Minimum Sum-of-Squares Clustering ($GKMSSC$) is studied. The problem is formulated as a DC (Difference of Convex functions) program for which a new algorithm based on DC programming and DCA (DC Algorithm) is developed. The related DCA is original and very inexpensive. Numerical simulations show the efficiency of DCA and its superiority with respect to K-mean, a standard method for clustering.

Keywords: clustering, kernel function, DC programming, DCA.

1 Introduction

Clustering, which aims at dividing a data set into groups or clusters containing similar data, is a fundamental problem in unsupervised learning and has many applications in various domains. In recent years, there has been significant interest in developing clustering algorithms to massive data sets (see e.g. [1,2,3,4] and the references therein). Two main approaches have been used for clustering: statistical and machine learning based on mixture models (see e.g. [1]) and the mathematical programming approach that considers clustering as an optimization problem (see e.g. [3,4,11,12,13,14,15] and the references therein).

The general term "clustering" covers many different types of problems. All consist of subdividing a data set into groups of similar elements, but there are many measures of similarity, many ways of measuring, and various concepts of subdivision (see [2] for a survey). Among these criteria, a widely used one is the minimum sum of squared Euclidean distances from each entity to the centroid of its cluster, or minimum sum-of-squares (MSSC) for short, which expresses both homogeneity and separation.

An instance of the partitional clustering problem consists of a data set $\mathcal{A} := \{a^1, ...a^m\}$ of m entities in $\mathbb{R}^n$, a measured distance, and an integer k; we are to choose k members x^i $(i = 1, ...k)$ and assign each member of $\mathcal{A}$ to its closest

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 331–340, 2012.

centroid. Among many criteria used for cluster analysis, the minimum sum-of-squares (or MSSC in brief) is one of the most popular since it expresses both homogeneity and separation. MSSC consists in partitioning the set $\mathcal{A}$ into k clusters in order to minimize the sum of squared distances from the entities to the centroid of their cluster. This problem may be formulated mathematically in several ways, which suggest different possible algorithms. The two most widely used models are a bilevel programming problem and a mixed integer program. Many early studies were focused on the well-known K-means algorithm ([16]) and its variants (see [17] for a survey).

In this paper,we are interested in the bilevel formulation of MSSC which was given by Vinod [18]:

$$\min \quad \Big\{ \sum_{j=1..m} \min_{i=1,\ldots,k} \left\|x^i - a^j\right\|^2 : x^i \in \mathbb{R}^n, i = 1, \ldots, k \Big\}$$

We introduce a Gaussian Kernel version of MSSC, called ($G\mathcal{K}MSSC$), and develop a new algorithm for solving the problem. Our approach is based on DC (Difference of convex functions) programming and DCA (DC Algorithms), an efficient approach in nonsmooth, nonconvex programming that has been successfully applied to many large-scale (smooth or nonsmooth) nonconvex programs in various domains of applied sciences, (see [6,8,9] and references therein), in particular in data analysis and data mining (see e.g. [3,4,5]). A so-called DC program is that of minimizing a DC function over a convex set. According to the theory of DC programming, it is easy to show that ($G\mathcal{K}MSSC$) model is a DC program. We then suggested using DC programming approach and DCA to solve the problem. Preliminary numerical simulations on real-world databases, show the robustness, the efficiency and superiority of the appropriate DCA with respect to the K-means algorithm in terms of quality of solutions.

The paper is organized as follows. The DC programming and DCA are briefly presented in Section 2. Section 3 deals with a customized DCA for the underlying Gaussian Kernel MSSC problem. A (VNS) algorithm for finding a good starting point for the main algorithm DCA is discussed in Section 4. Computational results are reported in the last section.

2 An Introduction to DC Programming and DCA

In this section, we give a brief introduction of DC programming and DCA to give the reader an easy understanding of these tools and our motivation to use them for solving Problem ($G\mathcal{K}MSSC$).

DC (Difference of Convex functions) Programming and DCA(DC Algorithms), being introduced by Pham Dinh Tao in 1985 and extensively developed by Le Thi Hoai An and Pham Dinh Tao since 1994, constitute the backbone of smooth/nonsmooth nonconvex programming and global optimization. They address DC programs of the form

$$\alpha = \inf\{f(x) := g(x) - h(x) : x \in \mathbb{R}^p\} \quad (P_{dc}) \tag{1}$$

where g, h are lower semicontinuous proper convex functions on $\mathbb{R}^p$. Such a function f is called DC function, and $g - h$, DC decomposition of f while g and h are DC components of f. Recall the natural convention $+\infty - (+\infty) = +\infty$ in DC programming, and that DC program with a closed convex constraint set $C \subset \mathbb{R}^p$

$$\beta = \inf\{\varphi(x) - \phi(x) : x \in C\}$$

can be rewritten in the form of (P_{dc})

$$\beta = \inf\{g(x) - h(x) : x \in \mathbb{R}^p\},$$

where $g := \varphi + \chi_C$, $h := \phi$ and χ_C stands for the indicator function of C : $\chi_C(x) = 0$ if $x \in C$, $+\infty$ otherwise. Let

$$g^*(y) := \sup\{\langle x, y\rangle - g(x) : x \in \mathbb{R}^p\}$$

be the Fenchel conjugate function of g. Then, the following program is called the dual program of (P_{dc}):

$$\alpha_D = \inf\{h^*(y) - g^*(y) : y \in \mathbb{R}^p\}. \quad (D_{dc}) \tag{2}$$

One can prove that $\alpha = \alpha_D$, (see e.g. [8]) and there is the perfect symmetry between primal and dual DC programs: the dual to (D_{dc}) is exactly (P_{dc}).

DCA is based on the local optimality conditions of (P_{dc}), namely

$$\partial h(x^*) \cap \partial g(x^*) \neq \emptyset \tag{3}$$

(such a point x^* is called critical point of $g - h$ or generalized KKT point for (P_{dc})), and

$$\emptyset \neq \partial h(x^*) \subset \partial g(x^*). \tag{4}$$

The necessary local optimality condition (4) for (P_{dc}) is also sufficient for many classes of DC programs.

The idea of DCA is quite simple. Each iteration of DCA approximates the second DC component h by its affine minorization (that corresponds to taking $y^l \in \partial h(x^l)$) and solves the resulting convex program (P_l) (that is equivalent to computing $x^{l+1} \in \partial g^*(y^l)$).

DCA Scheme

Initialization: Let $x^0 \in \mathbb{R}^p$ be a best guest, $l \leftarrow 0$.
Repeat
Calculate $y^l \in \partial h(x^l)$
Calculate $x^{l+1} \in \arg\min\{g(x) - h(x^l) - \langle x - x^l, y^l\rangle : x \in \mathbb{R}^p\}$ (P_l)
$l \leftarrow l + 1$
Until convergence of x^l.

Convergence properties of DCA and its theoretical basis can be found in [6,7,8,9]. For instance it is important to mention that

- DCA is a descent method (the sequences $\{g(x^l)-h(x^l)\}$ and $\{h^*(y^l)-g^*(y^l)\}$ are decreasing) without linesearch;
- If the optimal value α of the problem (P_{dc}) is finite and the infinite sequences $\{x^l\}$ and $\{y^l\}$ are bounded, then every limit point x^* (resp. y^*) of the sequence $\{x^l\}$ (resp. $\{y^l\}$) is a critical point of $g-h$ (resp. h^*-g^*).
- DCA has a linear convergence for general DC programs and has a finite convergence for polyhedral DC programs.

For a detailed convergence analysis of DCA, see [10].

DCA's distinctive feature relies upon the fact that DCA deals with the convex DC components g and h but not with the DC function f itself. Moreover, a DC function f *has infinitely many DC decompositions (and there are as many DCA as there are equivalent DC programs and their DC decompositions) which have crucial implications for the qualities* (speed of convergence, robustness, efficiency, globality of computed solutions,...) of DCA. Finding an appropriate equivalent DC program and a suitable DC decomposition is consequently important from the algorithmic point of view. For a complete study of DC programming and DCA the reader is referred to [6,7,8,9] and references therein. The solution of a nonconvex program by DCA must be composed of two stages: the search of both a fit DC program and its relevant DC decomposition, and the choice strategy of a good initial point, taking into account the specific structure of the nonconvex program.

3 New Model of Kernel Minimum Sum-of-Square Clustering

To simplify related computations in DCA we will work on the vector space $\mathbb{R}^{k\times n}$ of $(k\times n)$ real matrices. The variables are then $X\in\mathbb{R}^{k\times n}$ whose i^{th} row X_i is equal to x^i for $i=1,...,k$. The Euclidean structure of $\mathbb{R}^{k\times n}$ is defined with the help of the usual scalar product

$$\mathbb{R}^{k\times n}\ni X\longleftrightarrow(X_1,X_2,\ldots,X_k)\in(\mathbb{R}^n)^k,\ X_i\in\mathbb{R}^n,(i=1,..,k),$$
$$\langle X,Y\rangle\ :=Tr(X^TY)=\sum_{i=1..k}\langle X_i,Y_i\rangle$$

and its Euclidean norm $\|X\|^2:=\sum\limits_{i=1..k}\langle X_i,X_i\rangle=\sum\limits_{i=1..k}\|X_i\|^2$ (Tr denotes the trace of a square matrix).

Let $\phi:\mathbb{R}^m\longrightarrow\mathcal{H}$ be a transformation that maps each data a^i from the input space $\mathbb{R}^m$ to a new space $\mathcal{H}$, being a Hilbert space, where the given algorithm can be used. Such transformation is done implicitly by means of a kernel function [20] $\mathcal{K}$, satisfying:

$$\mathcal{K}(a^i,a^j)=\langle\phi(a^i),\phi(a^j)\rangle.$$

Mercer's theorem guarantees that as long as the corresponding matrix $\mathcal{K}$ of kernel function is positive definite, the algorithm implicitly operates in a higher

dimensional space. This kernel trick saves the algorithm from the computational expense of explicitly representing all of the features in a higher-dimensional space.

With a kernel function ϕ the objective function of (MSSC) becomes

$$F(X) := \sum_{j=1..m} \min_{i=1,\dots k} \left\|\phi(x^i) - \phi(a^j)\right\|^2. \tag{5}$$

Since

$$\begin{aligned}\left\|\phi(x^i) - \phi(a^j)\right\|^2 &= \langle\phi(x^i), \phi(x^i)\rangle - 2\langle\phi(x^i), \phi(a^j)\rangle \ + \ \langle\phi(a^j), \phi(a^j)\rangle \\ &= \mathcal{K}(x^i, x^i) - \ 2\mathcal{K}(a^j, x^i) + \mathcal{K}(a^j, a^j),\end{aligned} \tag{6}$$

the kernel version of (MSSC) takes the form

$$(\mathcal{K}MSSC) \left\{ \min_X \left\{ f(X) := \sum_{j=1..m} \min_{i=1,\dots k} \left(\mathcal{K}(x^i, x^i) - \ 2\mathcal{K}(a^j, x^i) + \mathcal{K}(a^j, a^j)\right) \right\}.$$

This is named "Kernel Minimum Sum-of-Square Clustering" problem which will be called briefly, in the sequel, $\mathcal{K}MSSC$ problem.

Using a given kernel function we will get its corresponding $\mathcal{K}MSSC$ model. Obviously, from numerical points of view, the degree of difficulty of $\mathcal{K}MSSC$ problems varies from each to other model. Among several kernel functions we consider in this paper the Gaussian Kernel. We will see in the next that this choice is interesting: we can investigate simple and efficient algorithm based on DC programming and DCA for solving the corresponding $\mathcal{K}MSSC$ problem.

By replacing the Gaussian kernel function $k(x, y) := \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ in the objective function of $(\mathcal{K}MSSC)$ we have:

$$\begin{aligned}F(X) := \sum_{j=1..m} \min_{i=1,\dots k} \Big[&\exp\Big(-\tfrac{\left\|a^j - a^j\right\|^2}{2\sigma^2}\Big) - 2\exp\Big(-\tfrac{\left\|x^i - a^j\right\|^2}{2\sigma^2}\Big) \\ &+ \exp\Big(-\tfrac{\left\|x^i - x^i\right\|^2}{2\sigma^2}\Big)\Big].\end{aligned}$$

Then the optimization model of the Gaussian Kernel MSSC is written as

$$(G\mathcal{K}MSSC) \quad \min \quad \left\{ F_G(X) := \sum_{j=1..m} \min_{i=1,\dots k} \left[-\ 2\exp\left(-\frac{\left\|x^i - a^j\right\|^2}{2\sigma^2}\right)\right] \right\}.$$

We will show in next section how to investigate DCA for solving the above $G\mathcal{K}MSSC$ problem.

4 Solving Gaussian Kernel Minimum Sum-of-Square Clustering Problems by DCA

DC Formulation of the Gaussian Kernel MSSC Problem. First of all, we note that by bounding the variable X in the box $\mathcal{T} := \prod_{i=1\dots k} \left[\alpha^i, \beta^i\right]$ with

$\alpha^i = \alpha = (\alpha_l)_{l=1\ldots n}$, $\alpha_l := \min_{j=1\ldots m} a_l^j$, and $\beta^i = \beta = (\beta_l)_{l=1\ldots n}$, $\beta_l := \max_{j=1\ldots m} a_l^j$. We can express the unconstrained optimization problem $(G\mathcal{K}MSSC)$ as a constrained optimization problem whose the feasible set is $\mathcal{T}$.

In an elegant way, we introduce a nice DC reformulation of $(G\mathcal{K}MSSC)$ for which the resulting DCA is explicitly determined via a very simple formula. Such a DC decomposition of f is inspired by the following result.

Let us denote

$$\varphi(x) = -2\exp\left(-\frac{\|x-a\|^2}{2\sigma^2}\right). \tag{7}$$

Theorem 1. *There exists $\rho > 0$ such that the function $h(x) = \frac{\rho}{2}\|x\|^2 - \varphi(x)$ is a convex function on $x \in [\alpha, \beta]$.*

Proof. We have: $\nabla\varphi(x) = \frac{2}{\sigma^2}\exp\left(-\frac{\|x-a\|^2}{2\sigma^2}\right)(x-a)$. Hence,

$$\nabla^2\varphi(x) = \frac{2}{\sigma^2}\exp\left(-\frac{\|x-a\|^2}{2\sigma^2}\right)\left[I - \frac{1}{\sigma^2}(x-a)(x-a)^T\right].$$

We have

$$\lambda_n(\nabla^2\varphi(x)) = \frac{2}{\sigma^2}\exp\left(-\frac{\|x-a\|^2}{2\sigma^2}\right) \leq \frac{2}{\sigma^2}$$

where λ_n denotes the largest eigenvalue of $\nabla^2\varphi(x)$. So, if $\rho \geq \frac{2}{\sigma^2} \geq \lambda_n(\nabla^2\varphi(x))$ then $\rho I - \nabla^2\varphi(x)$ is semi-definite positive. Thus, $h(x) = \frac{\rho}{2}\|x\|^2 - \varphi(x)$ is a convex function. ■

Let us denote

$$f_{ij}(X) = -2\exp\left(-\frac{\|x^i-a^j\|^2}{2\sigma^2}\right). \tag{8}$$

Using the theorem above, we get the DC decomposition of $f_{ij}(X)$:

$$g_{ij}(X) = \tfrac{\rho}{2}\|x^i\|^2, h_{ij}(X) = \tfrac{\rho}{2}\|x^i\|^2 - f_{ij}(X) \tag{9}$$

with $\rho \geq \frac{2}{\sigma^2}$. It follows that, for $j = 1..m$,

$$\begin{aligned} f_j(x) &:= \min_{i=1..k} f_{ij}(X) = \min_{i=1..k}\left(\tfrac{\rho}{2}\|x^i\|^2 - \left(\tfrac{\rho}{2}\|x^i\|^2 - f_{ij}(X)\right)\right) \\ &= \tfrac{\rho}{2}\sum_{i=1..k}\|x^i\|^2 - \max_{i=1..k}\left(\tfrac{\rho}{2}\sum_{l=1..k, l\neq i}\|x^l\|^2 + \tfrac{\rho}{2}\|x^i\|^2 - f_{ij}(X)\right) \\ &= \tfrac{\rho}{2}\|X\|^2 - \max_{i=1..k}\left(\tfrac{\rho}{2}\|X\|^2 - f_{ij}(X)\right). \end{aligned} \tag{10}$$

Consequently, the objective function of $(G\mathcal{K}MSSC)$ can be now written as

$$F_G(X) = \sum_{j=1..m} f_j(X) = G(X) - H(X), \tag{11}$$

where $G(X) := \frac{m\rho}{2}\|X\|^2$ and $H(X) := \sum_{j=1..m}\max_{i=1..k}\left(\frac{\rho}{2}\|X\|^2 - f_{ij}(X)\right)$ are convex functions. Hence, we can recast Problem $(G\mathcal{K}MSSC)$ as a DC program as follows

$$\min\left\{G(X) - H(X) : X \in \mathcal{T}\right\}. \tag{12}$$

DCA Applied to DC Program (12). According to Section 2, determining the DCA scheme applied to (12) amounts to computing the two sequences $\{X^{(l)}\}$ and $\{Y^{(l)}\}$ in $\mathbb{R}^{k\times n}$ such that $Y^{(l)} \in \partial\left(H(X^{(l)})\right)$ and $X^{(l+1)}$ by solving the following convex quadratic program

$$\min\left\{G(X) - \langle Y^{(l)}, X\rangle : X \in \mathcal{T}\right\} = \min\left\{\frac{m\rho}{2}\|X\|^2 - \langle Y^{(l)}, X\rangle : X \in \mathcal{T}\right\}$$

whose optimal solution can be explicitly determined in a very simple way. Indeed, the solution of above problem can be computed as

$$(X^i)^{(l+1)} = P_{[\alpha,\beta]}\left((Y^i)^{(l)}/(m\rho)\right), \text{for i=1..k,} \tag{13}$$

where $P_{(.)}$ denotes the projection onto the set (.). Note that the projection of points onto a box is explicitly computed.

We compute a gradient of $H(X)$ as follows:

$$Y \in \partial H(X) \Leftrightarrow Y = \sum_{j=1..m} \partial h_j(X), \tag{14}$$

where $h_j(X) := \max_{i=1..k} h_{ij}(X)$ and $h_{ij}(X) = \frac{\rho}{2}\|X\|^2 - f_{ij}(X)$.

Let $I_j(X) := \{i = 1, \ldots, k : h_{ij}(X) = h_j(X)\}$. We have:

$$\partial h_j(X) = co\left\{\cup_{i\in I_j(X)} \partial h_{ij}(X)\right\},$$

where *co* stands for the convex hull. Hence $\partial h_j(X)$ is a convex combination of $\{\nabla(h_{ij})(X) : i \in K_j(X)\}$, i.e.,

$$\partial h_j(X) = \sum_{i\in I_j(X)} \lambda_i^{[j]} \nabla(h_{ij})(X) \text{ with } \lambda_i^{[j]} \geq 0 \text{ for } i \in I_j(X) \text{ and } \sum_{j\in I_i(X)} \lambda_i^{[j]} = 1. \tag{15}$$

The gradient of $h_{ij}(X)$ is computed as (for $q = 1, ..., k$)

$$[\nabla(h_{ij})(X)]_q = \rho x^l - \frac{2}{\sigma^2}.\exp\left(-\frac{\left\|X_i - a^j\right\|^2}{2\sigma^2}\right)(X_i - a^j) \text{ if } q = i, 0 \text{ otherwise.} \tag{16}$$

Finally, $Y \in \partial H(X)$ is determined via the formulas (14), (15) and (16). From the above computations, the DCA applied to problem (12) can be described as follows:

DCA-GKMSSC

- **Initialization:** Let $\epsilon > 0$ be small enough and $X^{(0)}$ be given. Set $l \leftarrow 0$; $er \leftarrow 1$.
- **While** $er > \epsilon$ **do**
  - Compute $Y^{(l)} \in \partial H(X^{(l)})$ via the formulas (14), (15) and (16).
  - Compute $X^{(l+1)}$ via (13).
  - $er \leftarrow \left\|X^{(l+1)} - X^{(l)}\right\| / \left(\left\|X^{(l+1)}\right\| + 1\right)$.
  - $l \leftarrow l + 1$.
- **Endwhile**

5 VNS for Initializing DCA

For finding a good starting point of DCA we use a recent, simple and effective metaheuristic (or framework for heuristics) called Variable Neighborhood Search (VNS) [19].

The principle of VNS is to change and randomly explore neighborhoods with an appropriate local search routine. Contrary to other metaheuristics, e.g., simulated annealing or Tabu search , VNS does not follow a trajectory but explores increasingly distant neighborhoods of the current incumbent solution, and jumps from there to a new one if and only if an improvement has been made, through the local search.

Let us denote a finite set of preselected neighborhood structures with $\mathcal{N}_l$ $(l = 1, \ldots, l_{max})$ and with $\mathcal{N}_l(x)$ (and preferably such as $\mathcal{N}_l \subset \mathcal{N}_{l+1}$) the set of solutions in the l^{th} neighborhood of x.

Steps of a basic (VNS) heuristic, applied to the problem $\min\{f(x) : x \in S\}$, are the following:

◇ **Initialization :** Select the set of neighborhood structures $\mathcal{N}_l$ and $l = 1, \ldots, l_{max}$, that will be used in the search; find an initial solution x; choose a stopping condition;

◇ **Repeat the following until the stopping condition is met:**

- Set $l \leftarrow 1$
- Until $l = l_{max}$, repeat the following steps:
  - Shaking. Generate a point $X^{'}$ at random from the l^{th} neighborhood of $X(X^{'} \in \mathcal{N}_l(X))$;
  - Local search. Apply some local search method with $X^{'}$ as initial solution; denote with $X^{''}$ the so-obtained local optimum;
  - Move or not. If this local optimum is better than the incumbent, move there $(X \leftarrow X^{''})$, and continue the search with $\mathcal{N}_1(l \leftarrow 1)$; otherwise, set $l \leftarrow l + 1$;

◇ **End**

Note that (VNS) uses only one parameter l_{max} (except for computer time allocated, e.g., the stopping condition), which can often be disposed of, e.g., by setting it equal to the size of the vector X considered.

For all the problems considered the neighborhood structure $\mathcal{N}_l(X)$ is defined by the Hamming distance γ between solutions X and $X^{'}$ (i.e., the number of components in which these vectors differ):

$$\gamma(X, X^{'}) = l \Longleftrightarrow X^{'} \in \mathcal{N}_l(X).$$

The local search routine is two step of K-Means algorithm on the neighborhood $\mathcal{N}_l(X)$.

6 Numerical Experiments

In our experiments, we compare the performance of DCA-GKMSSC and the standard K-means (with Gaussian Kernel [21] and without kernel) on datasets taken from UCI Machine Learning Repository. The information about datasets

is summarized in Table 1. In order to evaluate the influence of the procedure of finding a good starting point of DCA, we compare two variants of DCA-GKMSSC (with and without VNS procedure):

– DCA-GKMSSC-IP1: Randomly chosen k centres in $[\alpha, \beta]^n$.
– DCA-GKMSSC-IP2: Apply some iterations of the VNS algorithm from k centres randomly chosen.

We have implemented the algorithms in the V.S C++ v6.0 environment and performed the experiments on a Intel Duo Core 3.06GHz, with 4Go of RAM. We choose the Gaussian kernel with $\sigma \in [0.015, 0.1]$.

Table 1. Datasets

Dataset	Points	Dimension	Number of classes
Pima	768	8	2
Yeast	1484	8	10
ADN	3186	60	3
Vote	435	16	2
Lympho	148	18	4
Wave	5000	40	3
Papillon	23	4	4
Iris	150	4	3

Table 2. Comparaison of three algorithms

Dataset	DCA-GKMSSC-IP1			DCA-GKMSSC-IP2			K-Means			GK K-Means		
	Cost	Time	PWPO	Cost	Time	PWPO	Cost	Time	PWPO	Cost	Time	PWPO
Pima	542.1	0.2	80.2	**521.1**	0.4	**84.3**	532.2	**0.1**	80.2	538.2	12	76.5
Yeast	45.7	0.15	36.6	**40.0**	0.20	**50.4**	46.8	**0.08**	33.3	44.8	34	39.8
ADN	**224e3**	0.5	58.2	225e3	0.7	**71.2**	225e3	**0.3**	59.9	225e3	102	58.7
Vote	1645	0.1	83.8	**1634**	0.2	**88.5**	1642	**0.03**	84.3	1645	3.4	85.3
Lympho	245	0.15	49.2	**221**	0.4	**55.4**	241	**0.1**	50.3	236	1.2	53.2
Wave	229e3	0.5	63.2	228e3	0.6	**77.6**	**227e3**	**0.2**	51.4	228e3	97	58.2
Papillon	274	3e-3	89.1	**265**	5e-3	**94.5**	272	**1e-3**	87.2	8e-2	0.12	83.2
Iris	82.1	0.01	87.3	78.9	0.03	**92.3**	**78.8**	**0.005**	88.7	86.3	7.5	65.7

We summarize, in Table 2, the average percentage (of ten executions) of well classified points (PWPO) obtained by each of four methods. We also report CPU time in seconds and the obtained overall quadratic deviation between data points and corresponding cluster centers, termed "cluster cost", say $\sum_{j=1}^{m} \min_{i=1,\ldots,k} \left\|x^i - a^j\right\|^2$. From the computational results we observe that DCA-GKMSSC-IP2 gives the best results (PWPO) on all datasets.

Conclusion. In this paper we have proposed a new algorithm for solving the Gaussian Kernel MSSC problem. The choice to use the Gaussian kernel method is to facilitates not only the formulation of the problem, but also the implicit projection of the data in a space where the representation is richer and the resolution of the problem is simpler. To solve this problem, a very simple and inexpensive DCA scheme is developed, and for initializing DCA a local search procedure VNS is used. Numerical results on several real datasets showed the robustness, the effectiveness and the superiority of the DCA-GKMSSC-IP2 with respect to the K-means algorithmS.

References

1. Arora, S., Kannan, R.: Learning mixtures of arbitrary Gaussians. In: Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, pp. 247–257 (2001)
2. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31(3), 264–323 (1999)
3. An, L.T.H., Belghiti, T., Tao, P.D.: A new efficient algorithm based on DC programming and DCA for Clustering. Journal of Global Optimization 37, 593–608 (2007)
4. An, L.T.H., Minh, L.H., Tao, P.D.: Optimization based DC programming and DCA for Hierarchical Clustering. European Journal of Operational Research 183, 1067–1085 (2007)
5. An, L.T.H., Minh, L.H., Tao, P.D.: Fuzzy clustering based on nonconvex optimisation approaches using difference of convex (DC) functions algorithms. Journal of Advances in Data Analysis and Classification (2), 1–20 (2007)
6. An, L.T.H., Tao, P.D.: Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. Journal of Global Optimization 11(3), 253–285 (1997)
7. An, L.T.H., Tao, P.D.: The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems. Annals of Operations Research 133, 23–46 (2005)
8. Tao, P.D., An, L.T.H.: Convex analysis approach to d.c. programming: Theory, Algorithms and Applications. Acta Mathematica Vietnamica, Dedicated to Professor Hoang Tuy on the Occasion of his 70th Birthday 22(1), 289–355 (1997)
9. Tao, P.D., An, L.T.H.: DC optimization algorithm for solving the trust region subproblem. SIAM J. Optimization 8, 476–505 (1998)
10. An, L.T.H., Ngai, N.V., Tao, P.D.: Exact Penalty and Error Bounds in DC Programming. Submitted to Journal of Global Optimization Dedicated to Reiner Horst (2011)
11. Bradley, B.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Proceedings of the 15th International Conferences on Machine Learning (ICML 1998), San Francisco, California, pp. 82–90 (1998)
12. Brusco, M.J.: A repetitive branch-and-bound procedure for minimum within-cluster sum of squares partitioning. Psychometrika 71, 347–363 (2006)
13. Peng, J., Xiay, Y.: A Cutting Algorithm for the Minimum Sum-of-Squared Error Clustering. In: Proceedings of the SIAM International Data Mining Conference (2005)
14. Hansen, P., Jaumard, B.: Cluster analysis and mathematical programming. Mathematical Programming 79, 191–215 (1997)
15. Sherali, H.D., Desai, J.: A global optimization RLT-based approach for solving the hard clustering problem. Journal of Global Optimization 32, 281–306 (2005)
16. Fisher, D.: Knowledge acquisition via incremental conceptual clustering. Machine Learning 2, 139–172 (1987)
17. Hartigan, J.A.: Clustering Algorithms. Wiley, New York (1975)
18. Vinod, H.D.: Integer programming and the theory of grouping. J. Amer. Stat. Assoc. 64, 506–519 (1969)
19. Mladenović, N., Hansen, P.: Variable neighborhood search. Comput. Oper. Res. 24, 1097–1100 (1997)
20. Herbrich, R.: Herbrich, Learning kernel classifiers. MIT Press (2002)
21. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recognition 41, 176–190 (2008)

DB Schema Based Ontology Construction for Efficient RDB Query

Hyun Jung Lee[1] and Mye Sohn[2,*]

[1] Graduate School of Business
Sogang University
35 Baekbeom-ro (Sinsu-dong), Mapo-gu, Seoul, Korea
hjlee5249@yahoo.com
[2] Department of Industrial Engineering
Sungkyunkwan University
300 Cheoncheon-dong, Jangan-Gu, Suwon, Korea
myesohn@skku.edu

Abstract. Relational database (RDB) is an adequate tool for rendering concepts, their attributes and relations between them that relate to a target domain. However, RDB has a certain limitation that does not fully allow it to render the semantics or meanings within concepts and their relations even though it has many advantages in representing the relations between concepts. In this paper, we propose a framework which can automatically construct an ontology from an RDB schema. We try to help understand about data structure by clearly identifying the semantic relations between data through the ontology construction. The ontology is constructed on the better understanding about data structure and acts as an assistant tool to efficiently query the data from RDB.

Keywords: Relational database (RDB), Schema, Ontology, Automated ontology construction.

1 Introduction

The methodologies for automated ontology construction are still on the floor even though ontologies are a key to implementing the semantic web and searches. This is why knowledge engineers have to engage in the entire process of ontology construction. Furthermore, it is difficult to collect and refine domain knowledge that is unstructured and scattered. As a result, an ontology is one of the bottlenecks in implementing the semantic web and searches. Lots of research is underway to remove the bottleneck. [2, 3, 5, 9] propose methods to develop a new ontology through the integration of existing ontologies, and [1, 7, 9] adopt the reverse engineering technique that reuses and analyzes the existing RDB to construct the ontology. In addition, tools for supporting the ontology construction such as protégé or swoop are developed. In addition, several tools such as protégé or swoop are developed to

* Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 341–350, 2012.

support the ontology construction [4, 6]. Even though the effort to use RDB as a knowledge source for the ontology construction is a good attempt, the fact that the ontology that is constructed using RDB schema is still far from perfect draws our attention. We are able to confirm the incompleteness of the ontology essentially on the ground that most research has focused so far not on the application of the constructed ontology but just on construction methods.

In this paper, we develop a set of rules that can map the components of RDB schema in connection with the components of the ontology. If we find the relationships between Entities such as a→b and b→c, then we create a transitive relationship named *'derivedObjectProperty'* such as a→c. It shows that the relationships between the Entities can be easily found by using the ontology. Thereby we can apply the existing relationships among Entities on RDB to the properties on the ontology. It is convenient to understand the structure of RDB, and it supports the query for appropriate data from the RDB. In addition, the ontology can identify the semantics that are implied in the relationships between Entities. Finally, the constructed ontology is applied as an assistant tool to improving the effectiveness and efficiency of the query performance on RDB.

The outline of the paper is organized as follows. Section 2 presents the procedural architecture of our framework, which consists of two modules. In Section 3, our description includes illustrative examples to help understand the methodology of ontology construction. The method to improve the query performance is described in Section 4. Finally, we present the conclusion of our work in Section 5.

2 Procedural Architecture of Our Framework

Our framework consists of two modules: Module for Ontology Generation (MOG), Module for Query using the Ontology (MQO). MOG is a module that constructs the ontology from DB schema for RDB, and MQO is a supporting module for performing the query using the ontology. The overall architecture is depicted in Fig. 1.

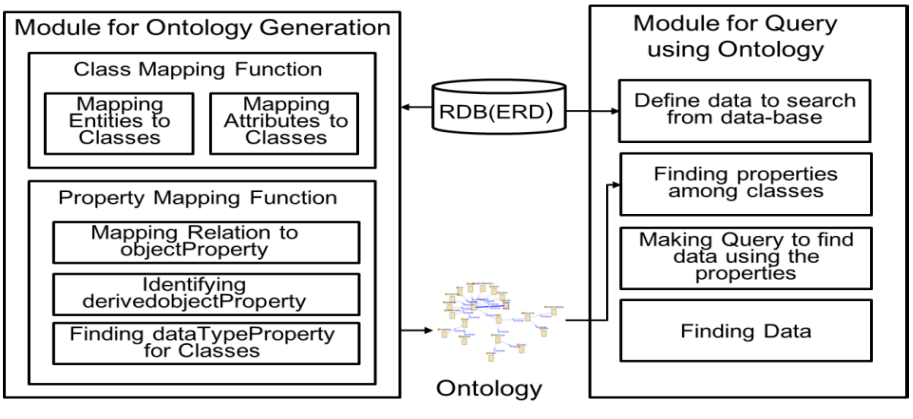


Fig. 1. Overall architecture of hybrid semantic matching using MOG and MQO

- MOG identifies classes and properties for the ontology. In order to map the RDB schema onto the ontology precisely, we develop various mapping rules like entity-to-class mapping rules, attribute-to-class mapping rules, and relation-to-property mapping rules, etc.
- MQO supports the creation of the query to retrieve data from RDB. At this time, MQO uses properties of the ontology. To generate query in RDB, our framework directly uses semantics within the ontology and its properties (not nested) instead of searching for the nested relationships among the entities.

3 Migration from RDB to Ontology with Additional Semantics

As depicted in Fig. 1, the construction process of the ontology using DB schema is composed of two phases: class identification phase, property identification phase (i.e., *subsumption relation, objectProperty, and derivedObjectProperty*). First, we describe the class identification phase. All entities and their attributes on DB schema are mapped onto classes on the ontology. At this time, record data within RDB and their identifier (primary key) are not imported into the ontology because the ontology represents the relationships between concepts only from structural perspective. Fig. 2 depicts the algorithm for mapping entities and attributes on DB schema onto classes on the ontology.

```
Procedure ClassMapping
  begin
    class ← NULL
    for each entity
        for each attribute
          class ← class ∪ attribute
        end for
     class ← class ∪ entity
    end for
  end
```

Fig. 2. Algorithm for Classes Identification from ERD

For instance, identified classes from the DB schema in Fig. 3 are summarized in Table 1. It is needless to say that we adopt the above algorithm to identify the classes.

Table 1. Illustrative example of classes mapping

ERD components	Elements to be mapped	Ontology component
Entity	Patient, Fpatient, Mpatient, Job, Disease, Diagnose, Parts Specialty, Doctor, WhiteColorJob, BlueColorJob	Classes
Attribute	Patient_ID_name, Pregnancy, Military, Job_Title, Campany_Name, Campany_Address, Professor, Researcher, plumber, Worker, Disease_name, Diagnose_name, Causal Relationship, Parts_Name, Doctor_ID_name	

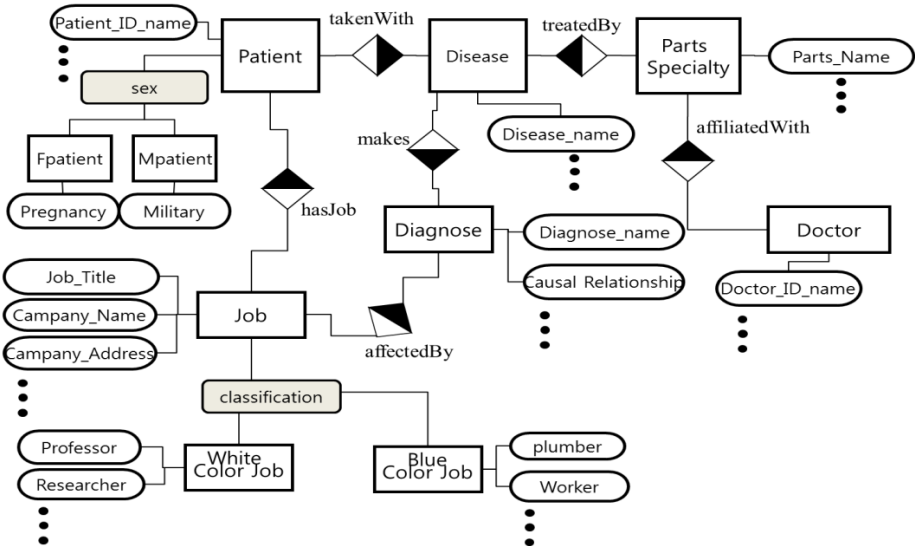


Fig. 3. Illustrative ERD

Fig. 3 depicts in entity-to-entity relations on DB schema. Entity-to-entity relations are mapped onto '*objectProperty*' using the algorithm as in Fig. 4. The entities that have many-to-many relationship between *'Patient'* and *'PartsSpecialty'* are mapped onto '*objectProperty*' after dividing the many-to-many relationships into two one-to-many relationships such as *'Patient'* and *'Disease,'* and '*Disease'* and *'PartsSpecialty.'*

```
Procedure objectPropertyMapping
    begin
      for each entity
          psi← entity
          phi←nil
          for each entity
              phi ← entity
              linked←CheckRelation&Property(psi, phi)
          end for
       endfor
    end

    CheckRelation&Property (a, b)
        begin
            if (∃ a,b | Link (a, b))
             then
                  Concatenate (a, b)
                  Add Property (a, b)
            else
                  Concatenate (a, Nil)
            endif
        end
```

Fig. 4. Algorithm for Identification of '*objectProperty*'

As a next step, a series of '*derivedObjectProperty*' is identified using the set of '*objectProperty*.' '*derivedObjectProperty*' is not stated on DB schema but derived in the process of searching for the entity-to-entity relationships. '*derivedObjectProperty*' is found as follows. If the codomian (*rdfs:range*) of an '*objectProperty*' and the domain (*rdfs:domain*) of another '*objectProperty*' are equivalent, we link the domain (*rdfs:domain*) of the former '*objectProperty*' to the codomian (*rdfs:range*) of the latter '*objectProperty*' and call it '*derivedObjectProperty*.' '*derivedObjectProperty*' can abundantly represent the entity-to-entity relationships in the RDB. As a result, queries on the RDB are conducted efficiently. Fig. 5 depicts the details of an algorithm for the identification '*derivedObjectProperty*'.

```
Procedure derivedObjectPropertyIdentifying
  begin
    for each entity
      delta ← nil
      psi ← nil
      phi ← entity

      for each entity except for phi, psi, delta
        psi←entity
        CheckDerivedObject(phi, psi, delta)
        for residue entity set is not empty
          phi←psi
          psi←entity
          CheckDerivedObject (phi, psi, delta)
        endfor
      endfor
    ensfor
  end

  CheckDerivedObject (a, b, c)
    begin
      for each entity except for phi, psi, delta
        linked ← CheckRelation&Property (delta, psi)
        derivedObjectProperty (delta) ← Concatenate(delta, linked)
      end for
    end
```

Fig. 5. Algorithm for Identification of '*derivedObjectProperty*'

Detail of the mapping processes are summarized in Fig. 6, '*dataTypeProperty*' can be simply identified by referring to a physical model of the RDB.

```
Procedure dataTypePropertyMapping
  begin
    for all entity
      ontology.rdf.datacardinality ← rdb.datacardinality
      ontology.rdf.datatype ←rdb.datatype
      …
    end for
  end
```

Fig. 6. Algorithm for Identification of '*dataTypeProperty*'

Furthermore, generalization-specialization relationships among entities on DB schema are directly mapped onto subsumption relationships (*subClassOf*) as in Table2. For instance, identified properties from entity-relation diagram using the algorithm in Fig. 4, 5, and 6 are summarized in Table 2.

Table 2. Illustrative example of property identification

ERD	Target elements	Mapped elements	Ontology
Entity-entity Relation	hasJob, takenWith, treatedBy, makes, affilitedWith, affectedBy		objectProperty
Transitive relation	Patient ←→ Job & Job ←→ Diagnose Patient ←→Disease & Disease ←→ Diagnose Disease ←→ PartSpecialty & PartSpecialty ←→ Doctor Patient ←→Disease & Disease ←→ PartsSpecialty Patient ←→Disease & Disease ←→ PartsSpecialty ←→Doctor	Patient ←→ Diagnose Disease ←→ Doctor Patient ←→ PartsSpecialty Patient ←→ Doctor	derivedObject Property
Subsumption relation	Patient ←→ Fpatient, Patient ←→ Mpatient Job ←→WhiteColor, Job←→BlueColor	Patient → Fpatient, Patient → Mpatient Job→WhiteColor Job→BlueColor	subClassOf Property
Entity-attribute relation		hasPaitentName, hasDoctorName, hasDiagnoseName, etc	userDefined Object Properties

4 Efficient Query on RDB Using Ontology

In this chapter, for showing of the efficiency the proposed framework, we compared the data search procedure using the RDB system only with the ontology and RDB-based hybrid system. As already mentioned, the framework is designed to support database engineers' efficient data search, who perform the query on the databases that do not model semantic relationships among data. So, we tried to map DB schema for RDB onto the migrated ontology using the proposed algorithm. The migrated ontology that is empowered by the semantics can be easily adopted to unfold the relationships and hided structure of DB schema.

Even though the DB schema is useful tool to express and order the huge size of data, there are some limitations such as covered data structure preventing database engineers' good understanding, difficulty of finding relationship among entities, absence of semantics, and so on. To overcome the limitation, we adopt ontology concept. Ontologies can help database engineers' understanding for the data structure

including relationship and for data semantics and efficient query to get needed data. Fig.7 shows the procedure to query on data from the RDB. In the RDB, it doesn't depict the unfolding data structure in DB Schema. So, whenever database engineers search needed data, the engineers should construct relationship using like foreign keys among the entities and find needed data structure. Even though the data structure is searched, it is difficult to find semantic relationship between entities, because it doesn't be specified. It also needs for data engineers to specify the needed data structure, which is depending on the experiences and specialty.

Procedure SearchingDataStructureOnRDB
begin
Define data to search from data-base
Identify needed data tables
Find relationships among the tables
Construct data structure using the relationship
Query data from the newly constructed data structure
end

Fig. 7. Procedure for searching data structure on RDB

Fig. 8 shows the procedure for searching data structure on ontology. The ontology maps all of relationships with semantics among classes, unlike DB Schema in RDB. Therefore, it is possible for data engineers to easily understand the data structure of usable data-base. In following procedure, the data engineer doesn't have to retrieve the data structure using RDB, because the ontology map shows all of relationships with semantics. Therefore, the engineer easily determines and finds the needed data without retrieving folding relationship difficultly.

Procedure SearchingDataStructureOnOntology
begin
Define data to search from data-base
Find properties among classes
Query data from the data-base using the properties
end

Fig. 8. Procedure for searching data structure on ontology

By Fig. 5, in the constructed ontology, the '*derivedObjectProperty*' depicts transitive relationship and contains the all of paths to construct the transitive relationship. For example, as in Table 2, if it is transitivity between a class *Disease* and a class *Doctor*, then the transitivity relationship on the proposed ontology contains all of the related paths between *Disease* and *Doctor* like *Disease*→*PartSpecialty*→*Doctor*. Therefore, data searching time is reduced when the data engineer uses the migrated ontology and the efficiency of data-base searching is improved. When we assumed that the each effort of each process is α as in Fig. 8, the effort is defined $3\alpha+2\beta$ as in Fig. 7. Finally, when

data-base engineers search data from the RDB, the engineers reduces efforts to search data structure more than 2β.

5 Performance Evaluation

To evaluate our framework, we review related research first. The related researches that construct the ontologies using the RDBs' physical or logical models are classified as mapping and transformation. The mapping the RDBs onto the ontologies is conducted when all the RDBs and the ontologies are existed [10, 11, 12]. A majority of the related research is categorized into the mapping. The transformation performs when the only RDBs exist [2]. So, series of transformation rules should be identified to construct the ontologies from the RDBs. In this context, our framework is categorized into the transformation. To demonstrate the superiority of our framework, we compare the resulting ontologies of [2] with ours that are constructed using the same RDB. The ERD that is adopted to construct the RDB is depicted in Fig. 3. Identified components of the resulting ontology are summarized in Table 3.

Table 3. Comparison of the resulting ontologies from our framework with Astrova, Korda, and Ka [2]

Ontology components	Our Framework	Astrova, Korda, and Ka [2]
class	Patient, Fpatient, Mpatient, Job, Disease, Diagnose, Parts Specialty, Doctor, WhiteColorJob, BlueColorJob Patient_ID_name, Pregnancy, Military, Job_Title, Campany_Name, Campany_Address, Professor, Researcher, plumber, Worker, Disease_name, Diagnose_name, Causal Relationship, Parts_Name, Doctor_ID_name	Patient, Fpatient, Mpatient, Job, Disease, Diagnose, Parts Specialty, Doctor, WhiteColorJob, BlueColorJob
objectProperty	hasJob, takenWith, treatedBy, makes, affilitedWith, affectedBy	Patient_ID_name, Job_Title, Campany_Name, Disease_name, Diagnose_name, Parts_Name, Doctor_ID_name
derived objectProperty (Transitive relation)	Patient ←→ Diagnose Disease ←→ Doctor Patient ←→ PartsSpecialty Patient ←→ Doctor	cannot be identified
Subsumption relationship	Patient ←→ Fpatient, Patient ←→ Mpatient Job ←→WhiteColor, Job←→BlueColor	cannot be identified

As demonstrated in Table 3, [2] transformed the entities (tables) in the RDB into classes. On the contrary, in our framework, classes are created by mapping of the entities (tables) and their attributes in the RDB. In case of *'objectProperty'* on the ontology, only primary and foreign keys are applied to, and other attributes do not take into account in create them [2]. However, considering the meaning of *'objectProperty,'* our framework, which uses the relations in the ERD to create *'objectProperty,'* contains more semantics than [2]. So, we can expect precise ontology by our framework. Furthermore, [2] cannot identify the subsumption relationships between classes because they did not consider the generalization-specialization in the RDB. But, we can identify the transitivity relationships among classes in addition to the subsumption relationships between classes. In this respect, the constructed ontology through adoption of our framework is more effective to represent of data structure and relationship considering on semantics than this of other methodology.

6 Concluding and Further Studies

In this thesis, we proposed the methodologies for automated ontology construction using RDB which contains embedded relationships among entities as well as entities, attributes, and relationships. The entities and attributes of RDB are mapped onto the classes of the ontology, and the relationships of RDB are derived to the properties of the ontology. In addition, we derived embedded relationship which is not specifically defined on RDB, which are defined as a *'derivedObjectProperty'* on the constructed ontology. For the effective data searching using query, the *'ObjectProperty,'* *'derivedObjectProperty,'* and *'userDefinedObejctProperty'* are used because the properties specifies all of relationships among classes on the ontology. Finally, it facilitates database engineers' data-base understanding using discovered relationships including semantics and contributes efficient database searching through the specified data structure. However, we proposed the conceptual framework for ontology construction using RDB and the applicability of query using the ontology using a particular case. In near future, we will process the verification of the proposed methodology and develop the automated ontology construction system to control RDB and ontology integrated.

Acknowledgments. This work is partially supported by Sogang University BK21 Development Group of Human Resources for Business Professional Services and is partially supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract UD080042AD, Korea.

References

1. Astrova, I.: Reverse Engineering of Relational Databases to Ontologies. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 327–341. Springer, Heidelberg (2004)
2. Astrova, I., Korda, N., Kalja, A.: Rule-Based Transformation of SQL Relational Databases to OWL Ontologies. In: Proceedings of the 2nd International Conference on Metadata & Semantics Research (2007)

3. Tomás, F.J., Martínez-Béjar, R.: A cooperative framework for integrating ontologies. International Journal of Human-Computer Studies 56(6), 662–717 (2002)
4. Kalyanpur, A., Parsia, B., Sirin, E., Grau, B.C.: Debugging OWL Ontologies Using Swoop (2005), http://www.mindswap.org/2005/debugging
5. Noy, N., Musen, A.: The PROMPT Suite: Interactive tools forontology merging and mapping. International Journal of Human-Computer Studies 59, 983–1024 (2003)
6. Protégé (2009), http://protege.stanford.edu
7. Trinh, Q., Barker, K., Alhajj, R.: RDB2ONT: A Tool for Generating OWL Ontologies From Relational Database Systems. In: AICT/ICIW-2006, pp. 170–178 (2006)
8. Stojanovic, L., Stojanovic, N., Volz, R.: Migrating Data-intensive Web Sites into the Semantic Web. In: Proc. of the 17th ACM Symposium on Applied Computing (SAC), pp. 1100–1107 (2002)
9. Gerd, S., Alexander, M.: FCA-MERGE: Bottom-Up Merging of Ontologies. In: Proc. IJCAI 2001, pp. 225–234 (2001)
10. Barrasa, J., Corcho, O., Shen, G., Gomez-Perez, A.: R2O: An extensible and semantically based database-to-ontology mapping language. In: SWDB 2004 (2005)
11. Laclavík, M.: RDB2Onto: Relational Database Data toOntology Individuals Mapping. In: Navrat, P., et al. (eds.) Tools for Acquisition, Organisation and Presenting of Information and Knowledge (2008)
12. Xu, Z., Zhang, S., Dong, Y.: Mapping between relational database schema and OWL ontology for deep annotation. In: WI 2006, IEEE/WIC/ACM International Conference on Web Intelligence (2006)

RF-PCA2: An Improvement on Robust Fuzzy PCA

Gyeongyong Heo[1], Kwang-Baek Kim[2], Young Woon Woo[3], and Seong Hoon Kim[4,⋆]

[1] Dept. of Electronic Engineering, Dong-Eui University, Korea
[2] Dept. of Computer Engineering, Silla University, Korea
[3] Dept. of Software Engineering, Kyungpook National University, Korea
[4] Dept. of Multimedia Engineering, Dong-Eui University, Korea

Abstract. Principal component analysis (PCA) is a well-known method for dimensionality reduction while maintaining most of the variation in data. Although PCA has been applied in many areas successfully, one of its main problems is the sensitivity to noise due to the use of sum-square-error. Several variants of PCA have been proposed to resolve the problem and, among the variants, robust fuzzy PCA (RF-PCA) demonstrated promising results, which uses fuzzy memberships to reduce noise sensitivity. However, there are also problems in RF-PCA and convergence property is one of them. RF-PCA uses two different objective functions to update memberships and principal components, which is the main reason of the lack of convergence property. The difference between two objective functions also slows convergence and deteriorates the solutions of RF-PCA. In this paper, a variant of RF-PCA, called improved robust fuzzy PCA (RF-PCA2), is proposed. RF-PCA2 uses an integrated objective function both for memberships and principal components, which guarantees RF-PCA2 to converge on a local optimum. Furthermore, RF-PCA2 converges faster than RF-PCA and the solutions are more similar to desired ones than those of RF-PCA. Experimental results with artificial data sets also support this.

Keywords: principal component analysis, noise sensitivity, fuzzy membership, convergence property.

1 Introduction

Principal component analysis (PCA) is a well-known and widely used dimensionality reduction method [1]. Although PCA has been used successfully in many applications, it is sensitive to outliers and limited to Gaussian distributions. Especially, the noise sensitivity originates from the objective of PCA, reconstruction error minimization. There have been several approaches to resolve the noise sensitivity and they can be divided roughly into two groups: subset-based methods and fuzzy methods. Subset-based methods utilize one or more subsets

⋆ Corresponding author.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 351–360, 2012.

of data to robustly estimate principal components (PCs) [2,3,4]. Although they showed successful results, they tend to suffer the small sample size problem and instability in calculating PCs.

To find robust PCs, fuzzy variants of PCA adopt fuzzy memberships and reduce the effect of outliers by assigning small membership values to outliers. Although fuzzy methods are efficient in reducing noise sensitivity, they bring about another problem, deciding membership values. Most fuzzy methods except robust fuzzy PCA (RF-PCA) consist of two steps [5], (1) estimating memberships and (2) finding PCs by building a fuzzy covariance matrix, and they put a focus on the first step [6,7]. Fuzzy methods estimate memberships by formulating objective functions and optimizing them, the basic limitation, however, lies in the formulation of objective functions which are based on the first PC. Although the first PC retains the largest portion of data variance, it can be easily affected by noise. Another problem is that the quantity optimized, memberships from fuzzy clustering for example, does not have a direct relationship with PCA.

RF-PCA, also belongs to the second group, extended the previous methods by using $k(k \geq 1)$ PCs simultaneously and minimizing the sum of reconstruction errors in the estimation of memberships. PCA also minimizes the sum of reconstruction errors, therefore, it is natural to use reconstruction errors in the estimation of memberships. By iteratively optimizing memberships and PCs, RF-PCA demonstrated better result than other methods did. However, two different objective functions for memberships and PCs result in the lack of convergence property. Although RF-PCA showed convergence empirically, only the convergence of membership values was shown. The difference between the two objective functions also slows the convergence and deteriorates the solutions of RF-PCA.

In this paper, a variant of RF-PCA, called improved robust fuzzy PCA (RF-PCA2) is proposed. RF-PCA2 uses an integrated objective function both for memberships and PCs, which enables it to converge on a local optimum. Furthermore, RF-PCA2 finds faster and better solutions than RF-PCA does.

In the next section, RF-PCA is briefly reviewed and RF-PCA2 is formulated in Section 3. Experimental results are given in Section 4 followed by a discussion.

2 Robust Fuzzy PCA (RF-PCA)

Let $X = \{x_1, \cdots, x_N\}$ be a data set and N is the size of it. RF-PCA finds k robust PCs with iterative calculation of PCs and memberships. Given memberships, RF-PCA calculates k robust PCs corresponding to k eigenvectors with k largest eigenvalues of a weighted covariance matrix defined as [5]

$$C_{RF-PCA} = \frac{1}{N} \sum_{i=1}^{N} u_i^2 (x_i - \mu_{RF-PCA})(x_i - \mu_{RF-PCA})^T. \tag{1}$$

The vector μ_{RF-PCA} is a weighted mean defined as

$$\mu_{RF-PCA} = \frac{\sum_{i=1}^{N} u_i x_i}{\sum_{i=1}^{N} u_i}. \tag{2}$$

1: Initialize a membership vector $U_0 = [u_1, \ldots, u_N]^T = [1, \ldots, 1]^T$, a counter $t = 0$, and an orthogonal basis matrix W_0 using PCA.
2: **repeat**
3: $t \leftarrow t + 1$.
4: Calculate a weighted mean vector (μ_{RF-PCA}) and a weighted covariance matrix (C_{RF-PCA}).
5: Build an orthogonal basis matrix W_t using the eigenvectors of C_{RF-PCA}.
6: Calculate the memberships U_t using Eq. (5).
7: **until** $\max |U_{t-1} - U_t| < \varepsilon$ or $t > t_{max}$
8: **return**

Fig. 1. RF-PCA algorithm

Given robust PCs, memberships were updated by optimizing

$$J_1 = \sum_{i=1}^{N} u_i e(x_i) + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i), \tag{3}$$

where σ is a regularization parameter and $e(x_i)$ is a reconstruction error defined as [8]

$$e(x_i) = ||(x_i - \mu_{RF-PCA}) - WW^T(x_i - \mu_{RF-PCA})||^2. \tag{4}$$

In Eq. (4), W is a matrix having k eigenvectors of C_{RF-PCA} as columns. The second term of Eq. (3) is a regularization term, which also consists of two terms. The first one is negative entropy which is widely used to make fuzzy clustering algorithms noise-robust [9] and the second term is added to avoid a trivial solution. By taking a partial derivative of Eq. (3) with respect to u_i, one can obtain the update equation of memberships:

$$u_i = \exp\left(-\frac{e(x_i)}{\sigma^2}\right). \tag{5}$$

By iteratively updating PCs and memberships using Eqs. (1) and (5), respectively, RF-PCA can find k PCs robust to noise. The RF-PCA algorithm is summarized in Fig. 1, where ε is a pre-defined constant, t_{max} is the maximum number of iterations, and $\max |U_{t-1} - U_t| = \max_{i=1,\ldots,N} |u_i^{t-1} - u_i^t|$ represents the maximum of element-wise differences between two vectors.

Although RF-PCA outperformed the previous methods, it has a problem in formulation. It can be shown that, given μ_{RF-PCA} and memberships, the eigenvectors of Eq. (1) optimizes

$$\begin{aligned} J_2 &= \sum_{i=1}^{N} u_i^2 ||(x_i - \mu_{RF-PCA}) - WW^T(x_i - \mu_{RF-PCA})||^2, \\ &= \sum_{i=1}^{N} u_i^2 e(x_i), \end{aligned} \tag{6}$$

which is different from Eq. (3). Although the difference is u_i and u_i^2, it results in the violation of the formal definitions of weighted mean and weighted covariance, which can be written as [10]

$$\mu = \sum_{i=1}^{N} p(x_i) x_i, \tag{7}$$

$$C = \sum_{i=1}^{N} p(x_i)(x_i - \mu)(x_i - \mu)^T. \tag{8}$$

However, in RF-PCA, $p(x_i)$ is not defined in a consistent way, $u_i \neq u_i^2$. It also can be shown that μ_{RF-PCA} optimizes Eq. (3) not Eq. (6). This problem slows the convergence of RF-PCA and deteriorates the quality of solutions.

3 Improved Robust Fuzzy PCA (RF-PCA2)

In this section, the objective function of improved robust fuzzy PCA is defined by unifying Eqs. (3) and (6) and the update equations for μ_R, C_R, and U are derived. The objective function of RF-PCA2 can be written as

$$\begin{aligned} \arg\min_{W} \quad J &= \sum_{i=1}^{N} u_i ||(x_i - \mu_R) - WW^T(x_i - \mu_R)||^2 \\ &\quad + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i), \\ &= \sum_{i=1}^{N} u_i e(x_i) + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i), \\ \text{s.t.} \quad & W^T W = I, \end{aligned} \tag{9}$$

where I is an identity matrix. RF-PCA2 tries to find k orthonormal basis vectors minimizing a weighted sum of reconstruction errors with regularization on memberships. The optimal orthogonal basis vectors minimizing Eq. (9) correspond to PCs robust to noise. The objective function consists of two terms. The first term measures the sum of membership-weighted reconstruction errors and the second one is a regularization term to make the estimated PCs noise-robust.

Alternating optimization can be used to optimize the objective function, in which $U = [u_1, u_2, \cdots, u_N]^T$ and W are optimized iteratively. By taking a partial derivative of Eq. (9) with respect to u_i, one can obtain the update equation for memberships,

$$u_i = \exp\left(-\frac{e(x_i)}{\sigma^2}\right), \tag{10}$$

which is equal to Eq. (5). Similarly, by taking a partial derivative with respect to μ_R, one can obtain the update equation for mean,

$$\mu_R = \frac{\sum_{i=1}^{N} u_i x_i}{\sum_{i=1}^{N} u_i}. \tag{11}$$

1: Initialize a membership vector $U_0 = [u_1, \ldots, u_N]^T = [1, \ldots, 1]^T$, a counter $t = 0$, and an orthogonal basis matrix W_0 using PCA.
2: **repeat**
3: $t \leftarrow t + 1$.
4: Calculate a weighted mean vector μ_R and a weighted covariance matrix C_R.
5: Build an orthogonal basis matrix W_t using the eigenvectors of C_R.
6: Calculate the memberships U_t using Eq. (10).
7: **until** $J^{t-1} - J^t < \varepsilon$ or $t > t_{max}$
8: **return**

Fig. 2. RF-PCA2 algorithm

Using the mean obtained, data can be translated to have zero mean, i.e., $\sum_{i=1}^{N} u_i x_i' = 0$, where $x_i' = x_i - \mu_R$. Then finding an optimal W minimizing Eq. (9) is equivalent to finding W minimizing

$$J' = \sum_{i=1}^{N} u_i ||x_i' - WW^T x_i'||^2. \tag{12}$$

The matrix W minimizing Eq. (12) under the constraint $W^T W = I$ can be obtained by finding k eigenvectors having k largest eigenvalues of a weighted covariance matrix, which is defined as

$$C_R = \frac{1}{N} \sum_{i=1}^{N} u_i x_i' x_i'^T. \tag{13}$$

The RF-PCA2 algorithm is summarized in Fig. 2, which looks similar to Fig. 1 except two things. First, the definition of a weighted covariance matrix is different. Furthermore, Eqs. (11) and (13) are derived from the objective function in Eq. (9) and follow the formal definitions in Eqs. (7) and (8).

Another difference is a stopping criterion. RF-PCA uses the maximum difference of membership values. Figure 3 shows examples of maximum membership differences as a function of iteration number. As shown in Fig. 3, the maximum membership difference does not decrease monotonically, although it converges to zero eventually. Furthermore, the difference converges to zero faster in RF-PCA2 than in RF-PCA on average.

RF-PCA2 uses the difference of objective function values as a stopping criterion. Figure 4 shows examples of the objective function value in Eq. (9) after each update step (step 5 and 6 in Figs. 1 and 2) in RF-PCA and RF-PCA2. As each step optimizes different objective function in RF-PCA, sometimes the objective function value in Eq. (9) increases and the zigzag pattern clearly shows the problem. RF-PCA2 uses an integrated objective function and the objective function value decreases monotonically as shown in Fig. 4(b).

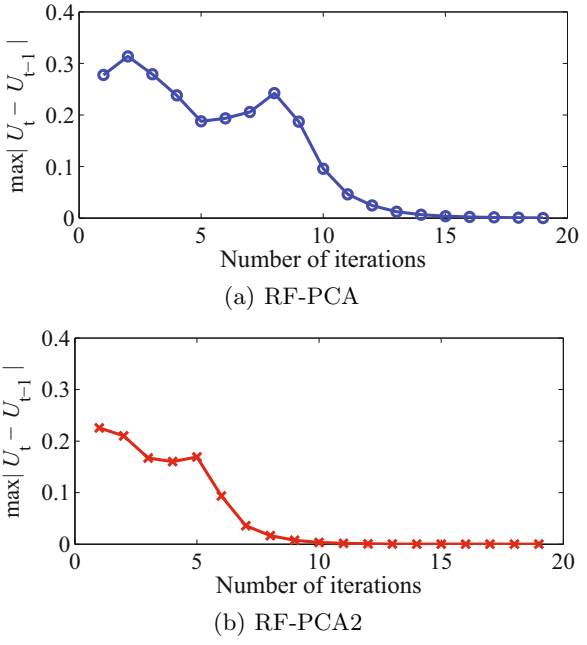


Fig. 3. Maximum difference of membership values in (a) RF-PCA and (b) RF-PCA2

Table 1. Mean number of iterations

Criterion	RF-PCA	RF-PCA2	
1	5.276	4.725	$\max \lvert U^{t-1} - U^t \rvert < \varepsilon$
2	7.737	6.811	$J^{t-1} - J^t < \varepsilon$

4 Experimental Results

To investigate the effectiveness of the proposed method, RF-PCA and RF-PCA2 were implemented and tested using Matlab. First, the properties of RF-PCA and RF-PCA2 were compared using an artificial data set with $k = 1$. Figure 5 shows the data set and the first PCs found by PCA and RF-PCA2. RF-PCA found almost the same first PC with the PC found by RF-PCA2, which is not shown in Fig. 5. The data consist of 110 randomly generated points, 100 points from a Gaussian distribution and 10 noise points from another Gaussian distribution. As is clear from Fig. 5, PCA found a skewed PC due to noise, but RF-PCA2 found an unskewed one because the noise points have small membership values and are negligible in the calculation of the first PC.

The first set of experiments were to compare convergence speed. Figures 6 and 7 show the histograms of the number of iterations for RF-PCA and RF-PCA2 to converge on a local optimum over 1,000 runs and Table 1 summarizes the average numbers of iterations. Two stopping criterions were used: one is the criterion used in RF-PCA, the maximum difference of membership values, and

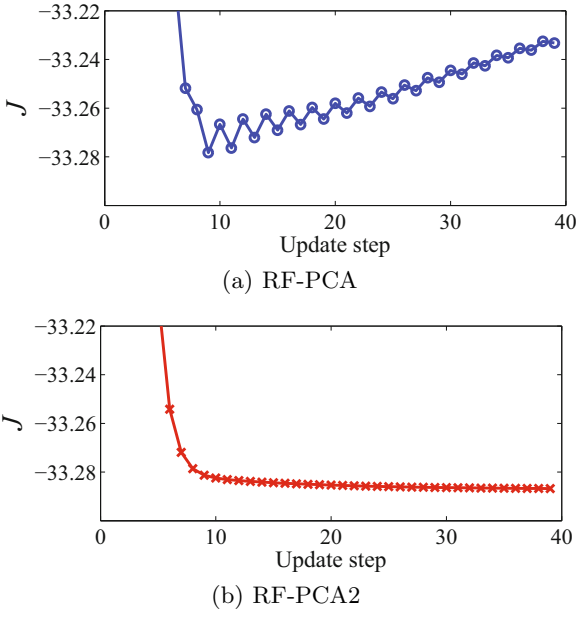


Fig. 4. Objective function values after each update step in (a) RF-PCA and (b) RF-PCA2

the other is the one proposed in this paper, the difference of objective function values. The threshold ε and the maximum number of iterations t_{max} were set $\varepsilon = 0.0001$ and $t_{max} = 20$, respectively. As is clear from these experiments, RF-PCA2 converges faster than RF-PCA under the two criterions.

The next experiments were to compare the quality of solutions. Let w_0 be the first PC found by PCA using a clean data set that is equal to the data used in the previous experiments without noise points. The vectors w_1 and w_2 are the first PCs found by RF-PCA and RF-PCA2 using the noisy data set. Figure 8 represents the histograms of angles between pairs of vectors – w_0 vs. w_1 and w_0 vs. w_2 – over 1,000 runs. The mean angle between w_0 and w_1 is 3.623^o and that between w_0 and w_2 is 2.058^o, which means that PCs found by RF-PCA2 are more similar to the desired PCs than those of RF-PCA and that RF-PCA2 is more noise resistant than RF-PCA.

As is clear from the previous experiments, RF-PCA2 converges faster than RF-PCA. Furthermore the eigenvectors found by RF-PCA2 are more noise-resistant than those by RF-PCA. As shown in Section 3, the basic limitation of RF-PCA is the different objective function in each update step. RF-PCA2, however, uses an integrated objective function, which results in the performance improvement.

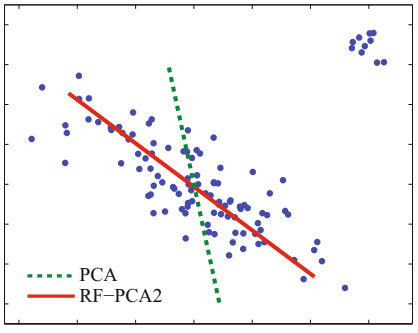
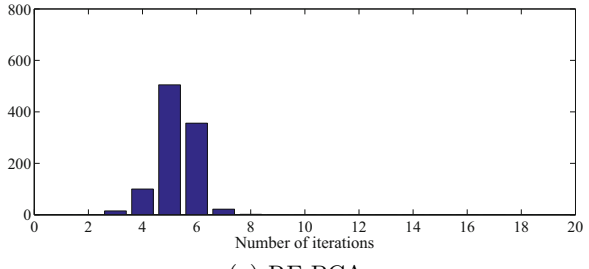
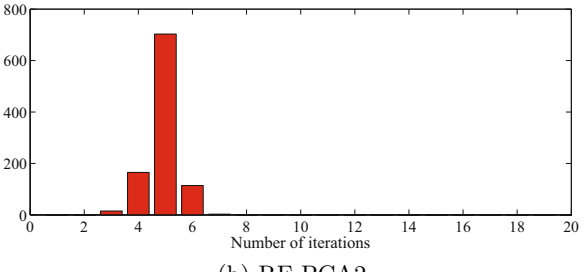


Fig. 5. Principal components found by PCA and RF-PCA2

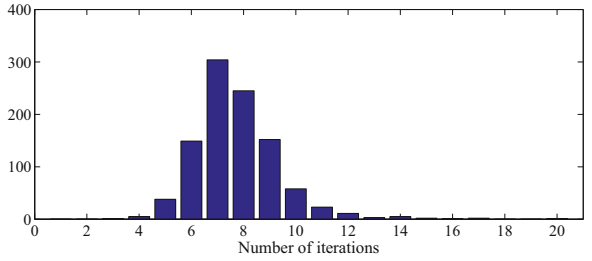


(a) RF-PCA

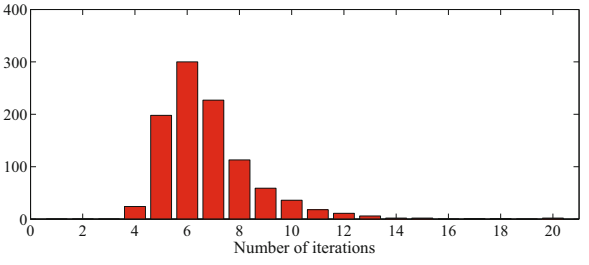


(b) RF-PCA2

Fig. 6. Number of iterations using the criterion of $\max |U^{t-1} - U^t| < \varepsilon$ in (a) RF-PCA and (b) RF-PCA2

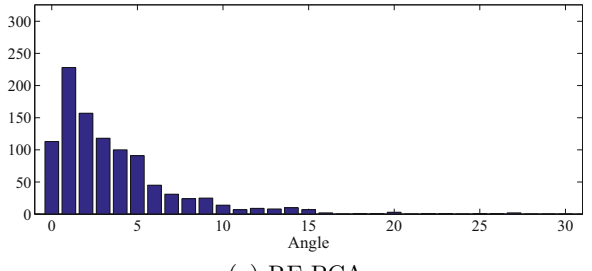


(a) RF-PCA

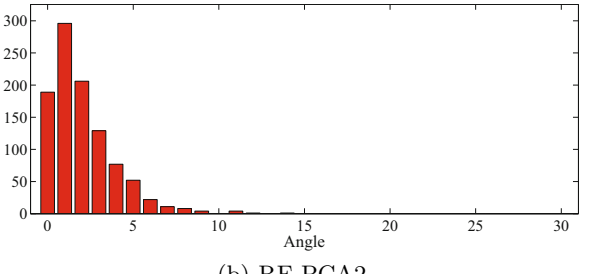


(b) RF-PCA2

Fig. 7. Number of iterations using the criterion of $J^{t-1} - J^t < \varepsilon$ in (a) RF-PCA and (b) RF-PCA2



(a) RF-PCA



(b) RF-PCA2

Fig. 8. The angles between two principal components using (a) RF-PCA and (b) RF-PCA2

5 Discussion

A variant of noise-robust fuzzy principal component analysis is introduced. Among the various variants of PCA, robust fuzzy PCA (RF-PCA) demonstrated the best result. However, RF-PCA cannot be guaranteed to converge on a local optimum due to the use of two different objective functions. The proposed algorithm, improved robust fuzzy PCA (RF-PCA2), is based on an integrated objective function, which guarantees the convergence of the algorithm. Experimental results with artificial data sets show the superiority of the proposed method to previous methods.

Although RF-PCA2 is better than previous methods, it still can be improved further. First of all, RF-PCA2 cannot accommodate non-Gaussian distribution due to the Gaussian assumption on PCA. This problem can be relaxed by the introduction of kernel methods. Computational complexity is another problem in RF-PCA2. To reduce the complexity, a different measure from the reconstruction error together with an incremental modification of RF-PCA2 is under investigation.

Acknowledgement. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (Ministry of Education, Science and Technology) [NRF-2010-355-D00052].

References

1. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, Heidelberg (2002)
2. Rousseeuw, P.: Multivariate estimation with high breakdown point. Mathematical Statistics and Applications B, 283–297 (1985)
3. Lu, C.-D., Zhang, T.-Y., Du, X.-Z., Li, C.-P.: A robust kernel PCA algorithm. In: Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 3084–3087 (2004)
4. Lu, C., Zhang, T., Zhang, R., Zhang, C.: Adaptive robust kernel PCA algorithm. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. VI 621–VI 624 (2003)
5. Heo, G., Gader, P., Frigui, H.: RKF-PCA: Robust kernel fuzzy PCA. Neural Networks 22(5-6), 642–650 (2009)
6. Yang, T.-N., Wang, S.-D.: Fuzzy auto-associative neural networks for principal component extraction of noisy data. IEEE Transaction on Neural Networks 11(3), 808–810 (2000)
7. Cundari, T.R., Sarbu, C., Pop, H.F.: Robust fuzzy principal component analysis (FPCA). A comparative study concerning interaction of carbon-hydrogen bonds with molybdenum-oxo bonds. Journal of Chemical Information and Computer Sciences 42(6), 1363–1369 (2002)
8. Heo, G., Gader, P.: Fuzzy SVM for noisy data: A robust membership calculation method. In: Proceedings of the 2009 IEEE International Confernce on Fuzzy Systems, pp. 431–436 (2009)
9. Ichihashi, H., Honda, K., Tani, N.: Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization. In: Proceedings of the 4th Asian Fuzzy Systems Symposium, pp. 217–221 (2000)
10. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press (1990)

Structures of Association Rule Set

Anh Tran[1], Tin Truong[1], and Bac Le[2]

[1] University of Dalat, Dalat, Vietnam
{anhtn,tintc}@dlu.edu.vn
[2] University of Natural Science Ho Chi Minh, Ho Chi Minh, Vietnam
lhbac@fit.hcmuns.edu.vn

Abstract. This paper shows a mathematical foundation for almost important features in the problem of discovering knowledge by association rules. The class of frequent itemsets and the association rule set are partitioned into disjoint classes by two equivalence relations based on closures. Thanks to these partitions, efficient parallel algorithms for mining frequent itemsets and association rules can be obtained. Practically, one can mine frequent itemsets as well as association rules just in the classes that users take care of. Then, we obtain structures of each rule class using corresponding order relations. For a given relation, each rule class splits into two subsets of basic and consequence. The basic one contains minimal rules and the consequence one includes in the rules that can be deducted from those minimal rules. In the rest, we consider association rule mining based on order relation min. The explicit form of minimal rules according to that relation is shown. Due to unique representations of frequent itemsets through their generators and corresponding eliminable itemsets, operators for deducting all remaining rules are also suggested. Experimental results show that mining association rules based on relation min is better than the ones based on relations of minmin and minMax in terms of reduction in mining times as well as number of basic rules.

Keywords: association rule, basic rule, consequence rule, generator, eliminable itemset, equivalence relation, order relation.

1 Introduction

Firstly introduced and researched by Agrawal et al. [1], association rule mining is one of the important problems in data mining. Traditional approach solves this problem in two phases of (1) to extract frequent itemsets whose the occurrences exceed minimum support in the data, and (2) to generate association rules from them with the given minimum confidence. The cardinalities of frequent itemset class $\mathcal{FS}$ and association rule set $\mathcal{ARS}$ can also grow unwieldy. The traditional algorithms (such as AIS [1], Apriori [2], Ap-genrules [2]) for finding those sets generate many candidates and check many sufficient conditions. Then, the running time and the memory capacity are usually enormous. Moreover, it is too complicated for users to understand and manage the results whose sizes are so big. Recently, some researchers have proposed a new framework for mining association rules. Firstly, frequent closed itemsets are mined by the efficient algorithms such as Charm-L [12], Closet [8]. The number of those itemsets is usually less than the one of all frequent itemsets. Based on them, a small set

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 361–370, 2012.

of useful association rules (basic rules) is obtained. Zaki [11] considered the most-general rules and indicated how to find them. However, his method generates many candidates. Furthermore, there are no algorithms for finding the remaining rules (see [10]). Pasquier et al. [7] proposed the algorithms for mining the minimal rules and deriving the other ones from them. However, those algorithms miss some rules and waste much time to generate rules as well as to delete the repeated ones (see [9]).

In order to overcome these disadvantages, understanding structures of frequent itemsets and association rule set is essential. Based on frequent closed itemset lattice, we use two appropriate equivalence relations on $\mathcal{FS}$ and $\mathcal{ARS}$ to partition them into disjoint classes $\mathcal{FS}$(L) and $\mathcal{AR}$(L, S), where (L, S) is a pair of two nested non-empty frequent closed itemsets. All itemsets in each itemset class have the same closure so the same support. All rules in each rule class have the same support and confidence. Without loss of the generality, we only need to investigate each class independently. For each itemset class $\mathcal{FS}$(L), we show that the generators [3] are even minimal itemsets according to an appropriate order relation. In order to generate all remaining itemsets of the same class, we only need to add eliminable itemsets [9] to generators of L. To avoid the duplications, a simple sufficient condition related to generators and eliminable itemsets is checked. For each pair (L, S), based on different order relations over each class $\mathcal{AR}$(L, S), we show different structures of association rule set. According to given order relation, this set splits into subsets of basic and consequence. All rules in basic one are minimal. The consequence one contains non-minimal rules. Those rules are sufficiently deduced by adding, deleting or moving appropriate eliminable itemsets in both sides of the basic ones. A sufficient condition is verified to avoid the duplications. Experiments will figure out that mining association rules based on order relation min is better than the ones based on order relations of minMax and minmin in terms of reductions in number of basic rules as well as in mining times.

The paper is organized as follows. Section 2 reminds some elementary concepts of concept lattice, frequent itemsets, association rules and results of partitioning itemset class, generators and eliminable itemsets. It also shows unique representations of itemsets with the same closure based on equivalence relation on itemsets, their generators and eliminable itemsets. Section 3 partitions association rule set by a different equivalence relation and figures out its structures using different order relations over each equivalence rule class. Sections 4 and 5 show the experimental results and conclusion.

2 Preliminaries

2.1 Concept Lattice, Frequent Itemset and Association Rule

Given non-empty sets $\mathcal{O}$ containing objects (or transactions), $\mathcal{A}$ containing attributes (or items) related to objects $o \in \mathcal{O}$. Let $\mathcal{R}$ be a binary relation on $\mathcal{O}x\mathcal{A}$. Consider two set functions: $\lambda: 2^{\mathcal{O}} \rightarrow 2^{\mathcal{A}}$, $\rho: 2^{\mathcal{A}} \rightarrow 2^{\mathcal{O}}$ defined as follows: $\forall A \subseteq \mathcal{A}$, $O \subseteq \mathcal{O}$: $\lambda(O)=\{a \in \mathcal{A} \mid (o, a) \in \mathcal{R}, \forall o \in O\}$, $\rho(A)=\{o \in \mathcal{O} \mid (o, a) \in R, \forall a \in A\}$, where $2^{\mathcal{O}}$ and $2^{\mathcal{A}}$ are the classes of all subsets of $\mathcal{O}$ and $\mathcal{A}$. Assign that, $\lambda(\varnothing)=\mathcal{A}$, $\rho(\varnothing)=\mathcal{O}$, $h=\lambda_o\rho$, $h'=\rho_o\lambda$. Sets of $h'(O)$ and $h(A)$ are in turn called the closures of O and A. Itemsets A and O are closed iff $h(A)=A$ [3] and $h'(O)=O$. If $A=\lambda(O)$ and $O=\rho(A)$, the pair $C=(O, A) \in \mathcal{O}x\mathcal{A}$ is

called a concept. In the class of concepts $C = \{C \in \mathcal{O}x\mathcal{A}\}$, if defining order relation $\prec$ as relation $\supseteq$ between subsets of $\mathcal{O}$, then $L \equiv (C, \prec)$ is a concept lattice [3]. On $\mathcal{O}$, consider σ-field [5] $\mathcal{F}_{max}=2^{\mathcal{O}}$ including all subsets of $\mathcal{O}$. Let $\mathcal{P}$ be a countable probability measure: $\mathcal{P}(O)=|O|/|\mathcal{O}|$, $\forall O \subseteq \mathcal{O}$. We have probability space $(\mathcal{O}, \mathcal{F}_{max}, \mathcal{P})$. A set of items containing at least one transaction is called an *itemset*. The class of all itemsets according to $\mathcal{R}$ is denoted as $\mathcal{IC}$, formerly $\mathcal{IC} := \{A \subseteq \mathcal{A} \mid \exists o \in \mathcal{O}: (o, a) \in \mathcal{R}\}$.

Let s_0 and c_0 be minimum support and minimum confidence. For any itemset $S \in \mathcal{IC}$, the probability $\mathcal{P}(\rho(S)) = |\rho(S)| / |\mathcal{O}|$ is called the support of S, denoted by supp(S). An itemset S is frequent iff $supp(S) \geq s_0$ [1]. Let $\mathcal{CS}$, $\mathcal{FS}$ and $\mathcal{FCS} = \mathcal{CS} \cap \mathcal{FS}$ be respectively the classes of all closed itemsets, all frequent itemsets and all frequent closed itemsets. For every non-empty, strict subset L from S ($\emptyset \neq L \subset S$), $S \in \mathcal{FS}$ and R=S\L, denote r: L→R as the rule created by L, R (or L, S). The conditional probability of ρ(R) given ρ(L): $c(r) \equiv \mathcal{P}[\rho(S)|\rho(L)] = \mathcal{P}[\rho(L) \cap \rho(R)]/\mathcal{P}(\rho(L)) = |\rho(S)|/|\rho(L)|$ is called the confidence of r. The rule r is called an association rule iff $c(r) \geq c_0$ [1]. Let $\mathcal{ARS}$ be the set of all association rules corresponding with s_0 and c_0. For two non-empty itemsets G, A: $\emptyset \neq G \subseteq A \subseteq \mathcal{A}$, G is called a generator [7] of A iff $h(G)=h(A)$ and $(\forall G': \emptyset \neq G' \subset G \Rightarrow h(G') \subset h(G))$. Let $\mathcal{G}(A)$ be the class of all generators of A numbered by 1, 2, …: $\mathcal{G}(A) = \{A_i, i \in I = \{1, 2, \ldots, n_A\}, n_A \leq |A|\}$.

2.2 Partition of the Class of All Itemsets, Generators and Eliminable Itemsets

An equivalence relation based on the closures of itemsets partitions $\mathcal{FS}$ into disjoint equivalence classes. Using an appropriate order relation, we figure out that generators are minimal elements (in term of subset relation) over each class.

Definition 1 [9]**.** Closed mapping $h: 2^{\mathcal{A}} \to 2^{\mathcal{A}}$ generates a binary relation $\sim_h$ in class $2^{\mathcal{A}}$: $\forall A, B \subseteq \mathcal{A}$: $A \sim_h B$ *iff* $h(A) = h(B)$.

Theorem 1 [9] (*Partition of itemset class*)**.** $\sim_{\mathcal{A}}$ is an equivalence relation. It partitions $\mathcal{IC}$ into disjoint equivalence classes. All itemsets in each class have the same closure so the same support. We have the following result, where: [A] denotes the equivalence class containing A and "+" denotes union operator of two disjoint sets.

$$\mathcal{IC} = \sum_{A \in \mathcal{CS}} [A] = \sum_{A \in \mathcal{CS}} \{\emptyset \neq X \subseteq A \mid h(X) = A\} \quad or \quad \mathcal{FS} = \sum_{A \in \mathcal{FCS}} [A]$$

Definition 2. Consider order relation $\prec_{\mathcal{A}}$ (over the set $2^{\mathcal{A}} \setminus \{\emptyset\}$) defined as follows: $\forall A, B: \emptyset \neq A, B \subseteq \mathcal{A}$: $A \prec_{\mathcal{A}} B$ *iff* $(A \subseteq B$ *and* $h(A) = h(B))$.

Proposition 1. $\prec_{\mathcal{A}}$ *is a partial order relation. A minimal element G of* equivalence class [A] *is a generator of A* and $\mathcal{G}(A) \neq \emptyset$.

Definition 3 (*Eliminable itemsets*) [9]**.** In $2^{\mathcal{A}}$, a subset *R is called eliminable itemset in S iff* $R \subset S$ *and* $\rho(S) = \rho(S \setminus R)$. Denote the class of all eliminable itemsets in S by $\mathcal{N}(S)$ and assign that $\mathcal{N}^*(S) := \mathcal{N}(S) \setminus \{\emptyset\}$.

Proposition 2 (*Recognizing an eliminable itemset*)**.** $\forall R \subset S$, we have:

a) $R \in \mathcal{M}(S) \Leftrightarrow \rho(S \backslash R) \subseteq \rho(R) \Leftrightarrow h(S) = h(S \backslash R) \Leftrightarrow supp(S) = supp(S \backslash R) \Leftrightarrow \mathcal{P}(\rho(S \backslash R) \backslash \rho(S)) = 0$.

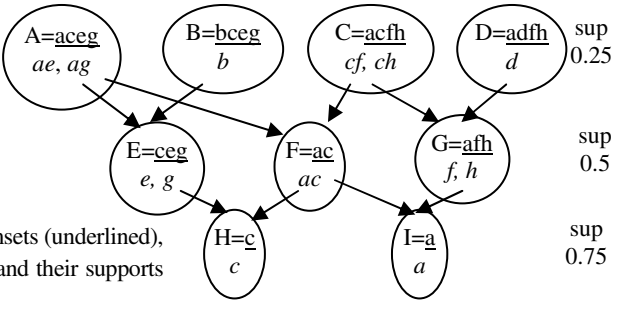
b) $\mathcal{M}(S) = \{A: A \subseteq S \backslash G_0, G_0 \in \mathcal{G}(S)\}$ [9].

c) $\mathcal{M}(S) = \bigcup_{G_0 \in G(S)} \mathcal{M}(S, G_0)$, *where* $\mathcal{M}(S, G_0) = \{A: A \subseteq S \backslash G_0\}$.

Example 1. Figure 1 contains database T, the lattice of closed itemsets on T and the corresponding generators. This figure is used in examples in the rest of the paper. $\mathcal{A}$ = {a, b, c, d, e, f, g, h} is partitioned into nine disjointed equivalence classes: [A], [B] ... [I]. With X=aceg (i.e. {a, c, e, g}) we have: $\mathcal{G}$(X) = {ae, ag}, [X] = {ae, ag, ace, acg, aeg, aceg} (the supports of all itemsets in [X] are equal to 0.25) and $\mathcal{N}$*(X, ae) = {cg, c, g}, $\mathcal{N}$*(X, ag)={ce, c, e}, $\mathcal{N}$*(X) = $\mathcal{N}$*(X, ae) ∪ $\mathcal{N}$*(X, ag) = {cg, c, g, e, ce}.

Trans	Items
1	a, c, e, g
2	a, c, f, h
3	a, d, f, h
4	b, c, e, g

(a) Database T



(b) The lattice of closed itemsets (underlined), their generators (italicized) and their supports (outside numbers)

Fig. 1. Database T, the lattice of closed itemsets of T and their generators

2.3 Unique Representation of Itemsets with the Same Closure

We show unique representations of itemsets of the same closure with and without constraints that play an important role in deriving non-repeatedly consequence rules.

Proposition 3. For every itemset X: $\emptyset \neq X \subseteq \mathcal{A}$, denote *equivalence class restricted on X* as follows: $\lfloor X \rfloor = \{X' \subseteq X \mid X' \neq \emptyset, h(X') = h(X)\}$ or $\lfloor X \rfloor = \{X' \subseteq X \mid X' \in [X] \backslash \{\emptyset\}\}$. The following statements hold:

a) $\lfloor X \rfloor \subseteq [X]$; $\forall X' \in \lfloor X \rfloor, \mathcal{G}(X') \subseteq \mathcal{G}(X)$; and $X \in \mathcal{CS} \Leftrightarrow \lfloor X \rfloor = [X]$.

b) *Closure-preserved property of eliminable itemset*:
$\forall X' = X_0 + Y \in \lfloor X \rfloor$, where $X_0 \in \mathcal{G}(X)$, $Y \subseteq X \backslash X_0$, then: $\forall X'' \subseteq X' \backslash X_0, X' \backslash X'' \in \lfloor X \rfloor$.
$\forall X' \in \lfloor X \rfloor, X'' \subseteq X \backslash X' \Rightarrow X' + X'' \in \lfloor X \rfloor$.

c) *Representation of itemsets*: $X' \in \lfloor X \rfloor \Leftrightarrow \exists X_i \in \mathcal{G}(X), X'' \in \mathcal{M}(X, X_i)$: $X' = X_i + X''$.

Itemset X could have different generators, so itemsets generated in $\lfloor X \rfloor$ by proposition 3.c could be repeated. Theorem 2 shows unique representation of them.

Theorem 2 (*Unique representation of itemsets*). $\forall \emptyset \neq X', X \subseteq \mathcal{A}$, let $X_U = \bigcup_{X_i \in \mathcal{G}(X)} X_i$, $X_{U,i} = X_U \backslash X_i$, $X\_ = X \backslash X_U$, $\mathcal{IS}(X) = \{X' = X_i + X'_i + \tilde{X} \mid X_i \in \mathcal{G}(X), \tilde{X} \subseteq X\_, X'_i \subseteq X_{U,i}, i=1$ or $(i>1: X_k \not\subset X_i + X'_i, \forall k: 1 \le k < i)\}$ and denote $\mathcal{FS}(X) = \mathcal{IS}(X)$, *if* $supp(X) \ge s_0$:

a) *All itemsets of* $\mathcal{IS}(X)$ *are non-repeatedly generated.* **b)** $\lfloor X \rfloor = \mathcal{IS}(X)$.

Proof: (a) Assume that there exist i, k such that $i>k\ge 1$ and $X_i + X'_i + \tilde{X}_i \equiv X_k + X'_k + \tilde{X}_k$, where $X_i, X_k \in \mathcal{G}(X)$, $\tilde{X}_i, \tilde{X}_k \subseteq X\_$, $X'_i \subseteq X_{U,i}$, $X'_k \subseteq X_{U,k}$. Since $X_k \cap \tilde{X}_i = \emptyset$, $X_k \subset X_i + X'_i$ (the equality does not occur because $X_i, X_k \in \mathcal{G}(X)$). It contradicts to the selection of the index i! Thus, all itemsets of $\mathcal{IS}(X)$ are non-repeatedly generated.
(b) **"$\subseteq$":** If $X' \in \lfloor X \rfloor$, by proposition 3.c, assume that i is the minimum index such that $X_i \in \mathcal{G}(X)$, $X''_i \subseteq X \backslash X_i$ and $X' = X_i + X''_i$. Let $X'_i = X''_i \cap X_U$, $\tilde{X} = X''_i \backslash X_U = X' \backslash X_U$, then $X'_i \subseteq X_{U,i}$, $\tilde{X} \subseteq X\_$ and $X' = X_i + X'_i + \tilde{X}$. Assume that there exists the index k such that $1 \le k < i$, $X_k \in \mathcal{G}(X)$, $X_k \subset X_i + X'_i$. Then $X' = X_k + X''_k$, where $X''_k = X'_k + \tilde{X}$ and $X'_k = (X_i + X'_i) \backslash X_k \subseteq X \backslash X_k$, $\tilde{X} \subseteq X \backslash X_k$. Therefore, $X''_k \subseteq X \backslash X_k$: It is absurd!
"$\supseteq$": It is easy to prove.

Example 2. Consider class [X], where X=aceg, $\mathcal{G}(X) = \{X_1=ae, X_2=ag\}$. Then, X_U=aeg, $X_{U,1}$=g, $X_{U,2}$=e and X_=c. By theorem 2, X'=aceg$\in \mathcal{IS}(X)$ and X''=cg$\in \mathcal{N}^*(X)$ are uniquely generated: $X' = X_1 + X'_1 + \tilde{X}$ and $X'' = X'_1 + \tilde{X}$, where $X'_1 = g \subseteq X_{U,1}$, $\tilde{X} = c \subseteq X\_$. By proposition 3.c, X' has two duplicate representations: X' = ae+cg = ag+ce. If the condition $(i>1: X_k \not\subset X_i + X'_i, \forall k: 1 \le k < i)$ is absent, duplicate X' is generated once again: $X' = X_2 + X'_2 + \tilde{X}$, where $X'_2 = e \subseteq X_{U,2}$. Hence, all itemsets in $\lfloor X \rfloor = [X] = \mathcal{IS}(X) = \{$ae, aeg, aegc, aec, ag, agc$\}$ are non-repeatedly generated.

For two itemsets X, Y: $\emptyset \ne X$, $Y \subseteq \mathcal{A}$ and $X \cap Y = \emptyset$, we denote the set of itemsets (with constraint X) Y' contained in Y that the closure of Y'+X is the same with the one of Y+X as follows: $\lfloor Y \rfloor_X := \{Y' \subseteq Y \mid h(X+Y') = h(X+Y)\}$. Let $Y_{min,X} = \text{Minimal}\{Y_k \equiv Z_k \backslash X \mid Z_k \in \mathcal{G}(X+Y)\}$ be the class containing all minimal sets (in term of relation "$\subseteq$") of $\{Z_k \backslash X \mid Z_k \in \mathcal{G}(X+Y)\}$, $Y_{X,U} = \bigcup_{Y_k \in Y_{min,X}} Y_k$, $Y_{X,U,k} = Y_{X,U} \backslash Y_k$, $Y_{X\_} = Y \backslash Y_{X,U}$, $\mathcal{IS}(Y)_X = \{Y' = Y_k + Y'_k + \tilde{Y} \mid Y_k \in Y_{min,X}, \tilde{Y} \subseteq Y_{X\_}, Y'_k \subseteq Y_{X,U,k}$, k=1 or (k>1 and $Y_j \not\subset Y_k + Y'_k, \forall j: 1 \le j < k)\}$ and denote $\mathcal{FS}(Y)_X = \mathcal{IS}(Y)_X$, *if* $supp(X+Y) \ge s_0$. Theorem 3 shows unique representation of itemsets with constraint X.

Theorem 3 (*Unique representation of itemsets with constraint*). $\forall \emptyset \ne X$, $Y \subseteq \mathcal{A}$ and $X \cap Y = \emptyset$,

a) $Y' \in \lfloor Y \rfloor_X \Leftrightarrow \exists Z_0 \in \mathcal{G}(X+Y), Y'' \in \mathcal{N}(X+Y): Y' = (Z_0 + Y'') \backslash X$.

b) $\lfloor Y \rfloor_X = \mathcal{IS}(Y)_X$. **c)** *All itemsets of* $\mathcal{IS}(Y)_X$ *are non-repeatedly derived.*

3 Structures of Association Rule Set

This section partitions rule set into disjoint rule classes using an equivalence relation based on the closure L of left-handed side and the one S of two-sided union of rules.

Due to this relation, we only independently consider each class $\mathcal{AR}$(L, S) of the rules in the form: r:L'→S'\L', where ∅≠L'⊂S' and L'∈[L], S'∈[S]. Thus, they also have the same confidence. The size of $\mathcal{AR}$(L, S) can be still big. Then, we can just mine the set of basic rules. When it is necessary, remaining rules can be generated from it. Pasquier et al. [7] and Zaki [11] considered mining basic rules in forms of minimal and most general. In [9], [10], they are also shown in the forms of minMax, minmin. By the viewpoint based on order relation, we see that there are different forms of basic rules such as MaxMax, Maxmin and min. This section proposes five order relations in order to obtain five pairs of sets of basic and consequence. For a given relation, the basic one contains minimal rules and the consequence one includes in the rules that can be deducted from them. In the rest, we show the explicit form of minimal rules according to relation min. Thanks to unique representations of itemsets, we propose how to derive non-repeatedly all remaining rules of $\mathcal{AR}$(L, S).

3.1 Partition of Association Rule Set by Equivalence Relation

Definition 4 (*Equivalence relation on association rule set*) [9]. Let $\sim_r$ be a binary relation on $\mathcal{ARS}$ defined as follows: ∀L',S',Ls,Ss ⊆ $\mathcal{A}$, ∅≠L'⊂S', ∅≠Ls⊂Ss, r:L'→S'\L', s:Ls→Ss\Ls: $s \sim_r r$ *iff (Ls ∈ [L'] and Ss ∈ [S']).*

Theorem 4 (*Partition of the association rule set*) [9]. Relation $\sim_r$ is an equivalence relation. It *partitions* $\mathcal{ARS}$ *into disjoint equivalence rule classes* $\mathcal{AR}$(L, S):

$$\mathcal{ARS} = \sum_{(L,S):\, sup(S)/sup(L) \geq c_0} \mathcal{AR}(L,S).$$

All rules in each class have the same support and confidence.

Then, we need only to independently investigate structure of each equivalence rule class $\mathcal{AR}$(L, S) = {r:L'→R' | L'∈[L], L'+R'∈[S]}. For example, $\mathcal{AR}$(ceg, aceg) = {e→acg, e→a, e→ac, e→ag, ec→ag, eg→ac, egc→a, ec→a, eg→a, g→ace, g→a, g→ac, g→ae, gc→ae, gc→a} and $\mathcal{AR}$(ceg, ceg) = {e→cg, e→c, e→g, g→ce, g→c, g→e, eg→c, ec→g, gc→e}. Obviously, $\mathcal{AR}$(L, L)=∅, ∀L∈ $\mathcal{FCS}$: L∈ $\mathcal{G}$(L). Then, when considering $\mathcal{AR}$(L, L), we always suppose that L∉ $\mathcal{G}$(L).

3.2 Basic and Consequence Sets According to Different Order Relations

Definition 5 (*Order relations over each rule class*). Consider binary relations over $\mathcal{AR}$(L, S) defined by: ∀r_j: $L_j \to R_j \in \mathcal{AR}$(L, S), $S_j = L_j + R_j$, j=1,2:

a) $r_1 \prec_{MinMax} r_2$ *iff* $(L_1 \subseteq L_2$ *and* $R_1 \supseteq R_2)$. **b)** $r_1 \prec_{minmin} r_2$ *iff* $(L_1 \subseteq L_2$ *and* $R_1 \subseteq R_2)$.
c) $r_1 \prec_{MaxMax} r_2$ *iff* $(L_1 \supseteq L_2$ *and* $S_1 \supseteq S_2)$. **d)** $r_1 \prec_{Maxmin} r_2$ *iff* $(L_1 \supseteq L_2$ *and* $R_1 \subseteq R_2)$.
e) $r_1 \prec_{min} r_2$ *iff* $(L_1 \supseteq L_2$ *and* $R_1 \supseteq R_2$, *if* $L \subset S)$; $(L_1 \subseteq L_2$ *and* $R_1 \supseteq R_2$, *if* $L = S)$.

It is easy to prove that the above binary relations are partial order relations over $\mathcal{AR}$(L, S). Basic rule set $\mathcal{B}_{name}$(L, S) contains minimal elements with respect to order

relation $\prec_{\text{name}}$ (name $\in$ {minmin, minMax, Maxmin, MaxMax, min}). Corresponding consequence rule set is denoted as $\mathcal{C}_{name}(L, S) := \mathcal{AR}(L, S) \setminus \mathcal{B}_{name}(L, S)$. It contains all non-minimal rules, i.e., for every consequence rule r_c, there exists a basic rule r_b such that: $r_b \prec_{\text{name}} r_c$. We need to figure out: (1) *how to determine basic rules* r_b, and (2) *how to derive non-repeatedly all consequence rules* r_c *of them*?

The works related to relations of minmin and minMax have been considered by Pasquier et al. [7] and Zaki [11]. Overcoming the weaknesses in their results, in [10], [11] we indicated the better mining algorithms based on those relations. In the rest of the paper we consider structure of association rule set based on relation min (for relations of Maxmin and MaxMax, one can get the similar results). We will show that mining rules based on this relation is better than the ones based on minmin relation and minMax one in terms of reductions in number of basic rules and mining times.

3.3 Generating Min Basic Rules and Deriving Non-repeatedly Consequence Ones

Theorem 5 shows explicit form of min basic rules (basic rules determined by $\prec_{\text{min}}$).

Theorem 5 (*Explicit form of min basic rules*). Let L $\subseteq$ S. Then

a) $\mathcal{B}_{min}(L, S) = \{r_b: L \rightarrow S \setminus L\}$, *if* $L \subset S$.

b) $\mathcal{B}_{min}(L, L) = \{r_b: L_i \rightarrow L \setminus L_i \mid L_i \in \mathcal{G}(L)\}$, *if* $L \notin \mathcal{G}(L)$.

For example, $\mathcal{B}_{\text{min}}$(ceg, aceg) = {ceg→a} and $\mathcal{B}_{\text{min}}$(ceg, ceg) = {e→cg, g→ce}. Using closure-preserved property of eliminable itemset, from each rule r_0, for generating consequence rules that belong to the same class, we need only delete or move eliminable itemsets in both sides of r_0. These operators are used (by users) for generating preserved-confidence and non-repeated consequence rules from each basic rule. For different rules, however, corresponding consequence sets could include duplicate rules. To avoid the duplications, we replace generating consequence rule set from each rule with the one from each rule set containing rules of the same left-handed or right-handed sides.

Definition 6 (*Operators for deriving non-repeatedly all consequence rules*)

a) L⊂S: $\mathcal{B}_{min\text{-}L\text{-}R}(L, S) = \{r_c: L' \rightarrow R' \mid R' \in \mathcal{FS}(S \setminus L)_L, L' \in \mathcal{FS}(L)$ *and* $(L' \subset L$ *or* $R' \subset S \setminus L)\}$,

$\mathcal{B}_{min\text{-}L+R}(L, S) = \{r_c: L'' \rightarrow R' + (L' \setminus L'') \mid L' \in \mathcal{FS}(L), R' \in \mathcal{FS}(S \setminus L)_L, L'' \in \mathcal{FS}(L') \setminus \{L'\}\}$.

b) L=S: $\mathcal{B}_{min\text{-}R+L}(L, L) = \{r_c: L_i + R'' \rightarrow R' \setminus R'' \mid L_i \in \mathcal{G}(L), \emptyset \neq R' \subseteq L \setminus L_i, \emptyset \neq R'' \subset R', i=1$ *or* $(i>1: L_k \not\subset L_i + R'', \forall k: 1 \leq k < i)\}$,

$\mathcal{B}_{min\text{-}R}(L, L) = \{r_c: L_i \rightarrow R' \mid \emptyset \neq R' \subset L \setminus L_i, L_i \in \mathcal{G}(L)\}$.

To illustrate the definition, considering (L, S) = (ceg, aceg), we have: $\mathcal{B}_{\text{min-L-R}}$(L, S) = {e→a, ce→a, eg→a, g→a, cg→a} and $\mathcal{B}_{\text{min-L+R}}$(L, S) = {e→ac, e→ag, g→ae, e→acg, ce→ag, eg→ac, g→ace, cg→ae, g→ac}. Theorem 6 shows structure of the set of all min consequence rules that are non-repeatedly derived from min basic ones.

Theorem 6 (*Deriving non-repeatedly all min consequence rules*)

a) L ⊂ S: i) All *rules in* $\mathcal{B}_{min\text{-}L\text{-}R}(L, S)$, $\mathcal{B}_{min\text{-}L+R}(L, S)$ *are non-repeatedly generated.*

ii) All *rules in* $\mathcal{B}_{min\text{-}L\text{-}R}(L, S)$, $\mathcal{B}_{min\text{-}L+R}(L, S)$, $\mathcal{B}_{min}(L, S)$ *are totally different.*

iii) $\mathcal{C}_{min}(L, S) = \mathcal{B}_{min\text{-}L\text{-}R}(L, S) + \mathcal{B}_{min\text{-}L+R}(L, S)$.

b) L = S: i) All *rules in* $\mathcal{B}_{min\text{-}R}(L, L)$, $\mathcal{B}_{min\text{-}R+L}(L, L)$ *are non-repeatedly generated.*
ii) All *rules in* $\mathcal{B}_{min\text{-}R}(L, L)$, $\mathcal{B}_{min\text{-}R+L}(L, L)$, $\mathcal{B}_{min}(L, S)$ *are totally different.*
iii) $\mathcal{C}_{min}(L, L) = \mathcal{B}_{min\text{-}R}(L, L) + \mathcal{B}_{min\text{-}R+L}(L, L)$.

Proof: (a)*"$L \subset S$":* (i), (ii): It is easy to prove them. (iii) *"$\supseteq$":* For every r_c:L'$\rightarrow$R'$\in \mathcal{B}_{\text{min-L-R}}$(L, S), where R'$\in \mathcal{FS}$(S\L)$_L$, L'$\in \mathcal{FS}$(L) and (L'$\subset$L or R'$\subset$S\L), then R'$\in \mathcal{FS}$(S\L)$_{L'}$, h(L'+R')=S and $r_c \in \mathcal{C}_{\min}$(L, S). For every r_c:L''$\rightarrow$R'+(L'\L'') $\in \mathcal{B}_{\text{min-L+R}}$(L, S), where L'$\in \mathcal{FS}$(L), L''$\in \mathcal{FS}$(L')\{L'}, R'$\in \mathcal{FS}$(S\L)$_L$, we have R'$\in \mathcal{FS}$(S\L)$_{L'}$, h(L''+R'+(L'\L'')) = h(L'+R') = h(L+R') = S and $r_c \in \mathcal{C}_{\min}$(L, S). *"$\subseteq$":* For every r_c:L''$\rightarrow$R''$\in \mathcal{C}_{\min}$(L, S): h(L''+R'')=S, h(L'')=L and (L''$\subset$L or R''$\neq$S\L), we have R''=R'$_0$+R', where R'$_0$=R''$\cap$L, R'=R''\L, L'=L''+R'$_0\supseteq$L'', L''+R''$\subseteq$ L$\cup$R''=L+R'$\subseteq$S. Then R'$_0$=L'\L'', h(L'')=h(L')=L and h(L+R')=S, i.e., L'$\in \mathcal{FS}$(L), L''$\in \mathcal{FS}$(L')$\subseteq \mathcal{FS}$(L), R'$\in \mathcal{FS}$(S\L)$_L$ and r_c:L''$\rightarrow$R'+(L'\L''). If R'$_0$=(L'\L'') = $\emptyset$, then (R'=R''$\neq$S\L or L''$\subset$L). Thus r_c:L''$\rightarrow$R'$\in \mathcal{B}_{\text{min-L-R}}$(L, S). If R'$_0$=(L'\L'') $\neq \emptyset$, then L''$\in \mathcal{FS}$(L')\{L'}, r_c:L''$\rightarrow$ R'+(L'\L'')$\in \mathcal{B}_{\text{min-L+R}}$(L, S).
(b) It is similarly proved.

Based on theorems 5 and 6, the fast algorithms for mining (directly) min basic rules and generating (non-repeatedly) all consequence ones are obtained.

4 Experimental Results

Four benchmark databases in [13] are used during these experiments: Pumsb contains 49046 transactions, 7117 items (P, 49046, 7117); Mushroom (M), 8124, 119; Connect (C), 67557, 129; Pumsb* (P*), 49046, 7117. Experiments will show the efficiency of mining rules using relation min with the ones using relations of minmin and minMax.

Consider Table 1 where: the number of all rules is shown in column #AR and the percent ratios of the cardinalities of basic sets of min, minmin and minMax to #AR are in turn shown in columns of B_m, B_{mm} and B_{mM}. It shows that *the cardinality of min basis is smaller than the ones of bases of mm and mM*. For the present experiments, the reduction in the number of basic rules ranges from a factor of 1.0 to 2.1 times.

Table 1. The sizes of basic sets (MS = min sup, MC = min confidence: %)

DB (1)	MS=MC (2)	#AR	B_m (%)	B_{mm} (%)	B_{mm}/B_m	B_{mM} (%)	B_{mM}/B_m
C	90	3640704	8.77	8.92	*1.0*	8.77	*1.0*
	80	326527774	1.21	1.23	*1.0*	1.21	*1.0*
M	20	19191656	0.15	0.31	*2.1*	0.18	*1.2*
	15	34505370	0.18	0.33	*1.8*	0.21	*1.2*
P	85	1408950	31.55	51.64	*1.6*	41.21	*1.3*
	80	28267480	13.19	27.47	*2.1*	19.46	*1.5*
P*	40	5659536	2.99	4.68	*1.6*	3.49	*1.2*
	30	311729540	0.88	1.67	*1.9*	1.14	*1.3*

Table 2 figures out that *the times for mining sets of basic rule and all rules based on relation min* (in columns TB_m, T_m) *are less than the ones based on relations of mm* (in columns TB_{mm}, T_{mm}) *and mM* (in columns TB_{mM}, T_{mM}). The reduction in the time for mining *basic* rules ranges from a factor of 1.1 to 4.4 times. The one in the time for mining *all* rules ranges from 1.0 to 1.9 times. Figures of 2 and 3 show the effect of minimum confidence on the mining times. The time to mine *basic* rules using relation min is less much than the one using relation mm. In comparison with relation mM, it is noteless. However, when mining *all* rules, the conversion comes. Hence, we can conclude that *mining rules using relation min is better*.

Table 2. The times for mining basic and all rules using relations of min, minmin and minMax

(1)-(2)	TB_m	TB_{mm}	TB_{mm} / TB_m	TB_{mM}	TB_{mM} / TB_m	T_m	T_{mm}	T_{mm} / T_m	T_{mM}	T_{mM} / T_m
C-90	1.31	1.66	*1.3*	1.41	*1.1*	49	53	*1.1*	53	*1.1*
C-80	18.16	21.50	*1.2*	19.53	*1.1*	2674	2878	*1.1*	3612	*1.4*
M-20	0.09	0.41	*4.4*	0.14	*1.5*	88	91	*1.0*	164	*1.9*
M-15	0.25	0.72	*2.9*	0.31	*1.2*	158	163	*1.0*	265	*1.7*
P-85	1.70	4.42	*2.6*	2.22	*1.3*	40	54	*1.4*	44	*1.1*
P-80	15.36	56.13	*3.7*	22.45	*1.5*	504	738	*1.5*	646	*1.3*
P*-40	0.68	1.59	*2.3*	0.84	*1.2*	51	57	*1.1*	71	*1.4*
P*-30	12.45	36.70	*2.9*	16.38	*1.3*	2597	2511	*1.0*	3625	*1.4*

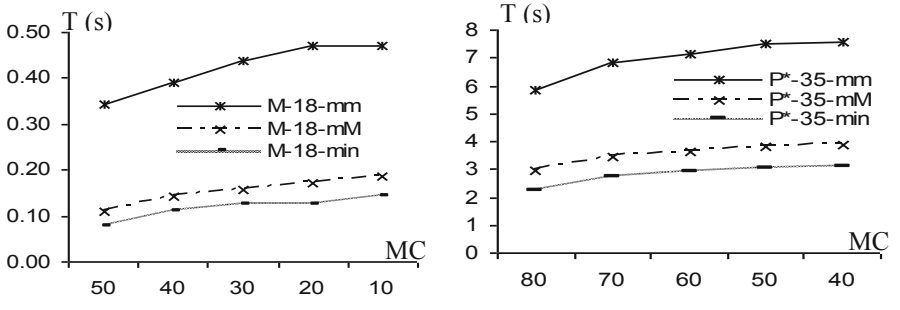


Fig. 2. The times for mining *basic* sets: M with MS = 18 and P* with MS = 35

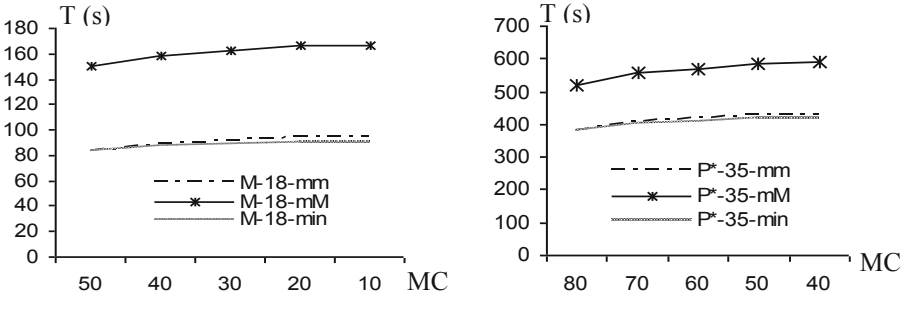


Fig. 3. The times for mining all rules: M with MS = 18 and P* with MS = 35

5 Conclusion

An equivalence relation partitions the class of itemsets into disjoint classes. Each class contains itemsets of the same closure so the same support. Their unique representations are figured out. The association rule set is also partitioned into disjoint rule classes by an appropriate equivalence relation. Each rule class contains all association rules having the same closures of left-handed sides and two-sided unions, so the same confidence. Each rule class splits into different sets of basic and consequence based on different order relations. According to relation min, the explicit form of basic rules is shown. Operators for deducing non-repeatedly all remaining rules are also obtained. The paper shows that mining association rules based on order relation min is better the ones based on relations of minmin and minMax.

References

1. Agrawal, R., Imielinski, T., Swami, N.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOID, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 478–499 (1994)
3. Bao, H.T.: An approach to concept formation based on formal concept analysis. IEICE Trans. Information and Systems E78-D(5) (1995)
4. Duquenne, V., Guigues, J.-L.: Famille minimale d'implications informatives re'sultant d'un tableau de données binaires. Math. Sci. Hum. 24(95), 5–18 (1986)
5. Feller, W.: An introduction to probability theory and its applications, vol. I. John Wiley & Sons. Inc., Chapman & Hall. Ltd., New York, London (1950)
6. Luxenburger, M.: Implications partielles dans un contexte. Math. Inf. Sci. Hum. 29(113), 35–55 (1991)
7. Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., Lakhal, L.: Generating a condensed representation for association rules. J. of Intelligent Information Systems 24(1), 29–60 (2005)
8. Pei, J., Han, J., Mao, R.: CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. In: Proceedings of the DMKD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 21–30 (2000)
9. Tin, T., Anh, T.: Structure of Set of Association Rules Based on Concept Lattice. In: Nguyen, N.T., Katarzyniak, R., Chen, S.-M. (eds.) Advances in Intelligent Information and Database Systems. SCI, vol. 283, pp. 217–227. Springer, Heidelberg (2010)
10. Tin, T., Anh, T., Thong, T.: Structure of association rule set based on min-min basic rules. In: Proceedings of the 2010 IEEE-RIVF International Conference on Computing and Communication Technologies, pp. 83–88 (2010)
11. Zaki, M.J.: Mining non-redundant association rules. Data Mining and Knowledge Discovery (9), 223–248 (2004)
12. Zaki, M.J., Hsiao, C.-J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. Knowledge and Data Engineering 17(4), 462–478 (2005)
13. Frequent Itemset Mining Dataset Repository (2009), `http://fimi.cs.helsinki.fi/data/`

Database Integrity Mechanism between OLTP and Offline Data

Muhammad Salman[1], Nafees Ur Rehman[2], and Muhammad Shahid[3]

[1] Graduate School of Engineering Sciences & Information Technology,
Hamdard University Karachi-75400, Pakistan
[2] Department of Computer & Information Sciences,
University of Konstanz, P.O. Box. D-188, 78457 Germany
[3] Department of Computer & Information Technology,
Pakistan International Airline Karachi-75200, Pakistan
{msalman_74,muhammadshahid}@hotmail.com,
nafees.rehman@uni-konstanz.de

Abstract. This paper describes integrity mechanism between OLTPs and offline data. Normally every RDBMS supports five Integrity Constraints (ICs) namely primary key or composite key, unique key, foreign key, not null and check constraints. Online database integrity is achieved through these five ICs. However, as per the retention period data is backed up and removed from the OLTPs for space and performance efficiency. But there is no standardized protocol on keeping integrity between offline data and data present in the OLTPs. Therefore, we present a solution to address the problem of offline data integrity by keeping a representative set of purged data & ICs in the online database to ensure data integrity between OLTPs and offline data. We further support our proposed solution with the help of two types of integrity tests i.e., sufficient and complete test.

Keywords: Integrity Constraints, OLTPs Performance, Offline Data.

1 Introduction

The word integrity is derived from Latin adjective integer (whole, complete). In this context, integrity is the sense of wholeness. Integrity is a concept of consistency of actions, values, methods, measures, principles, expectation and outcomes [1]. Similarly database integrity means keeping your data accurate, consistent and valid. It is achieved by preventing accidental or deliberate but unauthorized insertion, modification or destruction of data in a database. Online database integrity is achieved by integrity constraints. So in order to improve OLTPs performance we have to implement certain retention policy and offline backup will be taken as per the retention policy. And then that particular set of data will be removed from OLTPs. Unfortunately, not that much work has been done to maintain offline data integrity in such a scenario.

To explain the problem, consider a database schema containing tables namely Customer, Product, Order and etc. The size of data in Order table, let's assume, is 50GB. Therefore, we have to implement certain retention policy and offline backup will be taken as per the retention policy. Finally, this particular set of data will be

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 371–380, 2012.

deleted from online Order table. However, any DML operation on Order table would challenge the data integrity of the offline data as there is no communication with the running OLTPs and the backed up offline data.

2 Related Work

Let us present a summary of few related works in this section.

In [2], an integrity subsystem for RDBMS has been discussed, and it shows how integrity is different from the other important area of security, consistency, and reliability. The integrity subsystem reacts as guards that prevent database against unauthorized insertion, modification or destruction of data in a database, and data accurately stored in database.

[3] Introduces schema integration process. Schema integration is an important step in the database integration, and deal with integrity constraints problems. The knowledge about the correspondences between integrity constraints and extensional assertions must be known in the schema integration process, and handle all issues of database integrity at schema level in the distributed environment.

[4] Explains the active integrity constraints (AICs) with consistent database. An active integrity constraint is a special constraint which is maintained database integrity for all DML operations. DML operation would be generated constraints violation while data already available in database. Otherwise, DML operations are allowed.

[5] Maintains database integrity during updating process has been explained. It means that, multiple users access online database concurrently. If one user performs update query on particular record set during this process other users cannot perform update query on that particular record set until first user performs commit or rollback transaction on that particular record set.

[6] This paper describes database integrity patterns. Actually vendors of RDBMS (Relational Database Management System) are quite high, but number of available solutions to active database integrity is limited. The solutions to avoid RDBMS integrity problems can be formalized as a pattern language. Integrity patterns were defined in the form of constraints, locking, and transactions.

[7] Explains the process of database integrity for synthetic vision system (SVS). Synthetic vision system provides information to cockpit crew with whether a heads down display and heads up display. HDD and HUP containing aircraft state guidance, navigation information, and weather condition. So therefore all information depends upon database integrity. It is very essential for the safety.

In [8], a solution for maintenance of database integrity in the replication environment is presented. The previous version of this paper presented a method for replication through 1-copy serialability. The new version of this paper is suggested replication process through snapshot isolation (SI). SI controls all issues of database integrity.SI takes copy of data from active database and only updates that data which is changed in the active database. This method is improved performance of replication process and maintained database integrity. Most of modern all RDBMSs support SI through materialized view (stored snapshot of data).

In [9], authors discussed schema integration in the entity relationship model (ERM). Schema integration is achieved through three steps. That is, first, finds out ambiguities of integrity constraints in the integration process and solved; second, schemas are merged; third restructure integrated schema according to specific goal.

The above mentioned research papers describes solution of various online database integrity problems and contributed further enhancement in this direction. None of these directly deal with the problem of online database integrity with offline data. Now we present a solution to this problem in the following sections.

3 Method to Achieve Integrity

To achieve offline data integrity, the two following approaches have been proposed:

- Offline data integrity at record level
- Offline data integrity on batch mode

3.1 Offline Data Integrity at Record Level

Offline data integrity at record level is achieved through RDBMS trigger. In this technique, as per retention policy (i.e., when there are large amounts of data)) we took the online database backup and stored representative subset of backup data that consists of primary key or composite key, unique key and foreign key values and finally removed backup data from the OLTPs. Afterwards, before a new record is inserted or unique values of the existing records are updated, this is first verified for any integrity violation against the representative subset of backed up data stored in the online database system. In case of a violation, insertion or updation is rejected and integrity validation exception is raised. Otherwise, DML operations in the online database are allowed. Similarly in the foreign key case, when we insert new record into tables then first it is checked against online database subsets (foreign key) for integrity violations. And similarly if there is a violation it raises an exception and DML operation is disallowed. Otherwise, the change takes effect and DML operations are performed in the online database system.

3.2 Offline Data Integrity on Batch Mode

Integrity verification for each tiny change results in comparatively huge process and effect the performance of OLTPs. To control this problem, we suggest a method where we temporarily allow data in the OLTPs and ignore integrity violation just to provide for better performance in the OLTPs. Instead of checking integrity for each row, we do it in batch mode. It can be done on the off-hours of the day or at a particular time as deemed appropriate. The logic goes into a stored procedure to check integrity in batch mode. In this method, as per retention policy (i.e., when there are large amounts of data) we took the online database backup and stored representative subset of backup data that consists of primary key or composite key, unique key and foreign key values and finally removed backup data from the database. In this case, insertion of new

records into tables or updating of the existing values is allowed. Later, at the appropriate time, the stored procedure is invoked and is executed in a manner ignoring values with no conflicts but reporting records that are in integrity violation.

3.3 Logical Design of Integrity Mechanism between Online System and Offline Data

The Fig 1 explains the process of our solution. As per retention policy we take the backup of online database and store representative subset of backup data in the online database and finally remove backup data from the database. When a user performs insert or update query on the online database then on first step the query is validated by representative subset of data that is offline data. If a record or tuple primary or composite key, unique key values are matched with a representative subset of data then a transaction is aborted from the database. Otherwise, query is allowed on the online database. Similarly if the new updated value for foreign key in OLTP matches a foreign key value in the representative subset of data then the transaction is allowed. Otherwise, constraint violation exception is raised.

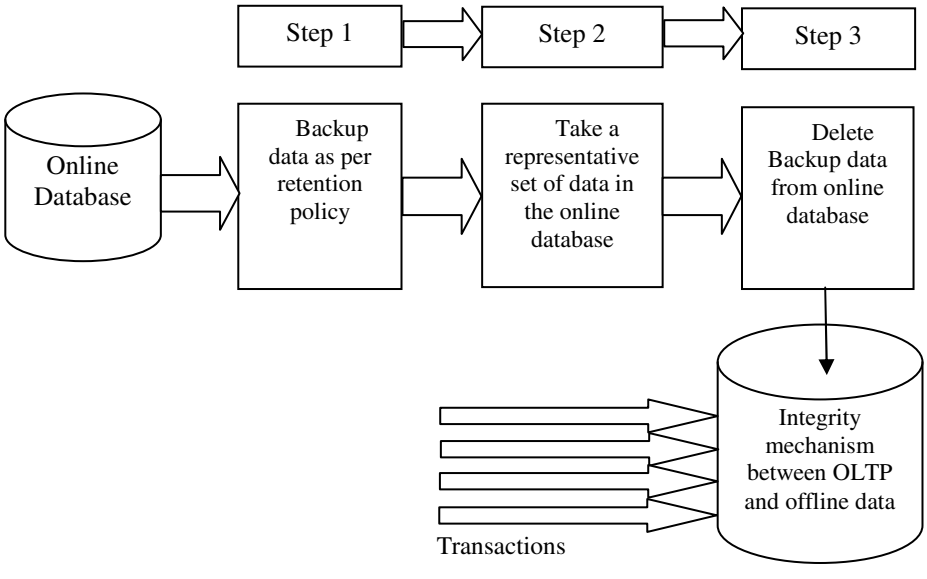


Fig. 1. Integrity mechanism between OLTPs & offline data

3.4 Program Code

3.4.1 Trigger Level Checking Integrity Constraints and Tests

In the trigger code, first we select all keys values (primary key, unique Key, foreign key) from offline data set which is available in the online database. When new transaction is inserted or updated then at time check value against the selected primary and unique keys values. If value is already available then exception is

generated otherwise transaction is performed. Similarly in the foreign key case, new transaction foreign key value is checked against the selected foreign key values. If current transaction value is matched then transaction is performed otherwise, transaction generated the exception.

```
Input: List of integrity constraints primary key PK,
unique key UK, foreign key FK from Representative set of
data (Offline Data)
Output: Valid transaction performed or aborts transaction
(if available in the offline data)
Begin
 Store the begin time in the Tri_Result table
 For each j in Off-Transaction loop
  Begin
  Invoke sufficient test
  If current_value.PK is equal to j.PK and
  Current_value.UK is equal to j.UK then
  Action: Abort current transaction from the database;
  Else
  Action: Complete test is performed & DML 'T' Done;
  End.
 For each k in Off-Transaction loop
  Begin
  Invoke sufficient test
  If current_value.FK is equal to k.FK then
  Action: DML 'T' Done;
  Else
  Action: Complete test is performed & Abort Current
  Transaction;
  End.
  Action: Store the end time in the Tri_Result table;
End.
```

3.4.2 Procedure Level Checking Integrity Constraints and Tests

In the procedure code, first we selected complete all keys values (primary, unique and foreign) on daily base from the online database. Second we selected complete all keys values from the offline dataset which is available in the online database. If daily base online database values matched with offline represented set of data then it report to an integrity violation.

```
Input: List of integrity constraints primary key PK,
unique key UK, foreign key FK from Representative set of
data (Offline Data)
Output: Valid transaction performed or aborts transaction
(if available in the offline data)
```

```
Begin
 Store the begin time in the Proc_Result table
 For each i in online-transaction loop
  For each j in off-transaction loop
  Begin
  Invoke sufficient test
  If i.PK is equal to j.PK and i.Uk is equal to j.UK then
  Action: Abort Current transaction from the database;
  Else
  Action: Complete test is performed & DML transaction
  Done;
  End.
 End.
 For each i in online-transaction loop
  For each j in off-transaction loop
  Begin
  Invoke sufficient test
  If i.FK is equal to j.FK then
  Action: DML performed on database;
  Else
  Action: Complete test is performed & Abort
  Transaction, T;
  End.
 End.
  Action: Store the end time in the Proc_Result table;
End.
```

4 Experimental Result Proof through McCarroll

Our approach has been developed to manage RDBMS integrity between OLTPs and offline data McCarroll introduces different level of integrity tests namely sufficient test, necessary test and complete test [10]. In the sufficient test, when update or insert operation is satisfied with respect to the integrity constraint and thus DML operation is performed on the table. In the necessary test, when insert or update operation is not satisfied with respect to the integrity constraint and thus DML operation abort the transaction from the database. Complete test have both sufficient and necessary tests properties.

When a user issues an update or insert query again the online database, the integrity constraints for the offline data are tested for any violation. If there is a violation, i.e. the update or inserted field value is in conflict with integrity constraints of offline data, then an exception is raised and transaction is aborted. Otherwise, update or insert query is executed and is allowed to make changes in the database. Referring to the below mention Table-2, ($\forall i \exists u \exists v \exists w \exists x \exists y \exists z$order ($i$, u, v, w, x, y, z)) is an example of sufficient test, which verify the existence of *Ord_id* values in the *Order_offline* table that is representative subset of data whereas i, u, v, w, x, y and z are constants.

When insert or update query is executed on an *Order* table. If *Ord_id* attribute value exists in an *Order_offline* table, then we conclude that the sufficient test and initial constraint, *I1*, is satisfied. If there is no record available with respect to an *Ord_id* attribute of *Order_offline* table, then further verification is performed by complete test and record is inserted or updated into *Order* table. Similarly we performed same test for *Card_no* attribute (which is unique key) in an *Order* table.

In the case of ($\forall u \exists i \exists v \exists w \exists x \exists y \exists z$ order (*u, i, v, w, x, y, z)*) is an example of sufficient test, which verified an existence of *Cid* attribute value in an *Order_offline* table, that is representative subset of data. When insert or update query is executed on an *Order* table. If *Cid attribute* value exists in an Order_*offline* table, then we conclude that the sufficient test property or initial constraint, *I2*, is satisfied. If there is no record available with respect to *Cid* attribute of *Order_offline* table, then further verification performed by complete test. Similarly we performed same test for *Pid* attribute (which is foreign key) in an Order table.

Every record is verified against the integrity constraint values between online database (*Order* table) and offline data (*Order_offline* table).

Table 1. Schematic Algebra

Schema: *Customer* (cid, cname, add, loc); *Product*(pid, pname, price, unit_price); *Order* (ord_id, cid, pid, card_no, ord_date, sup_date, tot_price); *Order_offline* (ord_id, cid, pid, card_no) Integrity Constraints:
Every record of *Ord_id* attribute in an *Order_offline* table exists in an *Order* table
I_1: $(\forall i \forall b \forall c \forall d \exists u \exists v \exists w \exists x \exists y \exists z)(Order\_offline(i, b, c, d) \rightarrow Order(i, \text{u}, \text{v}, \text{w}, x, y, z))$
Every record of *Cid* attribute in an *Order_offline* table exists in an *Order* table
I_2: $(\forall a \forall i \forall c \forall d \exists u \exists v \exists w \exists x \exists y \exists z)(Order\_offline(a, i, c, d) \rightarrow Order(\text{u}, \text{i}, \text{v}, \text{w}, x, y, z))$
Every record of *Pid* attribute in an *Order_offline* table exists in an *order* table
I_3: $(\forall a \forall b \forall i \forall d \exists u \exists v \exists w \exists x \exists y \exists z)(Order\_offline(a, b, i, d) \rightarrow Order(\text{u}, \text{v}, \text{i}, \text{w}, x, y, z))$
Every record of *Card_no* attribute in an *Order_offline* table exists in an *order* table
I_4: $(\forall a \forall b \forall c \forall i \exists u \exists v \exists w \exists x \exists y \exists z)(Order\_offline(a, b, c, i) \rightarrow Order(\text{u}, \text{v}, \text{w}, \text{i}, x, y, z))$

Table 2. Integrity Association

I	Insert and Update Template	Integrity Test
I_1	insert(*order*(*i, u, v, w, x ,y ,z*))	1. $(\forall i \exists x \exists y \exists z)(order\_offline(i, b, c, d))$ [1]
		2. $(\forall i \exists u \exists v \exists w \exists x \exists y \exists z)(order(i, u, v, w, x, y, z))$[2]
	update(*order*(*i, u, v, w ,x, y, z*))	3. $(\forall b \forall c \forall d)(\neg order\_offline(i, b, c, d))$ [1]
I_2	insert(*order*(*u, i, v, w, x, y, z*))	1. $(\forall x \exists i \exists y \exists z)(order\_offline(a, i, c, d))$ [1]
		2. $(\forall u \exists i \exists v \exists w \exists x \exists y \exists z)(order(u, i, v, w, x, y ,z))$[2]
	update(*order*(*u, i, v, w, x ,y, z*))	3. $(\forall a \forall c \forall d)(\neg order\_offline(a, i, c, d))$ [1]
I_3	insert(*order*(*u*, v, *i, w, x, y, z*))	1. $(\forall x \exists y \exists i \exists z)(order\_offline(a, b, i, d))$ [1]
		2. $(\forall u \exists v \exists i \exists w \exists x \exists y \exists z)(order(u, v, i, w, x, y, z))$[2]
	update(*order*(*u, v, i, w, x, y, z*))	3. $(\forall a \forall b \forall d)(\neg order\_offline(a, b, i, d))$ [1]
I_4	insert(order(u, v, w, i, y, z))	1. $(\forall w \exists x \exists y \exists i)(order\_offline(a, b, c, i))$ [1]
		2. $(\forall u \exists v \exists w \exists i \exists x \exists y \exists z)(order(u, v, w, i, x, y, z))$[2]
	update(*order*(*u, v, w, i, x, y, z*))	3. $(\forall a \forall b \forall c)(\neg order\_offline(a, b, c, i))$ [1]

5 Experimental Results

The system on which we ran our experiment is running Window XP, with 2GB RAM and 2GHz processor. This experimental result is done between online database (Order table) and offline data (*Order_offline table*).

5.1 Offline Data Integrity at Record Level

In Fig 2, shows experimental result for offline data integrity using Oracle 10g, SQL Server 2000 and IBM DB2 9.7 trigger. In this experiment, online database stored the representative set of data that is five million, seven million and up to 10 million and each valid transaction took process time is 1.600, 2.100 and up to 3.500 seconds for Oracle 10g in the online database. Similarly in the SQL Server 2000 case, each valid transaction took process time is two, three, four and five seconds. In the IBM DB2, each valid transaction took process time is 1.600, 2.200 and up to 3.400 seconds.

Table 3. Record level RDBMS Time Statistics

S.No	Offline Data Set	Oracle Time (Second)	SQL Server Time (Second)	IBM DB2 Time (Second)
1	5000000	1.600	2	1.600
2	7000000	2.100	3	2.200
3	9000000	2.400	4	2.500
4	11000000	3.500	5	3.400

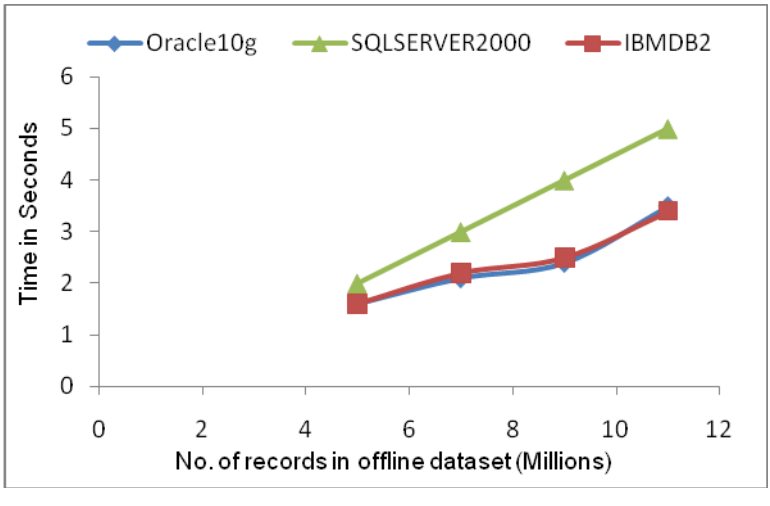


Fig. 2. Offline Data Integrity through Trigger

5.2 Offline Data Integrity on Batch Mode

This method improved OLTP performance as compare to record level approach because it is execute on batch mode. In each experiment, we stored a representative subset of data in the online database that is four millions, six millions, eight millions and 10millions and daily OLTPs (online transaction processing system) transactions on a table that is two thousand, four thousands, six thousand and ten thousand. Therefore, we calculated the stored procedure execution timing in different tools (Oracle10g, SQL Server 2000 and IBM DB2 9.7).

Table 4. Batch mode RDBMS Time Statistics

S.No	Daily OLTP Transaction	Oracle Time (Second)	SQL Server Time (Second)	IBM DB2 Time (Second)
1	2000	4	5	3
2	4000	4.100	6	4
3	6000	4.200	8	4.500
4	8000	4.400	12	5.100
5	10000	4.500	18	6.500

In Fig 3, shows experimental result for offline data integrity using Oracle 10g, SQL Server 2000 and IBM DB2 procedure. In this experiment, Oracle stored procedure execution took 4, 4.2 and up to 4.5 seconds against an OLTP daily transaction on the database that is 2000, 6000 and up to 10000 and reporting records that are in integrity violation. Similarly in the SQL Server 2000 case, stored procedure execution took five, six and up to eighteen seconds and reporting records that are in integrity violation. In the IBM DB2 case, stored procedure took execution three, four and up to 6 seconds and reporting records that are in integrity violation.

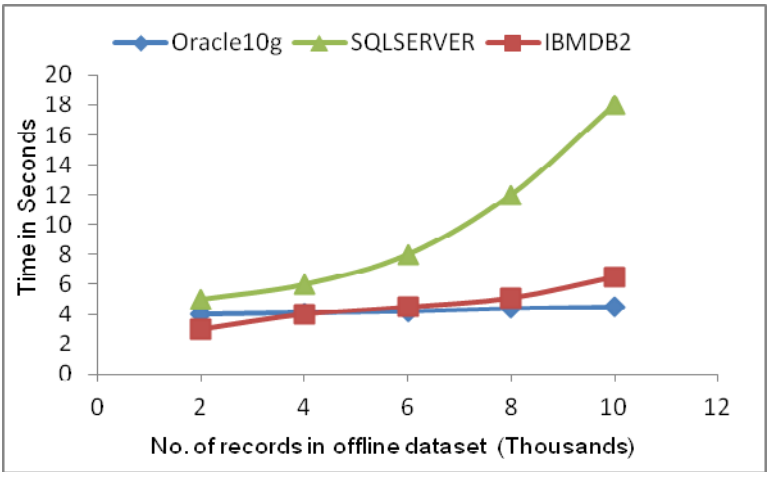


Fig. 3. Offline Data Integrity through procedure

6 Conclusions

We were able to introduce an integrity checking mechanisms between two disconnected systems, i.e. online database systems (OLTPs) and Offline data. We justified the offline data integrity through *sufficient* and *complete* tests. The paper presented results of various experiments conducted to support our work. These experiments were carried out on the major market products available like Oracle (10g) Microsoft SQL Server (2000) and IBM DB2 (9.7). The integrity mechanism presented was tested at two levels in the database systems i.e., at record level and in batch level. The integrity verification at level is implemented using a trigger and checks the integrity of the new record with the offline data. In case of a violation, the record insertion is rejected. The integrity verification in batch mode tolerates any integrity inconsistencies of the updated or newly inserted data for temporary time period to avoid performance degradation of OLTP while it is running the business. However, at an appropriate time, all updations and insertions can be checked for any integrity violation with the offline data.

References

1. The American Heritage Dictionary of the English Language, El-shaddai, p. 2000 (2009)
2. Eswaran, K., Chamberlin, D.: Functional Specifications of a Subsystem for Database Integrity. ACM (1975)
3. Türker, C.: Consistent Handling of Integrity Constraints and Extensional Assertions for Schema Integration. In: Eder, J., Rozman, I., Welzer, T. (eds.) ADBIS 1999. LNCS, vol. 1691, pp. 31–45. Springer, Heidelberg (1999)
4. Caroprese, L., Greco, S., Zumpano, E.: Active Integrity Constraints for Database Consistency Maintenance. IEEE Transactions on Knowledge and Data Engineering, 1042–1058 (2009)
5. Mayol, E., Teniente, E.: A Survey of Current Methods for Integrity Constraint Maintenance and View Updating. Advances in Conceptual Modeling, 67–73 (1999)
6. Mircea Petrescu, A., Rotaru, O.P.: A Database Integrity Pattern Language. ACM (2008)
7. Uut de Haag, M., Sayre, J.: Terrain Database Integrity Monitoring for Synthetic Vision Systems. IEEE Transactions on Aerospace and Electronic Systems, 386–406 (2005)
8. Lin, Y., Kemme, B.: Snapshot Isolation and Integrity Constraints in replicated Databases. ACM (2009)
9. Batini, Carlo.: A Methodology for Data Schema Integration in the Entity Relationship Model. IEEE Transactions on Software Engineering, 650–664 (2009)
10. McCarroll, N.F.: Semantic Integrity Enforcement in Parallel Database Machines PhD Thesis, University of Sheffield, UK (1995)

Novel Criterion to Evaluate QoS of Localization Based Services

Juraj Machaj[1], Peter Brida[1], and Norbert Majer[2]

[1] University of Zilina, Faculty of Electrical Engineering, Department of Telecommunications and Multimedia, Univerzitna 8215/1, 010 26 Zilina, Slovakia
{Juraj.Machaj,Peter.Brida}@fel.uniza.sk
[2] Research Institute of Posts and Telecommunications, Banská Bystrica, Slovakia
noro.majer@gmail.com

Abstract. This paper describes novel criterion to evaluate QoS of Localization Based Services (LBSs). Architecture of LBSs and communication between particular parts are described. Localization systems based on different technologies are briefly introduced. Parameters required to evaluate QoS of location based services are described. Sufficient range of values for each of the described parameters are introduced and used to evaluate QoS of different location based services based on their demands.

Keywords: QoS, Positioning, Localization, Services.

1 Introduction

In past few years, large number of location based services (LBSs) was developed. Basic need for these services is known position of mobile user. This can be achieved using one of existing localization systems.

The most widely used localization systems are GNSS (Global Navigation Satellite Systems) like GPS (Global positioning system) developed by US Army, GLONASS (Global Navigation Satellite System) developed by Russia (former Soviet Union), or GALILEO which is developed by European union. Drawback of these systems is that direct visibility to the satellites must be achieved to estimate position accurately. This can be problem in dense urban environment and in indoor environment.

GNSS is not very accurate in dense urban and indoor environment, therefore alternative localization systems had to be developed. These systems are based on radio networks. These systems are very popular for example in emergency situation positioning. Most localization systems based on radio networks are based on GSM/UMTS networks for urban environment and on WiFi networks for indoor positioning. Main advantage of these localization systems is that almost every device (mobile phone, smart phone, PDA…) is equipped with receiver for mentioned radio networks.

Many research teams try to develop a localization system that will be able to provide for ubiquitous positioning services. The system can be designed as modular

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 381–390, 2012.

systems, i.e. it consists of various localization subsystems based on different technologies to achieve best possible localization accuracy. It seems to be optimal solution how to fill different QoS demands of various services. There are more QoS parameters of location based services and it is clear that QoS values are different for each service. Generally, it is necessary to know QoS parameters provided by localization systems.

In this paper novel concept and criterion to evaluate QoS of localization systems will be proposed. With use of this criterion one from available localization subsystems can be chosen to mobile user positioning based on service demands.

Rest of the paper is organized as follows, in Section 2 the LBS communication model will be presented, components of LBS will be described in Section 3, Section 4 describes localization systems which can be used to estimate mobile user position, QoS parameters for localization systems will be defined in Section 5 and criterion for QoS evaluation will be proposed in Section 6.

2 LBS Communication Model

Communication of LBS can be described using three layer communication model, which consists of localization layer, middleware layer and application layer [1]. This model can be seen in Fig. 1.

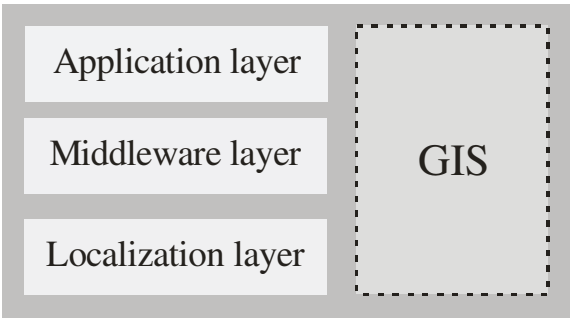


Fig. 1. Communication model of LBS

Function of localization layer is to estimate position of mobile device using Position Determination Equipment (PDE) and geospatial data stored in Geographic Information System (GIS). PDE estimates position of mobile device in network and GIS is used to convert data from PDE to geographic position. Geographic position is then sent thru location gateway to the middleware platform or directly to the application [1].

Middleware layer represents abstract layer between localization and application layers. Middleware layer is used to simplify integration of services, because it is connected to mobile network and to services provided by operator of network [1]. It can also be used to control localization services that will be provided. Simple

integration provided by middleware layer is important to operators because of mass access to localization data.

Application layer is represented by application which uses localization data to provide geospatial based or geospatial information to the user of mobile device.

3 LBS Components

Services provided by LBS are based on transfer of information between its components. LBS can be divided into these 5 basic components [2]:

- Mobile device
- Communication network
- Localization component
- Application and service provider
- Content and data provider

Mobile device is the equipment used by user to request information he needs. Response of system can be represented for example by text or picture. Mobile devices are mostly represented by cell phone, PDA or even GPS navigation device in car.

Communication network is used to transfer requests from the mobile device to a service provider and also provide data from a service provider to the mobile device.

Since provider of services needs to know position of mobile device to provide related information and services, localization component is used to automatically estimate position of mobile device. This component can be represented by one of localization systems which will be described in next section.

Service provider offers different services to mobile users and is responsible for their handling. These services can be represented by position estimation, finding routes, search in yellow pages based on user position or search of information about points of interest.

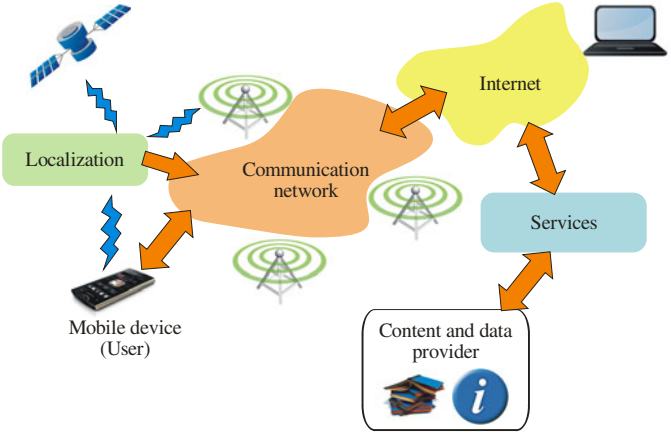


Fig. 2. Communication between LBS components

Since service provider does not store all of the required information requested by the users, content and data provider is needed to provide required data. Basic geographic data are mostly requested from companies which work with them or from business partners, for example mapping agencies or transport companies.

In Fig. 2 communication between the LBS components is shown. At first user decides what information he needs and sends request using mobile device. In case that requested function is activated, position of mobile user is achieved from localization server, or can be provided by the mobile device. Request from user together with estimated position is sent to the gateway using a communication network.

Gateway is used to transfer data between communication network and internet network. In gateway addresses of different application servers are stored, so gateway can send request directly to the specific application server.

Application server activates appropriate service based on request from mobile device. Service then analyzes request and decides which information are needed to provide response to mobile device. Additional information can be requested from content and data provider.

When requested information is available then is sent from application server to gateway and from gateway thru communication network to the mobile device.

4 Localization Systems

Localization systems can be divided based on different properties, in this paper division based on network infrastructure will be used to describe widely used localization systems.

4.1 Satellite Navigation Systems

Satellite navigation systems are mostly used to estimate position of mobile user in outdoor environment. Advantage of these systems is their wide availability (one system can cover whole world) and high accuracy especially in urban environment [3]. Achievable accuracy of these localization systems is in area of 5m -10m [4]. On the other hand, drawback of these localization systems is high sensitivity on signal fading, which cause problems in dense urban and indoor environments.

4.2 Localization Systems Based on Cellular Network

Localization systems based on cellular networks are mostly based on GSM or UMTS networks. In this case position of mobile device is estimated using signals from base stations [3]. Accuracy of these systems is very low. In dense urban environment accuracy of 150m can be achieved but in urban environment localization error can grows to 1km or more.

Mobile providers can use more localization techniques to improve positioning accuracy and availability of positioning information, based on application requirements. Advantage is that mobile devices do not need any improvements and position is estimated using existing network in some cases.

4.3 Localization Systems for Indoor Environment

Localization systems for indoor environment can be based on different radio networks like IEEE 802.11 (WiFi), UWB, ZigBee, RFID and Bluetooh, but there are also systems based also on infrared or ultrasound signals. Advantage of these localization systems lies in their high accuracy in range of 1m - 5m [4, 5]. Drawback of indoor positioning system is in small coverage areas.

5 QoS Parameters

Parameters defined in standards can give us just the basic overview on quality of services provided by LBS [6, 7]. QoS of localization systems is affected by morphology of environment, structure of localization system or radio channel properties.

Basic three QoS parameters are horizontal accuracy, vertical accuracy and response time. Since there was not high attention given to QoS of location based services, other QoS parameters needs to be defined.

Almost everybody tries to create LBS that will have best possible values of all QoS parameters. This is not important in al cases and drawback of this approach is mostly in economic point of view. It is more effective to use localization based systems which are able to achieve QoS needed for the desired service.

In this section, parameters which have the biggest impact on QoS of localization based services will be introduced.

5.1 Availability of Service

Availability is the most important parameter for LBS. Availability can achieve just two values – available and unavailable. These values are mostly given by used localization system. Availability depends on possibility to estimate position of mobile device. In proposed criterion availability will be used as percentile of time for which the service is available.

5.2 Localization Accuracy

Localization accuracy is also given by capabilities of localization system. It represents how accurate is the position estimate and it is mostly given in meters [8, 9]. Accuracy depends mostly on environment and radio signal properties [10]. Accuracy can be changed because of these factors:

- Dynamic changes in environment (signal fluctuations, multipath).
- Topology of network (number and position of reference nodes).
- Weather and morphology of environment.

Accuracy can be divided into two parts – horizontal and vertical accuracy. Vertical accuracy is important in indoor environment and also in cases when also vertical position is needed (for example in parking houses and multi level crossroads).

Accuracy of localization based services can be given by range of feasible errors, which are acceptable by the application. This is due to fact that different applications have different demands on localization accuracy.

5.3 Time of Response

This parameter, same like the two before, depends on localization system properties. Response time is the time from localization request send from the mobile device till response reception on the mobile device and is given in seconds [9, 11].

This parameter is not given only by computational complexity of localization algorithm, but also by time which is needed to transfer data from mobile device to relevant part of the system and back to the mobile device.

5.4 Position Report Frequency

Position report frequency is very important parameter in LBS [11]. For tracking applications it will be very good if position of mobile device is estimated in short periods of time, which means with high frequency. High frequency of position reports can cause overload of network because of data transfer between mobile device and localization server. It is important to find tradeoff between available network resources and position report frequency, which will be high enough for specific application requirements.

6 Criterion to Evaluate QoS of LBS

In this section criterion k which was proposed to evaluate QoS of LBS will be introduced. For this purpose the following QoS parameters have been chosen horizontal accuracy, vertical accuracy, position report frequency, response time and availability of localization technology.

It is important to notice that each one of chosen parameters has different importance for different LBS. This is the reason why weights for the parameters will be different for different services. Since availability is the most important parameter, its weight will be set to $W_1=1$.

Importance will be assigned to each of the previously introduced parameters based on services demands. Since we have four parameters, importance value will be x will achieve values from 1 to 4. The most important parameter will have importance $x=4$ and less important parameter will have importance $x=1$.

Importance will be assigned based on service demands and also based on type of user. Importance of parameters will differ in cases when user is pedestrian and user is in vehicle. Weights will be represented by these symbols:

W_2 – weight of horizontal accuracy parameter,
W_3 – weight of vertical accuracy parameter,
W_4 – weight of position report frequency parameter,
W_5 – weight of response time parameter.

Weights of QoS parameters will be calculated based on its importance using:

$$W_i = \frac{x_i}{\sum_i x}, \; i = 2, 3, 4, 5 \tag{1}$$

where W_i represents weight of i-th parameter and x_i [-] represents importance of i-th QoS parameter. Weights of QoS parameter can be seen in Table 1. As can be seen from Table 1, weights of QoS parameters are defined for different types of LBS services based on division introduced in [4]. Weights are chosen based on assumed importance of given parameter.

Table 1. Weights and importance of QoS parameters for different services

Service	User type		Horizontal accuracy	Vertical accuracy	PRF	Response time
Emergency	**Pedestrian**	***W***	0,4	0,1	0,2	0,3
	Vehicle	***W***	0,3	0,1	0,4	0,2
Navigation	**Pedestrian**	***W***	0,4	0,3	0,2	0,1
	Vehicle	***W***	0,4	0,3	0,2	0,1
Information	**Pedestrian**	***W***	0,3	0,2	0,1	0,4
	Vehicle	***W***	0,2	0,1	0,3	0,4
Tracking and management	**Pedestrian**	***W***	0,4	0,1	0,2	0,3
	Vehicle	***W***	0,4	0,1	0,3	0,2
Billing	**Pedestrian**	***W***	0,2	0,1	0,3	0,4
	Vehicle	***W***	0,2	0,1	0,3	0,4

Now when weights for all parameters and services were defined it is important to define range of parameters. This defines range of values which are sufficient for each parameter based on service requirements [12].

These parameters need to be defined for pedestrian and also for vehicle, but for vehicle there can be different values in urban environment and outside the city. This is due to differences in speed limitations [13, 14]. Ranges of values are shown in Table 2 and Table 3 for pedestrians and vehicles respectively. Range of values for different parameters were chosen based on values introduced in [1], [12-14].

In next step it is important to normalize values from Table 2 and Table 3. For this purpose percentiles will be used in calculations, so values 0 - 100% can be achieved. For

Table 2. Range of QoS parameters for pedestrian

Service	Horizontal accuracy	Vertical accuracy	PRF	Response time
Emergency	0-20m	0-10m	0-14,5 s	0-1s
Navigation	0-50m	0-10m	0-1,2 min	0-5s
Information	0-200m	0-10m	0-2,4 min	0-5s
Tracking and management	0-50m	0-10m	0-36 s	0-3s
Billing	0-300m	0-10m	0-3,6 min	0-5s

Table 3. Range of QoS parameters for vehicles

Service	Horizontal accuracy	Vertical accuracy	PRF		Response time
			Urban environment	Outside city	
Emergency	0-20m	0-10m	0-1,4 s	0-0,9 s	0-1s
Navigation	0-100m	0-10m	0-21,6 s	0-3,5 s	0-5s
Information	0-500m	0-10m	0-36 s	0-22,5 s	0-5s
Tracking and management	0-200m	0-10m	0-14,4 s	0-9 s	0-3s
Billing	0-500m	0-10m	0-36 s	0-22,5 s	0-5s

example maximum allowed localization error for rescue services is 20m which represents 100%, so 100% will represent maximum admissible value for each parameter.

Criterion k to evaluate QoS of location based services can be calculated using this formula:

$$k = W_1 \cdot p_1 \cdot \left(W_2 \cdot p_2 + W_3 \cdot p_3 + W_4 \cdot p_4 + W_5 \cdot p_5\right), \tag{2}$$

where W_1 - W_5 stands for weights assigned to the parameters based on provided service, a represents availability of LBS, b represents horizontal accuracy, c stands for vertical accuracy, d represents PRF and e is response time.

Values of coefficients $p_2 \div p_5$ will be inserted into formula as number between 1 and 3. If coefficient is equal to 1 it represents ideal condition, for example 0 m = 0 % error. On the other hand, coefficient equal to 2 represents maximum allowed error. When error of parameter is in rage of sufficient errors, then coefficients can be computed as:

$$p_i = 1 + \frac{E_R}{E_M}, i = 2,3,4,5, \tag{3}$$

where p is one of parameters $p_2 \div p_5$, E_R stands for real error and E_M represents maximum error which is sufficient for the given LBS.

In case that error is higher than maximum error sufficient for LBS, value of parameter will be set to 3 which mean that parameter is not feasible for the provided service [15].

Parameter p_1 availability of LBS represents specific case. This parameter can achieve values from 0% to 100%, means that localization technology is available or not. With lower availability of technology, percentile of parameter decrease as well. Worst case is 0% ant it means that technology is not available at all. In this case it represents insufficient parameter [15]. When availability $p_1 = 1$ the localization technology is available at 100% of time, $p_1 = 2$ will represent case when technology is available just for 1% of time and $p_1 = 3$ means that localization technology is not available at all.

Computation of parameter k will be shown in the model situation [15]. It is general modeled situation. Values of particular parameters of positioning systems are also

modeled, but they are in correlation with real values. In this model situation, we assume LBS can achieve position estimate from three different localization techniques. Provided service will be emergency service for vehicles, for example information about car accident. In Table 4 requested values of QoS parameters are shown together with QoS parameters achieved by different localization systems.

Table 4. QoS parameters achieved by localization systems

		Availability	Horizontal accuracy	Vertical accuracy	PRF	Response time
Requested QoS parameters		100%	0-20 m	0-10 m	0-1,4 s	0-1 s
Real QoS parameters	**GPS**	99%	7 m	10 m	0,7 s	0,5 s
	GSM	99%	300 m	10 m	2 s	2 s
	WiFi	30%	15 m	7 m	2 s	2 s

In Table 1, it can be seen that weights for QoS parameters are $W_2=0,3$; $W_3=0,1$; $W_4=0,4$ and $W_5=0,2$. Then using formula (2) we can calculate QoS criterion k for each available localization system.

$$k_{GPS} = 1 \cdot 1,01 \cdot (0,3 \cdot 1,35 + 0,1 \cdot 2 + 0,4 \cdot 1,5 + 0,2 \cdot 1,5) = 1,52005$$
$$k_{GSM} = 1 \cdot 1,01 \cdot (0,3 \cdot 3 + 0,1 \cdot 2 + 0,4 \cdot 3 + 0,2 \cdot 3) = 3,045$$
$$k_{WiFi} = 1 \cdot 1,7 \cdot (0,3 \cdot 1,75 + 0,1 \cdot 1,7 + 0,4 \cdot 3 + 0,2 \cdot 3) = 4,195$$

From results of criterion k can be seen that best localization system available in model situation is GPS. WiFi and GSM localization systems suffer from low PRF and high response time. WiFi localization system achieved the worst results because of low availability.

7 Conclusion and Future Work

In this paper novel criterion to evaluate QoS of different LBSs was proposed. Proposed criterion is based on assumption that each service has different demands on accuracy and response time and also frequency of position report.

It is also clear that different demands on LBSs are given by user speed. This is the reason why different values of parameters were introduced for pedestrian and vehicles.

In future criterion will be tested on real localization systems and then can be implemented and used to help to choose the localization technology which will be sufficient for provided services.

Acknowledgements. This work was partially supported by the Slovak Research and Development Agency under contract No. LPP-0126-09 and by the Slovak VEGA grant agency, Project No. 1/0392/10.

References

1. Schiller, J., Voisard, A.: Location – Based Services, p. 266. Morgan Kaufann Publishers, San Francisco (2004) ISBN 1-55860-929-6
2. Steinger, S., Neum, M., Edwardes, A.: Foundations of Location Based Service, p. 28. University of Zurich, Zurich (2011)
3. Kupper, A.: Location-based Services: Fundamentals and Operation, p. 386. John Wiley & Sons Ltd., England (2005) ISBN-13 978-0-470-09231-6
4. Shek, S.: Next – Generation Location – based Services for Mobile Devices, p. 66 (2010)
5. Machaj, J., Brida, P.: Performance Comparison of Similarity Measurements for Database Correlation Localization Method. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS, vol. 6592, pp. 452–461. Springer, Heidelberg (2011) ISBN 978-3-642-20041-0
6. Filjar, R., Bušić, L., Dešić, S., Huljenić, D.: LBS Position Estimation by Adaptive Selection of Positioning Sensors Based on Requested QoS. In: Balandin, S., Moltchanov, D., Koucheryavy, Y. (eds.) NEW2AN 2008. LNCS, vol. 5174, pp. 101–109. Springer, Heidelberg (2008)
7. Filjar, R., Busic, L., Pikija, P.: Improving the LBS QoS through Implementation of QoS Negotiation Algorithm, Croatia, Zagreb, p. 4 (2008)
8. Kolodziej, K.W., Hjelm, J.: Local Positioning Systems: LBS Applications and Services, p. 462. Taylor & Francis Group, London (2006)
9. Hongying, Y.: Location Based Service: T-109.551. In: Research Seminar on Telecommunications Business II. Univ. of Technology, Helsinki (2002)
10. Krejcar, O.: User Localization for Intelligent Crisis Management. In: Proceedings of the 3rd IFIP Conference on Artificial Intelligence Applications and Innovation (AIAI 2006), Athens, Greece, pp. 221–227 (2006)
11. Busic, L., Filjar, R.: The Role of Position Reporting Frequency in LBS QoS Establishment, Zagreb, Croatia, p. 5
12. Beinat, E., Diaz, E.: Location Services and Accuracy: An analysis for field work applications, p. 23. Vrije Universteit, Amsterdam (2004)
13. ASCOM: Aspects of Geo-Location in Mobile Networks. White Paper, p. 11 (November 2010)
14. Sahinoglu, Z., Gezici, S., Guvenc, I.: Ultra-wideband Positioning Systems, p. 263. Cambridge University Press, New York (2008) ISBN-13 978-0-511-43816-5
15. Tatarova, B.: Implementation of Location Based Services to Intelligent Transport Systems. Diploma thesis, University of Zilina, Slovakia (2011) (in Slovak)
16. Mikulecky, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, pp. 523–528. Univ. Houston, Houston (2009)
17. Tucnik, P.: Optimization of Automated Trading System's Interaction with Market Environment. In: Forbrig, P., Günther, H. (eds.) BIR 2010. LNBIP, vol. 64, pp. 55–61. Springer, Heidelberg (2010)
18. Pindor, J., Penhaker, M., Augustynek, M., Korpas, D.: Detection of ECG Significant Waves for Biventricular Pacing Treatment. In: The 2nd International Conference on Telecom Technology and Applications, ICTTA 2010, Bali Island, Indonesia, March 19-21, vol. 2, pp. 164–167. IEEE Conference Publishing Services, NJ (2010), doi:10.1109/ICCEA.2010.186, ISBN 978-0-7695-3982-9

Proposal of User Adaptive Modular Localization System for Ubiquitous Positioning

Jozef Benikovsky, Peter Brida, and Juraj Machaj

University of Zilina, Faculty of Electrical Engineering,
Department of Telecommunications and Multimedia, Univerzitna 1, 010 26 Zilina, Slovakia
{jozef.benikovsky,peter.brida,juraj.machaj}@fel.uniza.sk

Abstract. The paper deals with general concept of modular and portable localization system that utilizes existing radio network infrastructure. The modularity means multiple independent collections of algorithms and technologies that allow determining geographical position in various radio and geographical environments. The portability of the system is provided by implementation into small pocket-sized device. The environments are characterized mostly by the parameters of available radio systems such as Global System for Mobile Communications (GSM), Institute of Electrical and Electronics Engineers (IEEE) 802.11 standard-based ones or Global Navigation Satellite Systems (GNSS). Suitable software and utilization of a portable device equipped with necessary hardware can turn the system into the provider of ubiquitous positioning service.

Keywords: Localization, Localization system, Modular localization system, Positioning, Ubiquitous positioning.

1 Introduction

Localization represents the process that determines position of an object. The position is usually estimated with some precision and probability. There are many reasons why emphasis is put on localization nowadays. Firstly, it is important part of life-saving activities by emergency calls or accidents. Apart from that, there are numerous industrial, commercial, intelligent or entertainment services based on localization continuously created. The services do not use only position itself but attempt to be aware of object position, context or situation and allow obtaining wide range of information anywhere and anytime. Localization service has evolved into value added service (VAS). This is enabled by progress of ubiquitous computing (or pervasive computing) and portable devices.

There are many localization algorithms such as cell (sector) identification or cell of origin, centroid, trilateration, triangulation, multilateration, traversing, fingerprinting, statistical methods, numerical methods, range-free methods etc. The algorithms use various inputs, mostly data measured by some gauge. The common measurements are *Received Signal Strength* (RSS), *Time of Arrival* (ToA), *Time Difference of Arrival* (TDoA) or *Angle of Arrival* (AoA) [1], [2]. There are many others like *Phase of*

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 391–400, 2012.

Arrival (PoA), *Bit Error Rate* (BER), *Roundtrip Time of Flight* (RToF), *Enhanced Observed Time Differences* (E-OTD), *Observed Time Difference of Arrival* (OTDoA), *Idle Period Downlink* (IPDL), *Advanced Forward Link Trilateration* (AFLT) *Enhanced Forward Link Trilateration* (EFLT), *Multipath Power Delay Profile* (MPDP) or *Channel Impulse Response* (CIR) [3].

The aforementioned algorithms and measurements require radio communication networks that allow actually performing the measurements. There are many of them available beginning from *Global Navigation Satellite Systems* (GNSS), cellular networks (GSM), IEEE 802.11, ad hoc networks, *Ultra-Wide Band* (UWB), *Radio Frequency Identification* (RFID), *World Interoperability For Microwave Access* (WiMAX), *Wireless Sensor Networks* (WSN), *Indoor GPS* (iGPS) etc.

The principal requirement to allow seamless provisioning of localization service is the ability to localize an object in many different environments and in environments where people spend lot of time such as cities, villages, buildings or parking areas. It is annoying and even application critical for some services to have some areas, where user cannot be localized. For instance, applications such as position tracking or navigation highly rely on availability of the localization service. There are some systems that allow localization in multiple environments with some drawbacks such as:

- *High Sensitivity GPS* [4] - precision decreases when satellites are not in line-of-sight
- *Rosum TV-GPS* or *Rosum TV-GPS Plus* [5] – require extra cost for radio network infrastructure and integration of custom measurement module (*Rosum TV Measurement Module*)
- *Real Time Location Server* [6] – based on indoor positioning only, but with extra device equipped with GPS and capable of running Ekahau Client® software can work outdoor as well
- *Aeroscout* [7] – requires extra infrastructure (RFID, location receivers and exciters)
- *Locata* [8] – requires extra infrastructure (pseudolites)

For these reasons this paper proposes modular localization system, which is designed to maximize coverage of the localization service by being reachable over large area together with being portable and with utilization of existing infrastructure.

The paper is organized as follows. Section 2 explains the term Modular Localization System and defines the basic architecture of the system. Section 3 deals with requirements for implementation and shows how to perform the actual implementation. Section 4 explains the importance of testing the system and proposes parameters to take into consideration and testing environments. The paper is finished with acknowledgments, conclusion and references sections.

2 Modular Localization System

System modularity is provided by components called localization modules. A localization module allows positioning by utilization of particular radio communication technology such as GSM, IEEE 802.11 wireless networks or GNSS systems. The

modules are managed by decision-making core of the system that recognizes current parameters of the environment such as availability and quality parameters of present radio communication networks. Afterwards, it processes this data and determines the best localization procedure.

Practically, the data are measured by mobile station, which in this case, represents a device capable to measure data from one or more radio networks. If the device cannot determine the location by itself (e.g. GNSS is not available or its accuracy is poor), it transfers the measurements to localization server and waits until the server responds with estimated position. The configuration is referred as mobile-assisted positioning. The example of such configuration is shown in Fig. 1.

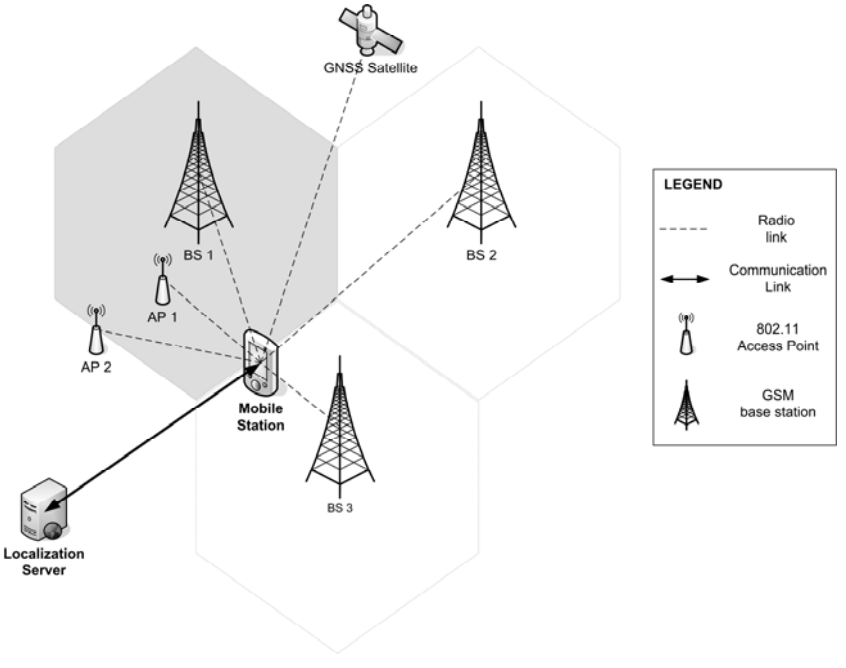


Fig. 1. Example of modular localization system in environment with one reachable GNSS satellite, three GSM base stations (*BS 1*, *BS 2*, *BS 3*) and two 802.11 access points (*AP 1*, *AP 2*)

The modular localization system can be then defined as system of self-reliant localization modules managed by the decision-making module in order to achieve reasonable localization quality in the majority of environments. The obvious question is how to define the quality. Because localization system is software-based the ISO 9126 model [8] for software quality seems appropriate. It consists of six main characteristics such as efficiency, functionality, maintainability, portability, reliability and usability. These are further explained in section 4.1.

The users of the system are expected to be people that want to localize themselves as well as location-based service (LBS) providers.

The security of the proposed system can be divided into two layers. Firstly it is access to service using authentication and authorization. Only the users that have valid account and the account has permissions to perform localization can be localized. Secondly, it is security of transmission of sensitive data, which user position certainly represents, provided by encrypted communication channel. Extra security is achieved via communication model that always initiates connection from the user. This prevents malicious service providers from determining user position without their perception. Because the system can be implemented by multiple service providers there can be a malicious one that would try to misuse the position of users that asked for it. There are models which can overcome this problem, such as non-cooperative model, centralized trust third party model or Peer-to-peer cooperative model [9].

The system is composed of two basic components – *Mobile Station* (MS) and *Localization Server* (LCS). Mobile station represents end user device that is equipped with necessary hardware and localization software and cooperates with LCS in order to facilitate localization. LCS represents decision-making core of the system that contains necessary communication interfaces, user interface for maintenance, localization algorithms and databases with measurements and radio maps necessary for the algorithms. LCS can consist of one or more servers as shown in Fig. 2, mainly for performance and scalability reasons.

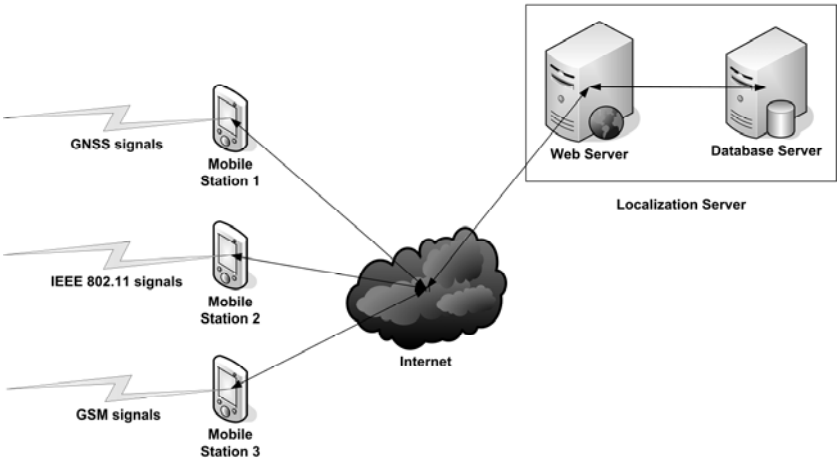


Fig. 2. Modular localization system architecture displaying mobile stations that are detecting various radio signals and communication between mobile stations and localization server via internet

The system is designed to utilize fingerprinting-based algorithms. The algorithms require so called radio map, which is database of signal parameter measurements paired with position where measurement was taken. These data that are stored in database (on database server) and are divided into three main entities – spot, measurement and reference stations. The relations between entities are shown in Fig. 3.

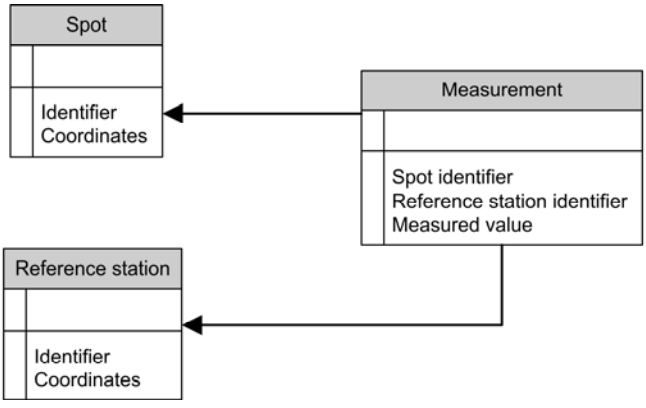


Fig. 3. Relations between basic entities for radio map

Spot represents the particular position where a measurement was taken. Reference station represents station with known position that transmits radio signals used for localization (e.g. base station in GSM or access point in IEEE 802.11 network). Measurement contains measured signal parameters such as RSS or ToA paired with spot where measurement was taken and reference station for which the signal parameters were measured.

3 Implementation

3.1 Hardware Platform

In order to support localization in any environment, the system would have been equipped with all kinds of radio signal receivers to measure data from any kind of radio network. However, only electromagnetic waves shorter than 1 meter (very high frequency and higher) are used. Moreover, the price of such device would be tremendous. Therefore, the device that contains radio receivers for major popular radio networks and is available to potential users without any extra costs is smart phone.

The smart phone nowadays is usually equipped with GNSS receiver, mobile phone network receiver, IEEE 802.11 network receiver and possibly others. The mentioned radio networks are also the best candidates for localization purposes, because they are widely used. The mostly used GNSS system nowadays is GPS, however other systems are coming e.g. Glonass. The localization using GNSS is part of modular localization system however algorithm that calculates the position is part of device hardware and is used by the system. The best candidate from mobile phone networks is GSM because of its about 80% market share [10]. Networks based on IEEE 802.11 standards are also good candidate because of their wide use in buildings as well as outdoors. The number of GSM base stations and IEEE 802.11 networks access points in range for the inhabited areas is usually sufficient in order to obtain reasonable localization results.

There is a great set of smart phones equipped by all these radio technologies however the operating system that they have installed and interoperability between devices of different device vendors is not possible for some of them (Apple iOS, RIM Blackberry OS). The greatest market share in first half of 2011 belongs to Android with about 50% [11], [12]. Android provides useful programming interface and allow taking advantage of low level hardware functions which are harder to use or even blocked in other operating systems. Therefore smart phones designed for Android operating systems with necessary hardware are the best candidates for implementation of the end user device.

The other part of localization system is localization server. Because there is only one in the system, it is not important on which hardware or software it is based. However, it must provide communication interface capable of handling different client software versions and hardware capabilities.

3.2 Algorithms

When usual algorithms for position estimation are used e.g. triangulation or trilateration, the results are significantly influenced by radio channel properties such as path loss, reflections, multipath or delay spread. These factors can be eliminated by using fingerprinting (also called database correlation or pattern matching). Moreover, it uses current network infrastructure and does not require extra costs for modifications. The measured parameters can be Base Station Identifier (BSI), Received Signal Strength (RSS), Signal to Noise Ratio (SNR), Link Quality Indicator (LQI), power and Response Rate (RR) [13].

4 Testing

One of the key components of the system development lifecycle is testing. It can be divided into test environment and test scenarios. Test environment represents set of devices (base stations, mobile stations), set-up parameters (transmission power, interference), data (input data for algorithms such as radio map) and other factors prepared before system is tested. Test environments should be as close as possible to real environment; otherwise the results may not be applicable in practice because of many unexpected factors. Even one such factor can cause total uselessness of the entire system. Test scenarios represent activities in particular order created to verify whether system works as expected. The definition of expected results may be part of system proposal or ad hoc. The activities should provide statistically same results when repeated in test environments. The results should be somehow comparable in various environments as well as with existing solutions or algorithms.

In order to meet aforementioned criteria, we have defined some expected values for results and proposed basic environments and activities necessary to verify the system described in sections below.

4.1 ISO Quality Parameters

This section contains set of ISO quality characteristics created to measure quality of the modular localization system. They are shown in Table 1. Some of the expected values are given to fulfill national or international regulatory criteria for localization for emergency calls. Value of average localization error is given to match most precise criteria given by European Union for city and indoor areas [14] and thus matching the criteria for rural and suburban areas. Circular Error Probability is defined to fulfill the requirements in United States of America [15]. There are no other requirements for localization service precision in other countries, so the modular localization system should be applicable worldwide after meeting these two requirements.

Table 1. Quality parameters

ISO characteristic	Measured parameter	Expected value or verification criteria
Efficiency	Time to first fix	< 10 seconds
	Average localization time	< 1 second
	Minimum number of localizations performed by server per second	100
Functionality	Average localization error	10 – 50 m
	Circular Error Probability	67% for 50m circle radius; 95% for 150m circle radius
	Security	User without permissions cannot use any localization function of Localization server
Maintainability	Mobile Station can easily setup server connection	User enters network address or hostname of localization server
	Localization algorithm parameters can be configured	Localization server administrator can change behavior of algorithms by setting necessary parameters
Portability	Database portability	Localization server administrator can change database server any time
	Installability	Mobile Station software can be installed on currently the most used Android version and newer.
Reliability	Invalid localization request handling	Localization requests with invalid data are ignored (do not execute localization algorithms)
	Error handling	When error occurs, user is clearly notified what happened and what to do
Usability	Understandability	User interface is intuitive

4.2 Test Environments

The localization services are usually tested in three basic environments according to building density – rural areas, suburban areas and cities. However, there are other factors which influence localization such as weather, time, number of moving objects such as people or cars, materials from which a building is constructed, interference and others related to environment [16, 17]. It is also radio map parameters such as age of measurements, density of measurements etc. All of these factors have to be taken into account when testing.

4.3 Test Activities

The basic test activity is self-localization workflow. It represents set of activities that system user has to perform in order to localize himself. They are:

- user executes localization application in mobile station
- user configures connection to localization server, if not already configured before
- users enters request for localization
- user is being notified of current progress
- user gets localization result in the most understandable way (e.g. earth map) or message explaining what happened if error occurs

Besides the workflow above, ISO quality characteristics from Table 1 have to be checked thoroughly. There are also many other activities to check, however these are the most important.

5 Conclusion

The proposed modular localization system provides solution for localization with GPS, GSM and IEEE 802.11 standard-based networks. It is also ready for the future modifications by possibility to add new radio networks as new modules or easily modify algorithms that are deployed on server. Moreover, the paper contains the proposal how to implement and test modular localization system. The idea of modular localization system has been explained together with reasons why it is needed nowadays, who are the possible users and what has to be investigated when someone wants to make solution with similar implementation. The ideas how to make system widely usable are also explained.

Acknowledgements. This work was partially supported by the Slovak VEGA grant agency, Project No. 1/1027/12 “The research of modular positioning system” and by the Slovak Research and Development Agency under the contract No. LPP-0126-09 and by

Centre of excellence for systems and services of intelligent transport II., ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

References

1. Sahinoglu, Z., Gezici, S., Guvenc, I.: Ultra-wideband Positioning Systems: Theoretical Limits, Ranging Algorithms, and Protocols. Cambridge University Press, Cambridge (2008)
2. Akgul, F.O., Heidari, M., Alsindi, N., Pahlavan, K.: Monitoring and Surveillance Techniques for Target Tracking. In: Mao, G., Fidan, B. (eds.) Localization Algorithms and Strategies for Wireless Sensor Networks, pp. 54–95. IGI Global, London (2009)
3. Brida, P., Cepel, P., Duha, J.: Mobile Positioning in Next Generation networks. In: Kotsopoulos, S.A., Ioannou, K.G. (eds.) Handbook of Research on Heterogeneous Next Generation Networking, Innovations and Platforms, pp. 223–252. IGI Global, London (2009)
4. Lachapelle, G.: GNSS Indoor Location Technologies. Journal of Global Positioning Systems 3(1-2), 2–11 (2004)
5. Rosum, http://www.rosum.com
6. Ekahau, http://www.ekahau.com
7. Aeruscout, http://www.aeroscout.com
8. International Organization for Standardization: ISO/IEC 9126-1:2001, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22749
9. Rodden, T., Friday, A., Muller, H., Dix, A.: A Lightweight Approach to Managing Privacy in Location-Based Services, http://www.cs.bris.ac.uk/Publications/Papers/2000743.pdf
10. GSM Association: Market Data Summary (Q2 2009) - Connections by Bearer Technology, http://www.gsmworld.com/newsroom/marketdata/market_data_summary.html
11. Android Market Share Growth Accelerating, Nielsen Finds, http://www.pcworld.com/article/226339/android_market_share_growth_accelerating_nielsen_finds.html
12. Gartner: Android market share to near 50 percent, http://news.cnet.com/8301-13506_3-20051610-17.html#ixzz1Zw5tDXgS
13. Kjærgaard, M.B.: A Taxonomy for Radio Location Fingerprinting. In: Hightower, J., Schiele, B., Strang, T. (eds.) LoCA 2007. LNCS, vol. 4718, pp. 139–156. Springer, Heidelberg (2007)
14. ETSI TS 102 650 V1.1.1: Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Analysis of Location Information Standards produced by various SDOs, http://webapp.etsi.org/action/PU/20080805/ts_102650v010101p.pdf

15. FCC: Guidelines for Testing and Verifying the Accuracy of Wireless E911 Location Systems, http://www.fcc.gov/Bureaus/Engineering_Technology/Documents/bulletins/oet71/oet71.pdf
16. Krejcar, O.: Testing and Evaluating of Predictive Data Push Technology Framework for Mobile Devices. In: Cai, Y., Magedanz, T., Li, M., Xia, J., Giannelli, C. (eds.) Mobilware 2010. LNICST, vol. 48, pp. 276–283. Springer, Heidelberg (2010)
17. Deligiannis, N., Kotsopoulos, S.: Mobile positioning based on existing signaling messaging in GSM networks. In: Proc. of 3rd International Mobile Multimedia Communications Conference (MSAN), Nafpaktos, Greece, August 27-29 (2007)
18. Mikulecky, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, pp. 523–528. Univ. Houston, Houston (2009)
19. Tucnik, P.: Optimization of Automated Trading System's Interaction with Market Environment. In: Forbrig, P., Günther, H. (eds.) BIR 2010. LNBIP, vol. 64, pp. 55–61. Springer, Heidelberg (2010)
20. Pindor, J., Penhaker, M., Augustynek, M., Korpas, D.: Detection of ECG Significant Waves for Biventricular Pacing Treatment. In: The 2nd International Conference on Telecom Technology and Applications, ICTTA 2010, Bali Island, Indonesia, March 19-21, vol. 2, pp. 164–167. IEEE Conference Publishing Services, NJ (2010), doi:10.1109/ICCEA.2010.186, ISBN 978-0-7695-3982-9

User Adaptivity in Smart Workplaces

Peter Mikulecky

Faculty of Informatics and Management, University of Hradec Králové,
50003 Hradec Králové, Czech Republic
peter.mikulecky@uhk.cz

Abstract. The area of smart workplaces can be considered as one of the most challenging areas for Ambient Intelligence applications. Special focus here can be detected on such typical smart workplaces as smart offices or smart classrooms. In our paper we present some related works specifying more the area of smart workplaces, with accent on their adaptability features. Further on, some approaches to possibilities of user adaptivity concept implementation in smart environments, especially in some types of smart workplaces, will be presented and briefly discussed.

Keywords: Ubiquitous environments, Ambient intelligence, User adaptivity, Smart workplaces, Smart offices.

1 Introduction

Smart workplaces belong to one of the most interesting areas for Ambient Intelligence (*AmI*) applications. These applications aim to enhance the original workplaces by ubiquitously working information and communication technology hidden (or disappeared) in the environment and naturally and intuitively used by the users surrounded by the environment to help the in solving routine but also very peculiar problems. Such a vision cannot be realistic without achieving certain degree of adaptivity of the environment to the users' needs.

According to [1], *AmI* systems represent a new generation of systems showing the following characteristics. *AmI* systems are:

- *invisible*, i.e., embedded in things like clothes, watches, glasses, etc.,
- *mobile*, i.e., being carried around,
- *context aware*, i.e., equipped with sensors and wireless communication interfaces making it possible to scan the local environment for useful information and spontaneously exchange information with similar nodes in their neighborhood,
- *anticipatory*, i.e., acting on their own behalf without explicit request from a user,
- *communicating naturally* with potential users by voice and gestures instead by keyboard, mouse and text on a screen,
- *adaptive*, i.e., capable of reacting to all kinds of abnormal exceptional situations in a flexible way without disruption of their service.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 401–410, 2012.

In our paper we wish to present some related works specifying more the area of smart workplaces. Further on, an approach to the user adaptivity concept implementation in smart environments, especially in smart workplaces, will be presented. It is based on the concept of *"ad hoc collaborating groups of humans and agents"* introduced by Misker and others [2]. Some other possible approaches, which are still not fully elaborated, will be mentioned shortly.

2 Smart Workplaces

One of the most challenging areas for Ambient Intelligence applications seem to be smart workplaces, with special focus on smart offices but also on smart classrooms. Cook, Augusto, and Jakkula note in [3], that while *Ambient Intelligence* incorporates aspects of context-aware computing, disappearing computers, and pervasive / ubiquitous computing into its sphere, there is also an important aspect of intelligence in this field. As a result, *AmI* incorporates artificial intelligence research into its purview, encompassing contributions from machine learning, agent-based software, and robotics. *AmI* research can therefore include work on hearing, vision, language, and knowledge, which are all related to human intelligence, and there is where *AmI* differs from ubiquitous computing. By drawing from advances in artificial intelligence, *AmI* systems can be even more sensitive, responsive, adaptive, and ubiquitous, concludes Cook and her colleagues [3].

According to [4], a special focus in *AmI* seems to be put on *Smart Offices* and *Smart Decision Rooms*. As mentioned already in various papers (e.g., [1], [3], or [4]), decision making is one of the noblest activities of the human being. Most of times, decisions are made involving several persons (group decision making) in specific spaces (e.g. meeting rooms). On the other hand many scenarios demand fast decision making to achieve success.

Smart offices are defined in [5] as an environment that is able to adapt itself to the user needs, release the users from routine tasks they should perform, to change the environment to suit to their preferences and to access services available at each moment by customised interfaces. Smart offices handle several devices that support everyday tasks. They may anticipate user intentions, doing tasks on his behalf, facilitating other tasks, etc.

As it is stressed in [4], smart offices and decision rooms contribute to reducing the decision cycle, and offer connectivity wherever the users are, aggregating the knowledge and information sources. The topics as smart offices and intelligent meeting rooms are well studied and they intend to support the decision making activity, however, they have received a new perspective from the *AmI* concept. This concept enables a different way to look at traditional offices and decision rooms, where it is expected that these environments support their inhabitants on a smart way, promoting an easy management, efficient actions and, most importantly, to support the creation and selection of the most advantageous decisions.

There is already a number of interesting applications and other results in the area of smart workplaces that tackle various important problems necessary for further

development of this important field. We try to mention some of them, e.g., multi-criterial decision making, moving smart environment, or some smart educational applications later in this chapter.

Bühler [6] is using the name *Ambient Assisted Working* (*AAW*) for an *AmI* strategy that provides a flexible approach towards workplace adaptation for all, including people with disabilities and older people in the workforce. He also pointed out, that people in the workforce develop growing expertise and different abilities over time. Therefore they need tailored support systems at work keeping the efficiency and effectiveness and elements of prevention or adjustment to changing abilities.

In order to use *AAW*, the process has to start much earlier in a more inclusive way. Without knowing the exact demands of a future worker, the system needs to be designed. The flexible networking character of *AmI* provides the required flexibility. In [6] two scenarios are presented aiming to illustrate the *AAW* approach. *AmI* in a work environment can support very different user needs and can help to create working environments for all.

Macindoe and Maher in [7] bring an original view on motivation of agents in *intelligent environments* (*IE*). As modeling of intelligent environments via multi-agent systems seems to be the best way for studying their architectures and functionalities, their approach could be beneficial with its new ideas.

The authors of [7] argue, that *IEs* should adapt to, and be useful for, ordinary everyday activities; they should assist the user, rather than requiring the user to attend to them; they should have a high degree of interactivity; and they should be able to understand the context in which people are trying to use them and behave appropriately. An *IE* is essentially an immobile robot, but its design requirements differ from those of normal robots in that it ought to be oriented towards maintaining its internal space rather than exploring or manipulating an external environment [7].

Combining the ideas in *IEs* with motivated learning agents leads thus to a model for an *intrinsically motivated intelligent room*. Motivation can play a valuable role in the agent model for an intelligent room generally, not just in learning, because it provides a model for the pro-active characteristics that are desirable in *IEs*. In [7], a motivated agent model for an intelligent room is presented, that is motivated by novelty to learn and by competency to act.

With a somehow similar aim our papers [8] and [9] are focused on decision making of agents in multi-agent environments, with a special accent on multi-agent based modeling of smart environments. The ongoing research related to that is oriented on further study of *Multi-Criterial Decision Making* in autonomous decision making, especially when multiple entities (users or agents) are present at the same time. Solution of conflicts, negotiation, settings of user preferences, multiple objectives, setting priorities, etc. are the main areas of interest in our further research.

When the necessity of user preferences appears, usually new location-based services can be adapted to accomplish this task [10]. For this the ubiquitous system needs to know user profiles, likings, and habits. But in the case when the user moves, this information must be made available at the new location of the user. Either the user carries the data on wearable or portable computers or the smart environment takes responsibility for transporting them. Related to this, Bagci and others propose in

their paper [11] that a smart environment takes care for storing and sending the personal information. The person in this approach is always accompanied by a mobile virtual object in the smart environment. So location based services adapted to personal profiles can be offered. The paradigm of mobile agents, used in this approach, ideally fits into the decentralized approach. The mobile agent constitutes a virtual reflection of the user and carries personal information which enables the agent to perform various services for the user. Additionally the mobile agent can use the environmental information which is provided by the local ubiquitous system. Moreover, the movement of the mobile agent should be in this approach faster than the movement of the person. This fact helps to solve a couple of related problems.

When using scenarios in design of new technologies, the services and applications described in the various beneficial scenario elements comprise a wide range of functionalities, ranging from the automation of standard office tasks, like automatic meeting protocols to complete new services, like adaptive office environments. Röcker in [12] brings a survey of state-of-the-art application scenarios for smart office environments. Based on an analysis of ongoing research activities, representative functionalities and services of future office systems were extracted. The results of the analysis show, that the vision of smart office environments is not as vague and unclear as often argued, as current technological developments revolve around a few, clearly identifiable topics.

An idea of recognition of a current situation and behaviour of a user, as well as an unobtrusive satisfaction of his needs underlies the *Ambient Intelligence*. Integration of diverse computation, information and communication resources into a united framework is one of the important issues at design of ambient intelligence and it identifies the modern tendency to transition from smart devices to an ambient intelligent space. Multimodal interfaces provide natural and intuitively comprehensible interaction between a user and intellectual devices, which are embedded into the environment. All the means should be hidden, thus the user can see only the results of intellectual devices activities and concentrate attention on her/his work. Rondzhin and Budkov [13] describe a development of an intelligent meeting room as a distributed system with the network of intelligent agents (software modules), actuator devices, multimedia equipment and audio-visual sensors. The main aim of the room is providing of meeting or lecture participants with required services based on analysis of the current situation. Awareness of the room about spatial position of the participants, their activities, role in the current event, their preferences helps to predict the intentions and needs of participants. Context modelling, context reasoning, knowledge sharing are stayed the most important challenges of the ambient intelligent design of this kind of rooms.

One of challenging applications is without any doubts any *AmI* application bringing new ideas and approaches into educational process at every level of education. Educational environment certainly is a rather specific workplace deserving a special attention and focus. One of these applications is the *Smart Classroom* project [14]. It aims to build a real-time interactive classroom with tele-education experience by bringing pervasive computing technologies into traditional distance learning. The goal of *Smart Classroom* project is to narrow the gap between

the teacher's experience in tele-education and that in the traditional classroom education, by means of integrating these two currently separated education environments together. The used approach was to move the user interface of a real-time tele-education system from the desktop into the 3D space of an augmented classroom (called *Smart Classroom*) so that in this classroom the teacher could interact with the remote students with multiple natural modalities just like interacting with the local students.

A more general overview of the *AmI* possibilities in education brings our recently published book chapter [15]. The objective of the chapter is to identify and analyze key aspects and possibilities of *AmI* applications in educational processes and institutions (universities), as well as to present a couple of possible visions for these applications. A number of related problems are discussed there as well, namely agent-based *AmI* application architectures. Results of a brief survey among optional users of these applications are presented as well. The conclusion of this research was that introduction of *Ambient Intelligence* in educational institutions is possible and can bring us new experiences utilizable in further development of *AmI* applications.

Among the other interesting, but somehow non-traditional problems related to the smart offices area it is worthwhile to mention paper [16] where a music recommendation agent in a smart office to recommend music for users was proposed. The idea was that by collecting and analyzing the contextual information of users, the agent can automatically senses the mood of users and recommends music. This is a nice example of a higher degree of a smart environment adaptation towards the user needs. Another, in a sense "funny" application, is described in [17]. It is devoted to a description of a proactive user adapted application for pleasant wake-up. Nevertheless, there are vast of workplaces where "pleasant wake-up" can assure to provide further activities of a freshly awoken worker in a more safe way.

3 User Adaptivity in Smart Workplaces

Paramythis [18] explains the term "adaptation" and its various meanings in the literature as follows:

- *adaptation* in general refers to the idea of having a system that can be tailored to one's individual needs;
- *self-adaptation*, adding the capability on the part of the system to perform the tailoring itself;
- *adaptability*, incorporating the notion of the system's being able to carry out by itself most of the steps required to decide upon and effect adaptation; and
- *adaptivity*, denoting, in addition to the above, that the system is capable of acquiring the user model, and performing non-trivial mapping between the contents of the said model and the range of possible forms of tailoring at runtime.

When concerning the concept of user adaptivity in relation with smart environments, we should speak more about adaptability of the smart environment, as it is expected that the technologies behind are capable to carry out by itself most of the steps

required for an effective adaptation of the whole system to the user's needs. It is necessary to take into account that in the case of a smart workplace it is usually supposed that its user is a person whose main activity is or could be decision making, and that the system should support all the user's activities leading to a good, if not optimal decision.

In the case of ubiquitous decision supporting system the smart workplace should, among its other features:

- ensure broad but focused and personalized access to relevant information and knowledge resources, supporting thus both learning needs of the manager as well as creation of his/her decisions that must be of highest quality;
- offer as much relief from stress as possible by avoiding all the usual stressful situations (or more precisely their potential sources);
- ensure broad and up to date technical support for all technically based activities in the workplace.

In the ubiquitous environment, it is expected that higher interaction will be accomplished between the environment and its user, and the decision-making methods necessary to perform the managing activities in such an environment will also change according to the changes in the work processes and methods [19]. As a result, the roles of the decision support system that supports decision-making in the ubiquitous workplace will diversify and attain greater importance [20].

A very significant result seems to be that the connectivity and context-awareness function, the key natures of the ubiquitous environment, are helpful for decision-making of decision makers who are actually using the underlying ubiquitous decision supporting system (UDSS) in such a smart environment. It is expected that more meaningful and notable results will be drawn when the competencies of the UDSS become more diverse and personalized. This requirement directly implies the need for a higher degree of user adaptivity of the smart workplace.

Ambient intelligence environments may be considered as strongly related with multi-agent systems in that they can be adequately modeled. One of the most promising approaches of achieving higher user adaptivity is here based on the idea of so-called ad hoc agent environments.

An *ad hoc agent environment* [2] is a way for users to interact with an ambient intelligent environment. Agents are associated with every device, service or content. The user interacts with his environment as a whole, instead of interacting with individual applications on individual devices. Devices and services in the environment have to be more or less independent, which fits well with the notion that agents are autonomous. The research was oriented on how the user is able to interact with the environment in such a way that he/she would have the control over collaborating agents. Some experiments showed certain tension between the user being in control and the autonomy of agents. Therefore the notion of cooperating groups was introduced [2] as a way for users to gain control over which agents collaborate. Users can then establish connections between devices and content that are meaningful to them, in the context of their task.

When representing a smart environment as a multi-agent systems consisting of heterogeneous agents with various level of their autonomy, within such agent environment there can be groups of agents that cooperate and/or share a purpose. For example, we could identify

- an information providing agent and a display agent;
- a knowledge seeking and providing agent, a display agent, and a personal security assuring agent;
- the personal agents of a group of users together with a coordinating and tasks optimizing agents;
- a set of agents evaluating the user's learning needs and providing the necessary resources through a display agent;
- a set of agents that represent one user's personal devices; etc.

Within such a group agents can communicate on the base of the role they perform in the group. It is expected that the notion of a group can help users gain a better grip on their interactive environment by allowing them to make explicit the combinations of devices, services and content that are relevant to them. The tasks and activities a user is currently involved in are each represented by a different group. We believe that this architecture based on ad hoc groups of collaborating agents can be rather effective in providing user adaptivity of smart environments, especially smart workplaces surrounding managers on various levels.

The achievements of ambient intelligence postulate an adequate shift in thinking that concerns also managerial work [21]. Managers, in order to be able of producing the best possible strategic decisions, should have the right information in the right time. However, without having the appropriate knowledge the production of good decisions would not be easy, if not impossible. It is, therefore, quite sensible to think about such a managerial workplace, where the manager would have the best possible working conditions in various meanings of this formulation. An analysis of managerial needs for a qualitatively higher working environment, based on the idea of ambient intelligence, oriented on intelligent environment design, can be seen in [22].

In order to stress the role of knowledge and of its management in such intelligent environments of various kinds, we have to take into account that such environments inevitably need to be rich of knowledge; therefore a synergy of approaches and techniques from ambient intelligence as well as from knowledge management is necessary here. An analysis of what were the basic common features of intelligent (and therefore knowledge rich) environments, what were they relations to ambient intelligence approaches, and what other serious problems could arise, can be found in our paper [23].

As a matter of fact, a workplace needs not to be limited by walls of a room. Some ideas related to the concept of large-scale ambient intelligence implementation were formulated in [24]. The presented concept considers a large network of intelligent devices (ambient artifact) distributed in a number of places in large natural environment (forests, rivers, etc.) ubiquitously communicating each to others, enabling thus monitoring weather conditions, rivers level, rain intensity, and other possible sources of living environment instability unceasingly. The ideas are focused

on initial concepts in this area only; still it is not possible to present final solutions. However, as large network of intelligent devices appropriately configured in an outdoor environment can be understood as a specific workplace (e.g. for foresters, or water managers), the presented ideas certainly could be elaborated in more detail for particular outdoor applications.

We mentioned first attempts in this direction already long ago, in [25, 26]. The area of water management, or river basin management as its special case, can be a good application field for outdoor adaptive smart workplaces. As we mentioned already, one of the important tasks in assuring higher user adaptivity of the respective smart workplace is his/her localization. This can be extremely important especially in outdoor applications of the just mentioned type. One new method in this direction has been proposed by Kotzian and others [27]. In such applications mobile devices can be used extensively [28]. However, further intensive research in this direction will be necessary.

4 Conclusions

In the paper, we tried to present some approaches towards user adaptivity concept implementation in smart environments, especially the smart workplaces, namely of the type of smart offices, but not only these. The concept based on Misker's and others [2] approach of ad hoc groups of collaborating users and agents seems to be applicable, however it still needs some more practical research, evaluation, and testing. We believe that the near future brings promising results.

Acknowledgements. This research has been partially supported by the Czech Scientific Foundation Project No. P403/10/1310 *SMEW - Smart Environments at Workplaces.*

References

1. Weber, W., Rabbaey, J., Aarts, E. (eds.): Ambient Intelligence. Springer, Berlin (2005)
2. Misker, J.M.V., Veenman, C.J., Rothkrantz, L.J.M.: Groups of Collaborating Users and Agents in Ambient Intelligent Environments. In: Proc. ACM AAMAS 2004, pp. 1318–1319. ACM, New York (2004)
3. Cook, D.J., Augusto, J.C., Jakkula, V.R.: Ambient Intelligence: Technologies, Applications, and Opportunities. Pervasive and Mobile Computing 5, 277–298 (2009)
4. Ramos, C., Marreiros, G., Santos, R., Freitas, C.F.: Smart Offices and Intelligent Decision Rooms. In: Handbook of Ambient Intelligence and Smart Environments, pp. 851–880. Springer Science+Business Media, Berlin (2010)
5. Marsá-Maestre, I., de la Hoz, E., Alarcos, B., Velasco, J.R.: A Hierarchical, Agent-based Approach to Security in Smart Offices. In: International Conference on Ubiquitous Computing: Applications, Technology and Social Issues, CEUR Workshop Proceedings, vol. 208, Madrid, Spain (2006)
6. Bühler, C.: Ambient Intelligence in Working Environments. In: Stephanidis, C. (ed.) UAHCI 2009, Part II. LNCS, vol. 5615, pp. 143–149. Springer, Heidelberg (2009)

7. Macindoe, O., Maher, M.L.: Intrinsically Motivated Intelligent Rooms. In: Enokido, T., Yan, L., Xiao, B., Kim, D.Y., Dai, Y.-S., Yang, L.T. (eds.) EUC-WS 2005. LNCS, vol. 3823, pp. 189–197. Springer, Heidelberg (2005)
8. Tučník, P.: Multicriterial Decision Making in Multiagent Systems – Limitations and Advantages of State Representation of Behavior. In: Data Networks, Communications, Computers, DNCOCO 2010, pp. 105–110. WSEAS Press, Athens (2010)
9. Tučník, P., Mikulecký, P.: Multicriteria Adaptation Mechanism of Agent Environment Behavior in Ambient Intelligence Services. In: Proc. of the International Conference on Applied Computer Science, pp. 401–405. WSEAS Press, Athens (2010)
10. Machaj, J., Brida, P.: Performance Comparison of Similarity Measurements for Database Correlation Localization Method. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS (LNAI), vol. 6592, pp. 452–461. Springer, Heidelberg (2011)
11. Bagci, F., Schick, H., Petzold, J., Trumler, W., Ungerer, T.: The reflective mobile agent paradigm implemented in a smart office environment. Pers. Ubiquit. Comput. 11, 11–19 (2007)
12. Röcker, C.: Services and Applications for Smart Office Environments - A Survey of State-of-the-Art Usage Scenarios. In: Proceedings of the International Conference on Computer and Information Technology (ICCIT 2010), Cape Town, South Africa, pp. 387–403 (2010)
13. Ronzhin, A.L., Budkov, V.Y.: Multimodal Interaction with Intelligent Meeting Room Facilities from Inside and Outside. In: Balandin, S., Moltchanov, D., Koucheryavy, Y. (eds.) ruSMART 2009. LNCS, vol. 5764, pp. 77–88. Springer, Heidelberg (2009)
14. Shi, Y., Qin, W., Suo, Y., Xiao, X.: Smart Classroom: Bringing Pervasive Computing into Distance Learning. In: Handbook of Ambient Intelligence and Smart Environments, pp. 881–910. Springer Science+Business Media, Heidelberg (2010)
15. Mikulecký, P., Olševičová, K., Bureš, V., Mls, K.: Possibilities of Ambient Intelligence and Smart Environments in Educational Institutions. In: Mastrogiovanni, F., Chong, N.-Y. (eds.) Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives, ch. 29, pp. 620–639. Information Science Reference (2011)
16. Guan, D., Li, Q., Lee, S., Lee, Y.: A Context-Aware Music Recommendation Agent in Smart Office. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 1201–1204. Springer, Heidelberg (2006)
17. Krejcar, O., Jirka, J.: Proactive User Adaptive Application for Pleasant Wakeup. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS (LNAI), vol. 6592, pp. 472–481. Springer, Heidelberg (2011)
18. Paramythis, P.: Adaptive Systems: Development, Evaluation, and Evolution. PhD Dissertation, Johannes Kepler University Linz (2009)
19. Kwon, O., Yoo, K., Suh, E.: UbiDSS: a proactive intelligence decision support system as an expert system deploying ubiquitos computing technologies. Expert Systems With Applications 28, 149–161 (2005)
20. Chung, N., Lee, K.C.: Effect of Connectivity and Context-Awareness on Users' Adoption of Ubiquitous Decision Support System. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS (LNAI), vol. 6592, pp. 502–511. Springer, Heidelberg (2011)
21. Bureš, V., Čech, P.: Complexity of Ambient Intelligence in Managerial Work. In: ITiCSE 2007: 12th Ann. Conference on Innovation and Technology in Computer Science Education, p. 325. ACM Press, New York (2007)
22. Mikulecký, P.: Ambient Intelligence in Decision Support. In: Proceedings of 7th Intl. Conference on Strategic Management and Its Support by Information Systems, pp. 48–58. VSB-Tech. Univ., Ostrava (2007)

23. Mikulecký, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: Recent Advances in Applied Mathematics and Computational and Information Sciences, Proc. of the 15th American Conference on Applied Mathematics and Proc. of the International Conference on Comp. and Information Sciences, Athens, vol. I, II, pp. 523–528 (2009)
24. Mikulecký, P.: Large Scale Ambient Intelligence – Possibilities for Environmental Applications. In: Ambient Intelligence Perspectives II. Ambient Intelligence and Smart Environments, vol. 5, pp. 3–10. IOS Press, Amsterdam (2010)
25. Mikulecký, P., Ponce, D., Toman, M.: A Knowledge-Based Decision Support System for River Basin Management. In: River Basin Management II, pp. 177–185. WIT Press, Southampton (2003)
26. Mikulecký, P., Ponce, D., Toman, M.: A Knowledge-Based Solution for River Water Resources Management. In: Water Resources Management II, pp. 451–458. WIT Press, Southampton (2003)
27. Kotzian, J., Konecny, J., Krejcar, O.: User Perspective Adaptation Enhancement Using Autonomous Mobile Devices. In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS (LNAI), vol. 6592, pp. 462–471. Springer, Heidelberg (2011)
28. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking, Article ID 802523, 8 pages (2009), doi:10.1155/2009/802523

Adaptive Graphical User Interface Solution for Modern User Devices

Miroslav Behan and Ondrej Krejcar

University of Hradec Kralove, FIM, Department of Information Technologies,
Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic
Miroslav.Behan@uhk.cz, Ondrej.Krejcar@ASJournal.eu

Abstract. Researchers along the whole world developed many User Adaptive solutions for various devices include mobile devices. The focus is often targeted to one device or group of devices with closer hardware parameters. Developing of universal adaptive GUI Solution is however still unfinished due to developing of new modern devices with many of new parameters and possibilities which is difficult to cover all. Our paper present a solution for "only" one platform – Android, which is however possible to run on various hardware platforms from small mobile Smartphones to large multimedia TV centers. Solution is based on the well-known vector graphics as well as on model view controller (MVC) design pattern. Solution was successfully tested on two user application for Android Market.

Keywords: Adaptive interface, GUI, user device, Android platform, intelligent solution.

1 Introduction

Mobile devices as well as end users devices markets pass through big changes during last two decades. Mobile phone with small black and white display is over. SMS message is not only one way to communicate with others different to talk. Operation system for these devices is not usually made especially for each product version, but in most cases for product series (with small differences).

Modern mobile devices are known as Smartphones, but what we can imagine under the name – Smart Phone? How much RAM memory has the device? What wireless technology is used to communicate in local areas? How long we can talk using such device? Answer is not as easy as these questions.

Concept Smart includes every hardware as well as software pieces which creates such intelligent device to help his user improve his life. In the best case the Smartphone must be prepared to serve you everything you mentioned before you need it in real (e.g. you will enter area with direct sun and expect the display of Smartphone will be visible – this will need adaptive adjustment of backlight) [8], [9].

The one of interesting "problems" of current mobile devices market is uncountable set of various devices with various sizes (namely display sizes and resolutions). This fact is of course not a problem for users, however it put strong pressure to software

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 411–420, 2012.

and developer to create a various version of software application for each different device. For the display resolution and size we have two main possibilities how to develop an application: using raster or vector graphics. But is it really needed?

1.1 iPhone vs. Android

Apple iPhone has a standard screen size 3,5" in all versions, so the user cannot select anything different. Is it pros or cons of such solution [11]?

Display size of 3,5" is ideal for holding of device by only one hand, because user can use his inch to control all part of such display without need of other hand. This fact is sometimes very useful (e.g. writing sms, map moving and scrolling etc.). This screen size can be however sometimes uncomfortable for some multimedia cases (e.g. wide HD movie watching, eBook reading, web browser, etc.) [10].

Fig. 1. Apple iPhone and Samsung Galaxy SII size comparison [5]

Except the display size it is however needed to take care about the screen resolution (e.g. WVGA resolution 800x480 pixels) [Fig. 2].

Apple develops their iPhone series only with one aspects ratio (5:3). This fact allow to developer create an application in one static GUI design, where only resolution of pictures as part of GUI is increased. It is not needed to change a lay-out of application. There is only one exception for iPad due to his screen rotation possibility.

Google has selected other way for his OS Android, where it allows various screen resolution from version 1.6. Elements lay-out can be changed in various ways. Developer needs to take care of all empty spaces on screen or e.g. best fit align of icons.

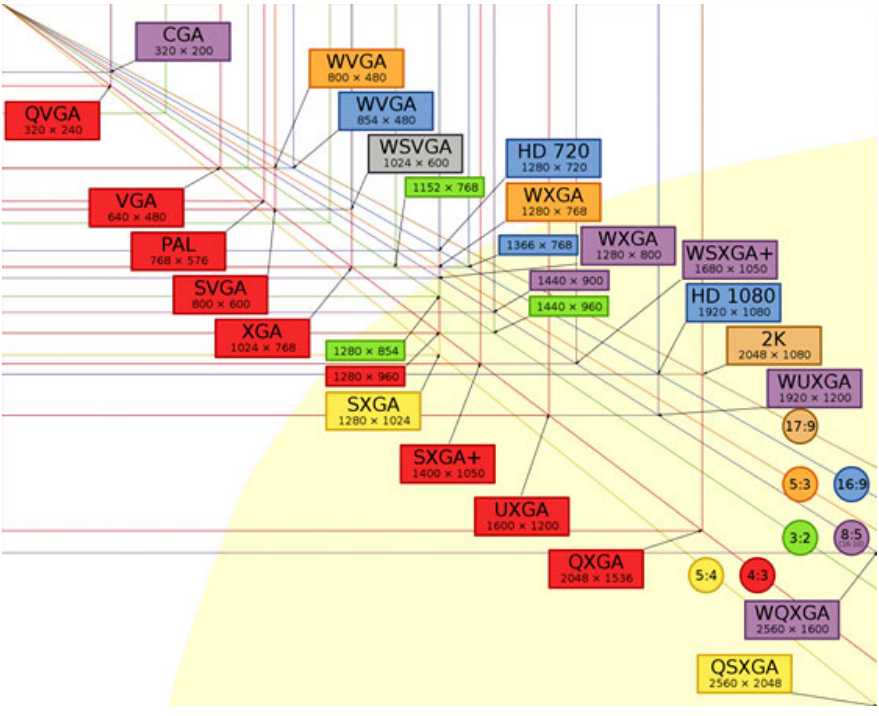


Fig. 2. Screen resolution chart with aspects ratios [5]

By this fact the Android OS need to be designed to run on several different assemblies from different manufacturers. Vice versa the iPhone OS is designed only for one hardware solution which results in much better optimalization.

So that's why when you designing android app you should all the time take in consideration how layout would look like. There are two possible ways for workout with layout elements. The first is based on raster technology which is historically mainstream for handling visualization of data. Google support well-formed programming approach which is based on categorization of layout and graphical elements which are located in different folders and stored in raster files png, jpg ect. This designing technique mode provided by android eclipse plug-in in eclipse studio automatized the process of development application. See more details in development documentation [6]. Therefore android application recognizes output resolution of device and attach correct layout if it is defined otherwise runs with default layout. However we recognized that raster technology is not as good as using of vector because a second one vector approach is more far potentially effective for its smaller byte-code size and much more effective development [17], [18], [19].

The vector based solution is more dependable on computing device capacity but in common evolutionary process where we suppose constant growing technology factor this issue will disappear in time. Advantage of vector design approach can be seen namely in scalability on device size screen in variable ratio between width and height. Of course there are some issues with embedded rasters which we considered as a

"must be included" in layout and therefore if application could have any scale then could deform some of these layout elements. For device screens with such variable ratio from default application layout would be also other option - the application layout logic resolver.

2 Problem of Layout Design Methods

The application layout logic resolver is effective approach for developing applications for various display resolutions. Nowadays development approach of designing application layout put the challenge more further on application logic and also push graphical user interface (GUI) step more forward into vector based graphical resources handling in terms of layout design. One of the design standard is scalable vector graphics (SVG) resources handling where graphical layout components are represented by script code which covers definition of mathematical abstraction for object visualization. However there are more than one stream existing for mobile application development. Far in this chapter we will describe some of these main streams according to their provider [12], [13].

2.1 Apple iPhone Platform

The supported solution from Apple company is based on one vendor controlled way of development and designing applications for all kind of apple devices. Advantage for app development with Apple solution is mainly in the power of environment which is supportive to transitivity between different device types. Although the developing language COCOA [1] supporting inter object message communication not support scalable vector graphics. The supported interface for vector graphics approach is portable document format (PDF) or direct API functions defined by mathematics functions whereas development is more about theoretical definitions instead of graphical designing tools production [22].

2.2 Windows Mobile Platform

However Microsoft approach for mobile application development comes up with different idea based on touchable areas and dimensions. Designing layout is within the standard of physical conditions which are aiming to improve user experience with mobile touchable screens. Also there is no native support of SVG resources and therefore you consider directX or scalable image objects with raster and SilverLight for vector approach.

2.3 Android Platform

The third one stream Android as mobile device platform solution is widely recognized as synonymous for future mainstream. As open source policy proves that the mass or crowd-computing as successful approach for coding therefore the open vendor device policy is key factor for universalize application coding [21].

The idea of independent mobile system based on Java will increase its potential in time [2]. At the end we outline the pros and cons.Google use magical keyword android for mobile system environment which due its open strategy for vendors and Java based programing language predicted the extremely effective approach for developers. Basically graphical layout designing would be generated by vector, raster graphics or combination of both. Also as on other platforms there are build-in functions for vector defined layout but for external SVG resources you have to include from open source community java libraries which provide scalable vector graphic processing [3], [23].

Table 1. Pros and cons layout design methods

Method	Positive	Negative
Raster resources	Fast computing External design	App resource size Less scalable Graphical workload
Vector resources	Scalable Low data size External design	More computing time Not all graphics expressions are possible
Combination R/V	All graphics expressions in small amount of data	Extends workload Internal/External design
Internal API	Fast Zero-Data based	Hard to code Internal/Math design

3 Solution Design

This chapter summarizes android application development solution in terms of design architecture. For better comprehension as overview of android architecture a next image is attached [Fig. 3] which is divided into access layers. The environment is built upon Linux kernel where *Display drivers* are connection between hardware and software OS calls. Upper layer from graphical user interface point of view in 2D is SGL library which is native build-in C++ lib coded as an access point for upper layer. And lastly on this schema the resource manager handles all access to resources stored in SVG vector files [14], [15].

Application architecture is based on MVC concept where model or domain are represent in classes and all user data are stored in serialized java object which are saved when application is paused, stopped or destroyed and visa-verse loaded when application is started. Other possible approach for data management would be over JPA- java persistent annotation which is development technique for storing java objects to database over annotation interface. The container for data in java is used POJO – which is plain old java object where getter and setter are appropriate to use and became portable to any database engine supported remotely or locally. For example with SQLite [4]. But in this case it is more effective just to load or save binary object image to local file because there are no data for network sharing.

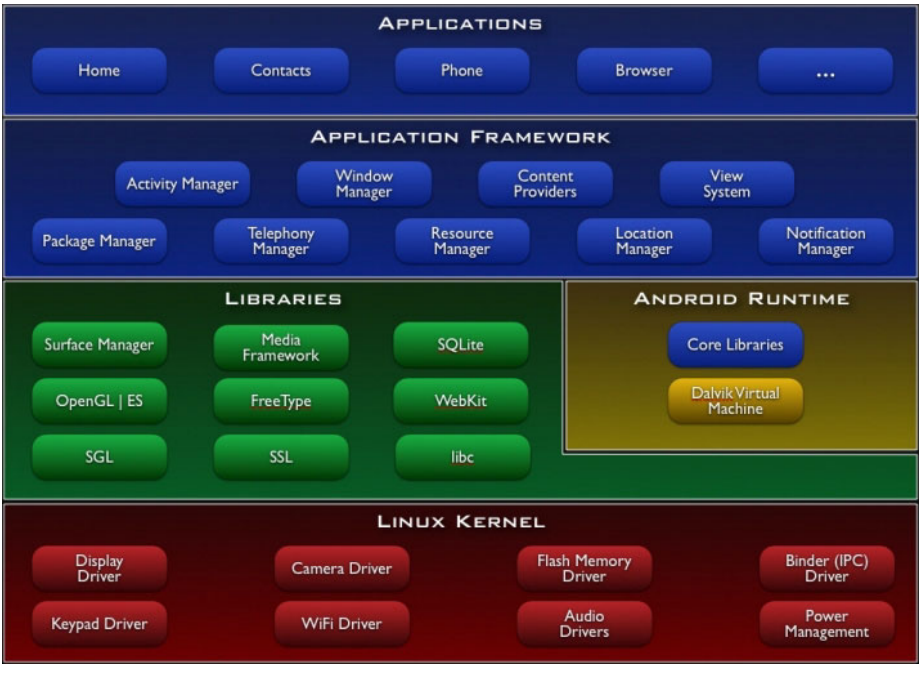


Fig. 3. Android architecture [7]

View in terms of MVC concept we called as a layout which consist of widgets or other layouts and views based on XML configuration file. In eclipse android plug-in there is a standard palette of views, layouts and widgets. Also developers could define their own views where it is override *onDraw* method of *View* class. All graphical resources have to be assigned in *res* folder of project specified by type of resolution for concrete device screen. That is in the case of standard raster based approach where developer spends plenty of time to prepare all images in required screen resolution. The vector approach is over own customized views and instead of raster image it is using SVG files as a graphical resources. Vector images are handled as a classic raster images but with one advantage. Developer could scale them anyhow without any influence on layout output quality.

Implementation for scalable vector graphics in android project is possible due to open-source `com.larvalabs.svgandroid.*` library imported as an external JAR to referenced libraries. The solution is based on SVG parser which translate vector file to drawable object which view components could work with. The vector file would be included as raw file into android project in resource directory because fast and easy implementation. In rendering method *onDraw* it is necessary to define drawing area which will be filled by vector image content.

Conclusion for vector based approach for designing android applications is to encourage it for developers as a future universal way to build high scalable apps. Of-course there are some white gaps in between which has to be covered like implementation of some expressions with shadowed color vector graphics which looks differently in Adobe Illustrator (as designing tool) and on the screen of android mobile.

4 Pilot Testing Applications

The *Roulette Predictor* is a first pilot project for vector based android application which we created. The user has a possibility to store roulette history data and has an online independent statistics over them. Visualization of data are as you could see on screen shots [Fig. 4] in customized views like accelerometer or circle graph. All of them they are inherited from *View* class with overrided *onDraw* method. There are defined mathematics functions existing which with draw primitives creating complex single widget.

For example the accelerometer widget with small hand and percentage of distributive successful probability is implemented as *sin* and *cos* function which calculate

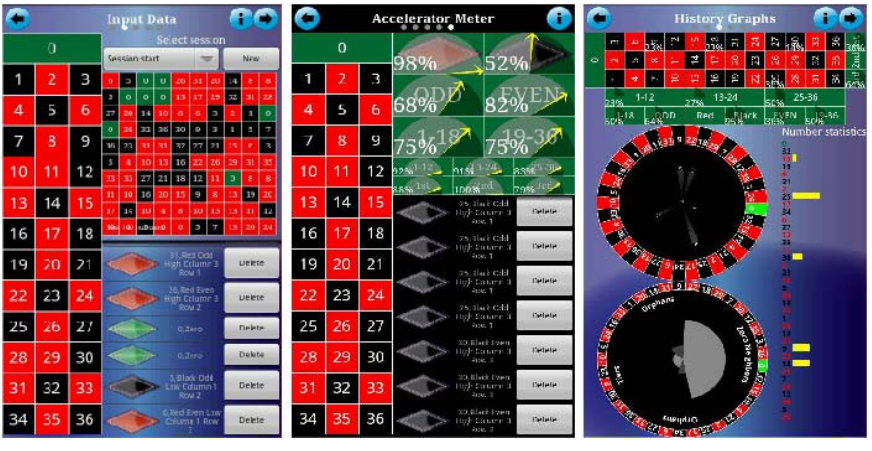


Fig. 4. Roulette Predictor Application screen shots

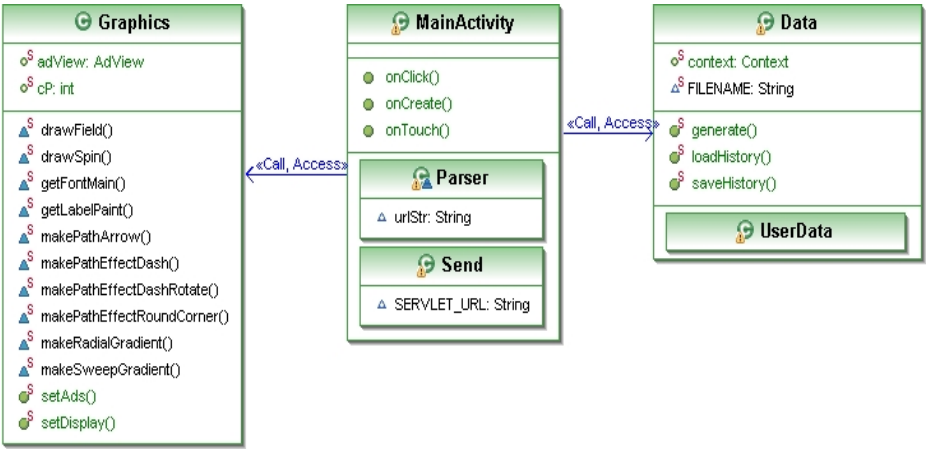


Fig. 5. UML diagram of core app classes

direction of "*LineTo*" object. Next figure [Fig. 5] provide overview on class model generated by eUML directly from eclipse. There is a graphics static pack of functions which predefined behavior of drawing elements like paths and fonts [16].

The second application called *Scratch Cards* [Fig. 6] is designed for user finger touch to scratch image on device screen. This application has all resources in SVG files whereas other graphical patterns are defined in native android draw primitives.

Fig. 6. Screen shots ScratchCards

Test pilot applications were proved the proof of concept of using SVG based application design approach on android devices. The results are dependable on computation device power in term of application performance. In terms of scalability the vector app would be fit-able for any size of screen even TV without any application size impact. The applications were tested on virtual test devices as well as one physical tablet device. For physical test purposes where you could not use virtual test due touch testing was used low price end tablet device from china market witch labeled as 7inch Android 2.2 8650 epad tablet PC 256M 2GB. Although the device was cheap and battery capacity insufficient for testing purposes it was the best choice in terms of the worst possible testing conditions predict better performance in the other cases. The device was not hundred percent accurate in aiming of touch sensors in some rare cases for finger, but with pen or other pointer resolve the accuracy of aim much better. Other test with graphical output and performance outages was quite interesting. The vector graphics with static overlay was quite fast also in raster mode, but dynamic changing content for vector content was faster than raster which concludes to use openGL for the best graphics output performance. As a testing results would be announced that the virtual and physical devices are in accurate correlation and cheap products are in some cases not sufficient as a good graphical rendering option.

5 Conclusions

Within this project we were able to successfully practice working with Google Android platform on which we developed a new MVC based GUI solution for application creation. Created solution is suitable for various hardware platforms from small mobile Smartphones to large multimedia TV centers. Solution was successfully tested on the two user applications for Android Market. Future milestones of our project are based on increasing power of current and future modern Smartphones which bring to users previously absented hardware capabilities as multi CPU core or graphical co-processing units.

Acknowledgement. This work was supported by „SMEW – Smart Environments at Workplaces“, Grant Agency of the Czech Republic, GACR P403/10/1310.

References

1. Hillegass, A.: COCOA language, Cocoa Programming for Mac OS X, 3rd edn. Addison-Wesley (2008), Paperback, ISBN: 0-321-50361-9
2. http://www.learncomputer.com/android-future-mobile-computing/
3. http://code.google.com/p/svg-android/
4. SQLite, http://www.sqlite.org/
5. China Phone Reviews, Are you still puzzled at the resolution of different phone screens?, Posted by MJ (January 31, 2011), http://www.chinaphonereviews.com/iphone-clones/are-you-still-puzzled-at-the-resolution-of-different-phone-screens.html
6. Android Developers, Application resources, http://developer.android.com/guide/topics/resources/index.html
7. Android architecture, http://elinux.org/Android_Architecture
8. Mikulecky, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, pp. 523–528. Univ. Houston, Houston (2009)
9. Tucnik, P.: Optimization of Automated Trading System's Interaction with Market Environment. In: Forbrig, P., Günther, H. (eds.) BIR 2010. LNBIP, vol. 64, pp. 55–61. Springer, Heidelberg (2010)
10. Labza, Z., Penhaker, M., Augustynek, M., Korpas, D.: Verification of Set Up Dual-Chamber Pacemaker Electrical Parameters. In: The 2nd International Conference on Telecom Technology and Applications, ICTTA 2010, Bali Island, Indonesia, March 19-21, vol. 2, pp. 168–172. IEEE Conference Publishing Services, NJ (2010), doi:10.1109/ICCEA.2010.187, ISBN 978-0-7695-3982-9
11. Brida, P., Machaj, J., Benikovsky, J., Duha, J.: An Experimental Evaluation of AGA Algorithm for RSS Positioning in GSM Networks. Elektronika IR Elektrotechnika 8(104), 113–118 (2010)

12. Chilamkurti, N., Zeadally, S., Jamalipour, S., Das, S.K.: Enabling Wireless Technologies for Green Pervasive Computing. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 230912, 2 pages (2009)
13. Chilamkurti, N., Zeadally, S., Mentiplay, F.: Green Networking for Major Components of Information Communication Technology Systems. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 656785, 7 pages (2009)
14. Liou, C.-Y., Cheng, W.-C.: Manifold Construction by Local Neighborhood Preservation. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 683–692. Springer, Heidelberg (2008)
15. Liou, C.-Y., Cheng, W.-C.: Resolving Hidden Representations. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 254–263. Springer, Heidelberg (2008)
16. Zelenka, J.: Information and Communication technologies in tourism – influence, dynamics, trends. E & M Ekonomie a Management 12(1), 123–132 (2009)
17. Skorupa, G., Katarzyniak, R.: Conditional Statements Grounded in Past, Present and Future. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part III. LNCS (LNAI), vol. 6423, pp. 112–121. Springer, Heidelberg (2010)
18. Popek, G., Katarzyniak, R.P.: Measuring Similarity of Observations Made by Artificial Cognitive Agents. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS (LNAI), vol. 4953, pp. 693–702. Springer, Heidelberg (2008)
19. Juszczyszyn, K., Nguyen, N., Kolaczek, G., Grzech, A., Pieczynska, A., Katarzyniak, R.: Agent-Based Approach for Distributed Intrusion Detection System Design. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006, Part III. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)
20. Bodnarova, A., Fidler, T., Gavalec, M.: Flow control in data communication networks using max-plus approach. In: 28th International Conference on Mathematical Methods in Economics 2010, pp. 61–66 (2010)
21. Thompson, T.: The Android mobile phone platform - Google's play to change the face of mobile phones. Dr Dobbs Journal 33(9), 40–+ (2008)
22. Shih, G., Lakhani, P., Nagy, P.: Is Android or iPhone the Platform for Innovation in Imaging Informatics. Journal of Digital Imaging 23(1), 2–7 (2010), doi:10.1007/s10278-009-9242-4
23. Hii, P.C., Chung, W.Y.: A Comprehensive Ubiquitous Healthcare Solution on an Android (TM) Mobile Device. Sensors 11(7), 6799–6815 (2011), doi:10.3390/s110706799

Visualizing Human Genes on Manifolds Embedded in Three-Dimensional Space

Wei-Chen Cheng

Institute of Statistical Science, Academia Sinica, Republic of China
r93108@csie.ntu.edu.tw

Abstract. This work provides a visualization tool for researchers to explore the geometry of the distribution of human protein-coding DNA in three-dimensional space by applying various manifold learning techniques, which preserve distinct relations among genes. The simulations suggest that the relations hidden among genetic sequences could be explored by the manifolds embedded in Euclidean space. Operating this software, users are able to rotate, scale and shift the three-dimensional spaces in an interactive manner.

1 Introduction

Viewing the population of the genes of a single creature, we can study the sequential DNA (deoxyribonucleic acid) in different aspect [9]. In manifold space, dimension reduction [7] can display meaningful relationships among patterns. Foundations for various data manifolds have been set down for factorial components [12] and for generalized adaline [23]. The principal component analysis (PCA) and multidimensional scaling (MDS) [21] are well established linear models that have been developed for such reduction. Nonlinear reduction algorithms have been devised with varying degrees of success. The Isomap [20] and the conformal C-Isomap [18] extend MDS by using the geodesic distance to construct the nonlinear manifold. The distance invariant self-organizing map (DISOM) [13] [2] uses the relative distances among local patterns to construct the manifold. Those embedded manifold coordinates in two and three dimensions are capable of mapping the whole distribution of sequence patterns to a perceptible space.

Visualizing the genes of Homo sapiens is useful for organizing the gene ontology data [8]. There have been multiple organizations and projects progressing sequencing and annotating human genome. Those bioinformatics groups are developing automatically annotation techniques and making the information of genome to be available. Esembl and the National Center for Biotechnology Information (NCBI) has developed computational processes to annotate vertebrate genomes [6] [16]. Both of them predict genes based on the interpretations of gene prediction programs. Havana group at the Wellcome Trust Sanger Institute and the Reference Sequence group at the NCBI manually annotate the sequences. Due to the unlike annotations of different projects, the consensus coding sequence

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 421–430, 2012.

project has done the effort to identify a set of human protein coding regions that are consistently annotated [17]. The project aims at providing a complete set of protein-coding regions that agree at the start codon, stop codon, and splice junctions, and for which the prediction meets quality assurance benchmarks. The project assigns a unique, tracked accession (CCDS ID) to identical coding region annotations. The population of the high-quality annotated sequences is suitable for being visualized by manifold learning techniques so that the relations among genes can be observed. Few remarkable genes, which is highly suspected to be related with the evolution of human, are marked out in the results.

In this work, I implement the visualization tool in Java codes so that the result can be shown through internet on mobile device or personal computer for researchers. The user can manipulate the results in real time and specifying the genes that they are interested. The designate genes are then downloaded from on-line website and marked in the space with small balls by different color. The text of the name of genes will be attached to the small balls nearby their location.

2 Methods

Suppose there are P genes in the nucleotide space, $X = \left\{\mathbf{x}^j \,|j = 1, \ldots, P\right\}$. Each gene $\mathbf{x}^j$ is a nucleotide sequence and has a corresponding mapped cell, $\mathbf{y}^j$, in the manifold space. The positions of the cells in the manifold space are $Y = \left\{\mathbf{y}^j \,|j = 1, \ldots, P\right\}$. Each cell position, $\mathbf{y}^j$, is a M-dimensional column vector. The distance between $\mathbf{y}^p$ and $\mathbf{y}^q$ in the Euclidean space in which the manifold space embedded is denoted by $d_Y\left(\mathbf{y}^p, \mathbf{y}^q\right)$. In the nucleotide space, giving a gene $\mathbf{x}^p$, the set of those genes whose distances to $\mathbf{x}^p$ are less than r are included in the set $U\left(p, r\right)$, where r denotes the radius of neighborhood region in the nucleotide space. The notation $\left|U\left(p, r\right)\right|$ denotes the number of genes in the set $U\left(p, r\right)$.

In the nucleotide space, there are ways which can be computed by dynamic programming to measure the similarity or dissimilarity of any two sequences. Longest common subsequence is longer while two genes have larger portion of DNA in common. Which means the genes are more similar to each other. Let the distance function $d_X^{LCS}\left(\mathbf{x}^p, \mathbf{x}^q\right)$ denotes the inverse length of the longest common subsequence between gene $\mathbf{x}^p$ and gene $\mathbf{x}^q$. Minimum edit distance calculates the minimum number of operations to alter one nucleotide sequence to match another. There being operations includes insertion, deletion and replacement. Let the distance function $d_X^{Edit}\left(\mathbf{x}^p, \mathbf{x}^q\right)$ represent the minimum edit distance between gene $\mathbf{x}^p$ and gene $\mathbf{x}^q$. Small value of the function indicates the two sequences have small differences. The relation between the length of the longest common subsequence and the edit distance has been discussed in [19],

$$d_X^{Edit}\left(\mathbf{x}^p, \mathbf{x}^q\right) = \left\|\mathbf{x}^p\right\| + \left\|\mathbf{x}^q\right\| - 2\frac{1}{d_X^{LCS}\left(\mathbf{x}^p, \mathbf{x}^q\right)}. \tag{1}$$

The $\left\|\mathbf{x}^p\right\|$ denotes the length of the sequence $\mathbf{x}^p$.

Supposing that short distance is more reliable than long distance, consider the local distance invariant energy [2],

$$E(r) = \frac{1}{4} \sum_{p} \sum_{\mathbf{x}^q \in U(p,r)} \left(d_Y \left(\mathbf{y}^p, \mathbf{y}^q \right)^2 - d_X \left(\mathbf{x}^p, \mathbf{x}^q \right)^2 \right)^2, \tag{2}$$

where the d_X is either d_X^{LCS} or d_X^{Edit} in this work. In each iteration the algorithm uses the relations between a sequence and its neighboring sequences to improve the location of its corresponding cell. The pseudocode to reduce the energy, $E(r)$, is as follow:

1. Initially assign Y by MDS
2. $\check{r} \longleftarrow \min_{i<j} d_X \left(\mathbf{x}^i, \mathbf{x}^j \right)$, $\hat{r} \longleftarrow \max_{i<j} d_X \left(\mathbf{x}^i, \mathbf{x}^j \right)$
3. For epochs $t = t_0$
4. Adjust the radius of neighborhood region: $r(t) \longleftarrow \check{r} + (\hat{r} - \check{r}) \times \exp\left(-4 \frac{(t-t_0)}{(t_1-t_0)} \right)$.
5. for $i = 1$ to P
6. $\Delta \mathbf{y}^i = \sum_{\mathbf{x}^j \in U(i,r(t))} \left(d_Y \left(\mathbf{y}^i, \mathbf{y}^j \right)^2 - d_X \left(\mathbf{x}^i, \mathbf{x}^j \right)^2 \right) \left(\mathbf{y}^i - \mathbf{y}^j \right)$
7. Update the cells: $\mathbf{y}^i \longleftarrow \mathbf{y}^i - \eta \Delta \mathbf{y}^i$.
8. End For
9. End For

This algorithm constructs local distance invariant manifold of the data in a low-dimensional Euclidean space that best preserves the local distance information in a deepest descent way. Genes in nucleotide space can therefore be displayed in three-dimensional space. The algorithm seeks a distribution Y in the manifold space that maximally resembles the distribution of X in the nucleotide space. The desired distribution is accomplished when the minimum value of $E(r)$ is reached. The reduction can be improved by a hybrid version of Gauss-Newton method and gradient descent method [10] [15] while dealing with large dataset [3]. I adopt the hybrid version to construct the manifold for genes in my simulations.

Considering another assumption that each existed genes is a node on the path of evolution. The process that a gene evolves to another gene is not arbitrary but is restricted by physical laws. Therefore, the paths that pass existed genes are reasonable approximations to evolution process. The following algorithm [20] accomplish this idea.

1. The first step is that to determines the neighborhood genes on the manifold M based on the distance $d_X \left(\mathbf{x}^i, \mathbf{x}^j \right)$ between pairs of sequence $\mathbf{x}^i$ and $\mathbf{x}^j$ in the gene pool. The set of neighbors for a given gene is the K nearest neighbors. These neighborhood relations are represented as a weighted graph G whose edges between neighboring sequences $\mathbf{x}^i$, $\mathbf{x}^j$ are the weight $d_X \left(\mathbf{x}^i, \mathbf{x}^j \right)$. In this case, vertices in the graph have degree equals to K.

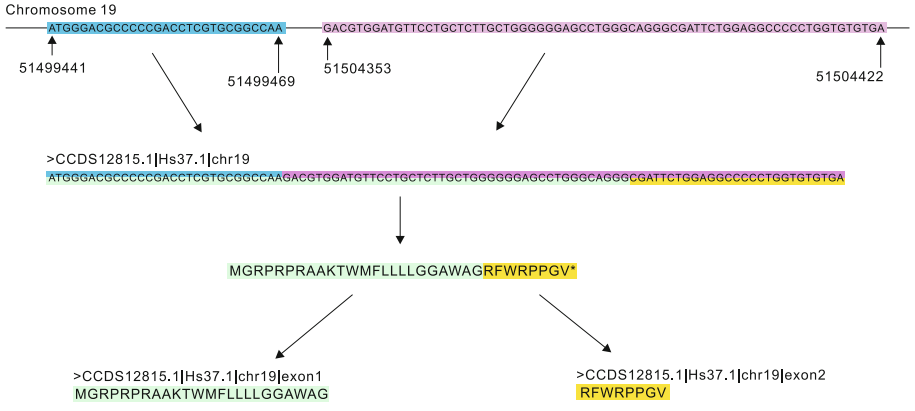


Fig. 1. The translation from DNA to protein sequences

2. The second step is that to computes the shortest path lengths $d_G\left(\mathbf{x}^i, \mathbf{x}^j\right)$ in the graph G to estimates the geodesic distances $d_M\left(\mathbf{x}^i, \mathbf{x}^j\right)$ between all pairs of genes on the manifold. The matrix of values $D_G = \{d_G(i,j)\}$ contains the lengths of the shortest paths between all pairs of vertices in G.
3. The final step applies classical MDS to the matrix of graph distances D_G, while constructing an embedding of the genes in a three-dimensional Euclidean space that preserves the intrinsic geometry of the manifold. The coordinate vectors $\mathbf{y}^i$ for points in Y are chosen to minimize the cost function

$$E = \|\tau(D_G) - \tau(D_Y)\|_{L^2} \tag{3}$$

where D_Y denotes the matrix of Euclidean distances $\left\{d_Y\left(\mathbf{y}^i, \mathbf{y}^j\right) = \|\mathbf{y}_i - \mathbf{y}_j\|\right\}$ and $\|A\|_{L^2}$ is the matrix norm $\sqrt{\sum_{i,j} A_{ij}^2}$. The τ operator converts distances to inner products, which uniquely characterize the geometry of the data in a form that supports efficient optimization.

The approach is capable of discovering the nonlinear degrees of freedom that underlay complex natural observations. The resulting embedding is quasi-isometric.

3 Data and Graphics Description

All genes of Homo sapiens acquired from the CCDS website [17] are collected to be the gene pool. There are $23,754$ genes in the version of HsGRCH37.1. The longest gene has $26,394$ base pairs while the shortest gene has 78. Each gene, which is piecewise consecutive, consists of several fragments of the DNA sequences. The sequence of joint fragments is transcribed to mRNA and then translated to protein. Figure 1 shows one of the examples in the data. The gene is from chromosome 19 and consists of two fragments. The length of the first

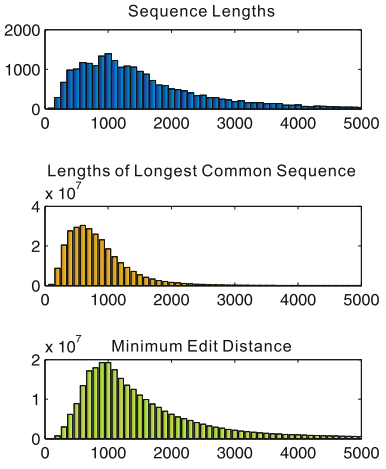


Fig. 2. The histograms of the length and distance information

fragment is 29 base pairs and the second is 70 base pairs. The sequence of the joint fragments CCDS12815.1 is translated to the protein CCDS12815.1 which has two exons, namely exon1 and exon2. This work analyzes the relationship among the sequences of joint DNA fragments.

To investigate the distances, we calculate the histogram of the length and the distance information in Figure 2. The lengths of $23,754$ genes are divided into 250 bins whose size is 100. The first bin is in the range $[0, 100)$ and the second bin is in the range $[100, 200)$ et. al. The distributions have very long tails in the histogram. The tail of the plot beyond the 5000 is cut off so that the distribution of short sequences can be displayed clearly.

The result is plotted by Java 3D API version 1.5.2 so that it can be displayed in browsers or by Java Virtual Machine on different platform (operation system). The code is compiled by Java Development Kit (JDK) 1.6.0 into Java bytecode. There are five spotlights and a point light to illuminate the balls. Each ball represent a gene and its corresponding point (location) in the space. The coordinate of all points are normalized into the range $[-1, 1]$, hence the balls can be lighted up correctly. The setting of spotlights is listed in Table 1. The coordinate of the point light is $(0.1, -0.4, 0.2)$. The attenuation of all lights is quadratic to the distance.

Table 1. The setting of lighting

Light	Position	Direction	Spread Angle	Concentration
1	$(0, 2, 0)$	$(0, -1, 0)$	$\frac{\pi}{9}$	50
2	$(1, 2, 1)$	$(-0.3, -1.2, 0.5)$	$\frac{\pi}{18}$	50
3	$(-1, 2, 0)$	$(0, 0, 0.3)$	$\frac{\pi}{18}$	50
4	$(0.3, 0.3, 1.5)$	$(0.3, 0.3, -1.5)$	$\frac{\pi}{3}$	50
5	$(-0.3, -0.3, 1.5)$	$(-0.3, -0.3, -1.5)$	$\frac{\pi}{3}$	50

Table 2. Information of the selected genes

Gene	CCDS ID	Length	Reference
FOXO3A	5068.1	2022 bps	Lin [11]
ADCYAP1	11825.1	531 bps	Wang [22]
DTNBP1	4534.1	1056 bps	Burdick [1]
KLK8	12813.1	783 bps	Lu [14]

4 Results

The result of MDS [21], which is a linear dimension reduction method, is ploted in Figure 3 and Figure 4. Originating from psychometrics [21], MDS was proposed to help understand people's judgments of the similarity of members of a set of objects. We can use MDS that takes a matrix of inter-gene distances to create a configuration of points. Those points are constructed in low dimensions, and the Euclidean distances between them approximately reproduce the original distance matrix. The figures shows the spiral shape of the result derived from MDS in two aspects. The distance function d_X is set to d_X^{Edit}. Short genes are on one side and the long genes on the other side.

I marks out several significant genes to see the location of the genes in the three-dimensional space. Firstly, gene DAF-16a is known to regulate longevity, stress response and dauer diapause [11]. The human homologue of DAF-16 includes four FOXOs: FOXO1, FOXO3, FOXO4 and FOXO6. It has been found that genetic variation within the FOXO3A gene was strongly associated with human longevity. The forkhead box protein O3, whose accession number is NP_963853.1 and CCDS number is CCDS5068.1, of Homo sapiens is acquired from the NCBI database. Secondly, a molecule PACAP, which regulates the generation and transmission of neuron signal evolves fast in human than other animals, has been found by Bing Su from Kunming Institute of Zoology, Chinese Academy of Sciences. Figure 3 and Figure 4 label the text of the gene ADCYAP1, which encodes the molecule PACAP, in the space. The gene is possibly related to the formation of memory and learning of human [22]. Thirdly, the Kallikrein 8 (KLK8), suggested by experiments [14], is a human-specific splice form of KLK8 (type II) with preferential expression in the human brain among diverse primate species. Finally, gene DTNBP1 is implicated in schizophrenia has been reported. The three later genes are associated with human intelligence. All those genes are marked in the Figure 3 and Figure 4. Their information is listed in Table 2.

To minimize the cost function (3), the bottleneck is to find the shortest path among all genes. When k is small, the weighted matrix G is very sparse. The most efficient algorithm is proposed by [5] while comparing to Floyd-Warshall algorithm, Bellman-Ford algorithm and Dijkstra algorithm. The software [4] implemented those algorithms to find the length of shortest paths among all pairs of vertices. The computational complexity of [5]'s method is $O(VD\log(V))$ where

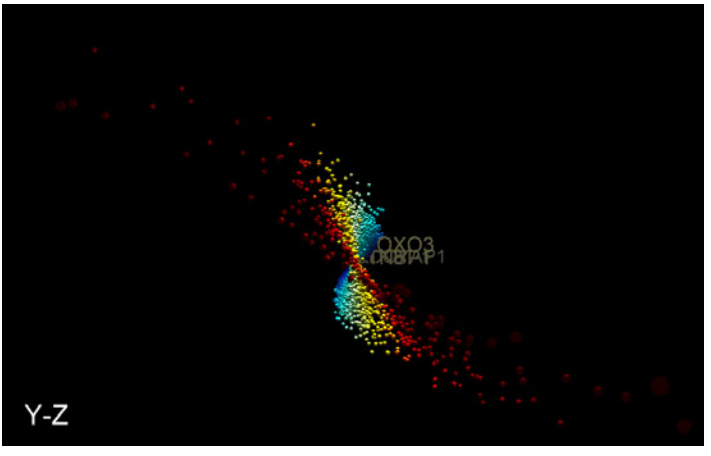


Fig. 3. The y-z plane of the MDS result by using the inverse lengths of LCS between gene pairs to generate the distance matrix. The color indicate the length of the genes. Red color indicates long gene and blue indicates short.

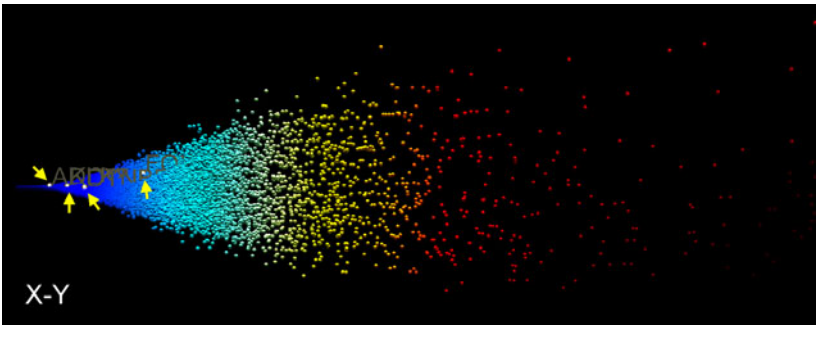


Fig. 4. The x-y plane of MDS result by using the inverse length of LCS

V is the number of points (genes) and the D denotes the number of edges. It take 48.58 minutes to perform the algorithm when $k = 20$. The distance function is d_X^{Edit} and the result whose neighborhood size K equals to 32 is ploted in Figure 5. It shows that short genes are near the core and the long genes are far from the core. Short genes tend to cluster together because they take fewer mutations to evolve to each others. The four designate genes are also marked in Figure 5.

Figure 7 and Figure 6 plot the result of minimizing (2) by using the inverse length of LCS, d_X^{LCS}, as the distance function. The result resembles a sphere. Observing Figure 7, a bridge crosses the hole of the red area. The density of blue area is lower than the red area. The four designate genes are inside the sphere and the FOXO3A is close to the center of the sphere.

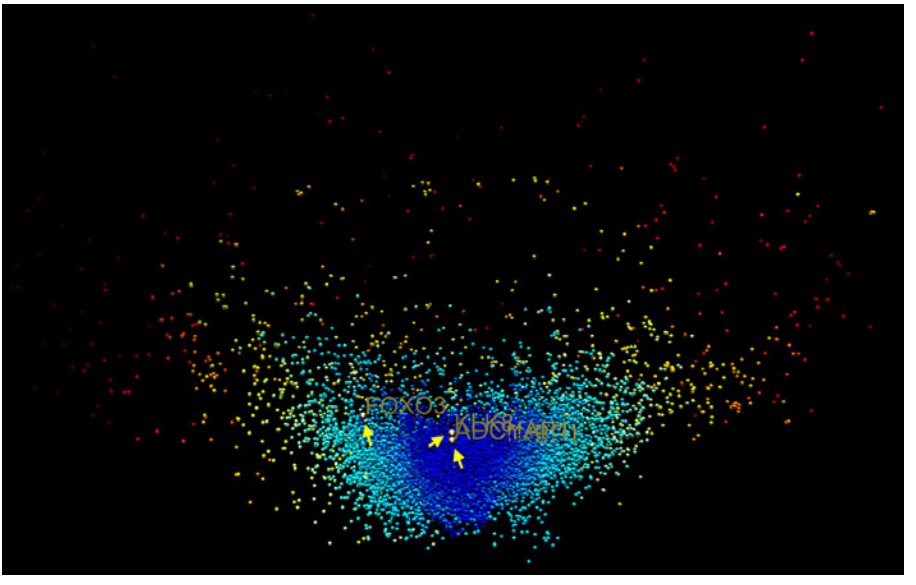


Fig. 5. The result of Isomap using the minimum edit distance as the distance matrix. Red color indicates long gene and blue indicates short.

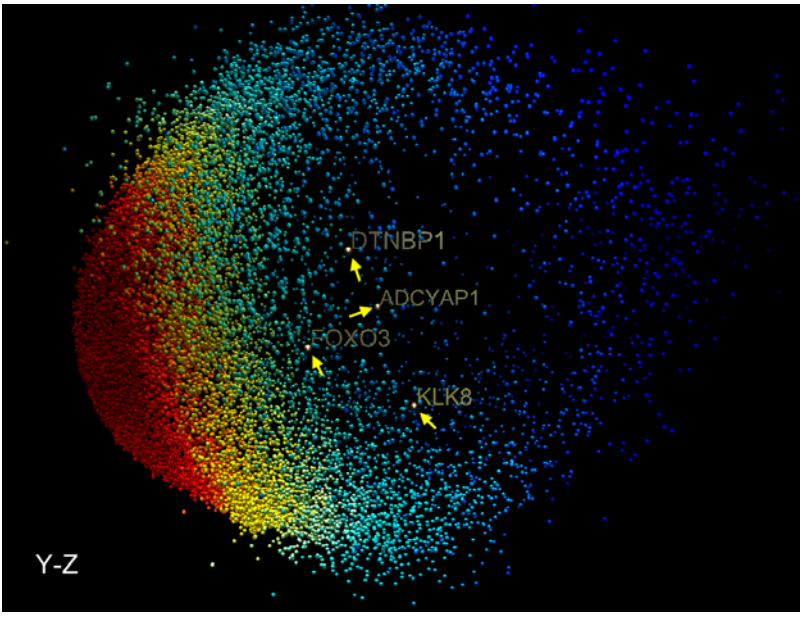


Fig. 6. The y-z plane of DISOM result by using the inverse length of LCS to be the distance. Red color indicates long gene and blue indicates short.

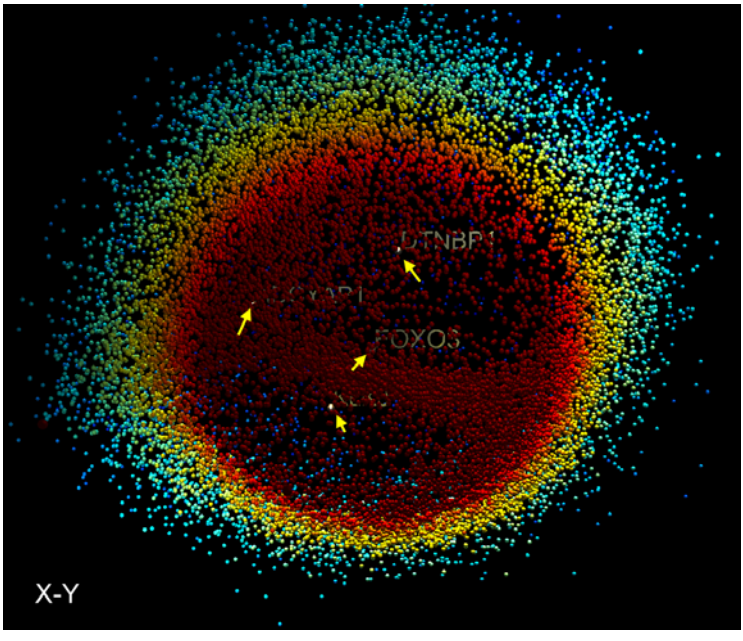


Fig. 7. The x-y plane of DISOM result by using the inverse length of LCS to be the distance

5 Conclusion

The simulations demonstrate the way to analyze the DNA sequences by manifold learning techniques. We find the genes that related to life and evolution of intelligence is located inside spherical result by DISOM. This means those genes are less similar to other genes. Normal users can upload their own genome sequence into the software and view their own genetic information. The sequences of the genome, which are matched in the gene database, can be used for generating biosignature.

References

1. Burdick, K.E., Lencz, T., Funke, B., Finn, C.T., Szeszko, P.R., Kane, J.M., Kucherlapati, R., Malhotra, A.K.: Genetic Variation in DTNBP1 Influences General Cognitive Ability. Human Molecular Genetics 15, 1563–1568 (2006)
2. Cheng, W.-C., Liou, C.-Y.: Manifold Construction Based on Local Distance Invariance. Memetic Computing 2, 149–160 (2010)
3. Cheng, W.-C., Liou, C.-Y.: Visualization of Influenza A Protein Segments in Distance Invariant Self-organizing Map. International Journal of Intelligent Information and Database Systems
4. Gleich, D.F.: Models and Algorithms for PageRank Sensitivity. Thesis in Stanford University (2009)

5. Johnson, D.B.: A Priority Queue in which Initialization and Queue Operations Take O(log log D) Time. Mathematical Systems Theory 15, 295–309 (1982)
6. Kitts, P.: Genome Assembly and Annotation Process. In: The NCBI Handbook, National Library of Medicine, Bethesda, MD (2002), http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch14
7. Kohonen, T.: Self-organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43, 59–69 (1982)
8. Krejcar, O., Janckulik, D., Motalova, L.: Complex Biomedical System with Biotelemetric Monitoring of Life Functions. In: Proceedings of the IEEE Eurocon, pp. 138–141 (2009)
9. Krejcar, O., Janckulik, D., Motalova, L.: Complex Biomedical System with Mobile Clients. In: IFMBE Proceedings of the World Congress on Medical Physics and Biomedical Engineering 2009. IFMBE Proceedings, vol. 25/5, pp. 141–144. Springer, Munich (2009)
10. Levenberg, K.: A Method for the Solution of Certain Non-linear Problems in Least Squares. The Quarterly of Applied Mathematics 2, 164–168 (1944)
11. Lin, K., Dorman, J.B., Rodan, A., Kenyon, C.: Daf-16: An HNF-3/forkhead Family Member that can Function to Double the Life-span of Caenorhabditis elegans. Science 278, 1319–1322 (1997)
12. Liou, C.-Y., Musicus, B.R.: A Separable Cross-entropy Approach to Power Spectral Estimation. IEEE Transactions on Acoustics, Speech and Signal Processing 38, 105–111 (1990)
13. Liou, C.-Y., Cheng, W.-C.: Manifold Construction by Local Neighborhood Preservation. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 683–692. Springer, Heidelberg (2008)
14. Lu, Z.X., Huang, Q., Su, B.: Functional Characterization of the Human-specific (type II) Form of Kallikrein 8, a Gene Involved in Learning and Memory. Cell Research 19, 259–267 (2009)
15. Marquardt, D.: An Algorithm for Least-squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics 11, 431–441 (1963)
16. Potter, S.C., Clarke, L., Curwen, V., Keenan, S., Mongin, E., Searle, S.M., Stabenau, A., Storey, R., Clamp, M.: The Ensembl Analysis Pipeline. Genome Research 14, 934–941 (2004)
17. Pruitt, K.D., Harrow, J., Harte, R.A., et al.: The Consensus Coding Sequence (CCDS) Project: Identifying a Common Protein-coding Gene Set for the Human and Mouse Genomes. Genome Research 19, 1316–1323 (2009)
18. de Silva, V., Tenenbaum, J.B.: Global versus Local Methods in Nonlinear Dimensionality Reduction. In: Advances in Neural Information Processing Systems, vol. 15, pp. 705–712 (2002)
19. Sankoff, D., Kruskal, J.B. (eds.): Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison. Addison-Wesley (1983)
20. Tenenbaum, J., Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290, 2319–2323 (2000)
21. Torgerson, W.S.: Multidimensional Scaling, I: Theory and Method. Psychometrika 17, 401–419 (1952)
22. Wang, Y.-Q., Qian, Y.-P., Yang, S., Shi, H., Liao, C.-H., Zheng, H.-K., Wang, J., Lin, A.A., Cavalli-Sforza, L.L., Underhill, P.A., Chakraborty, R., Jin, L., Su, B.: Accelerated Evolution of the Pituitary Adenylate Cyclase-activating Polypeptide Precursor Gene during Human Origin. Genetics 170, 801–806 (2005)
23. Wu, J.-M., Lin, Z.-H., Hsu, P.H.: Function Approximation using Generalized Adalines. IEEE Transactions on Neural Networks 17, 541–558 (2006)

Two-Step Analysis of the Fetal Heart Rate Signal as a Predictor of Distress

Robert Czabanski[1], Janusz Wrobel[2], Janusz Jezewski[2], and Michal Jezewski[1]

[1] Silesian University of Technology, Institute of Electronics, Akademicka 16, 44-101 Gliwice, Poland
[2] Biomedical Signal Processing Dept., Institute of Medical Technology and Equipment, Roosevelta 118, 41-800 Zabrze, Poland
{robert.czabanski@polsl.pl}

Abstract. Cardiotocography is a biophysical method of fetal state assessment based on analysis of fetal heart rate signal (FHR). The computerized fetal monitoring systems provide a quantitative evaluation of FHR signals, however the effective methods for fetal outcome prediction are still needed. The paper proposes a two-step analysis of fetal heart rate recordings that allows for prediction of the fetal distress. The first step consists in classification of FHR signals with Weighted Fuzzy Scoring System. The fuzzy inference that corresponds to the clinical interpretation of signals based on the FIGO guidelines enables to designate recordings indicating the fetal wellbeing. In the second step, the remained recordings are classified using Lagrangian Support Vector Machines (LSVM). The evaluation of the proposed procedure using data collected with computerized fetal surveillance system confirms its efficacy in predicting the fetal distress.

Keywords: Fetal heart rate monitoring, fuzzy systems, support vector machines, signal classification.

1 Introduction

Cardiotocography is a method of non-invasive monitoring of fetal state based on the Doppler ultrasound technique for recording of fetal heart rate (FHR) signal. Changes in the fetal heart activity provide the reliable information on fetal oxygenation. Therefore, the analysis of FHR is a crucial element of diagnostic process for evaluation of the fetal condition. The visual interpretation of FHR patterns is characterized by low inter- and intraobserver agreement [1]. In order to decrease the subjective nature of the visual assessment the automated computerized analysis was introduced. The computerized systems for fetal monitoring provide parameters describing quantitatively the FHR signals, but effective methods supporting medical decision are still the aim of research [3,7,8]. Numerous attempts were made to formalize the criteria for FHR signals assessment based on their quantitative analysis. Nevertheless, the clinical standards are in accordance with the recommendations of International Federation of Obstetrics

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 431–438, 2012.

and Gynecology (FIGO) [10]. According to FIGO guidelines the FHR signal can be assign to one of three classes, describing the fetal state as "normal", "suspicious" or "pathological". The assessment of the fetal state consists in identifying the characteristic features of the FHR signal including:

- baseline, that represents the changes of the so called basal level of the fetal heart rate,
- acceleration and deceleration patterns, that are defined as transient deviations from the baseline with established range of duration and amplitude,
- instantaneous variability, describing the changes of duration of cardiac intervals between consecutive heart beats.

There is no other noninvasive diagnostic method that would provide the actual fetal state with higher certainty and accuracy at the time of fetal monitoring. Therefore, the retrospective analysis is applied when assessing the quality of fetal state evaluation methods. Since, it can be assumed that the fetal state can not change rapidly during pregnancy, the fetal outcome, which defines the fetal condition just after the delivery, can be retrospectively assign to fetal state at the time of the FHR signal acquisition. The fetal distress is a term commonly used to describe fetal hypoxia being the result of an inadequate oxygenation. An improper level of hydronium ion activity (pH) in blood from the umbilical cord vessel is an objective symptom of fetal distress. The value of $\mathrm{pH} \geq 7.20$ indicates the fetal wellbeing, while $\mathrm{pH} < 7.10$ represents the fetal hypoxia. Values in between are usually interpreted as the possibility of fetal health risk.

2 Fuzzy Analysis of FHR Signals

The rules of fetal state assessment provided by FIGO are difficult to implement in a simple inference procedure. In order to facilitate the qualitative assessment of the FHR recording different point scales were proposed [4,5]. A point scale assigns a certain number of points to a specific range of values of the considered signal features. The resulting sum of points indicates the class of the recording. Since, there are no strict boundaries between the values of quantitative parameters of FHR signal that could really differentiate between the fetal wellbeing and fetal distress, the application of fuzzy set theory and fuzzy logic seems to be useful. Hence, we proposed the Weighted Fuzzy Scoring System (WFSS) which is a fuzzy model of a point scale for the fetal state assessment. As inputs for fuzzy system we used the quantitative parameters of FHR signals that are consistent with the FIGO guidelines i.e.: mean value of FHR baseline (mFHR), the number of identified acceleration (ACC), expressed as the number of patterns detected per one hour of recording, three types of decelerations ($\mathrm{DEC_A}, \mathrm{DEC_B}, \mathrm{DEC_C}$) also expressed as the number of patterns detected per one hour of recording, three types of oscillations ($\mathrm{OSC_0}, \mathrm{OSC_I}, \mathrm{OSC_{III}}$), expressed as a percentage of oscillation having amplitudes consistent with the ranges from guidelines, and the beat-to-beat variability index (STV). The formalized criteria for classification of antepartum FHR signals are shown in Table 1.

Table 1. The criteria for classification of FHR signals according to FIGO guidelines

Input parameter	Normal ($p_0 = -1$)	Suspicious ($p_0 = s$)	Pathological ($p_0 = +1$)
mFHR [bpm]	$[110, 150]$	$[100, 110)$ or $(150, 170]$	< 100 or > 170
ACC [1/h]	$> 12_{(13)}$	$\left(1.5_{(4)}, 12_{(13)}\right]$	$\left[0, 1.5_{(4)}\right]$
DEC [1/h]	$\mathrm{DEC_A} \in [0, 1.5)$ and $\mathrm{DEC_B} = 0$ and $\mathrm{DEC_C} = 0$	$\mathrm{DEC_A} \geq 1.5$ or $\mathrm{DEC_B} \in [0, 1.5)$ or $\mathrm{DEC_C} \in [0, 1.5)$	$\mathrm{DEC_B} \geq 1.5$ or $\mathrm{DEC_C} \geq 1.5$
OSC [%]	$\mathrm{OSC_0} = 0$ and $\mathrm{OSC_I} \in [0, 40)$ and $\mathrm{OSC_{III}} = 0$	$\mathrm{OSC_0} \in \left[0, 40_{(7)}\right)$ and $\mathrm{OSC_I} \geq 40$	$\mathrm{OSC_0} \geq 40_{(7)}$
STV [ms]	$\left[6_{(8)}, 14\right]$	> 14	$\left[0, 6_{(8)}\right)$

Fuzzy rules of the WFSS system correspond to the FIGO criteria allowing the evaluation of a single FHR signal feature. Consequently, a rule base of WFFS system consists of fuzzy rules in the form:

$$\underset{1 \leq i \leq I}{\forall} R^{(i)} : \textbf{if} \quad \left(x_{0j} \textbf{ is } A_j^{(i)}\right) \textbf{ then } y_0^{(i)} = p_0^{(i)}, \tag{1}$$

where: I is the number of rules that is equal to the number of ranges of quantitative parameters of FHR signal distinguish by FIGO (in our study $I = 25$), $A_j^{(i)}$ is the linguistic value (term) represented by fuzzy set defined with membership function $\mu_{A_j}^{(i)}(x)$, x_{0j} is the j-th parameter describing the FHR signal, and $p_0^{(i)}$ denotes the number of points assigned to a given range of the signal feature. We used the following point scale in WFSS system: $p_0^{(i)} = -1$ for a range corresponding to normal fetal state, $p_0^{(i)} = s, s \in [-0.5, 0.5]$ for a range corresponding to suspicious condition of the fetus, and $p_0^{(i)} = +1$ for a range corresponding to pathological fetal state. The number of points s assigned to the ranges indicating the suspicious fetal state allows for different interpretation of the ranges referring to class "suspicious". For $s < 0$ the "suspicious" range is interpreted towards indicating fetal wellbeing, while for $s < 0$ as the fetal distress. For $s = 0$ values of parameters evaluated to be suspicious do not affect the final result of the signal classification. Fuzzy sets $A_j^{(i)}$ in premises of the rules represent natural language terms "normal", "suspicious" or "pathological" corresponding to ranges of parameters that were distinguished according to FIGO. Each of these sets is given by the trapezoidal membership functions:

$$\mu_{A_j}^{(i)}(x) = \begin{cases} 0, & x \leq a, \\ (x-a)/(b-a), & a < x \leq b, \\ 1, & b < x \leq c, \\ (d-x)/(d-c), & c < x \leq d, \\ 0, & d < x, \end{cases} \tag{2}$$

The values of parameters a, b, c, d can be calculated on the basis of statistical analysis of the available set of quantitative parameters describing the FHR signals. The values of b and c are defined respectively as lower and upper quartile of quantitative parameter measurements in a given range. The a and d are calculated to fulfill the assumption that the membership of the values defining the boundary between classes of the FHR parameter should be the same for both classes and equal to 0.5. Consequently, the equations defining values of a and d are given as:

$$a = 2\,l - b;\ d = 2\,u - c, \tag{3}$$

where: l is lower, and u upper boundary corresponding to the particular range of quantitative parameters of FHR signal.

The crisp output value of WFSS is calculated as a weighted mean of points from all fuzzy rules. There are two types of weights introduced in the WFSS system. The first, represent the level of association between the value of the FHR quantitative parameter and the fuzzy rule. The level of association is defined with the degree of membership $\mu_{A_j}^{(i)}(x_{0j})$ of the given parameter in fuzzy set from the rule premise corresponding to a specific range of values. The second type of weights provides additional information about predictive capabilities of the particular quantitative parameters. These weights express the degree of certainty in assessing of the fetal state on the basis of the interpretation of a single FHR parameter only. Consequently, the equation defining the crisp output value of WFSS is defined as:

$$y_0 = \frac{\sum_{i=1}^{I} \mu_{A_j}^{(i)}(x_{0j})\ \ c_j\ p_0^{(i)}}{\sum_{i=1}^{I} \mu_{A_j}^{(i)}(x_{0j})}, \tag{4}$$

where c_j denotes the degree of certainty of FHR recording assessment based on the evaluation of j-th parameter of quantitative description of the FHR signal. Values of c_j can be calculated on the basis of the analysis of the relation of FHR signal features to the risk of fetal distress using receiver operating characteristics (ROC). The area under the ROC curve (AUC) is equal to the probability that the rank of the classifier of a randomly selected instance for abnormal fetal outcome X is higher than the rank of a randomly selected instance of fetal wellbeing Y. This relationship indicates the possibility of using the AUC values in evaluation of diagnostic capabilities of the particular FHR parameters. Hence, the normalized AUCs are used to define c_j:

$$c_j = \frac{\mathrm{AUC}_j}{\max\limits_{1 \leq k \leq m} (\mathrm{AUC}_k)}, \tag{5}$$

where: AUC_j denotes the area under ROC curve calculated for j-th parameter of quantitative description of FHR signal and m is the number of the analyzed parameters (in our study $m = 9$).

To estimate the performance of the system using standard indices of classification quality, and also to allow for the next step of the FHR signal analysis we assumed that the final result of fuzzy reasoning was the assignment of the FHR signal into two classes only, describing the wellbeing or distress. In the assumed point scale the sign of crisp output value of FSS system defines the assessment of the FHR recording: $y_0 \geq 0$ indicates the distress while $y_0 < 0$ represents the wellbeing.

3 Second Step of the FHR Signals Evaluation

The reassuring features of FHR signal are confirmed by normal fetal outcome in about 95% cases. Therefore, the fetal state assignment based on the evaluation of FHR signals works well as a screening method. The role of the WFSS is to model the diagnostic procedure designating the signals that confirm the fetal wellbeing with high certainty. However, the FHR features, which shall be considered as pathological characterize also fetuses whose condition is normal. Therefore, the evaluation of FHR recordings based only on the set of FIGO rules results in a high number of false positive assessments. To improve the classification efficacy at the second step, only recordings classified with fuzzy inference system as suspicious and pathological are evaluated. There are many classification algorithms that could be applied [3,7,8], nevertheless, the Support Vector Machines [2,11] are of special interest due to their high performance. In practical applications, low computational cost of classifier is of special interest, and therefore we applied the Lagrangian Support Vector Machines (LSVM) [9]. The primary LSVM method formulates a linear classifier, however to achieve the satisfying learning results we applied the nonlinear LSVM solution by transformation of input data using kernel functions.

4 Results and Discussion

The research material consisted of parameters of the quantitative description of FHR signals recorded during fetal monitoring sessions [6] and the corresponding measurement of pH level read from the newborn forms. The database comprised of 189 antepartum records from 51 patients. The fetal distress (pH < 7.10) was detected for seven fetuses from which a total number of 43 signals were recorded. To achieve the main goal of the fuzzy analysis, which was to identify the recordings indicating the fetal wellbeing with maximum certainty, we varied the number of points s assigned to the class of the suspicious fetal state in the range of $[-0.5, 0.5]$ with the step 0.25. For the purpose of FHR signals analysis using the LSVM algorithm, the entire dataset was divided into two equal parts: training and testing. The LSVM classifier was tested for 50 trials with random divisions

of the dataset. The parameters of LSVM algorithm providing the lowest misclassification rate of fetal assessment were determined on the basis of the grid search method for the first 10 random divisions. The set of parameters providing the lowest mean number of misclassification was chosen. The parameter γ, the parameter σ of the radial kernel as well as exponent n of the polynomial kernel were searched in the range $\left[10^{-3}, 10^{-2}, ..., 10^{5}\right]$. The LSVM algorithm was stopped if the maximum number of 100 iterations was achieved or if in the sequential iterations the change of the Lagrange multipliers was less than 10^{-5}. The data was normalized to the range $[-1, +1]$ before LSVM learning. The classification accuracy was evaluated using the number of correctly classified cases expressed as the percentage of the testing set size (CC). Since the classification accuracy is not adequate to describe the quality of the assessment in medical applications we calculated sensitivity (SE), specificity (SP), positive (PPV) and negative (NPV) predictive values. As the evaluation of the classifier when analyzing all the prognostic indices simultaneously is difficult we used also the index $\mathrm{QI} = \sqrt{\mathrm{SE} \cdot \mathrm{SP}}$.

In the first stage of experiments we classified the FHR signals using the fuzzy classifier. For the WFSS it was necessary to determine weights values c_j, defining the degree of certainty of the recording assessment on the basis of the evaluation of the particular single feature. The diagnostic values of the parameters were estimated using the ROC analysis. Mean values of AUC are presented in Table 2. The

Table 2. The values of the mean AUC for given parameter describing the FHR signal in relation to fetal distress

mFHR	STV	ACC	$\mathrm{DEC_A}$	$\mathrm{DEC_B}$	$\mathrm{DEC_B}$	$\mathrm{OSC_0}$	$\mathrm{OSC_I}$	$\mathrm{OSC_{III}}$
0.53	0.86	0.65	0.58	0.55	0.52	0.93	0.79	0.77

obtained results indicate high diagnostic values of the instantaneous FHR variability parameters (STV and oscillations). Nevertheless, the percentage of silent oscillations $\mathrm{OSC_0}$ can be distinguish as the best indicator of the fetal outcome.

Using the determined values of parameters we analyzed the FHR recordings with the WFSS. The classification results are presented in Table 3 (the first row). The obtained results indicate that we did not fulfill the objective of the first step of FHR signals analysis, which was the confirmation of the fetal well-being with the highest certainty. In order to improve the quality of the fuzzy assignment it was necessary to modify the boundaries for FHR quantitative parameters, previously determined in accordance with the FIGO guidelines (Table 1). The results of basic statistical analysis of parameters describing signals connected with fetal distress allowed for identifying new ranges of values for STV and ACC parameters. For recordings indicating fetal hypoxia the STV index did not exceed 8 ms, and the number of detected acceleration patterns 13 per hour.

Table 3. The results of FHR signals classification using WFSS

Prognostic index	QI [%]	SE [%]	SP [%]	PPV [%]	NPV [%]	CC [%]
Before modification ($s = -0.25$)	71.81	90.70	56.85	31.97	95.40	64.55
After modification ($s = -0.50$)	78.08	100.0	60.96	32.58	100.0	69.84

These values were assumed to be new boundaries of ranges representing the fetal wellbeing. Additionally, in the case of acceleration rate, we used the median of the ACC equal to 4 per hour as boundary between classes of suspicious and pathological FHR recordings. The additional statistical analysis of recordings characterized by the fetal wellbeing showed that in 95% of cases the value of OSC_0 in the entire recording does not exceed 7%. Therefore, we used this value as the boundary between suspicious and pathological classes of OSC_0. The new ranges of particular parameters are shown in Table 1 (presented in parenthesis as subscript of the original values).

The modification of ranges that correspond to particular classes of the fetal state assessment allowed to improve significantly the quality of fuzzy classification (Table 3, the second row). The increase of all prognostic indices was noticed. The resulted zero false-positives (SE = 100%) allowed for eliminating the FHR recordings that correspond to fetal wellbeing from the second step of the analysis using non-linear LSVM. The final results of the two-step analysis are presented in Table 4.

Table 4. The results of FHR signals analysis (mean values ± standard deviation [%])

	Radial kernel		Polynomial kernel	
	No first phase $\gamma = 3.000, \sigma = 0.6$	WFSS $\gamma = 0.300, \sigma = 0.9$	No first phase $\gamma = 0.007, n = 5.0$	WFSS $\gamma = 0.020, n = 3.0$
QI	83.3 ± 5.57	88.2 ± 4.31	83.4 ± 4.88	87.6 ± 4.75
SE	73.3 ± 10.04	82.4 ± 8.40	75.2 ± 9.62	82.1 ± 9.10
SP	95.2 ± 2.23	94.8 ± 2.09	93.0 ± 3.57	93.9 ± 2.35
PPV	81.9 ± 6.68	82.5 ± 5.78	76.7 ± 8.00	79.9 ± 6.20
NPV	92.6 ± 2.51	95.0 ± 2.24	93.0 ± 2.43	94.9 ± 2.43
CC	90.3 ± 2.19	92.0 ± 1.95	89.0 ± 2.48	91.3 ± 2.26

The obtained results show the benefits of the application of the initial assessment of FHR signals with the fuzzy inference. A combination of the WFSS and the LSVM for both types of kernel function allowed for increasing the classification quality, however the radial kernel provided slightly better results of the fetal distress prediction.

5 Conclusions

In the presented work we investigated the possibility of predicting the risk of fetal distress using the results of the fetal heart rate signal analysis. The FHR signals were evaluated on the basis of the two-step analysis using Weighted Fuzzy Scoring System and Lagrange Support Vector Machines. The application of fuzzy inference based on the FIGO guidelines allowed for indicating the FHR recordings that correspond to fetal wellbeing with the highest certainty. These recordings were eliminated from the second step of classification. The experiments showed high efficacy of the proposed procedure in the prediction of fetal distress.

Acknowledgments. This work was supported in part by the Polish Ministry of Sciences and Higher Education and by the Polish National Science Center.

References

1. Bernardes, J., Costa-Pereira, A., de Campos, D.A., van Geijn, H.P., Pereira-Leite, L.: Evaluation of interobserver agreement of cardiotocograms. International Journal of Gynecology & Obstetrics 57(1), 33–37 (1997)
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
3. Czabanski, R., Jezewski, M., Wrobel, J., Jezewski, J., Horoba, K.: Predicting the risk of low-fetal birth weight from cardiotocographic signals using ANBLIR system with deterministic annealing and ε-insensitive learning. IEEE Transactions on Information Technology in Biomedicine 14(4), 1062–1074 (2010)
4. Fischer, W.M., Stude, I., Brandt, H.: Ein vorschlag zur beurteilung des antepartalen kardiotokogramms. Geburtshilfe Perinatology 180, 117–123 (1976)
5. Hammacher, K.: Einfuhrung in die cardiotokographie, pp. 5953–1109. Hewlett Packard (1978)
6. Jezewski, J., Wrobel, J., Horoba, K., Kupka, T., Matonia, A.: Centralised fetal monitoring system with hardware-based data ow control. In: Proceedings of 3rd International Conference MEDSIP, Glasgow, pp. 51–54 (2006)
7. Jezewski, M., Czabanski, R., Wrobel, J., Horoba, K.: Analysis of extracted cardiotocographic signal features to improve automated prediction of fetal outcome. Biocybernetics and Biomedical Engineering 30(4), 39–47 (2010)
8. Krupa, N., Ma, M., Zahedi, E., Ahmed, S., Hassan, F.: Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine. BioMedical Engineering OnLine 10(6), 1–15 (2011)
9. Mangasarian, O.L., Musicant, D.R.: Lagrangian support vector machines. Journal of Machine Learning Research 1, 161–177 (2001)
10. Rooth, G.: Guidelines for the use of fetal monitoring. International Journal of Obstetrics & Gyneacology 25(3), 159–167 (1987)
11. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)

Bio-inspired Genetic Algorithms on FPGA Evolvable Hardware

Vladimir Kasik[1], Marek Penhaker[1], Vilem Novak[2], Radka Pustkova[1], and Frantisek Kutalek[1]

[1] VSB - Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, Department of Cybernetics and Biomedical Engineering, 17. Listopadu 15. 70833 Ostrava, Czech Republic

[2] Faculty Hospital Ostrava, Child Neurology Clinic, 17 Listopadu 1790/5, Ostrava, Czech Republic

{vladimir.kasik,marek.penhaker,radka.pustkova, frantisek.kutalek}@vsb.cz, vilem.novak@fno.cz

Abstract. The results presented in this article introduce the possibility of software processing by image data from CT and MRI in clinical practice. It is important to work with the most accurate data in the diagnosis and further monitoring of the patient. Especially in case of the birth defects or post-traumatic conditions of head called Hydrocephalus, it is necessary to work with this data. A production increase of the cerebrospinal fluid, called cerebrospinal fluid (CSF), causes bring intracranial pressure up. The oppression of the brain tissue has resulted of this procedure. The determination of CSF ratio to the skull in medical practice is used to improve diagnosis and monitoring before and after surgery in patients with Hydrocephalus diagnosed. Software was implemented in Matlab2006b using Image processing Toolbox. Next, the article also describes the design of hardware solutions to these methods of real-time image processing using FPGA programmable logic and genetic algorithm.

Keywords: CSF, Skull, FPGA, Genetic Algorithm.

1 Introduction

For diagnosis and further monitoring of a patient it is necessary to analyze continuously the available data. Especially in congenital defects and traumatic conditions head, due to the overproduction of cerebrospinal fluid increases intracranial pressure, and thus repression of brain tissue.

The results presented in this article provide means of software data processing from CT and MRI in clinical practice. The main goal is to determine the ratio of CSF to intracranial volume, which is vital for determining the progression of therapy.

2 Methods

To calculate there are used the single-frame semi-automatic selection masks. Input data are in JPEG format of 512 x 512 PX. JPEG format is sufficient for this application.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 439–447, 2012.

From the images there are also date obtained as the number of frames in the series, the distance between cuts, the number of PX per mm and the calculated conversion factor. These data are important for further processing. In the block diagram are represented by data.

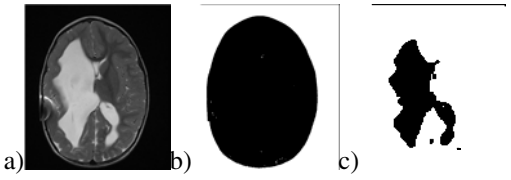


Fig. 1. a) Original MRI image, b) mask oft the skull and c) mask of the liquor

Some tissues share the same luminance value even though they are not in the selected choice, must be removed from the image. Conversely, some tissues, especially the cranial bones in pediatric patients were not due to lack of calcification and displays must be manually added. Individual functions and methods for proper selection of masks were processed in MATLAB. In the scheme of this procedure constitutes a block correction. The output is a series of mask data for cerebrospinal fluid and skull mask. Method of processing input images is shown in Fig.2).

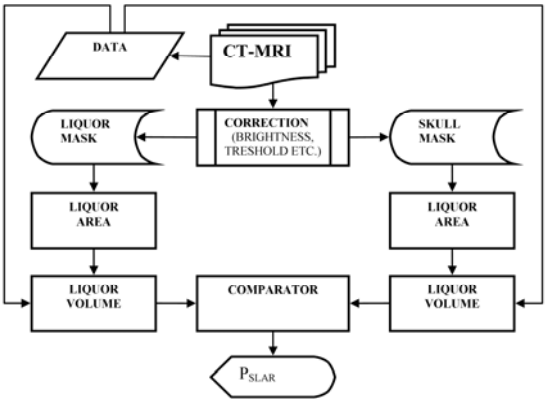


Fig. 2. Calculating scheme of the ratio of CSF to the skull

Input data were first regulated in the block luminance brightness. The removing noise from the input data is in charge of block noise correction. There was also needed a manual correction perform. Now the data are ready for segmentation of objects in the picture and create and save the masks.

The blocks liquor area and skull area are sorted data from cerebrospinal fluid and skull masks. Content area masks, calculated for the area of cerebrospinal fluid in one section, are determined by the number of black points in the mask. Entered by equation (1):

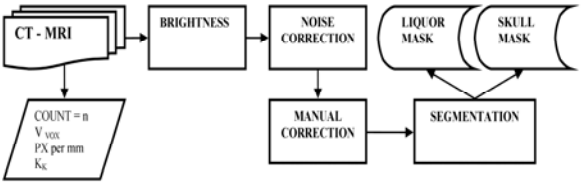


Fig. 3. Diagram shows the individual mask preparation

$$S_L = \sum_{i=0}^{i}\sum_{j=0}^{j} x_{i,jPRK} + \sum_{i=0}^{i}\sum_{j=0}^{j} x_{i,jNPRK} \quad . \tag{1}$$

The area is represented by a point in the masks of images. These points are overlap each other, when comparing the two images, marked $x_{i,jPRK}$ for points respectively S_{PRK} for the area, or not overlap, marked $x_{i,jNPRK}$ respective $S_{N\,PRK}$. This information is important for the calculation rounded. The scheme Fig.3) shows the function block Liquor area. Ranking data for talc CSF, as is the scheme for overlapping and the overlapping area is responsible for block comp. The procedure is applied to all data in the series. Then performed to calculate the volume represented by block Liquor volume.

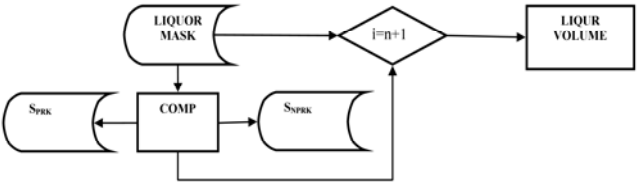


Fig. 4. Function of liquor area diagram block

To calculate the volume it is needed to know the distance v_{vox} of individual cuts. It was found analysis of input data. The resulting volume V_L , calculated for the volume of CSF is the sum of the volumes. Shows equation (2):

$$V_L = \sum_{n=1}^{n} \left(S_{LNPRK} \cdot K_k + S_{LPRK} \right) \cdot v_{vox} \quad . \tag{2}$$

The calculation should be calculated separately with data for the overlapping area and not overlapping areas. The not overlapping area is included in the calculation of the conversion factor K_k , which aims to refine the calculation of the radius edges.

The principle of calculating the volume of liquor is shown in Fig.4).

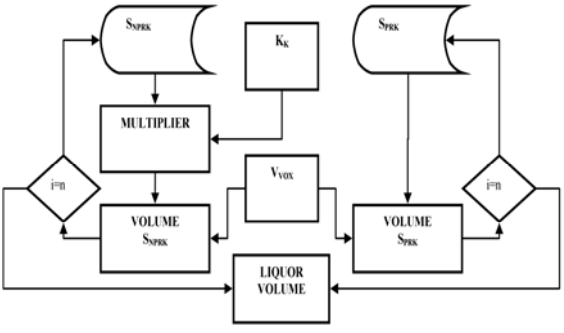


Fig. 5. Diagram shows the principle of the computation of the cerebrospinal fluid volume

The input data include information on the overlapping and not overlapping areas of two compared images. In case of no overlapping image is required overlapping area multiplied by the conversion factor. When using the conversion factor is the final volume of distorted error 6.86%. Provides block multiplier. Using aliasing filters can then move the value of distortion $\Delta V_L = 0{,}62 \pm 0{,}36\%$.

Multiply distance of each area v_{vox} is the task blocks volume S_{PRK} and volume $S_{N\,PRK}$. All data in the series applied. The function liquor volume block is findings the final volume of liquor. The same procedure is applied to the skull mask.

Data are ready to compare the different volumes. Compare two volumes between as it is an Comparator block, see. Fig.1), tasks. Determination of the ratio of CSF to the skull P_{SLar} is the most important output of the software since it indicates the current status of the patient. Entered by equation (3):

$$P_{SLar} = \left(\frac{V_L}{V_S} \right) \cdot 100\% \quad . \tag{3}$$

Where V_S is the volume of the skull. The V_L is the cerebrospinal fluid volume.

The resulting ratio is automatically calculated by the program.

3 FPGA Evolvable Hardware

The solution of the proposed methods in real-time leads to rapid FPGA hardware structures, which can perform some operations in parallel. For this purpose, the unit was designed for image processing with adaptive adjustment using genetic algorithm. The unit is built on the Xilinx FPGA architecture and uses a PowerPC processor inside the Virtex-II Pro, see Fig. 6.

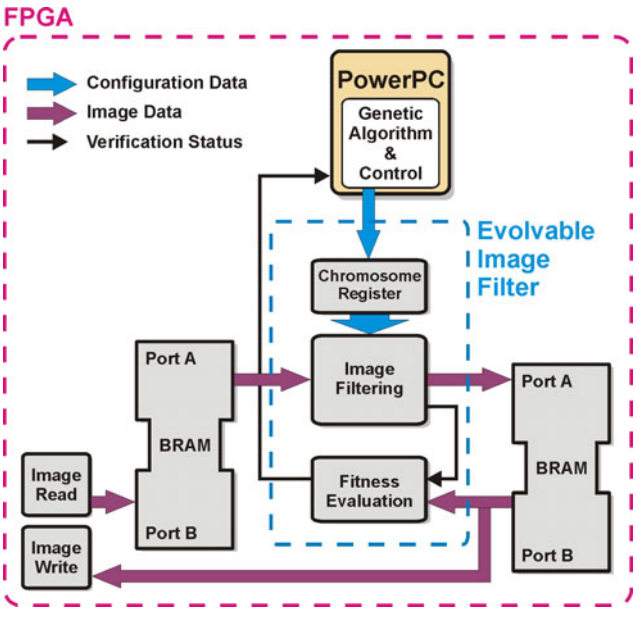


Fig. 6. Architecture of FPGA Evolvable Hardware

The design is based on Evolvable Image Filter, which performs matrix calculations of raster image. Image Filtering Block is configurable by the configuration bits, called chromosome. It is designed especially for convolution filtering with several types of masks. Type of the mask (kernel), type of filtering and their coefficients are parts of the information in the chromosome. A quality of the image filtering evaluates the Fitness Evaluation block, which monitors both the resulting image and continuous status information from the image filter. The whole process of chromosome evolution controls in the feedback the PowerPC processor using the genetic algorithm. For image storage are used two fast dual-port Block RAMs inside the FPGA. Image Read and Write Image blocks provide data conversion and communication with host system.

Hardware implementation of discrete convolution filter is solved in parallel, as shown in Fig. 7. For the elementary mathematical operations there are used fast multipliers and adders, so the resultant pixel is calculated in one clock cycle. This procedure utilizes significant amount of FPGA resources, especially the hardware multipliers.

However, at clock frequency of 100MHz, this allows the VGA image processing in just milliseconds. Such speed is necessary for a large number of generations using the genetic algorithm. The same demands are put on the speed of fitness evaluation. In this case, the evaluation has done in the actual process of filtering, when some parameters of processed image are transferred directly to a Fitness Evaluation block by individual data channel designated as Verification Status. The actual genetic algorithm is then implemented in software for the PowerPC processor on the FPGA chip.

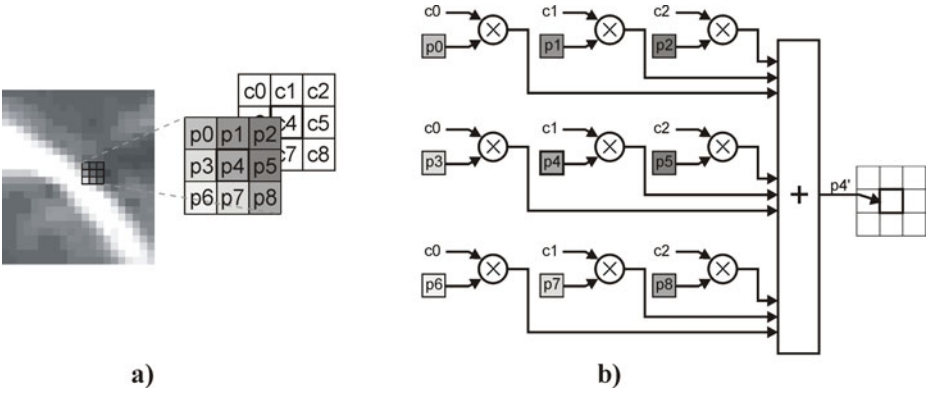


Fig. 7. Principle of the discrete convolution filter with 3x3 pixel mask. a) 3x3 pixel block p0 – p8 with mask c0-c8, b) hardware implementation of the filter for new p4' pixel calculation.

Data paths in Figure 7 represent buses with a width corresponding to the images bit depth. In the case of CT and MR images there is mainly an 8-bit grayscale format. Multipliers are used as standard 18-bit ones, while the adder can be created in any width from basic FPGA logic elements. Used wide data paths are well adapted to BRAM memories.

Table 1. Logic utilization of the Virtex-II Pro FPGA

Logic Utilization	Resources		
	Used	**Available**	**Utilization**
Number of Slices	3385	13696	25%
Number of Slice Flip Flops	5121	30816	17%
Number of 4 input LUTs	5480	30816	18%
Number of Block RAMs	92	136	68%

For large images processing can be obtained using the appropriate interface to use an external of memory such as SDRAM. For the processing of large images there can be used also external memories such as SDRAM with the corresponding interface. The image filter circuit also includes the logic for the filter mask feeding for each pixel, storing the resulting pixel in the memory and calculation of some statistical parameters of images for fitness evaluation.

4 Results Presentation

Presented solution was tested on a set of patient images of children with congenital hydrocephalus, a total of ten.

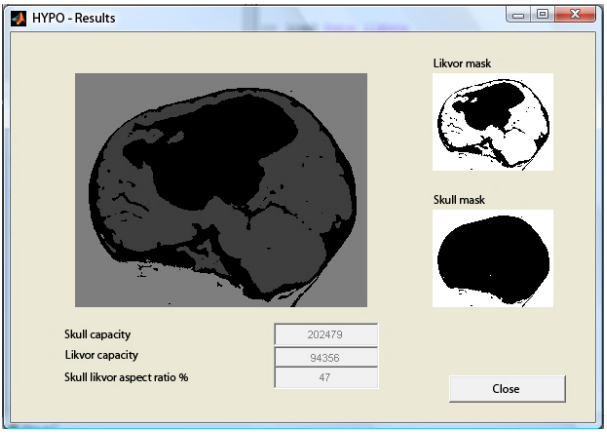


Fig. 8. Display ratio of CSF to mask skull cut before surgery

A data of patients were tested before and after surgery at intervals of several years with respect to increasing the volume of the skull in pediatric patients due to natural growth. The tested data showed change in the volume of cerebrospinal fluid before and after therapy. From the records that were taken at intervals of four years after the operation shows that the volume before $9{,}3cm^3$ and after $8{,}4cm^3$ surgery reduced the overall CSF volume by 10%.

This in the case of subjective assessment of the doctor could not be determined. In all cases the proposed algorithm successfully used and tested.

Table 2. Real data of kind patient with Hydrocephalus used by the same area before and after surgery. This data are before calculating.

cut	2004 before surgery		2008 after surgery	
	Liquor (PX)	**Skull (PX)**	**Liquor (PX)**	**Skull (PX)**
1	374341	2832128	345482	2596897
2	461426	2821788	436422	2542564
3	668928	2814429	517746	2524663
4	713404	2774935	672970	2450053
5	757437	2666283	633301	2411929
6	647723	2598683	636403	2401602
7	593565	2479986	620638	2343522
8	587979	2335049	562504	2256987
9	578149	2136652	541435	2210403
10	535405	2832128	532907	2035211
11	484295	1901501	473203	1830127
12	374341	1677214	383150	1570817
13	272873	1394122	169350	1312851

To solve this problem there was designed and implemented software, to semi automatic selection of images extracted from the processed desired area. Software was developed in Matlab2006b using Image Processing Toolbox, and uses knowledge in the field of image processing. The open design allows the software solutions developed.

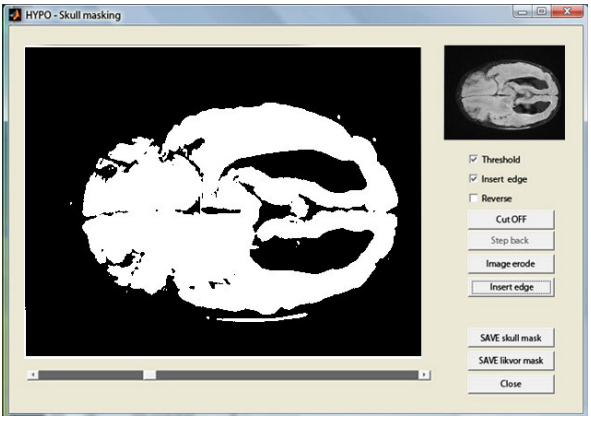


Fig. 9. Picture of the preparation of skull mask using liquor mask as input

Ratio determination of cerebrospinal fluid in the skull determined by the doctor based on subjective impression. The method used software processing more accurate. The results indicate that decreased CSF and at the same time to decreased the patient's.

5 Conclusion

It is a completely new method of calculating the volume of cerebrospinal fluid in the skull of the patients that is being developed.

Using the analogy to a sphere is to achieve better results than comparing individual images. It is better than without taking into account a rounding. Data processing is more automated, which eliminates error introduced by the user.

The results helped to make explicit decisions on the progression of treatment and the therapeutic techniques that were previously difficult objective.

Further improvements algorithms will eliminate errors in the semi automatic selection of masks, which currently can be up to 10%. Applications will aliasing filters, more precise calculation and reduce distortion.

Acknowledgments. The work and the contributions were supported by the project SV SP 2012/114 “Biomedical engineering systems VIII” and TACR TA01010632 “SCADA system for control and measurement of process in real time”. Also supported by project MSM6198910027 Consuming Computer Simulation and Optimization. The paper has been elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by Operational Programme 'Research and Development for Innovations' funded by Structural Funds of the European Union and state budget of the Czech Republic.

References

1. Cerny, M., Penhaker, M.: The HomeCare and circadian rhythm. In: Conference Proceedings 5th International Conference on Information Technology and Applications in Biomedicine (ITAB) in Conjunction with the 2nd International Symposium and Summer School on Biomedical and Health Engineering (IS3BHE), Shenzhen, May 30-31, vol. 1, 2, pp. 110–113 (2008) ISBN: 978-1-4244-2254-8
2. Penhaker, M., Cerny, M., Rosulek, M.: Sensitivity Analysis and Application of Transducers. In: 5th International Summer School and Symposium on Medical Devices and Biosensors, Hong Kong, Peoples R China, June 01-03, pp. 85–88 (2008) ISBN: 978-1-4244-2252
3. Kasik, V.: FPGA based security system with remote control functions. In: 5th IFAC Workshop on Programmable Devices and Systems, IFAC Workshop Series, Gliwice, Poland, November 22-23, pp. 277–280 (2001) ISBN:0-08-044081-9
4. Krejcar, O.: Large Multimedia Artifacts Prebuffering in Mobile Information Systems as Location Context Awareness. In: 4th International Symposium on Wireless Pervasive Computing, ISWPC 2009, Melbourne, Australia, February 11-13, pp. 216–220 (2009), doi:10.1109/ISWPC.2009.4800591, ISBN 978-1-4244-4299-7
5. Korpas, D., Halek, J.: Pulse wave variability within two short-term measurements. Biomedical Papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia 150(2), 339–344 (2006)
6. Bernabucci, I., Conforto, S., Schmid, M., D'Alessio, T.: A bio-inspired controller of an upper arm model in a perturbed environment. In: Proceedings of the 2007 International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 549–553 (2007) ISBN: 978-1-4244-1501-4
7. Garani, G., Adam, G.K.: Qualitative modelling of manufacturing machinery. In: Book - 32nd Annual Conference on IEEE Industrial Electronics, IECON 2006, vol. 1-11, pp. 1059–1064 (2006) ISSN: 1553-572X, ISBN: 978-1-4244-0135-2
8. Cerny, M.: Movement Activity Monitoringof Elederly People – Application in Remote Home Care Systems. In: Proceedings of 2010 Second International Conference on Computer Engineering and Applications, ICCEA 2010, Bali Island, Indonesia, March 19-21, vol. 2. IEEE Conference Publishing Services, NJ (2010) ISBN 978-0-7695-3982-9
9. Cerny, M.: Movement Monitoring in the HomeCare System. In: Dossel-Schleger (ed.). IFMBE Proceddings, vol. 25, Springer, Berlin (2009) ISBN 978-3-642-03897-6; ISSN 1680-07
10. Krištof, M., Hudák, R., Takáčová, A., Živčák, J., Fialka, L., Takáč, R.: Contact pressure measurement in trunk orthoses. In: ICCC-CONTI 2010 - Proceedings of IEEE International Joint Conferences on Computational Cybernetics and Technical Informatics, art. no. 5491304, p. 175, (2010), doi: 10.1109/ICCCYB.2010.5491304, ISBN: 978-142447433-2
11. Simsik, D., Galajdova, A., Majernik, J., Hrabinska, I., Želinsky, P.: The video analysis utilization in rehabilitation for mobility development. Lékař a technika. Česká republika, 4-5, Ročník 35, 87–92 (2004) ISSN 0301-5491
12. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking, Article ID 802523, 8 pages (2009), doi:10.1155/2009/802523

Influence of the Number and Pattern of Geometrical Entities in the Image upon PNG Format Image Size

Jiří Horák, Jan Růžička, Jan Novák, Jiří Ardielli, and Daniela Szturcová

VSB Technical University of Ostrava, Institute of Geoinformatics, 17. listopadu 15, 70833 Ostrava Poruba, Czech Republic
{jiri.horak,jan.ruzicka,jan.novak.st14,jiri.ardielli, daniela.szturcova}@vsb.cz

Abstract. The research is focused on the study of the impact of the number and the pattern of geometrical entities and colour models in map like drawings upon the file size of PNG format. The main outputs originate from the exploration of an extended sample set of PNG images generated by Web Map Services of selected European servers. Original images were subsequently transformed to different colour models of PNG (RGBA, RGB, Palette, interlaced, non-interlaced). The images were analysed according to the number of entities, style and size of lines and the number of used RGBA values. We found that the number of geometrical entities can be replaced by the ratio of foreground pixels in an image. The file size grows with higher image density according to different functions which are driven by colour models and line widths, partly also by patterns. In order to minimise files sizes it is recommended to transform images into appropriate palette models and to avoid anti-aliasing changes of transparency.

Keywords: PNG, file size, geometrical entity, colour model.

1 Introduction

PNG is a widespread format for various applications. A lossless compression, full colours, management of transparency, no licensing, and fast decompression are the main advantages. PNG and JPEG are also required for Web Map Services (WMS) compliant with INSPIRE directive (Infrastructure for Spatial Information in Europe, 2007/2/EC [3]). Utilization of WMS services grows rapidly. PNG is also frequently used for mobile devices. Mobile as well as mapping applications share a common critical issue – their performance strongly depends on a velocity of data transfers from distant sources. In case of poor conditions it is necessary to apply additional intelligence like pre-buffering techniques [7]. The velocity is significantly affected by file sizes. The objective of the study is to recommend suitable strategies for mapping applications how to shrink PNG files with respect to image complexity (number of geometrical entities, patterns, colours). Selected European geoportals serve as a source of map images however results may be applied for wide spectrum of applications where the effective management (control) of file sizes, according to the content, is crucial.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 448–457, 2012.

2 PNG Format

PNG (Portable Network Graphics) is a graphical format for lossless compression of raster graphics. It was created in order to improve and replace GIF format, because PNG offers 48-bit colour depth, better compression and eight-bit transparency (the alpha channel). This means that the image may be differently transparent in various parts. It can be delivered as palette-based images (with palettes of 24-bit RGB or 32-bit RGBA colours), grayscale images (with or without alpha channel), and RGB/A images (with or without alpha channel).

Using the transparency is necessary especially in cases of overlaying and merging outputs from several WMS sources [15].

The PNG file consists of a set of chunks which deliver elements of information concerning the applied palette (colour model), transparency, value range etc. PNG coding possibilities are completely described on the official page of the specification [11]. The structure can be analyzed using i.e. PNG Analyzer.

PNG is compressed using lossless algorithm Deflate (zlib library) which is a combination of LZ77 dictionary compression algorithm and Huffman coding [4]. Compression method is based on finding the same sequence of data in the pixel grid. A convolution filter is used before starting the compression process. This filter is looking for similarities among neighbouring pixels. It applies relevant operations with the purpose of increasing the probability that the data compression program finds more identical sequences and improves the compression ratio. Usually PNG generators offer five types of filters (None, Sub, Up, Average, Paeth). Principles of individual filters are explained in [11] or [8]. A possibility to use runlength-based image compression for PNG is discussed in [10].

PNG also offers an option to store interlaced images. In such files pixels are not stored point by point and line by line, but according to the algorithm Adam7 (two-dimensional) [12]. While downloading interlaced PNG and displaying, the image is immediately depicted on the screen. I.e. after 30% of the data depiction, it is possible to see and recognize the basic features of the image.

Many authors are interested in comparing the image file formats like PNG, JPEG, TIFF and BMP with each other ([1], [14], [5]). Less is known about the behaviour of filters for processing of map information, i.e. images generated by map servers. Such images typically contain simple linear or polygon patterns, supplemented by points in the objects of interest with limited number of colours and styles.

3 Methodology

Real objects are usually represented on digital maps by sets of basic discrete geometrical entities called features [9]. The complexity of maps influences file sizes due to worse compression behaviour.

First, the relationship between the number of features and the number of pixels was explored. Images were generated using ArcGIS 10 for different number of features (fig.1). The number of foreground (not-background) pixels was recorded. The high index of determination (ID=0.998 for linear features) supports the idea of using the

Fig. 1. Street network for testing images (left 800 features, right 1400 features)

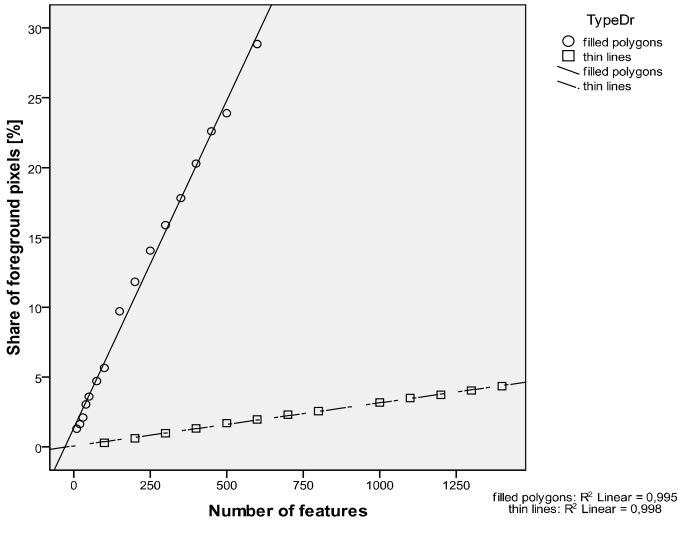


Fig. 2. Relationships between number of features and foreground pixels

share of foreground pixels in the image as an indirect indicator of image complexity (fig.2). Such indicator is suitable in cases when we study simple images namely with point or linear features. The application for cartoon images (relatively large filled polygons) may be misleading.

Further, we have explored 23 map services provided namely by national mapping agencies and selected 9 map portals (Estonian Land Board Geoportal, France Geoportal, German Federal Agency for Cartography and Geodesy, The Federal Institute for Research on Building, Urban Affairs and Spatial Development in Germany, Poland Geoportal, The Spatial Data Infrastructure of Spain, Directorate General for Cadastre - Spain, The Geoportal of Austrian Provinces, GeoNorge) for creating a sample set of map images. The sample set was generated using WMStester.

WMStester is a tool providing general random tests of WMS. The software is available at [http://sourceforge.net/projects/wmstester/].

WMStester requests various layers with linear features (including strictly linear features as well as "empty" polygons with linear borders like administrative units). We applied default settings provided by map servers including a standard graphical style of features (namely colour setting, linear pattern, line width, labelling) and a current setting of PNG originator (palette, transparency, interlacing). Following colour models occur: RGBA (32-bit RGBA colours), 8PLTE (8-bit palette) and 4PLTE (4-bit palette). The standard image size 800x600 pixels was requested [6].

For all obtained images we recorded the size, colour model and transparency, and the frequency of all unique combinations of colours and transparencies (unique sets of data in R, G, B, A channels) (hereafter refers as number of colours).

The background and foreground colours were distinguished and the share of foreground pixels was recorded. Usually setting of the alpha channel governs the background colour (full transparency for the selected colour – see table 1).

In case of several map servers we recognised a probable application of an anti-aliasing algorithm. I.e. the transparency in French road maps is smoothly modified to create buffer zones 3 pixels wide along 1 pixel wide road lines.

Table 1. Parameters of tested WMS map servers

Country	Topic	Background colour (R/G/B)	Background transparency	Model
France	hydrology	240/240/240	0	RGBA
France	administration	240/240/240	0	RGBA
France	transport	240/240/240	0	RGBA
France	transport, administration	0/0/0	0	RGBA
Germany	administration	255/255/255	0	RGBA
Germany	administration	234/245/188	255	RGBA
Austria	administration	0/0/0	0	RGBA
Spain	administration	183/216/247	0	4PLTE
Spain	cadaster	255/255/255	0	8PLTE
Estonia	transport	0/0/0	0	8PLTE
Norway	hydrology	0/0/0	0	8PLTE
Poland	cadaster	254/254/254		RGB

Finally, the sample images were externally converted and compressed into different colour models (applied SW ImageMagick 6.7.4 [http://www.imagemagick.org/script/index.php] [13] and optipng 0.6.5 [http://optipng.sourceforge.net/]). The goal is to estimate the impact of external optimised compression and impact of model change to file sizes.

4 Density and Different Types of Drawings

The sample set contains 822 images of different types and colour models (table 2).

Table 2. Colour model and type of drawing for sample images

Model	Type	Count of images
4PLTE	administrative units, irregular polygons, solid line	86
4PLTE	administrative units, solid line	41
8PLTE	hydrological network, thin solid line	54
8PLTE	transport network, combined solid line (symbols)	50
8PLTE	transport network, think solid line	40
8PLTE	cadaster, regular polygons, thin solid lines	7
RGB	cadaster, regular polygons, thin solid lines	32
RGBA	administrative units, irregular polygons, thick solid line	83
RGBA	administrative units, irreg. polygons, thick solid line, labels	68
RGBA	administrative units, irregular polygons, thin dashed line	84
RGBA	hydrological network, thin solid line	141
RGBA	road network, thin solid line	136

Thin lines are represented by width of 1 pixel, while thick lines are 2-3 pixels wide. Wider lines occasionally occur in map drawings.

The file size is highly influenced by density of a drawing, but dependences are different for different colour models (fig. 3). Due to different behaviours we analyse each colour model separately. In all cases the file size grows with higher image density. RGBA images demonstrate a slight increase in file sizes. They are organised in two separate branches of dependency described later. File sizes of 4PLTE images grow slightly faster with the density (quadratic polynomial function, ID=0.994). RGB images show also a mild increase in files size with the density (cubic polynomial function, ID=0.996). A steep increase is specific for 8PLTE images (linear dependency, ID=0.992). The position of this curve denotes significantly higher file sizes for 8PLTE than for other models with the same density of drawings which is curious. The explanation is mainly found in two following reasons. First, the filtering and compression applied on these servers (Norway, Estonia) are not optimised (after an external compression we obtain a reduction in file sizes between 15 and 25%). The second, these images contain high variation of transparency (creating an anti-aliasing effect) which causes unpleasant compression rates.

It is necessary to note that the colour model does not correspond fairly often to the real number of used colours (see next chapter). We explore the colour frequency in images, but no influence of the number of colours to the file size is recognised.

The regression analysis evaluates dependencies for all obtained combinations of colour models and drawings with a sufficient number of sample images (more than 60). All regression models are statistically significant ($p<0.01$) and pose the index of determination higher than 0.98. RGB images are excluded (only 32 images in several clusters). Table 3 summarizes outputs of regression analysis for each model according to the drawing density. The optimised (MLSE) regression functions are applied to calculate approximated file sizes for selected values of drawing density representing typical classes of map density.

The clear separation of drawing type for RGBA images is depicted in fig.4. RGBA thin lines are best represented by a cubic polynomial function (ID=0.992 and only 4%

of images are out of 95% confidence intervals). Thick or complex lines are well approximated by a linear function (ID=0.987) similarly to labelled thick or complex lines (ID=0.989). The influence of line width is quite clear – thick lines (3 pixels) enable to reach a better compression ratio (these files are significantly smaller than those ones for labelled or thin lines). In case of 5% image density the size increase is almost 50%.

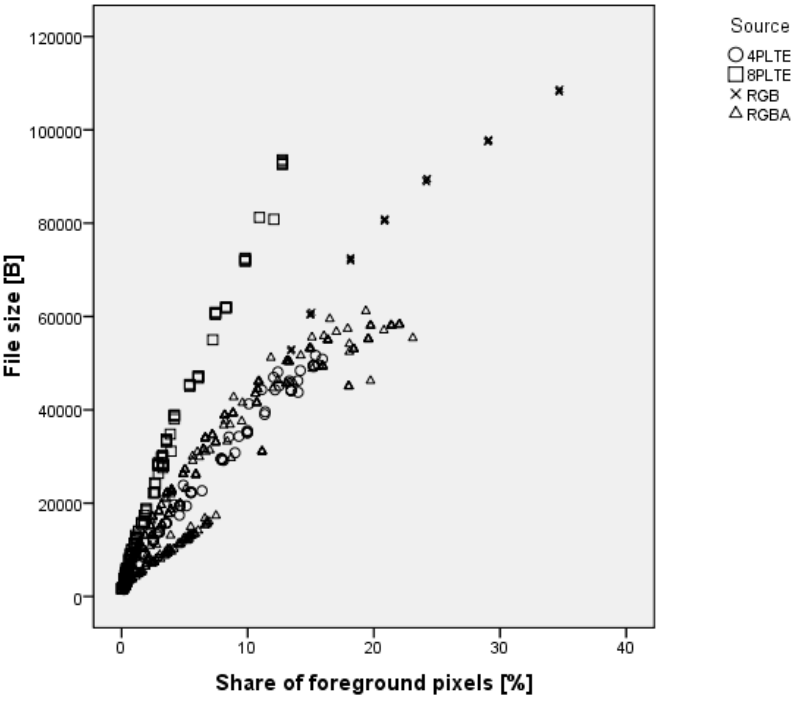


Fig. 3. Relationships between share of foreground pixels and the file size

Table 3. Approximated file sizes according density and PNG colour models

model:	4PLTE	8PLTE	RGBA	RGBA	RGBA
drawing type:	thin	*	thick	lab.thick	thin
Drawing density (%)	File sizes (kB)				
0,5	3	7	2	5	6
1	5	11	4	7	8
2	9	18	6	11	14
5	20	40	13	22	27
10	36	75	25	42	42
15	50	-	38	-	52

* Both thin and thick lines are presented, however with antialiasing and low compression.

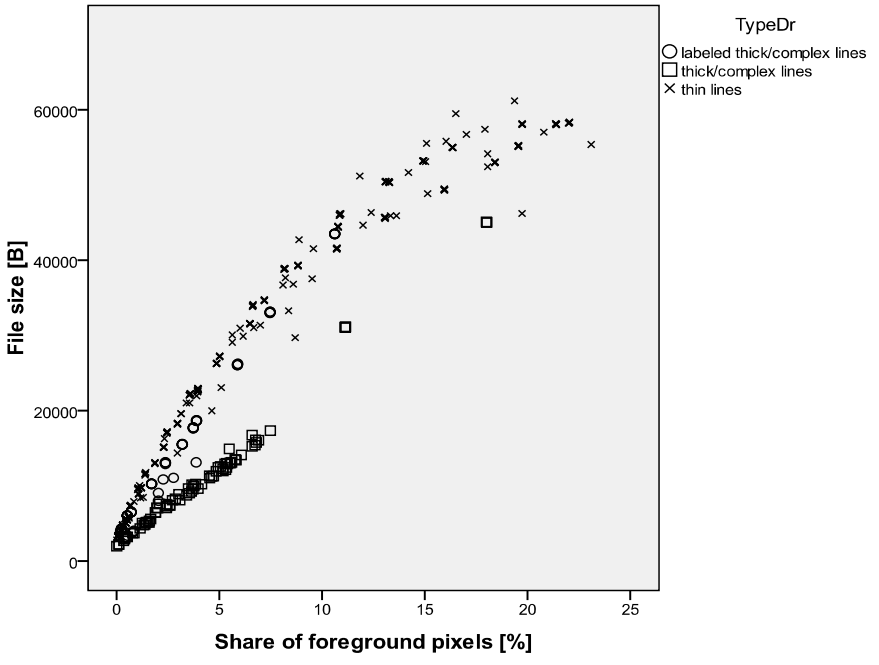


Fig. 4. Influence of drawing types to file sizes (only RGBA images)

It was assumed that regular polygons (bordered by straight lines) like cadastral maps should perform smaller file sizes due to easier compression. Nevertheless the exploration in behaviour of different topographic topics (administration, transport, and hydrology) does not show any significant relationship even for cadastral drawings. Only the study of separated thin lines shows small differences – road networks provide slightly smaller size of files then hydrological networks or administrative units.

5 Colour Models

A non-interlaced RGBA model (RGBAnI) dominates among default settings of map servers (table 4). The questions are whether it is sufficient and effective. The statistics of used colours (and transparency) (fig. 5) demonstrate that the number of used colours in RGBA is quite small, usually even smaller than counts in 8PLTE or RGB. 65% of RGBA images contain only 2 colours, 2% up to 16 colours, 25% between 16 and 256 colours, and only 8% more than 256 colours.

The testing of optimal additional filtration and compression using ImageMagick and optipng demonstrate possible significant reducing of file sizes. The optimal external processing provides about 40-50% decreasing of file sizes in case of transformation from RGBA to PLTE and 30-40% decreasing in case of changes from RGBA into RGB.

Table 4. PNG colour models of European WMS servers (default setting)

Colour model	Number of WMS servers
2/4/8bit, PLTE, no-interlace	5
8bit, RGBA, no-interlace	12
8bit,RGB, no-interlace	4
2bit, PLTE, interlace	2

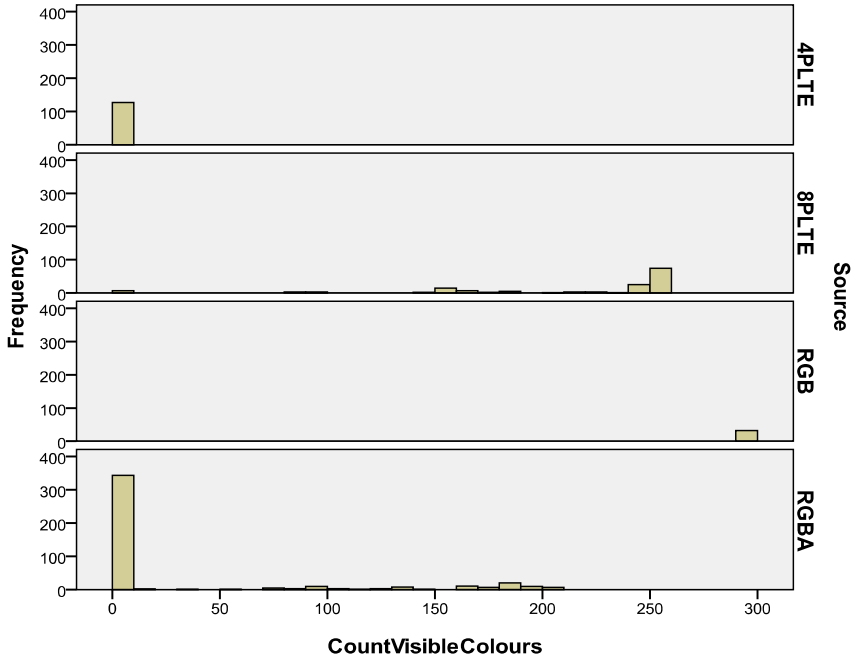


Fig. 5. Frequency of images according to the number of colours and models

The usage of transparency (alpha channel) increases the file sizes. Some authors warn the usage adds 33% to the size of RGB [12]. In case of map drawings we recognised about 6% increasing (manual testing in ArcGIS) and average 5% for Polish cadastral maps (the increment rises with the file size).

The more extended usage of transparency in the image will result in higher increase of file size. The impact of linear features anti-aliasing based on variable transparency is not measured in this study, thus we can only expect the increase is about 60%.

Also interlacing causes significant differences among file sizes. For the transformation from non-interlaced into interlaced RGBA we found important differences between thin and thick lines. Thin lines demonstrate about 10-20% increasing of file sizes due to interlacing, while thick or complex lines declare 30-45% increase (for file sizes above 5 kB).

6 Conclusions

The extended sample set of PNG images generated by European geoportals provides characteristics of real map images. It enables to focus on relevant applied strategies and to provide recommendations fitted for current needs. The study concentrates on map images with linear features and analyses both images provided by default settings as well as evaluate possible profits from improved external/internal filtering and compression.

52% of explored European map servers utilise RGBA model as default settings, nevertheless only 8% of their images contain more than 256 colours. The optimal solution should be to specify in all WMS requests an option for generating PNG image with the colour model appropriate to the data type. Such management may require 4PLTE or 8PLTE to be used instead of RGBA images for vector data. Another possibility is to change default setting directly to 8PLTE, because even for airborne or satellite images the visual difference between TrueColor PNG and Palette PNG is small [2].

If the size is critical issue, therefore anti-aliasing (locally variable transparency to create the effect of anti-aliasing) is not recommended. It leads to the substantial increase of file images (our expectation is more than 50%). It is more effective to use wider lines which enable better compression and smaller images.

Using interlaced images is a good option especially for slow networks. The penalty for using interlaced images is between 10 and 45% increase of file sizes depending on colour model and line width (the average of thick or complex lines is 38%).

The number of geometrical entities can be replaced by the ratio of foreground pixels in the image (drawing density). The file size grows with higher image density according to different function. The different relationships are driven by colour models and line widths, partly also by patterns (including labelling). The increase is smaller for RGBA and RGB than for 4PLTE or 8PLTE.

The influence of type of drawing was partly proved. The dependency of the file size on the line width was proved for RGBA images (for other models we did not collect enough reliable data). Thick lines (3 pixels) provide significantly better compression rate and resulted file sizes may reach only 50% of size for thin lines (1 pixel wide) (for 5% of drawing density). It was approved for small number of colours (up to 16). The anticipated influence of linear styles and type of drawings (regular lines x irregular lines) was not proved.

Acknowledgments. The research is supported by ESF CZ.1.07/2.2.00/15.0276 (Geocomputation).

References

1. Aguilera, P.: Comparison of different image compression formats. ECE 533 Project Report, Computer Aided Engineering (2006)
2. Comparison of png8 versus png24 - Mass GIS - CommonWiki, https://wiki.state.ma.us/confluence/display/massgis/comparison+of+png8+versus+png24

3. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community, INSPIRE (2007)
4. Feldspar, A.: An Explanation of the Deflate Algorithm, http://www.zlib.org/feldspar.html
5. Fränti, P., Hautamäki, V.: Compression of aerial images for reduced-color devices. In: Image and Video Communications and Processing, pp. 651–662. SPIE, Santa Clara (2003)
6. INSPIRE Commission Regulation (EC) No 976/2009 of 19 October 2009 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the Network Services, pp. 9–18 (2009)
7. Krejcar, O.: Localization by Wireless Technologies for Managing of Large Scale Data Artifacts on Mobile Devices. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 697–708. Springer, Heidelberg (2009)
8. Miano, J.: Compressed image file formats. Addison Wesley Longman, Massachusetts (1999)
9. Open GIS Consortium, Inc.: OpenGIS Simple Feature Specification for SQL. OGC (1999), http://portal.opengeospatial.org/files/?artifact_id=829
10. Oyvind, R.: Runlength-based processing methods for low bit-depth images. IEEE Transactions on Image Processing 18(9), 2048–2058 (2009)
11. Portable Network Graphics (PNG) Specification, 2 edn., http://www.w3.org/TR/PNG/
12. Roelofs, G.: PNG: The Definitive Guide. O'Reilly & Associates, Sebastopol (1999)
13. Still, M.: The Definitive Guide to Image Magick. Apress, Berkeley (2006)
14. Wiggins, R.H., Davidson, H.C., Harnsberger, H.R., Lauman, J.R., Goede, P.A.: Image File Formats: Past, Present and Future. Radio Graphics 21(3), 789–798 (2001)
15. Xuan, S.: Python for Internet GIS Applications. Computing in Science and Engineering 9(3), 56–59 (2007)

Threading Possibilities of Smart Devices Platforms for Future User Adaptive Systems

Ondrej Krejcar

University of Hradec Kralove, FIM, Department of Information Technologies,
Rokitanskeho 62, Hradec Kralove, 500 03, Czech Republic
Ondrej.Krejcar@ASJournal.eu

Abstract. Modern Smartphones bring to their users many advantages from well-known text prediction with T9 up to voice recognition or personal intelligent voice assistant. These new application features require more and more CPU power which makes a press to CPU producer to develop better and more powered CPUs. New trend can be recognized in multithreading as well as in multicore CPUs. This paper deals with threading possibilities of current versions of mobile devices with Windows Mobile or Android platform where we present a simple architecture to process data in hard or soft realtime with precise time base. We also deal with a future of Smartphones and possible trends to develop multithread application.

Keywords: Multicore CPU, Smart Device Platform, Android platform, threading, OpenCL framework.

1 Introduction

Modern digital world where we live is Smarter and Smarter, whereas information systems (including mobile devices) are more sophisticated, complex and Smarter. Magic word "Smart" is very famous for CPU and hardware manufacturers because of their strong CPU power needs. Operation systems can be nicer, smoothed and predict user needs what means "Smarter".

Current modern mobile phones are recognized as smart devices so they are called "Smart Phones" (Classic design of mobile phones is going to side track). There are several different operation systems existing for Smartphone devices as Windows Mobile, Apple iPhone iOS, Google Android, etc. However each one is trying to implement as much Smart capabilities as possible. Until this year (2011) only one CPU with one core was embedded to each one mobile device, so the CPU power was slightly limited and Smart capabilities was limited as well.

End of 2011 year bring new dimension of a free CPU power which will left unoccupied by operation system. This power can be used by various Smart capabilities to lift them to new dimensions of we recognize under Smart umbrella. The new free power is not created by classical way with increasing of semiconductor number, but it is done by increasing number of parallel CPUs or CPU cores. Table

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 458–467, 2012.

[Tab. 1] covers actual state of the art in this area where up to 64 CPU cores can be embedded to one small Smart Phone in very near future.

Table 1. Multi-Core CPUs for future Smart Devices [1], [2], [3], [4]

Model	Processor	Speed	Core	Note
Texas Instrument	TI OMAP 5	3x ARM (desktop PC power)	ARM Cortex A15 MPCore (up 2GHz)	60% power save, for Smart Phones, tablets
HTC Edge	Nvidia AP30 Tegra 3[1]	= Intel Core 2 Duo T7200	4x CORE 1,5 GHz	Android 4.0 Pocketnow.com
SamsungGalaxy S III	Samsung		4x CORE 2GHz	Phonearena.com Android 4.0???
Adapteva	Epiphany IV	46,667 x Apple A5 CPU	64 x CPU	Non SoC (System on Chip)

As it was mentioned one paragraph earlier the free CPU power is not done by classical way. It is a very important fact for working and using of this new free power in software meanings. We can say that everything will be parallelized or every software application will be parallelized or threaded. This fact will bring new possibilities, new designing needs, new developers' knowledge needs, etc. Threading possibilities of modern and future Smart Devices platforms will be discussed in the rest of this paper.

2 Parallelism as Future of Smart Devices

Several years ago when Intel comes with multicore CPU a small revolution was started at desktop PC platform. Now the same things coming to Smart devices platform. It is look like the history is repeated, but nothing is the same. Small mobile devices are not exactly PC platform or portable notebooks platform. Smartphones has several limitations and several different needs to be taken into the account. Of course the current time is quicker and business is harder. Now we can recognize several attempts to win this battle.

2.1 Nvidia Tegra Solution for Smartphones and Tablets

New chipset Tegra 3 from Nvidia company is suitable for Smartphones as well as for tablets, where it comes with a CPU power comparable to desktop PC platform. Tegra 3 is designed with four cores ARM CPU at frequency up to 2 GHz. Platform has also fifth core for power safe mode and power graphic card GeForce with 12 cores.

[1] CPU power is equivalent to Intel Core 2 Duo T7200, which is 5 years old processor with 2GHz.

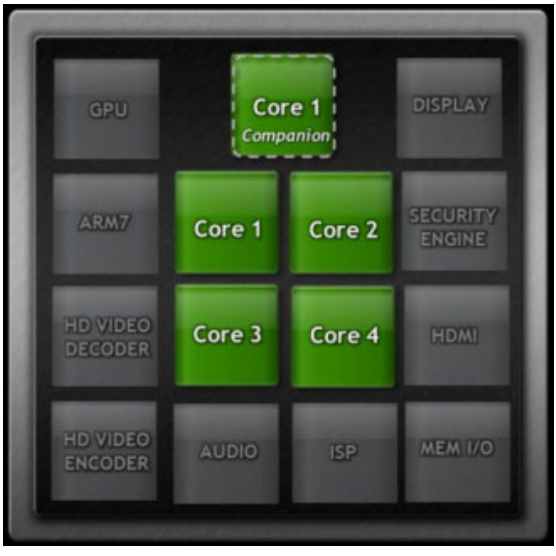


Fig. 1. Tegra 3 chipset overview. Power safe CPU named Companion [22]

CPU power is five times better than in older Tegra 2 case. Such great power is possible due to very wide scalable architecture:

a) Fifth core "Core 1 Companion" is used every time when there is no need of great power – e-g- in stand-by mode, during video playing, menu browsing, etc. Frequency is from 0 Mhz to 500 MHz. This core has another very interesting feature – it is hidden for operating system. This core is attached by CPU on demand.
b) If there is some need to higher power, CPU can wake up another from 4 cores in dependence on the exact power needs [Fig. 2], [Fig. 3].

Fig. 2. Tegra 3 CPU core load balance when watching movie [23]. White color mean enabled and used core, while green core means detached core.

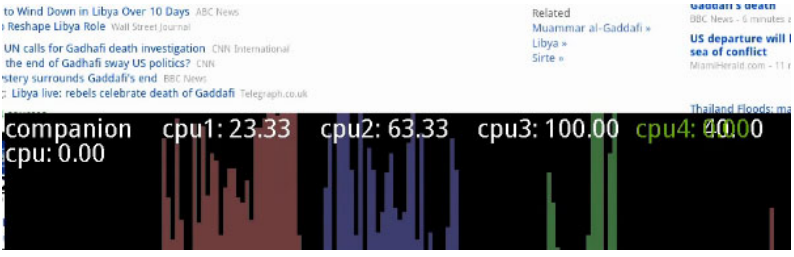


Fig. 3. Tegra 3 CPU core load balance when browsing on web pages [23]. Companion core is detached.

When common activities are made only one CPU core is used at frequency up to 1,4 GHz [Fig. 2]. CPU core activity history can be seen from opening movie and attaching and detaching of core 1 to 3.

More exacting task can take up to 3 cores as in case of web page browsing [Fig. 3]. Currently some games are in developing process especially for Tegra 3 chipset and GPU which can use provided CPU power and all features as rendering etc. Here we can see first developers issue – new or existing games need to be modified especially for this new platform. This is not only about games, but generally about all applications development. When we want to use all features, we need to use parallel architecture in our developed applications.

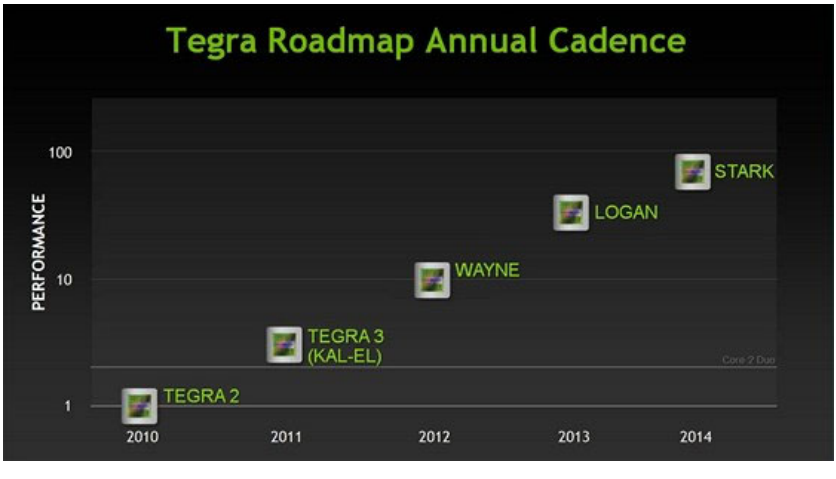


Fig. 4. Tegra Roadmap for future years [22]

For 2012 year a new generation of Tegra platform is scheduled under the code name “Wayne”, which will be 10 x more powered than Tegra 2 chipset. In case of “Stark” chipset the power will be 70 x more than Tegra 2. Can we imagine possibilities of such CPU power? This is a power of current clusters! But the question is how many cores will be embedded in chipset? Which design patterns will be effective for such great number of cores?

2.2 Adapteva Epiphany IV with 64 Cores CPU Unit

Previous subchapter deal with 4 core CPU for Smartphones. However there is a company Adapteva which develop chip Epiphany IV with incredible 64 cores. Epiphany is however slightly different of classical CPU architecture with System on Chip (SOC). It is more coprocessor unit than CPU. Such coprocessor is able to compute all necessary tasks which are needed by main CPU unit.

When all power is used, it can be app. 46,667 more power than current A5 CPU from Apple.

In previous subchapter we sad that such powered CPU will come from Nvidia in 2015, but from Adapteva it is possible to have it embedded in Smartphones earlier. Epiphany IV can scale up to 4096 cores for high-performance computers so the features of this platform are unbreakable.

3 Threading Principles of Selected Platforms

Future trends and future technologies which are coming now will put a strong press to developers to change their mind from single thread applications to multicore applications. Some of mentioned technologies can be covered by known principles or solutions for small amount of CPU cores, because of their equivalent in desktop solutions. However trend is clear [Fig. 4] targeted to tens of cores per one device. Such devices cannot be covered by traditional threading solutions, but need to be inspired from big room computers (like DeepBlue from IBM) or from mathematical coprocessor units (like used in graphical cards known NVidia CUDA technology).

What is the current state of the art at Smartphone area? We develop and tested realtime possibilities of two platforms Windows Mobile and Android. Somehow different platforms they use several common principles to run parallel tasks.

3.1 Realtime Data Processing on Windows Mobile 6.5 Platform

Realtime data processing on devices running Compact Framework v. 3.5 (All current devices with Windows Mobile 6.5) is solved on two levels. On first level processed data needs to be separated into different threads. On the second level the managed code needs to call some unmanaged libraries that manage Win32 named events. C# managed code has advantage in type safety. Beyond an automatic memory cleanup with help of garbage collector (it's method GC.Collect()) is disadvantage in realtime applications. Hard realtime functionality is possible to implement into an application running on .NET Compact Framework only by implementing Win32 unmanaged functions and events by means of P/Invoke platform invoke from managed C# code. Due to task separation it is necessary to watch code thread-safety.

Realtime Measuring/Processing

Data measuring phase is performed in realtime, but data analysis and presentation is performed within milliseconds after measuring had occurred. Then realtime

measuring is actually pseudo-realtime if we take data presentation and analysis into account. To have a non pseudo-realtime parallelism, we need to have multicore/multiCPU hardware which is actually coming with future Smart Phones.

Task Separation into Threads and Implementation

Threads must be synchronized and thread that collects measured data must have the highest priority. Thread performs the data collection into shared data buffer and signals other threads to perform their tasks. In the meantime the collecting thread is collecting new set of data into shared buffer.

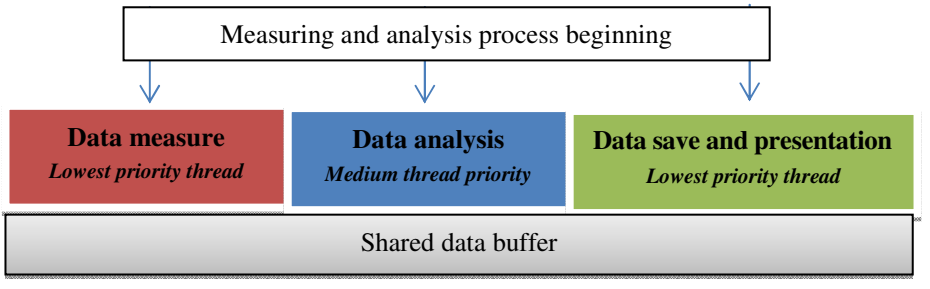


Fig. 5. Dividing of application to threads

Implementation itself in measuring application is made by EventWaitHandle class from namespace System.Threading, which by method WaitOne() call on its static instance and suspend all threads waiting for the data (consumer thread – data analysis, saving, presentation [Fig. 5]) until Set() method (from Producer) class is called on the same static instance which signals and resumes all suspended threads. Resumed threads in the order of set priority copies buffered data into own variables. During the period when buffer copying data same buffer needs to be locked by the lock statement. EventWaitHandle class is a simple semaphore synchronization implementation in .NET Compact Framework.

Thread Implementation Example

```
public    EventWaitHandle    processGvector    {get{return
this._processGvector;}}
this._mainWin.processGvector.WaitOne();
                              //Consumer is suspended
lock (this._gDataList)   //Lock of the shared buffer
{
this._gDataList.Add(new      double[3]{this._sensorData.X,
this._sensorData.Y, this._sensorData.Z});
}       //Data save into the shared buffer
this._processGvector.Set();
            //Producer signalization to consumer threads
```

Hard-Realtime Functionality Implementation

Platform invoke lets CLR (Common Language Runtime) managed environment call methods from unmanaged Win32 dynamic libraries. Set of native functions (CreateEvent, WaitForSingleObject, EventModify, CloseHandle), which lets native event management is key-note for hard-realtime functionality. During this event handling crated from native libraries no thread suspension from Garbage collector occurs.

Sample of Platform Invoke

```
using System.Runtime.InteropServices;
[DllImport("CoreDLL.dll", SetLastError = true)]
public static extern IntPtr CreateEvent(int alwaysZero,
int manualReset, int initialState, string eventName);
```

Timer Resolution Problem and Multimedia Timers

Especially for some kind of measuring application or user application where we need to process data in realtime as well as with absolutely exact time base a considerable problem will grow with using of standard timers. These timers from naspace System.Windows.Forms.Timer are not suitable for fine measurement, because of their tolerance. E.g. when we want to process data every 500 ms due to the systems messages front every processing is made every 400 to 600 ms. Error is +- 100 ms!

Solution can be found in using of multimedia timers, which can be adapted by using of OpenNETCF Smart Device Framework, which includes class (wrapper) Timer2 for working with these timers in OpenNETCF.Timers namespace.

3.2 Realtime Data Processing on Android Platform

Google Android platform is based on Linux OS, where a new Dalvik Virtual Machine (VM) is presented. There are several new projects developing a solution with multithreading [5]. Dalvik VM is a Java language VM. Realtime possibilities of Android platform are by this reality combined from capabilities of linux OS and virtual machine.

Dalvik VM can run several independent processes with own address areas and memory. Each Android application is mapped on linux process and at the same time it is capable use of his inter-process communication and synchronization.

Dalvik VM depend on core of linux OS in sense of task switching management. This mean that all threads are scheduled with priority SCHED_OTHER [6] and they are scheduled using CFS (Completely Fair Scheduler) Due to these limitations it is not possible to develop and run hard realtime application at Android platform but only soft realtime with the same principle as in the first case of Windows Mobile.

3.3 Testing of Realtime Data Processing

To compare a real hardware capabilities of current Smartphones we executed a several tests with implemented solutions presented in this paper.

For Windows Mobile platform we use HTC Athena device, where threads was implemented by the producer-consumer design pattern. We also test .NET timers and multimedia timers.

For second Android platform we select LG Optimus One with background worker threads scheduling in mode "Sensor delay game" or "Sensor delay fastest".

Table 2. Android and Windows Mobile realtime processing comparison

		Max. frequence [Hz]	Processing error [%]
WM 6	.NET Timers	13	10-15
	MM Timers	20	5-7
Android	Sensor Delay Game	40	7
	Sensor delay Fastest	22	20

3.4 OpenCL Framework

Open Computing Language (OpenCL) is a framework to develop parallel applications for many current CPU, GPU and other architectures. Language is based on C99 language [24]. OpenCL can be used to develop parallel application with task-based and data-based parallelism. Framework was born in 2008 at Apple Company, but now it is fully open (companies like Intel, AMD, Nvidia and Arm has adopted it).

This framework is then a cross platform solution to develop application for almost every processor on the market.

There are 19 research papers in Thomson ISI Wok database with only one targeted to mobile device platform [25]. It is evident that this fact will be changed in very near future because of need of power by new Smart application areas.

4 Conclusions

Threading or multitasking is well known in all mobile devices platforms. However until this year only pseudoparalelism can be used to resolve this problem. With end of 2011 year a new world come bringing several concepts with multicore CPUs or GPUs for mobile platforms. Traditional developers knowledge will need to be changed as well as new principles will be added.

We present some practical tests comparing two architectures Windows Mobile and Android and its suitability for realtime processing or measuring. In some Smart concepts a precise time base is very important to get required "intelligent and Smart" answers.

As we can predict a future of Smartphones the new framework for parallel computing will be used widely known as OpenCL allowing creation of application under one language for several CPU architectures.

Acknowledgement. This work was supported by „SMEW – Smart Environments at Workplaces“, Grant Agency of the Czech Republic, GACR P403/10/1310. We also acknowledge support from student Jakub Jirka in real application verification of proposed solutions.

References

1. TI Omap5 platform, http://www.ti.com/general/docs/wtbu/wtbuproductcontent.tsp?templateId=6123&navigationId=12862&contentId=101230&DCMP=OMAP5&HQS=Other+PR+wbu_omap5_pr_lp
2. Planck, S.: HTC Edge quad-core NFC phone coming March or April, http://www.nfcrumors.com/tag/nvidia-ap30-tegra-3-cpu/
3. Jacob, Samsung Galaxy S III with quad-core processor to land in early (2012), http://www.technobloom.com/samsung-galaxy-s-iii-with-quad-core-processor-to-land-in-early-2012/221699/
4. Ackerman, L.: Adapteva Announces 28nm 64-Core Epiphany-IV Microprocessor Chip - The Most Energy Efficient Microprocessor Solution in the World, http://www.businesswire.com/news/home/20111003005283/en/Adapteva-Announces-28nm-64-Core-Epiphany-IV-Microprocessor-Chip
5. Macario, G., Torchiano, M., Violante, M.: An in-vehicle infotainment software architecture based on google android. In: SIES, pp. 257–260. IEEE, Lausanne (2009)
6. Bodnarova, A., Fidler, T., Gavalec, M.: Flow control in data communication networks using max-plus approach. In: 28th International Conference on Mathematical Methods in Economics, pp. 61–66 (2010)
7. Brida, P., Machaj, J., Benikovsky, J., Duha, J.: An Experimental Evaluation of AGA Algorithm for RSS Positioning in GSM Networks. Elektronika Ir Elektrotechnika 8(104), 113–118 (2010) ISSN 1392-1215
8. Chilamkurti, N., Zeadally, S., Jamalipour, S., Das, S.K.: Enabling Wireless Technologies for Green Pervasive Computing. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 230912, 2 pages (2009)
9. Chilamkurti, N., Zeadally, S., Mentiplay, F.: Green Networking for Major Components of Information Communication Technology Systems. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 656785, 7 pages (2009)
10. Hii, P.C., Chung, W.Y.: A Comprehensive Ubiquitous Healthcare Solution on an Android (TM) Mobile Device. Sensors 11(7), 6799–6815 (2011), doi:10.3390/s110706799
11. Juszczyszyn, K., Nguyen, N.T., Kolaczek, G., Grzech, A., Pieczynska, A., Katarzyniak, R.P.: Agent-Based Approach for Distributed Intrusion Detection System Design. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 224–231. Springer, Heidelberg (2006)
12. Labza, Z., Penhaker, M., Augustynek, M., Korpas, D.: Verification of Set Up Dual-Chamber Pacemaker Electrical Parameters. In: The 2nd International Conference on Telecom Technology and Applications, ICTTA 2010, Bali Island, Indonesia, March 19-21, vol. 2, pp. 168–172. IEEE Conference Publishing Services, NJ (2010), doi:10.1109/ICCEA.2010.187, ISBN 978-0-7695-3982-9

13. Liou, C.Y., Cheng, W.C.: Manifold Construction by Local Neighborhood Preservation. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 683–692. Springer, Heidelberg (2008)
14. Liou, C.Y., Cheng, W.C.: Resolving Hidden Representations. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 254–263. Springer, Heidelberg (2008)
15. Mikulecky, P.: Remarks on Ubiquitous Intelligent Supportive Spaces. In: 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, pp. 523–528. Univ. Houston, Houston (2009)
16. Popek, G., Katarzyniak, R.P.: Measuring Similarity of Observations Made by Artificial Cognitive Agents. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS (LNAI), vol. 4953, pp. 693–702. Springer, Heidelberg (2008)
17. Shih, G., Lakhani, P., Nagy, P.: Is Android or iPhone the Platform for Innovation in Imaging Informatics. Journal of Digital Imaging 23(1), 2–7 (2010), doi:10.1007/s10278-009-9242-4
18. Skorupa, G., Katarzyniak, R.: Conditional Statements Grounded in Past, Present and Future. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part III. LNCS (LNAI), vol. 6423, pp. 112–121. Springer, Heidelberg (2010)
19. Thompson, T.: The Android mobile phone platform - Google's play to change the face of mobile phones. DR Dobbs Journal 33(9), 40-+ (2008)
20. Tucnik, P.: Optimization of Automated Trading System's Interaction with Market Environment. In: Forbrig, P., Günther, H. (eds.) BIR 2010. LNBIP, vol. 64, pp. 55–61. Springer, Heidelberg (2010)
21. Zelenka, J.: Information and Communication technologies in tourism – influence, dynamics, trends. E & M Ekonomie A Management 12(1), 123–132 (2009)
22. Mihu, F.: NVIDIA introduces the first Quad-core processor Tegra 3 (Kal-El), http://www.pocketdroid.net/nvidia-introduces-the-first-quad-core-processor-tegra-3-kal-el
23. NVIDIA Tegra 3: Fifth Companion Core, http://www.youtube.com/watch?v=R1qKdBX4-jc&feature=player_embedded
24. Stone, J.E., Gohara, D., Shi, G.C.: OpenCL: A Parallel Programming Standard for Heterogenous Computing Systems. Computing in Science & Engineering 12(3), 66–72 (2010)
25. Leskela, J., Nikula, J., Salmela, M.: OpenCL Embedded Profile Prototype in Mobile Device. In: IEEE Workshop on Signal Processing Systems (SIPS 2009), pp. 279–284 (2009)

Content Based Human Retinal Image Retrieval Using Vascular Feature Extraction

J. Sivakamasundari[1], G. Kavitha[2], V. Natarajan[1], and S. Ramakrishnan[3]

[1] Department of Instrumentation Engineering, MIT Campus, Anna University, India
[2] Department of Electronics Engineering, MIT Campus, Anna University, India
[3] Biomedical Engineering Group, Department of Applied Mechanics, IIT Madras, India
sivakamasundarij17@gmail.com

Abstract. In this work, an attempt has been made to analyze retinal images for Content Based Image Retrieval (CBIR) application. Different normal and abnormal images are subjected to vessel detection using Canny based edge detection method with and without preprocessing. Canny segmentation using morphological preprocessing is compared with conventional Canny without preprocessing and contrast stretching based preprocessing method. Essential features are extracted from the segmented images. The similarity matching is carried out between the features obtained from the query image and retinal images stored in the database. The best matched images are ranked and retrieved with appropriate assessment. The results show that it is possible to differentiate the normal and abnormal retinal images using the features derived using Canny with morphological preprocessing. The recall of this CBIR system is found to be 82% using the Canny with morphological preprocessing and is better than the other two methods. It appears that this method is useful to analyze retinal images using CBIR systems.

Keywords: Retinal image, Canny, retinal vessel, content based image retrieval, morphological preprocessing.

1 Introduction

The human retina consists of optic disc, fovea and blood vessels. Among all the structural components of eye, the retinal vessels play an important role in revealing the state of disease and also serve as landmarks for image-guided treatment. The retinal diseases such as glaucoma, age related macular degeneration, Diabetic Retinopathy (DR), hypertensive retinopathy and arteriosclerosis are assessed by observing the changes in blood vessel pattern [1]. Manifestations of several vascular disorders, cardio vascular disease and stroke depend on the analysis of the vascular networks.

Retinal fundus images of humans play an important role in the detection and diagnosis of many diseases of the eye [2]. Fundus images acquired in a non invasive manner are widely used by the medical community for large scale screening of patients. Early recognition of changes in the blood vessel pattern can prevent major vision loss [3].

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 468–476, 2012.

When the number of vessels are large or a large number of images are acquired, manual delineation and detection of vessels become tedious. The vessels in a retinal image are complex and have low contrast. This necessitates the need for a reliable automated method for extracting and measuring vessels in retinal images. The structural characteristics of blood vessels such as length, width, tortuosity and branching pattern not only provide information about pathological changes but also help to diagnose the disease and grade disease severity. Retinal images must be accurately segmented to extract sensitive objects present in the fundus image. Retinal vessel segmentation simplifies screening for retinopathy, aids micro aneurysm detection and delineates the optic disc and fovea [2].

Several methods have been carried out for vessel segmentation in retinal images. They are based on various algorithms which include intensity edges, matched filters, adaptive thresholding, intensity ridges and wavelets [4]. Many retinal image processing algorithms based on edge detection, classifiers and vessel tracking methods have been widely employed [5]. Recently, ACO method has also been used to segment blood vessels from retinal images [1].

Content Based Image Retrieval (CBIR) is a system for browsing, searching the query image and retrieving similar images from large databases. CBIR allows users to query and based on the image derived features, matching is carried based on automatically extracted primitive features such as color, shape, texture, and spatial relationships among objects [6]. CBIR based solutions have been explored for developing diagnosis aid in medical imaging solutions. CBIR has also been attempted for diagnosis of retinal diseases using large database. Retrieval based on statistical features of DR lesions is mapped on a semantic space corresponding to disease state using Fischer discriminant analysis [7]. An automated disease separation method to assist in CBIR for Stargardt's disease and age related macular degeneration has also been proposed [8]. Characterizing medical images using CBIR has also been carried out by building signatures from the distribution of wavelet transform coefficients [9]. Recently, detection of maculopathy and assessment of its severity level using symmetry based descriptor have been implemented using CBIR [3].

In this work, an automated CBIR system for the identification of diabetic retinopathy in retinal images is presented. The considered images are subjected to different types of enhancement namely, morphological and contrast stretching methods. Segmentation of blood vessels is carried out using Canny edge detection method with and without preprocessing and features are extracted. An effective feature vector called ratio of vessel to non vessel area is obtained. This measure is used along with statistical features such as entropy, energy, contrast, homogeneity for matching and retrieval.

2 Methodology

The architecture of the proposed CBIR system for retinal image assessment is shown in Fig.1. The proposed framework is divided into two main subsystems namely, the enrolment and the query subsystem. The enrolment subsystem is responsible for acquiring the information that will be stored in the database for later use.

The retinal images are taken from the publicly available image databases such as Digital Retinal Images for Vessel Extraction (DRIVE) and Diabetic Retinopathy database1 (DIARETDB1). These retinal images are acquired through high sensitive color fundus camera with constant illumination, resolution, field of view, magnification and dilation procedures. The different normal and abnormal retinal images from the image database are resized into 256 x 256. The resized images are preprocessed by morphological operator and segmented by Canny edge detection. Necessary feature vectors are extracted and stored in a feature vector database of this system.

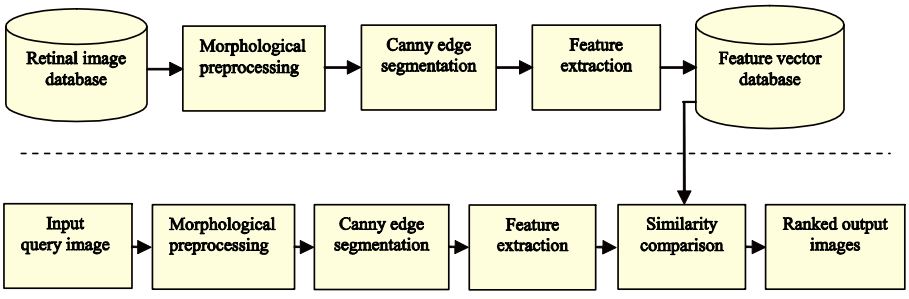


Fig. 1. CBIR framework

The query subsystem is responsible for retrieving similar images from the retinal image database according to the user's query image. The query subsystem receives an input query image from the patients. The query image is resized into 256 x 256 and preprocessed by morphological operator. Preprocessed image is segmented by Canny edge detection. Essential features are also extracted from the query image.

Mathematical morphological operators are used in applications related to image enhancement and segmentation [2].This system uses morphological and contrast stretching for preprocessing. The morphological enhancement is performed on the gray scale input image using the structuring element decomposition by disc-shaped opening. This enhancement is used to correct uneven background illumination from an image. The contrast stretching enhancement is performed to increase the contrast between the vessels and the background. Contrast Stretching is a piecewise linear transformation function which expands the dynamic range of intensity of the gray levels in the image. It is used to spread the details of an image from small range of pixel values to various degrees of gray levels to maintain the shapes [10], [11].

Edges characterize object boundaries and are useful for image segmentation, registration and identification of objects. In this work, the images are segmented using a Canny edge operator [12] to increase the separation between vessel and non vessel pixels.

Features are widely used in classification of textures and can also be applied to retrieval problems. Feature extraction methods are used for constructing combinations

of the variables and describe the given data accurately. Feature extraction algorithms and similarity measures used for image comparison underlie any CBIR system [13]. Medical images of different categories are distinguished using their homogeneousness characteristics [14].

Features such entropy, energy, contrast, homogeneity and maximum probability are computed by the following equations respectively.

$$\text{Entropy} = -\sum_i \sum_j C(i,j) \log C(i,j) \tag{1}$$

$$\text{Energy} = \frac{1}{MN} \sum_{i-1}^{M} \sum_{j-1}^{N} \{C(i,j)\}^2 \tag{2}$$

$$\text{Contrast} = \sum_i \sum_j (i-j)^2 C(i,j) \tag{3}$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{C(i,j)}{1+|i-j|} \tag{4}$$

$$\text{Maximum Probability} = \max C(i,j) \tag{5}$$

where, i and j are the pixel positions, C is pixel intensity and M, N correspond to row and column of the retinal image respectively. Other features such as minimum value, mean, standard deviation, median, range of intensity and ratio of vessel to non vessel area are also obtained.

Similarity measurement is obtained by performing matching between the features of query image and the database images. The features are combined to match the query image with the image database. A set of statistical parameter describing the retinal blood vessel image is matched with the stored database using Euclidean Distance (D) which is given by,

$$D_{\text{eucld}(r,s)} = \sqrt{\sum_{i=1}^{N} (r_i - s_i)^2} \tag{6}$$

where r_i and s_i represent the feature vectors of database image and query image respectively and N is the number of elements of the descriptors [11]. The shortest distance is considered as best matching image in the matching process. The CBIR system ranks the matched images based on the distance and then returns the best top four matched images.

The performance measure, recall is the fraction of the images that are relevant to the query successfully retrieved to the total number of relevant images in database [15] and is defined as:

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in database}} \tag{7}$$

3 Results and Discussion

Retinal images of normal (30) and abnormal (30) are considered for this analysis. The results of image processing performed on typical normal and abnormal retinal images are shown in Fig. 2(a) and Fig. 3(a).

(a) (b) (c)

Fig. 2. (a) Normal query image, (b) morphological preprocessed image, (c) Canny edge segmented image

(a) (b) (c)

Fig. 3. (a) Abnormal query image, (b) morphological pre-processed image, (c) Canny edge segmented image

The enhanced image using morphological transformation is shown in Fig.2 (b). The enhanced image subjected to Canny edge detection is shown in Fig.2 (c). From Fig. 2(b), it is observed that even thin blood vessels are enhanced by morphological preprocessing and are distinctly visible. Morphological operator is able to remove the uneven background illumination of the retinal vessels and linearly distributes the contrast value. It is visually observed that Canny segmentation method performs better in segmentation of blood vessels as shown in Fig.2(c). Similar observations for the abnormal images are shown in Fig.3 (b) and (c).

The average values of ratio of vessel to non vessel area for normal and abnormal images are obtained for various methods and the values are given in Table.1. The average value of ratio of vessel to non vessel area of abnormal images is distinctly high when compared to normal images. It is also observed that the average values of vessel to non vessel area obtained using segmentation by Canny with contrast stretch preprocessing and without preprocessing are almost similar. The images subjected to segmentation by Canny with morphological preprocessing showed clear difference from other two methods.

Table 1. Average ratio of vessel to non vessel area for different normal and abnormal retinal images

Images	Ratio of vessel to non vessel area		
	Canny segmentation without preprocessing	Canny segmentation with contrast stretch preprocessing	Canny segmentation with morphological preprocessing
Normal (N=20)	0.086	0.086	0.105
Abnormal (N=20)	0.098	0.099	0.119

The difference between average ratio of vessel to non vessel area for different normal and abnormal images using morphological operation based segmentation is high (0.014). This difference is considered as an important measure to differentiate normal and abnormal retinal images.

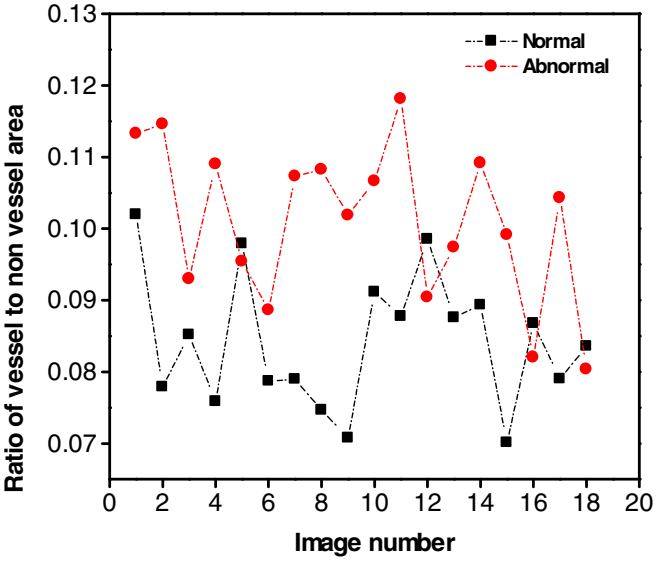
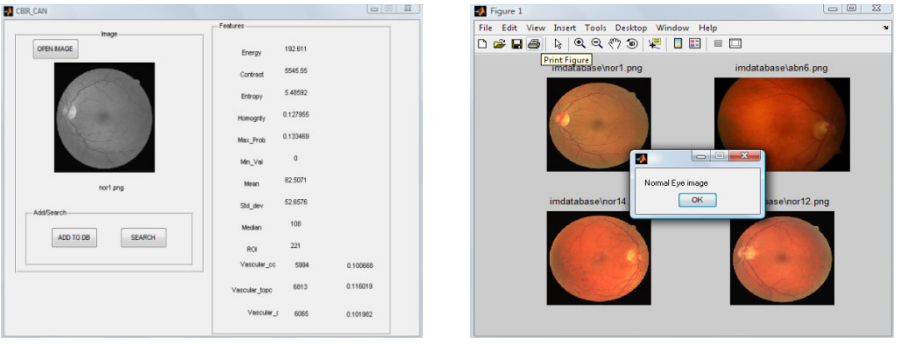


Fig. 4. Variation of vessel to non vessel area for different normal and abnormal images by Canny segmentation with morphological preprocessing

Fig. 4 shows the variation of the ratio of vessel to non vessel area for different normal and abnormal retinal images using Canny with morphological preprocessing. It is observed that the parameter, ratio of vessel to non vessel area is able to differentiate normal images from abnormal images using Canny with morphological preprocessing. These values are distinct with very less number of overlaps among the



(a) (b)

Fig. 5. (a) GUI of input query image with features, (b) GUI of output ranked images

normal and abnormal images. The ratio of vessel to non vessel area is used as a significant feature for further processing. Other features that are derived from morphological based Canny method include energy, contrast, entropy, homogeneity, maximum probability, minimum value, mean, standard deviation, median and range of intensity.

Fig.5 (a) and (b) shows the Graphical User Interface (GUI) of input with feature vectors and output retrieved images. The CBIR system ranks the images based on the minimum difference between the query and the retinal database images. For a given query image, the above mentioned features for Canny with morphological preprocessing are displayed as shown in Fig.5 (a) and best matches retrieved from the database are shown in Fig.5 (b).

Table 2. Variation of recall

Images	Recall for different methods (%)		
	Canny segmentation without preprocessing	Canny segmentation with contrast stretching preprocessing	Canny segmentation with morphological preprocessing
Normal	79.3	82.8	82.8
Abnormal	81.2	78.1	81.2
Average recall	80.3	80.5	82.0

The recall of this CBIR system is shown in Table 2. It is observed that the recall of this system using Canny with morphological preprocessing is high compared to Canny segmentation without preprocessing for normal images. The recall of Canny with morphological preprocessing is high compared to Canny segmentation with

contrast stretch preprocessing for abnormal images. It is also found that the average recall of the Canny segmentation using morphological preprocessing is very high for both normal and abnormal retinal images.

4 Conclusions

Automated analysis and interpretation of retinal images has become a necessary and an important diagnostic procedure in ophthalmology. In this proposed approach, the considered retinal images subjected to Canny segmentation method with and without preprocessing are analyzed. Canny segmentation with morphological preprocessing is found to provide better results than Canny without preprocessing and with contrast stretch preprocessing method.

Among the statistical and structural features extracted from the segmented images the feature, ratio of vessel to non vessel area differentiates the normal and abnormal images distinctly. The recall of this CBIR system is found to be high using Canny with morphological preprocessing compared to conventional Canny and contrast stretching based Canny method.

Further, the qualitative and quantitative analysis on normal and abnormal retinal images indicates that the proposed approach is effective and give better accuracy with less misclassification than the manual assessment.

References

1. Kavitha, G., Ramakrishnan, S.: Detection of Blood Vessels in Human Retinal Images using Ant Colony Optimisation Method. International Journal of Biomedical Engineering 5, 360–370 (2011)
2. Yong, Y., Shuying, H., Nini, R.: An Automatic Hybrid Method for Retinal Blood Vessel Extraction. International Journal of Applied Mathematics and Computer Science 18, 399–407 (2008)
3. Deepak, K.S., Gopal, D.J., Jayanthi, S.: Content-Based Retrieval of Retinal Images for Maculopathy. In: Proceedings of the 1st ACM International Health Informatics Symposium, New York, pp. 135–143 (2010)
4. Changhua, W., Gady, A.: Probabilistic Retinal Vessel Segmentation. In: SPIE Medical Imaging, vol. 6512, p. 651213 (2007)
5. Niall, P., Tariq, M.A., Thomas, M., Deary, I.J., Baljean, D., Robert, H.E., Kanagasingam, Y., Constable, I.J.: Retinal Image Analysis: Concepts, Applications and Potential. Progress in Retinal and Eye Research 25, 99–127 (2006)
6. Ramamurthy, B., Chandran, K.R.: Content Based Image Retrieval for Medical Images using Canny Edge Detection Algorithm. International Journal of Computer Applications 17, 32–37 (2011)
7. Tobin, K.W., Abdelrahman, M., Chaum, E., Govindasamy, V.P., Karnowski, T.P.: A Probabilistic Framework for Content Based Diagnosis of Retinal Disease. In: Annual International Conference of the IEEE Engineering EMBS, Lyon, France, pp. 6743–6746 (2007)

8. Acton, S.T., Soliz, P., Russell, S., Pattichis, M.S.: Content Based Image Retrieval: The Foundation for Future Case-based and Evidence-based Ophthalmology. In: IEEE International Conference on Multimedia and Expo ICME, Hannover, pp. 541–544 (2008)
9. Mathieu, L., Guy, C., Gwenole, Q., Lynda, B., Christian, R., Beatrice, C.: Content Based Image Retrieval Based on Wavelet Transform Coefficients Distribution. In: Conf. Proc. IEEE Engineering in Medicine and Biology Society, Lyon, France, vol. 1, pp. 4532–4535 (2007)
10. Abhir, B., Sarabjot, A., Ponnusamy, S.: Retinal Fundus Image Contrast Normalization using Mixture of Gaussians. In: IEEE 42nd Asilomar Conference on Signals, Systems and Computers, pp. 647–650 (2008)
11. Seyed, M.Z., Morteza, D., Hamid, R.P.: Retinal Vessel Segmentation Using Color Image Morphology and Local Binary Patterns. In: IEEE 6th Iranian Conference on Machine Vision and Image Processing (2010)
12. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 679–698 (1986)
13. Ryszard, S.C.: Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems. International Journal of Biology and Biomedical Engineering 1, 6–16 (2007)
14. Manikandan, S., Rajamani, V.: A Mathematical Approach for Feature Selection and Image Retrieval of Ultra Sound Kidney Image Databases. European Journal of Scientific Research 24, 163–171 (2008)
15. Henning, M., Nicolas, M., David, B., Antoine, G.: A Review of Content Based Image Retrieval Systems in Medical Applications-Clinical Benefits and Future Directions. International Journal of Medical Informatics 73, 1–23 (2004)

Potential Topics Discovery from Topic Frequency Transition with Semi-supervised Learning

Yoshiaki Yasumura, Hiroyoshi Takahashi, and Kuniaki Uehara

Shibaura Institute of Technology, Japan
Kobe University, Japan
yasumura@shibaura-it.ac.jp,
{takahashi,uehara}@ai.cs.kobe-u.ac.jp

Abstract. This paper presents a method for potential topic discovery from blogsphere. A potential topic is defined as an unpopular phrase that has potential to spread through many blogs. To discover potential topics, this method learns from topic frequency transitions in blog articles. Though this learning requires sufficient amount of labeled data, labeled data is costly and time consuming. Therefore this method employs a semi-supervised learning to reduce labeling cost. First, this method extracts candidates of potential topics from categorized blog articles. To detect potential topics from the candidates, a classifier is built from topic frequency transition data. Experimental results with real world data show the effectiveness of the proposed method.

Keywords: Web mining, potential topic, semi-supervised learning, topic frequency transition.

1 Introduction

Blog is a media that individual can easily provide commentary and news on a particular subject. Since the number of bloggers increases rapidly, a lot of blog articles are updated daily. Thus, blog articles reflect the trend of the real world. This fact enables us to analyze market trend by monitoring information on blogs[1,2,3].

So far, a lot of methods are proposed for monitoring and analyzing information on blogs. One of the most popular methods is detecting a burst of a word in a document stream of blogs [4,5,6,7,8,9]. Since burst words are viewed as hot topics in the blogsphere, detecting burst provides market analyzer the trend in the blogsphere. However, most burst topics are not valuable information from the view point of marketing because they are already popular in the blogsphere. Valuable topics are described in a few blogs and have a potential to spread through many blogs. We call such topics "potential topics". The system that can discover potential topics as early as possible is required for marketing.

For discovering potential topics, Kosaka et al. proposed a method that extracts potential topics from blogsphere by learning from topic frequency transition[10]. However, this method adopts supervised learning that requires sufficient amount

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 477–486, 2012.

of labeled data. Labeled data is costly and time consuming. In addition, it is difficult to label topics because we cannot know when the learner can recognize the topic is a potential topic.

In this paper, we develop a system for discovering potential topics from the blogsphere by semi-supervised learning. Semi-supervised learning requires a few labeled data and a lot of unlabeled data. This system first extracts candidates of potential topics by filtering general phrases. Next a predictor for detecting potential topics is built by semi-supervised learning from the data of topic frequency transition in the blogsphere.

2 Potential Topics Discovery

In this section, we present a method for discovering potential topics. First, we describe potential topics and their usefulness. Second, we present a method for categorizing blog articles and building a predictor for detecting potential topics.

2.1 Potential Topics

Valuable information for marketing can create or capture new demand. The system that can detect such information as early as possible is required for marketing. One of the systems is topic extraction by detecting a burst of a word. A burst of a word is defined as sharp increase in frequency of the word. However, most burst topics are not valuable information because they are already popular. Fig. 1 shows an example of burst detection. This graph charts the topic frequency transition of the phrase "subprime loan problem" in the blogsphere. From the graph, this phrase is first described in blogs in March, 2007. After that, the phrase increased in frequency gradually, and burst in October, 2007. If burst of the phrase is detected at that time, it is not valuable information. This is because subprime loan problem was already reported on TV and newspaper, and stock price began to fall at that time. However, the phrase "subprime loan problem" is valuable information if it was detected before bursting. In order to detect the topics earlier, we try to extract potentiality of the phrase by analyzing the topic frequency transition before bursting. To do this, we built a predictor for detecting potential topics. The predictor learns from the topic frequency transition of the old data of the bursted topics. For creating the predictor, this method learns from topic frequency transitions in blog articles. Though this learning requires sufficient amount of labeled data, labeled data is costly and time consuming. Besides it is difficult to label topics because we cannot know when the learner can recognize the topic is a potential topic. So this method employs a semi-supervised learning to reduce labeling cost.

In order to discover potential topics as early as possible, we extract specialists who can describe potential topics of the category earlier in their blogs. To extract them, we classify blog articles into some category. By analyzing the topic frequency transition in each category, we can detect potential topics earlier. Fig. 2 shows an example of topic frequency transition of the phrase "subprime loan

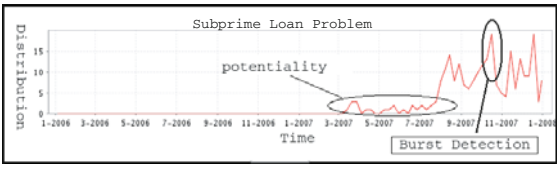


Fig. 1. An example of burst detection

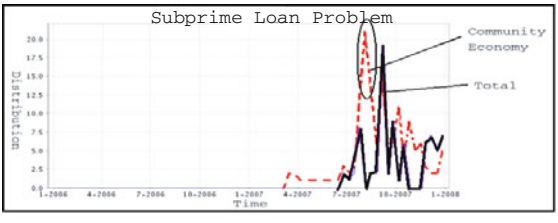


Fig. 2. An example of topic frequency transition in blog community

problem" in the economy category and the total of blogs. From Fig. 2, the bloggers in the economy category first described the phrase in the blogs and the burst in the economy category is detected before bursting in the total of the blogs. This example shows that blog categorization is effective for discovering potential topics earlier. This is because the bloggers in a category are the specialists of the category. Thus, we attempt to discover potential topics by using blog categorization.

From the above idea, we develop a system for discovering potential topics. We present an overview of our system in Fig. 3. The system discovers potential topics by the following procedure.

1. The system classifies blog articles into some categories to extract specialists in the category.
2. The system extracts topics as candidates of potential topics from the blogs in each category.
3. The system filters the extracted topics by using DF value (Document Frequency) because some extracted topics are too general such as "Christmas".
4. The system determines whether the candidate topics are potential topics or not. For this determination, a predictor is built by semi-supervised learning from the topic frequency transition.

3 Potential Topic Discovery by Semi-supervised Learning

In this section, we present a method for discovering potential topics by semi-supervised learning.

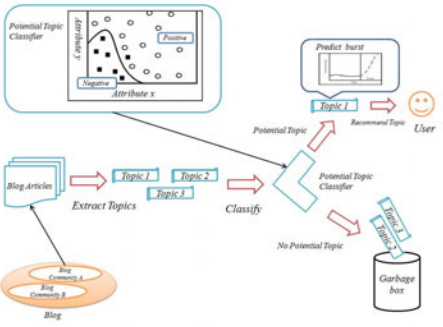


Fig. 3. Overview of the proposed method

3.1 Blog Categorization

Here, we describe blog categories and a method for categorization.

Bloggers describes commentary and news in their blogs. The contents of the blogs depend on the blogger's preferences. For example, the blogger who likes football often describes articles on football, and the articles usually contain details on football. Thus, the bloggers that have common preference is viewed as specialists of the category. For this reason, we classify blogs into some categories.

The classifier for this categorization is built by machine learning. The training data for this learning is manually created as the articles labeled their category. Table 1 shows the categories used in this system. In the system, a blog article can be labeled multi-category. The system uses Naive Bayes as a classifier for blogs.

In this system, we classify bloggers into categories according to their blogs. If the rate of the blogs classified into a particular category is over the threshold, the blogger is classified into the category.

3.2 Filtering General Term

In this system, candidates of potential topics are nouns and simple noun sequences. So the system collects all nouns and simple noun sequences as candidates of potential topics. However, the candidates are too many for discovering potential topics effectively. To reduce the candidates, the system eliminates general terms from the candidates. For example, Fig. 4 shows an example of the general phrase "fireworks event". Since fireworks events are usually held in summer, the phrase bursts every summer. However, detecting the burst of the phrase is not valuable information, because everyone knows the fact.

For eliminating general term, the system uses DF (Document Frequency) value. If a term has high DF value, it appears in many documents. So the system eliminates terms that have high DF value. By eliminating general terms, we can obtain unknown terms (e.g. subprime loan problem) as candidates of potential topics.

Table 1. Blog category

Parent Category	Subcategory
Politics Economy	politics, economy, news@
Sport	baseball, football, martial arts, golf
Music	pops, jazz, classical music
Food	restaurant, recipe, food
Entertainment	movie, TV program, entertainer, play
Art	literature, picture, fashion
Vehicle	car, train, motorcycle, bicycle, airplane
Gamble	horse racing, pachinko, mahjong
@ Life	school, family, love, business interior, travel, health
Pet	dog, cat, animals, gardening
Technology	computer, internet, science
Hobby	toy, military, cartoon, game

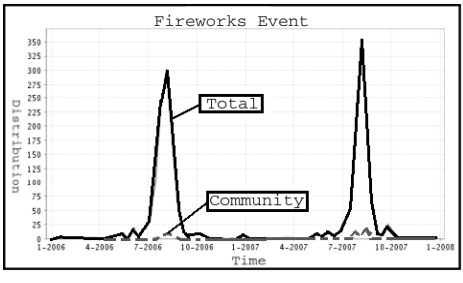


Fig. 4. Frequency transition of "fireworks event"

3.3 Potential Topics Predictor

From the candidates of potential topics, the system extracts the topics that have potential to burst. The candidates of potential topics extracted by above procedures consist of potential topics and low-frequency terms. A low-frequency term is a word that described in few blogs or in a small community of blogs such as "Federal Open Market Committee". Since low-frequency terms are not valuable for marketing, the system classifies the candidate terms into potential topics or low-frequency terms.

For the classification, we build a classifier for detecting potential topics from the topic frequency transition. The classifier is built by machine learning technique from manually labeled training data. For building the classifier, we choose attributes of training data. We can guess that a potential topic has the following features.

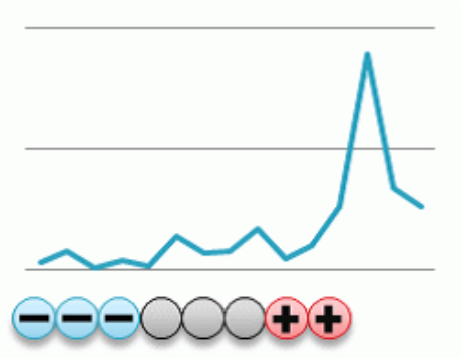


Fig. 5. Semi-supervised learning for potential topic detection

- The potential topic is described not only in a particular category but also in the other categories.
- The frequency of the potential topic increases gradually.
- The potential topic continuously appears in the blogs.
- It is not long from the first time the potential topics appears in the blogs.

Considering the above features of potential topics, we choose attributes for classification.

- The frequency of the topic in that day, the three days before, the seven days before, and thirty days before.
- The number of continuously appearance of the topic.
- The number of the day from the first appearance of the topic.
- The total number of the bloggers who described the topic.
- The number of the increased bloggers from the three days before, the seven days before, and the thirty days before.

Instances in the training data are labeled by detecting burst. If burst is detected, the instance is positive. If burst is not detected, the instance is negative. From the training data, the system builds classifier for detecting potential topics.

However, manually labeling training data is costly and difficult. Fig. 5 shows an example which is difficult to label. This graph shows topic frequency transition. We can label a few day before bursting as positive (potential topic). On the other hand, we can label days that have very low frequency as negative (not potential topic). However, it is hard to label the days between them. Hence we adopts semi-supervised learning for building a predictor of potential topics. In this learning, difficult instances are dealt as unlabeled instances. Using semi-supervised learning, labeling cost is reduced.

For semi-supervised learning, we adopt Tri-Training [11] that builds three classifiers and classifies by majority vote. This learning method first creates three datasets by bootstrap sampling of labeled data and builds classifiers from them. The built classifiers classify unlabeled instances one another. Then the classifiers

are built from the data that contains unlabeled instances labeled by the other classifiers. All unlabeled instances are labeled by repeating this procedure.

4 Experiment

We conducted evaluation experiments for assessing the effectiveness of our system. In this experiments, we evaluate our system by using actual blog data.

4.1 Experimental Setting

For this experiment, we collect blog data from January 2008 to December 2010. The blog data contains about 200,000 articles written by 2496 bloggers.

Training data is the manually labeled topics. The number of the positive topics is 40. The number of the negative topics is about 3500. The rest of the instances is unlabeled. In this experiment, we use the C4.5 decision tree as a base learner for detecting potential topics. We evaluate our method by 10-fold cross-validation. Since the data is divided into subsets based on topics for cross-validation, the instances of the same topic are included in the same subset.

4.2 Experimental Results

We evaluate our predictor with precision, recall and F-measure. Fig. 6 shows the result according to the size of positive instances. In the figure, 1 means that one day before bursting is labeled as positive, and 7 means that seven days before bursting is labeled as positive.

From the figure, recall is almost 0.9 in the case that the size of positive instances is over 10. This result show that the system can discover almost potential topics if sufficient size of positive instances are available. However, precision is almost 0.2 in the most cases. From this result, the system extracted some negative instances as potential topics. This mistake means two possibilities. One possibility is that the system simply mistakes in labeling. The other possibility is that the system extracts potential topics that never burst. The later case is rare, but has valuable information. For practical use, recall is more important than precision, because the system shows the users the potential topics and the users decide to use the potential topics at last.

Fig. 7, Fig. 8 and Fig. 9 show the results of the topic "monster hunter", which is famous game, in the case that the positive label size is one, ten and thirty, respectively. In Fig. 7, the prediction is scattered because the learning is not well. This caused by insufficiency of positive instances. In Fig 8, the system predicted well, and the positive instances are just before bursting. In Fig 9, the system predicted more instances as positive. From these results, the size of positive instances is important for predicting.

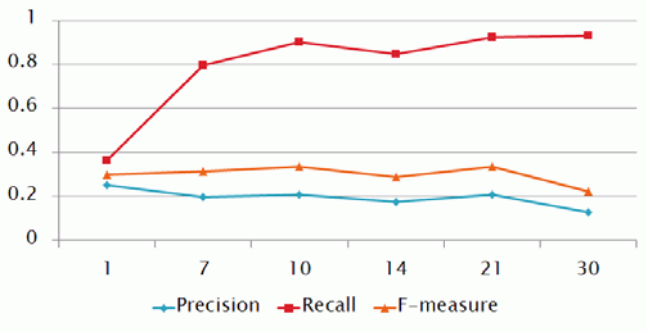


Fig. 6. Precision and recall of potential topics

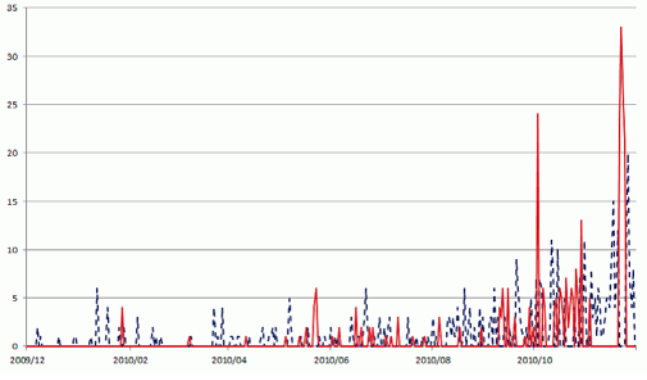


Fig. 7. The result example: positive size is 1

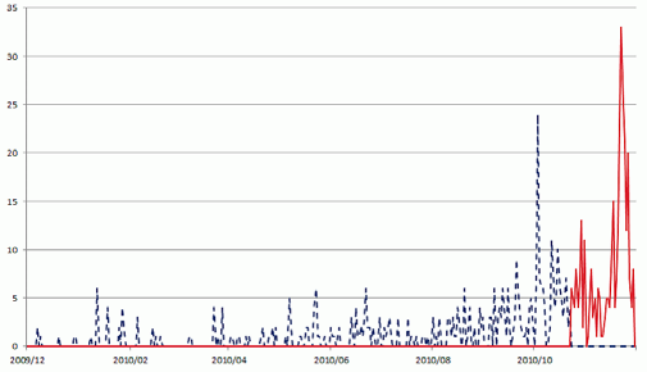


Fig. 8. The result example: positive size is 10

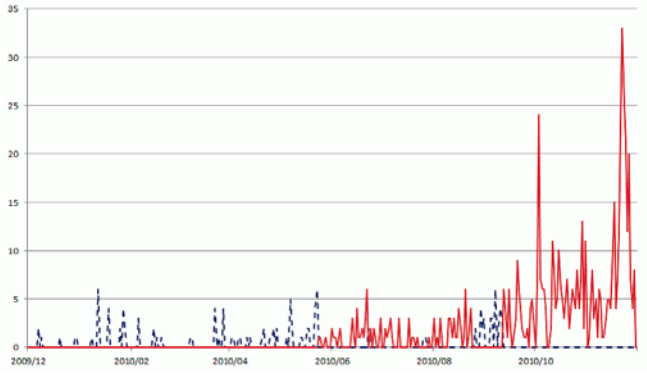


Fig. 9. The result example: positive size is 30

5 Conclusion

In this paper, we proposed a method for discovering potential topics earlier by semi-supervised learning. From the topic frequency transition in categories, the system predicts the burst of the potential topics by using small size of labeled data.

Experimental results using actual blog data show that our method is higher recall for predicting potential topics. This result indicates that our system is effective for potential topic discovery.

For future work, we try to raise the precision of potential topic discovery.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: Proc. of the International Conference on Web Search and Web Data Mining, pp. 207–218 (2008)
2. Cheng, Y., Qiu, G., Bu, J., Liu, K., Han, Y., Wang, C., Chen, C.: Model bloggers' interests based on forgetting mechanism. In: Proc. of the International Conference on World Wide Web, pp. 1129–1130 (2008)
3. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: Proc. of the International Conference on Web Search and Web Data Minig, pp. 231–240 (2008)
4. Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H.: Parameter free bursty events detection in text streams. In: Proc. of the International Conference on Very Large Data Bases, pp. 181–192 (2005)
5. Fujiki, T., Nanno, T., Suzuki, Y., Okumura, M.: Identification of bursts in a document stream. In: Proc. of the First International Workshop on Knowledge Discovery in Data Streams, pp. 54–64 (2004)
6. Bansal, N., Koudas, N.: BlogScope: a system for online analysis of high volume text streams. In: Proc. of the 33rd International Conference on Very Large Data Bases, pp. 1410–1413 (2007)

7. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proc. of the International Conference on Knowledge Discovery and Data Mining (2002)
8. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. In: Proc. of International World Wide Web Conference (2003)
9. Mane, K., Borner, K.: Mapping topics and topic bursts in PNAS. In: Proc. of National Academy of Sciences (2004)
10. Kosaka, Y., Yasumura, Y., Uehara, K.: Discovery of Potential Topics from Blogsphre Based on Blog Categorization. In: Proc. of the 4th International Conference on Knowledge, Information and Creativity Support System, pp. 55–60 (2009)
11. Zhou, Z.-H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Trans. Knowledge Data Engineering 17, 1529–1541 (2005)

Heuristic Energy-Efficient Routing Solutions to Extend the Lifetime of Wireless Ad-Hoc Sensor Networks

Nguyen Thanh Tung

Department of IT, NTT University
298A-300A Nguyen Tat Thanh Street, Ward 13, District 4, HCM City, Vietnam
nttung@ntt.edu.vn

Abstract. Sensor networks are deployed in numerous military and civil applications, such as remote target detection, weather monitoring, weather forecast, natural resource exploration and disaster management. Despite having many potential applications, wireless sensor networks still face a number of challenges due to their particular characteristics that other wireless networks, like cellular networks or mobile ad hoc networks do not have. The most difficult challenge of the design of wireless sensor networks is the limited energy resource of the battery of the sensors. This limited resource restricts the operational time that wireless sensor networks can function in their applications. Routing protocols play a major part in the energy efficiency of wireless sensor networks because data communication dissipates most of the energy resource of the networks. This paper studies the importance of considering neighboring nodes in the energy efficiency routing problem. After showing that the routing problem that considers the remaining energy of all sensor nodes is NP-complete, heuristics are proposed for the problem. Simulation results show that the routing algorithm that considers the remaining energy of all sensor nodes improves the system lifetime significantly compared to that of minimum transmission energy algorithms.

Keywords: Battery, Sensor, Routing Protocols, NP_Complete.

1 Introduction of Multihop Routing

There have been numerous studies on the energy efficiency of multi-hop routing in literature. These studies use the Dijkstra algorithm [3] or variants of this algorithm to calculate the shortest routes to a destination with different types of energy metrics. Unfortunately, only few studies mentioned about the remaining battery of sensor nodes, and it is difficult to apply the Dijkstra algorithm or its variants to the lifetime problem of wireless ad-hoc sensor networks (WASNs). In other words, it is preferred to use the variants of the Dijkstra algorithm as routing methods because the algorithm is Linear Programming (LP) problem, but the algorithm cannot provide an optimal routing solution for the lifetime problem of WASNs.

Wireless transmission is different to wire-line networks in that transmission from a source node to a destination node causes neighboring nodes to dissipate energy when they detect the transmission. Unfortunately, the energy dissipation of neighboring

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 487–497, 2012.

nodes may be comparable to the energy dissipation of the nodes in the path and can degrade the performance of the routing methods. It is shown that the reception energy of 802.11 products is at least 50 % that of the transmission energy [1, 2]. For example, Stemm and Katz measure the idle: receive: send energy consumption ratios of 1:1.05:1.4 [4]. This data measurement emphasizes the importance of considering reception energies by neighboring nodes in energy-efficient routing models. For example, Fig. 1 shows that the transmission from the source to the destination will be listened by other seven neighboring nodes.

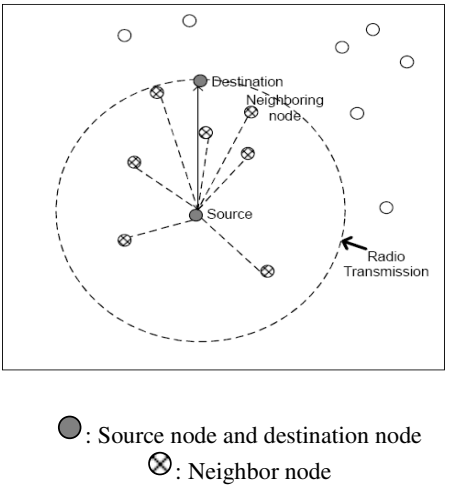


Fig. 1. Transmission from a source to a destination drains the energy of the source, the destination and neighbouring nodes

There are many studies in the literature to work out the best transmission power because the reduction of the transmission range will lead to the reduction of the energy consumption of neighboring nodes. The authors in [6] proposed a routing method considering the reception energy of neighboring nodes to control the transmission power. In [7], the authors also considered the reception energy usage in the selection of energy efficient paths. In [8], an analytical model for optimal transmission range for minimizing the total energy consumption was presented. Unfortunately, all of these papers are designed for mobile ad hoc networks but not sensor networks. Unlike sensor networks, battery constraint is not a major issue of mobile ad hoc networks. Therefore, only few papers in the literature consider the limited energy storage of nodes, which is the major challenge when designing sensor networks. For examples, [10, 11] mentioned about the control of sensor range to maximize the operation time of sensor networks under battery limits. This paper concentrates on multi-hop routing methods that prolong the operation time of practical sensor networks under the battery constraint of the sensors.

2 Formulating Routing Problem

Original routing problem:
Given a network of n sensors, in which any sensor node can connect to all other sensor nodes by adjusting its transmission power. Each sensor node i has the energy storage of e(i). A random source node s wants to transmit data to a destination node d. Obviously, there are many possible paths from s to d. Each path results in an energy reduction of all nodes on the path (including the nodes are within the transmission range of the data transmission). The routing problem is to find a path from s to d so that after the data transmission, the minimum remaining energy storage of all sensor nodes is maximized:

$$\text{Maximize: } \min(e(i)),\ \forall i \in n \tag{1}$$

where $e(i)$ is the remaining energy of Node i after the path is established.

Unfortunately, Problem (1) is NP-complete, therefore there is no polynomial time algorithm to find the energy efficient path. We will prove the NP-completeness of (1) using graph models and a well known NP-complete problems. Firstly, we give some preliminary results.

Graph problem:
A sensor network is modeled as $G(V, E)$, *where* V *is the set of nodes and* E *is the set of links between the nodes. Node* i *sets its power to zero or its power to* p_i^j *if Node* i *wants to transmit to Node* j, $\forall i, j \in V$. *Every node* i *has the remaining energy capacity of* $e(i)$. *Given a source node s and a destination node d, Find a path from* s *to* d *that maximize the minimum remaining energy of all nodes* $i \in V$.

$$\text{Maximize: } \min(e(i)),\ \forall i \in V \tag{2}$$

where $e(i)$ is the remaining energy of Node i after the path is established.

Problem (2) can be converted to a decision problem:

Decision problem (3):
A sensor network is modeled as $G(V, E)$. *Node* i *sets its power to zero or to* p_i^j *if Node* i *wants to transmit to Node* j, $\forall i, j \in V$. *Every node* i *has the remaining energy capacity* $e(i)$. *Given a source node s and a destination node d, find a simple path from* s *to* d *that* $e(i) \geq c$, $\forall i \in V$, *c is a constant.*

Let us consider a simple case of Problem (3), in which all nodes transmit at the same power:

Constant power problem (4):
A sensor network is modeled as $G(V, E)$. *All nodes can transmit at a constant* $p = P, \forall i \in V$. *In other words, a Node* i *transmits with power P or does not*

transmit. Every Node i *has the remaining energy capacity* $e(i)$. *Given a source node s and a destination node d, find a simple path from* s *to* d *that* $e(i) \geq c$ *for all nodes* $i \in V$, *c is a constant.*

The above problem can be polynomially reduced to the Path with Forbidden Pairs problem. This is a well known NP-complete problem. Details are given in [5, 9]. As a result, the simple constant power problem (4) is NP-complete and therefore, the original problem (1) is also NP-complete. From the above results, there are no polynomial algorithms to find a path to maximize the minimum residual energy of sensor nodes and hence we need to propose heuristic algorithms for the problem.

3 Heuristic Algorithms

Three heuristic energy-efficient routing methods are implemented to extend the lifetime of WASNs. A round of data transmission is defined as the duration of time a random source node transmits a unit of data to a random destination node. The lifetime of WASNs is defined as the total number of rounds sending data between sensor nodes until the first node run out of energy. The heuristic routing methods are summarized as below. The shortest path for these methods is calculated using the Dijkstra algorithm [3].

Shortest path of the energy dissipation (SP)

Given a source node s *and destination node* d, *find a simple path* Π *from* s *to* d *that minimizes the total energy dissipation by all nodes on the path.*

$$\sum_{i \in \Pi - d} (p_i + r) \text{ is minimized}$$

where r *is the reception energy consumption of any sensor node. (End of algorithm)*

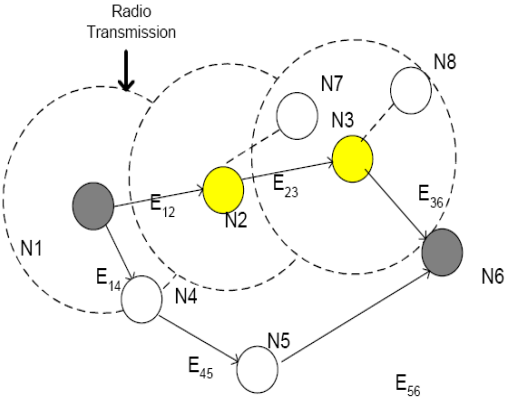


Fig. 2. The SP method

E_{ij} = Energy consumption to send data from Node i to Node j

E_j = Energy consumption to receive data from at Node j . $E_j = 1$ unit, $\forall j \in V$

Table 1. Energy metrics of Fig. 2

Energy usage	Unit
E_{12}	1
E_{23}	1
E_{36}	1
E_{14}	1
E_{45}	2
E_{56}	1

The total energy consumption for path (1, 2, 3, 6) (not including neighboring nodes) is:

$$E_{(1,2,3,6)} = E_{12} + E_{23} + E_{36} + E_2 + E_3 + E_6 = 1+1+1+1+1+1=6;$$

The total energy consumption for path (1, 4, 5, 6) is:

$$E_{(1,4,5,6)} = E_{14} + E_{45} + E_{56} + E_4 + E_5 + E_6 = 1+2+1+1+1+1=7;$$

Therefore, the SP algorithm will select the path (1, 2, 3, 6). Node 7 and Node 8 are not involved in the path selection process.

Shortest path of the energy dissipation including neighboring nodes (SP_N):
Given a source node s and a destination node d, Find a simple path Π from s to d that minimizes the total energy dissipation by all nodes participating in the data transmission.

$$\sum_{i \in \Pi - d} \left(p_i + \sum_{j \in N(i)} r_j \right) \text{ is minimized}$$

where N(i) is the set of neighboring nodes in the transmission range of Node i. (End of algorithm)

Fig. 3 shows that unlike SP algorithm, the SP_N algorithm considers nodes on a selected path and all neighboring nodes involved in the transmission. The total energy consumption for path (1, 2, 3, 6) is:

$$E_{(1,2,3,6)} = E_{12} + E_{23} + E_{36} + E_2 + E_3 + E_6 + E_4 + E_7 + E_8 + E_7 = 1+1+1+1+1+1 +1+1+1+1=10;$$

The total energy consumption for path (1, 4, 5, 6) is:

$$E_{(1,4,5,6)} = E_{14} + E_{45} + E_{56} + E_4 + E_5 + E_6 = 1+2+1+1+1+1=7;$$

Therefore, the SP_N algorithm will select the path (1, 4, 5, 6).

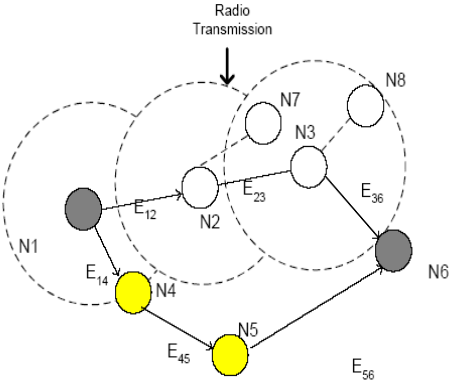


Fig. 3. Path calculation and selection from SP_N algorithm

E_{ij} = Energy consumption to send data from Node i to Node j

E_j = Energy consumption to receive data from at Node j . $E_j = 1$ unit, $\forall j \in V$

Table 2. Energy metrics of Fig. 3

Energy usage	SP
E_{12}	1
E_{23}	1
E_{36}	1
E_{14}	1
E_{45}	2
E_{56}	1

Shortest path of the remaining energy (SP_RE):
Let us define a weight for a link on any path as:

$$W(i) = \sum_{j \in N(i)} \frac{1}{e(j)}$$

where N(i) is the set of neighboring nodes in the transmission range of Node i. Given a source node s *and a destination node* d*, find a simple path* Π *from* s *to* d *that minimizes the total weight by all links participating in the data transmission.*

$$\sum_{i \in \Pi - d} W(i) \text{ is minimized}$$

(End of algorithm)

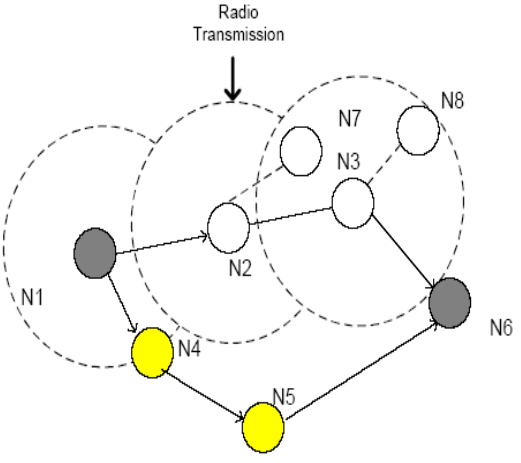


Fig. 4. Path calculation and selection from SP_RE algorithm

E_j = Remaining energy at Node j

Table 3. Energy metrics of Fig. 4

Remaining energy	SP
E_1	1
E_2	2
E_3	1
E_4	1
E_5	1
E_6	1
E_7	1
E_8	1

The total energy consumption for path (1, 2, 3, 6) is:

$$E_{(1,2,3,6)} = \frac{1}{E_2} + \frac{1}{E_3} + \frac{1}{E_6} + \frac{1}{E_4} + \frac{1}{E_7} + \frac{1}{E_7} + \frac{1}{E_8} =$$

$$0.5+1+1+1+1+1+1=6.5;$$

The total energy consumption for path (1, 4, 5, 6) is:

$$E_{(1,4,5,6)} = \frac{1}{E_4} + \frac{1}{E_5} + \frac{1}{E_6} = 1+1+1=3;$$

Therefore, the SP_RE algorithm will select the path (1, 4, 5, 6).

4 Simulation and Comparison

A number of simulators in C++ is developed to simulate the performance of SP, SP_N and SP_RE. The energy dissipation model used is given below. The total transmission energy of a message is calculated by:

$$E_t = kE_{elec} + \varepsilon_{amp} kd^n$$

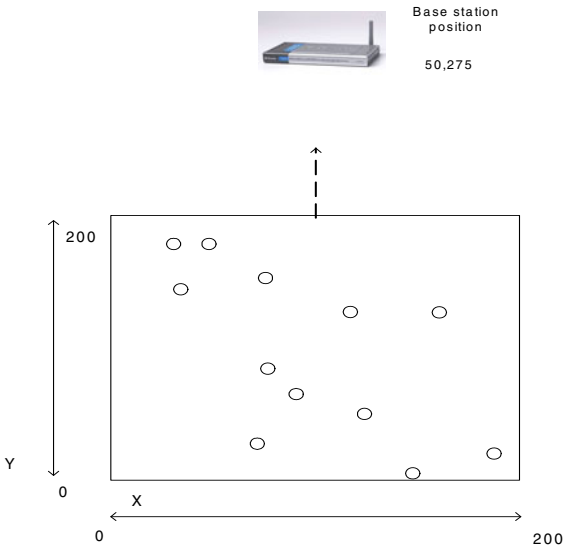
The reception energy is calculated by:

$$E_r = kE_{elec}$$

where E_{elec} is the energy dissipation of the electronic circuitry to encode or decode a bit, k is message size, ε_{amp} is the amplifier constant and d is the distance between the transmitter and the receiver.

The network settings for the simulations in the section are given below. This model is the same with the model in [9, 12].

Network size $(200m \times 200m)$
Base station $(50m, 275m)$
Number of sensor nodes: 100 nodes
Energy message: 20 bits
Position of sensor nodes: Uniform placed in the area
Energy model: $E_{elec} = 50*10^{-9}$ *J,* $\varepsilon_{fs} = 10*10^{-12}$ *J/bit/m*^{2} *and*
$\varepsilon_{mp} = 0.0013*10^{-12}$ *J/bit/m*^{4}
Broadcast ID message: 16 bits
Broadcast energy message: 32 bits



In the first set of simulations, the lifetime performance of the above routing methods is studied for the above 100 random 100-node sensor networks. Each node begins with 50 mJ of energy. The operation of each sensor network is divided into rounds. In each round, a random source node transmits a unit of data to a random destination node. The process is repeated until the first sensor node is out of energy and the lifetime for each routing method in each network topology is recorded. On average, SP, SP_N and SP_RE perform 268, 363, and 519 rounds respectively. These lifetimes are the time until the first sensor fails. The results are shown in Fig. 5.

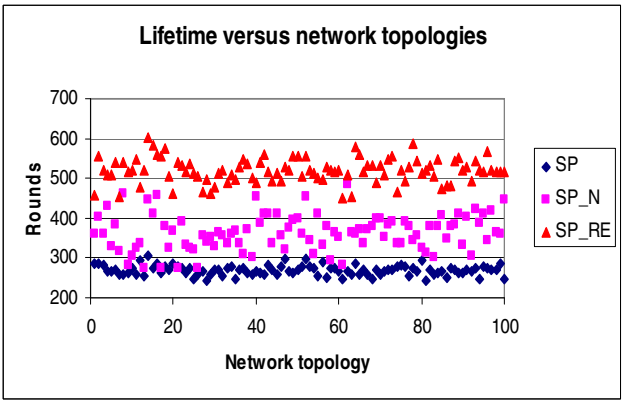


Fig. 5. Number of rounds over 100 random 100-node networks

Table 4. Results for Fig. 5

Number of rounds			
Protocol	SP	SP_N	SP_RE
Mean	268.9	363.3	519.3
Variance	12.5	45.5	30.7
90% confidence interval for the sample means	(267, 271)	(356, 371)	(514, 524)

5 Conclusion

It is shown that the problem of locating a simple path that maximizes the minimum remaining energy of all sensor nodes is NP-complete. Therefore, there is no polynomial time algorithm for the problem and heuristic solutions are required to achieve reasonable energy efficiency.

Three heuristic routing methods were implemented: (1) Dijsktra algorithm to minimize the total energy dissipation of nodes on a selected path (SP), (2) Dijsktra algorithm to minimize the total energy dissipation of nodes on a selected path including neighboring nodes (SP_N), and (3) the Dijskstra algorithm considering the remaining energy of all sensor nodes on a selected path (SP_RE). Simulation results show that SP_RE can double the lifetime of SP on average, while SP_N can minimize the total energy dissipation to half of SP.

References

1. Ilyas, M., Mahgoub, I.: Mobile Computing Handbook. CRC Press (2005)
2. Feeney, L., Nilsson, M.: Investigating the energy consumption of a wireless network interface in an ad hoc networking environment. In: IEEE INFOCOM (2001)
3. Dijkstra algorithm (September 23, 2011), http://en.wikipedia.org/wiki/Dijkstra's_algorithm
4. Stemm, M., Katz, R.H.: Measuring and reducing energy consumption of network interfaces in handheld devices. IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Science, Special Issue on Mobile Computing 80(8), 1125–1131 (1997)
5. Garey, M., Johnson, D.S.: Computers and Intractability. A Guide to the Theory of NP-completeness. Freeman (1979)
6. Liu, B.H., et al.: An energy efficient select optimal neighbor protocol for wireless ad hoc networks. In: Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN 2004), pp. 626–633. IEEE Computer Society, Washington, DC, USA (2004)

7. Shrestha, N., Mans, B.: Reception-Aware Power Control in Ad Hoc Mobile Networks. In: The Third International Conference on Innovative Applications of Information Technology for Developing World (Asian Applied Computing Conference (AACC 2005), Kathmandu, Nepal, December 10-12 (2005)
8. Chen, Y., et al.: On selection of optimal transmission power for ad hoc networks. In: 36th Annual Hawaii International Conference on System Sciences (HICSS 2003) - Track 9, Washington, DC, USA (2003)
9. Tung, N.T.: Energy-Efficient Routing Algorithms in Wireless Sensor Networks: PhD thesis, Monash University, Australia (July 2009)
10. Dhawan, A., Vu, C.T.: Maximum Lifetime of Sensor Networks with Adjustable Sensing Range. In: Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD 2006 (2006)
11. Badrinath, G.S., Gupta, P.: Maximum Lifetime Tree Construction for Wireless Sensor Networks. In: Janowski, T., Mohanty, H. (eds.) ICDCIT 2007. LNCS, vol. 4882, pp. 158–165. Springer, Heidelberg (2007)
12. Heinzelman, W.B., Chandrakasan, A.P.: An Application Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transaction on Wireless Communications 1(4), 660–670 (2002)

Formal Agent-Oriented Ubiquitous Computing: A Computational Intelligence Support for Information and Services Integration

Phan Cong Vinh

Department of IT, NTT University,
298A – 300A Nguyen Tat Thanh Street, Ward 13, District 4, HCM City, Vietnam
pcvinh@ntt.edu.vn
http://www.linkedin.com/in/phanvc

Abstract. Agent-based ubiquitous computing (AUC) is a form of distributed computing by which computational processes are executed concurrently by assigning each computational process to one of agents on a ubiquitous computing system (UCS). One of AUC goals is to support the seamless integration of information and services. Meeting this grand challenge of AUC requires that agent-orientation not tackled before is necessarily featured. To this end, this paper presents a firm formal development for featuring agent-orientation of ubiquitous computing to integrate smoothly information and services.

Keywords: Agent-orientation, Context-awareness, Information and services integration, Ubiquitous computing systems.

1 Introduction

Services based on ubiquitous computing systems (UCSs), so-called ubiquitous services, are currently booming. As web services and communications infrastructure become richer, as mobile devices become more powerful, and as consumer expectations evolve, we are faced with an array of challenges that affect how to integrate information and services efficiently. *Agent-based ubiquitous computing* (AUC) is an essential computing paradigm to keep such a seamless integration of information and services [3]. On such UCSs, AUC is only possible when ubiquitous agents autonomously interact and coordinate with each other to maintain properly the required information and services integration [6,5,4,2,3]. The essence of AUC is to enable the ubiquitous agents to execute concurrently assigned computational processes and manage themselves without direct intervention of human while interacting and coordinating with each other. Hence, for UCSs, one of major challenges is that how can agents self-manage in the face of changing location and context frequently? [1]. With this aim, we develop a firm formal approach in which the notions of *agent-orientation* and *processes types* are featured and specified in categorical structures. For UCSs, whose major features consist of heterogeneity, decentralization, nondeterminism and dynamicity,

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 498–508, 2012.

AUC is a form of distributed computing by which computational processes are executed concurrently by assigning each computational process to one of agents on a UCS. The major contribution of the paper is to propose some applied categorical structures for featuring agent-orientation to support for information and services integration.

The rest of this paper is organized as follows: In Section 2, we present clearly and exactly the notions of process, type of process and agents on UCSs. Section 3 presents concentrating on category of processes types. Extensional monoidal structure and symmetry monoidal structure of the category of processes types are constructed in Section 4 in order to consider the significant properties of agents on UCSs in Section 5. In Section 6, we briefly discuss a direction of further developments in future. Finally, a short summary is given in Section 7.

2 Agents on UCSs

For UCSs, processes on UCSs are defined when they interact with each other through input and output. To this end, we put some additional structure on the processes. Let X and Y be any sets of inputs and outputs, respectively; we define $X \longrightarrow Y$ as *type* of a process on UCS. A process a of the type $X \longrightarrow Y$ is called an *agent* denoted by $a : X \longrightarrow Y$ or $X \xrightarrow{a} Y$. This is read as the process a has inputs from the set X and outputs to the set Y and pictorially drawn as $X \longrightarrow$ [a] $\longrightarrow Y$.

As suggested by the arrow notation for types, agents can be composed. In fact, if $a : X \longrightarrow Y$ and $b : Y \longrightarrow Z$, then the sequential composition $a \centerdot b : X \longrightarrow Z$ is completely defined as follows:

Let A be the set of agents and IO the set of sets of inputs and outputs, then the sequential composition "$\centerdot$" is defined by the following functions:

$$
\centerdot \stackrel{def}{=} \left(\begin{array}{l}
\_ \centerdot \_ : A \times A \longrightarrow A \\
\qquad \{\text{Sequential composition function}\} \\
dom, cod : A \longrightarrow IO \\
\qquad \{\text{Domain and codomain functions}\} \\
id : IO \longrightarrow A \qquad \{\text{Identity function}\} \\
\\
\text{such that} \\
\\
dom\ (\_ \centerdot \_) = \{f, g : A \mid cod\ f = dom\ g\} \\
\forall f, g, h : A \mid cod\ f = dom\ g \wedge cod\ g = \\
\qquad dom\ h \bullet (f \centerdot g) \centerdot h = f \centerdot (g \centerdot h) \\
\forall f, g : A \mid cod\ f = dom\ g \bullet dom\ (f \centerdot g) = \\
\qquad dom\ f \wedge cod\ (f \centerdot g) = cod\ g \\
\forall X : IO \bullet dom\ (id\ X) = cod\ (id\ X) = X \\
\forall f : A \bullet id\ (dom\ f) \centerdot f = f \centerdot id\ (cod\ f) = f
\end{array} \right)
$$

Note that the identity function $id\ X$ defines an identity agent, denoted as id_X, in A. The identity agent $id_X : X \longrightarrow X$ is pictorially drawn as follows

$$X \longrightarrow \boxed{id_X} \longrightarrow X$$

Property 1. *Sequential composition "." is associative.*

Formally, this means that if $a : X \longrightarrow Y$, $b : Y \longrightarrow Z$ and $c : Z \longrightarrow T$ are agents with their types then, by repeated composition, we can construct an agent with type $X \longrightarrow T$ in two ways:

$$(a \centerdot b) \centerdot c : X \longrightarrow T \qquad \text{and}$$
$$a \centerdot (b \centerdot c) : X \longrightarrow T$$

So the following equation, also known as a coherence statement, must now hold

$$(a \centerdot b) \centerdot c : X \longrightarrow T = a \centerdot (b \centerdot c) : X \longrightarrow T$$

We can see that agents, as described so far to be morphisms, do not form a category of processes types yet under sequential composition "$\centerdot$". The reason is that there are no identity morphisms. However, we will obtain such a category in section 3 as below.

3 Category of Processes Types

By the defined structure of agents, we can construct **uPro** to be a category of processes types. In fact, **uPro** is constructed as follows:

• *Objects as the sets of inputs to or outputs from processes*: Let IO be the set, whose elements are the sets of inputs to or outputs from processes. That is,

$$IO = \{\text{X} \mid \text{X is a set of inputs to or outputs from processes}\} \qquad (1)$$

• *Morphisms as agents*: Let A be the set of agents. Then associated with each objects X in IO, we define morphism from X to X, called identity morphism on X, as an agent $X \xrightarrow{id_X} X$ in A or $\overset{id_X}{\circlearrowleft} X$.

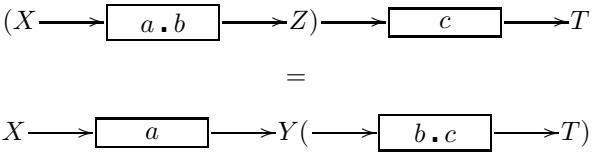
This means as $X \longrightarrow \boxed{id_X} \longrightarrow X$ and to each pair of agents $X \xrightarrow{a} Y$ and $Y \xrightarrow{b} Z$, there is an associated agent $X \xrightarrow{a \centerdot b} Z$, the composition of a with b. This means that if we have two agents $X \longrightarrow \boxed{a} \longrightarrow Y$ and $Y \longrightarrow \boxed{b} \longrightarrow Z$ then there exists the associated agent

$$X \longrightarrow \boxed{a \centerdot b} \longrightarrow Z = X \longrightarrow \boxed{a} \xrightarrow{Y} \boxed{b} \longrightarrow Z$$

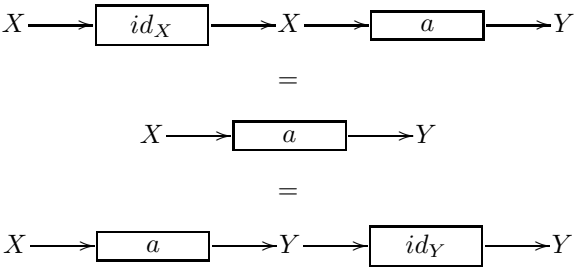
For every object X, Y, Z and T in IO and the agents $X \xrightarrow{a} Y$, $Y \xrightarrow{b} Z$ and $Z \xrightarrow{c} T$ in the set A of agents, then the following equations, also known as the coherence statements, hold:

Associativity: $(a \centerdot b) \centerdot c : X \longrightarrow T = a \centerdot (b \centerdot c) : X \longrightarrow T$
Identity: $id_X \centerdot a : X \longrightarrow Y = a : X \longrightarrow Y = a \centerdot id_Y : X \longrightarrow Y$

In other words, these coherence statements can be diagrammatically drawn such as



and



These are all the basic ingredients we need to form the category, named **uPro**, of processes types. Agents are closed under the sequential composition "$\centerdot$". Moreover, the agents such as $id_X : X \longrightarrow X$ act as identity morphisms for this composition.

Thus, category **uPro** has sets (denoted by X, Y etc, for example) as its objects, and agents (e.g., $a : X \longrightarrow Y$) as its morphisms.

Property 2. *Every agent $a : X \longrightarrow Y$ and $c : Z \longrightarrow T$ in* **uPro**, *there exist unique agents $x : Z \longrightarrow X$ and $y : Y \longrightarrow T$ in* **uPro** *such that the following equation holds*

$$x \centerdot a \centerdot y : Z \longrightarrow T = c : Z \longrightarrow T \tag{2}$$

It follows that the agents $id_X : X \longrightarrow X$ and $id_Y : Y \longrightarrow Y$ are identity agents, then the equation

$$id_X \centerdot a \centerdot id_Y : X \longrightarrow Y = a : X \longrightarrow Y \tag{3}$$

holds.

Property 3. *For all agents x and y in* **uPro**, *if $x \centerdot a \centerdot y = x \centerdot b \centerdot y$, then $a = b$*

These properties of agents in **uPro** lead to a categorical structure, called *extensional monoidal category* as below.

4 Extensional Monoidal Category of Processes Types

The composition operation "$\centerdot$" defines an extensional monoidal category on the category **uPro**. In fact, the extensional monoidal category is a category **uPro** equipped with the following additional structure:

- A multifunctor $\centerdot : \mathbf{uPro} \times \mathbf{uPro} \times \mathbf{uPro} \longrightarrow \mathbf{uPro}$ called the composition operation,
- For every agent $a : X \longrightarrow Y$, id_X and id_Y called left and right identity agent, respectively,
- The natural isomorphisms α, β, γ and λ subject to some coherence conditions expressing the fact that
  1. *The composition operation "$\centerdot$" is associative*: there are three natural isomorphisms α, β, γ, called associative ones, with components

$$\alpha(a,b,c,d,e) : (a \centerdot b \centerdot c) \centerdot d \centerdot e \longleftrightarrow a \centerdot (b \centerdot c \centerdot d) \centerdot e$$
$$\beta(a,b,c,d,e) : a \centerdot (b \centerdot c \centerdot d) \centerdot e \longleftrightarrow a \centerdot b \centerdot (c \centerdot d \centerdot e)$$
$$\gamma(a,b,c,d,e) : (a \centerdot b \centerdot c) \centerdot d \centerdot e \longleftrightarrow a \centerdot b \centerdot (c \centerdot d \centerdot e)$$

     Sometimes, the natural isomorphisms α, β, γ are also represented as

$$\alpha(a,b,c,d,e) : (a \centerdot b \centerdot c) \centerdot d \centerdot e \cong a \centerdot (b \centerdot c \centerdot d) \centerdot e$$
$$\beta(a,b,c,d,e) : a \centerdot (b \centerdot c \centerdot d) \centerdot e \cong a \centerdot b \centerdot (c \centerdot d \centerdot e)$$
$$\gamma(a,b,c,d,e) : (a \centerdot b \centerdot c) \centerdot d \centerdot e \cong a \centerdot b \centerdot (c \centerdot d \centerdot e)$$

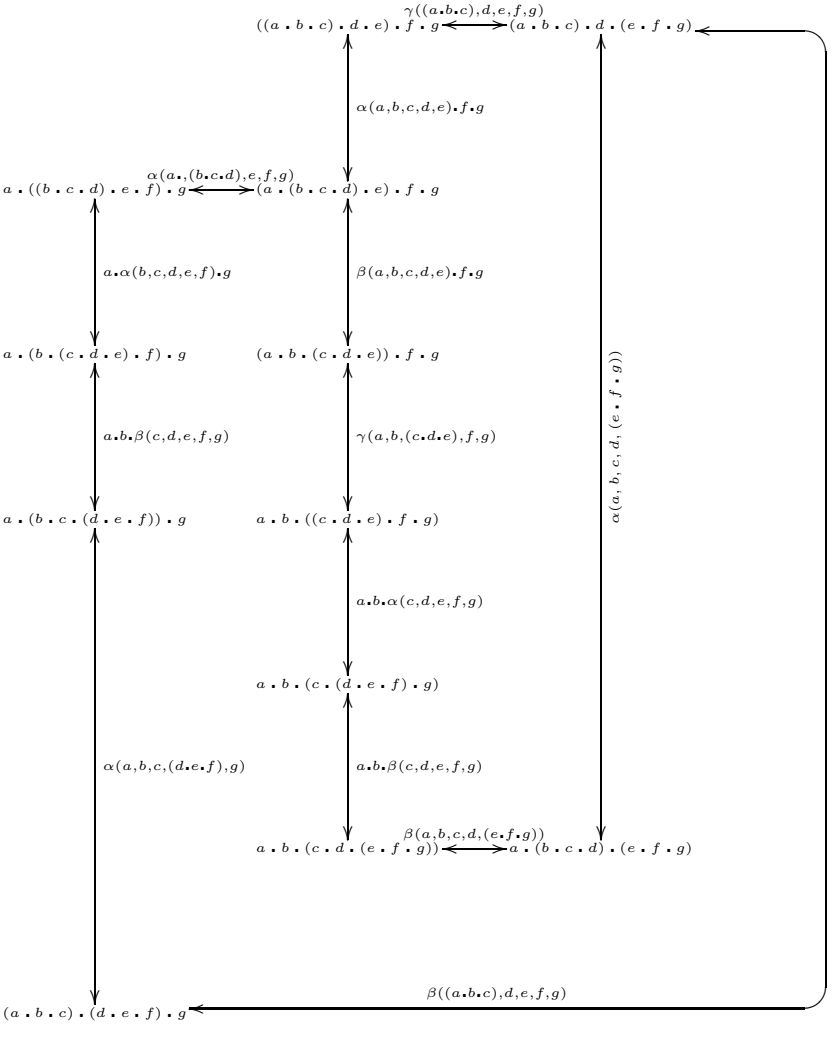
  2. *The composition operation "$\centerdot$" has id_X and id_Y as left and right identities of every agent $a : X \longrightarrow Y$*: there is a natural isomorphism λ, called identity one, with components $\lambda(a) : id_X \centerdot a \centerdot id_Y \longleftrightarrow a$ or $\lambda(a) : id_X \centerdot a \centerdot id_Y \cong a$

The coherence conditions for three natural isomorphisms α, β and γ are thought of as the diagram (4) commuting for all agents a, b, c, d, e, f and g in **uPro**.

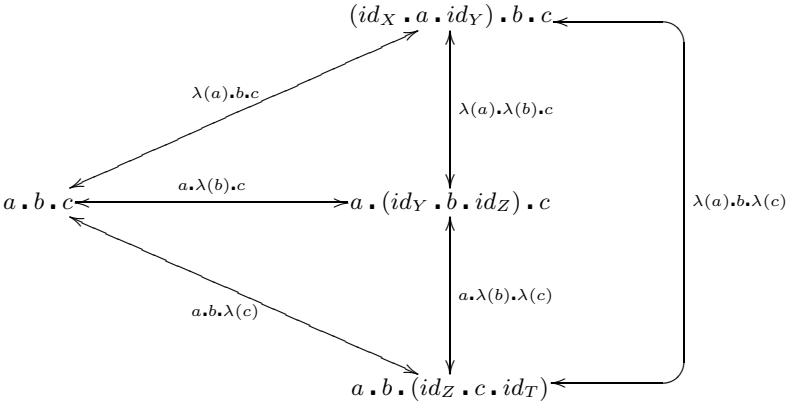
The coherence condition for the identity natural isomorphism λ is seen as the diagram (5) commuting for all agents $a : X \longrightarrow Y, b : Y \longrightarrow Z$ and $c : Z \longrightarrow T$ in **uPro**. The coherence condition expresses the statement that two or more natural isomorphisms between two given multifunctors are equal based on the existence of which is given or follows from general characteristics. Such situations are ubiquitous for agents on UCSs. Coherence conditions are formulated and studied in categorical structures of UCS agents known as a categorical approach of the AUC.

5 Concurrent Composition of Agents

What other operations, besides sequential composition "$\centerdot$", do we have on agents? To begin with, there is an obvious notion of concurrent composition. Let $X + X'$


((a . b . c) . d . e) . f . g
γ((a.b.c),d,e,f,g)
(a . b . c) . d . (e . f . g)
α(a,b,c,d,e).f.g
a . ((b . c . d) . e . f) . g
α(a.,(b.c.d),e,f,g)
(a . (b . c . d) . e) . f . g
a.α(b,c,d,e,f).g
β(a,b,c,d,e).f.g
a . (b . (c . d . e) . f) . g
(a . b . (c . d . e)) . f . g
α(a, b, c, d, (e . f . g))
a.b.β(c,d,e,f,g)
γ(a,b,(c.d.e),f,g)
a . (b . c . (d . e . f)) . g
a . b . ((c . d . e) . f . g)
a.b.α(c,d,e,f,g)
a . b . (c . (d . e . f) . g)
α(a,b,c,(d.e.f),g)
a.b.β(c,d,e,f,g)
β(a,b,c,d,(e.f.g))
a . b . (c . d . (e . f . g))
a . (b . c . d) . (e . f . g)
β((a.b.c),d,e,f,g)
(a . b . c) . (d . e . f) . g

(4)

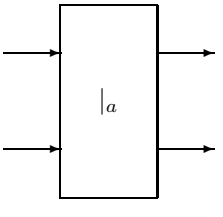


(5)

denote the disjoint union of sets X and X'. Then given two agents $a : X \longrightarrow Y$ and $b : X' \longrightarrow Y'$, we can define an agent $a \mid_a b : X + X' \longrightarrow Y + Y'$ as follows:

$$\mid_a \stackrel{def}{=} \left(\begin{array}{l} \_ + \_ : IO \times IO \longrightarrow IO \\ \_ \mid_a \_ : A \times A \longrightarrow A \\ id_{\mid_a} : A \\ \\ \text{such that} \\ \\ \forall f, g, h : A \bullet (f \mid_a g) \mid_a h = f \mid_a (g \mid_a h) \\ \forall f : A \bullet f \mid_a id_{\mid_a} = id_{\mid_a} \mid_a f = f \\ \forall X, Y : IO \bullet X + Y = dom(id\ X \mid_a id\ Y) \\ \forall X, Y : IO \bullet id\ X \mid_a id\ Y = id(X + Y) \\ \forall f, g, p, q : A \mid cod\ f = dom\ g \wedge cod\ p = \\ dom\ q \bullet (f \centerdot g) \mid_a (p \centerdot q) = (f \mid_a p) \centerdot (g \mid_a q) \end{array} \right)$$

A diagram of concurrent composition $\mid_a$ is pictorially represented as



(6)

The operation $\mid_a$ defines a symmetric monoidal structure on the category **uPro**. In fact, **uPro** equipped with the following bifunctor defines a symmetric monoidal category.

$$+ : \mathbf{uPro} \times \mathbf{uPro} \longrightarrow \mathbf{uPro} \tag{7}$$

which, called concurrent composition operation, is associative up to a natural isomorphism, and an empty set $\emptyset$ which is both a left and right identity for

the bifunctor "+", again, up to natural isomorphism. The associated natural isomorphisms are subject to some coherence conditions which ensure that all the relevant diagrams commute. We consider the facts in detail as below.

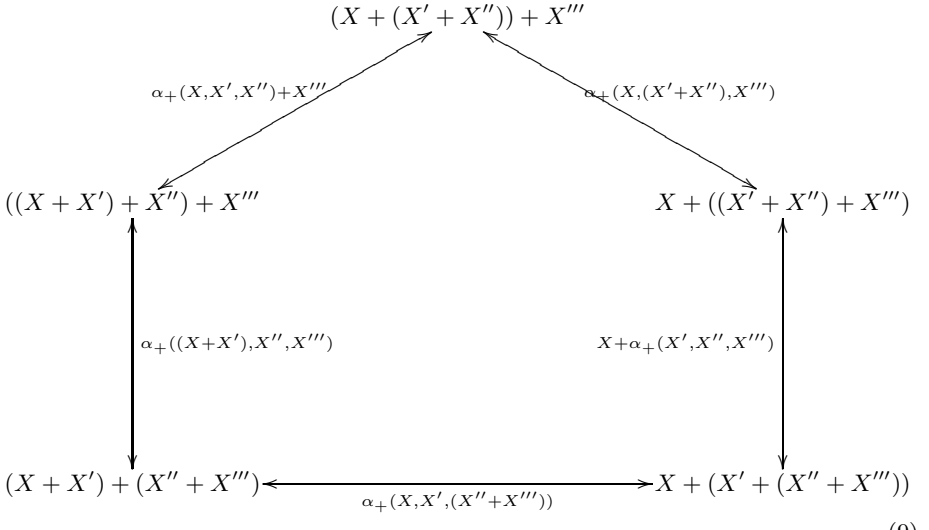
Property 4. *The concurrent composition operation "+" is associative up to a natural isomorphism α_+, called associativity, with components:*

$$\alpha_+(X, X', X'') : (X + X') + X'' \longleftrightarrow X + (X' + X'') \tag{8}$$

or, sometimes, this associativity is also written as

$$\alpha_+(X, X', X'') : (X + X') + X'' \cong X + (X' + X'')$$

The coherence condition for the associativity α_+ means that for all sets of inputs or outputs X, X', X'' and X''' in IO, the diagram (9) commutes.



(9)

Property 5. *The concurrent composition operation "+" has an identity $\emptyset$ up to left identity natural isomorphism λ_+ and right identity natural isomorphism ρ_+ with components:*

$$\lambda_+(X) : \emptyset + X \longleftrightarrow X \tag{10}$$

$$\rho_+(X) : X + \emptyset \longleftrightarrow X \tag{11}$$

or, sometimes, these left and right identity natural isomorphisms are also written as

$$\lambda_+(X) : \emptyset + X \cong X$$
$$\rho_+(X) : X + \emptyset \cong X$$

Moreover, the coherence conditions for the left and right identity natural isomorphisms λ_+ and ρ_+ mean that for all sets of inputs or outputs X and X' in IO, the diagram (13) commutes.

Property 6. *Given agents* $b : X \longrightarrow X'$, $c : X' \longrightarrow X''$, $p : Y \longrightarrow Y'$ *and* $q : Y' \longrightarrow Y''$, *then two following expressions are equal:*

$$(b \centerdot c) \mid_a (p \centerdot q) = (b \mid_a p) \centerdot (c \mid_a q) \tag{12}$$

where both left and right expressions define an agent, whose type is $X + Y \longrightarrow X'' + Y''$

$$\begin{array}{ccc} & (X + \emptyset) + X' & \\ {\scriptstyle \rho_+(X)+X'} \swarrow & & \uparrow {\scriptstyle \alpha_+(X,\emptyset,X')} \\ X + X' & & \\ {\scriptstyle X+\lambda_+(X')} \searrow & & \downarrow \\ & X + (\emptyset + X') & \end{array} \tag{13}$$

Property 7. *Given the identity agents* $id_X : X \longrightarrow X$ *and* $id_{X'} : X' \longrightarrow X'$, *then*

$$id_X \mid_a id_{X'} = id_{X+X'} \tag{14}$$

where $id_{X+X'}$ *is the identity agent* $id_{X+X'} : X + X' \longrightarrow X + X'$.

In the category **uPro**, we have $X + X' = X' + X$ for the sets of X and X' in IO. In particular, if the agent $id_{X+X'} : X + X' \longrightarrow X + X'$ then it is equivalent to $id_{X+X'} : X + X' \longrightarrow X' + X$ called a symmetric agent between $X + X'$ and $X' + X$. This means that **uPro** is what is called a symmetric monoidal category.

Here, as a practice offered to readers, let us try to represent the coherence condition for the symmetric agent $id_{X+X'}$. In other words, we have to draw commutative diagram(s) for $id_{X+X'}$ in **uPro**.

Formally, for X and X' in IO, the symmetric agent $id_{X+X'}$ is the natural isomorphism between $X + X'$ and $X' + X$, which is defined as follows:

$$id_{X+X'} \stackrel{def}{=} \left(\begin{array}{l} id\_+\_ : IO \times IO \longrightarrow A \\ id : IO \longrightarrow A \\ \\ \text{such that} \\ \\ \forall X, X' : IO \bullet dom(id_{X+X'}) = \\ X + X' \wedge cod(id_{X+X'}) = X' + X \\ \forall X, X' : IO \bullet id_{X+X'} \centerdot id_{X'+X} = \\ id\ (X + X') \\ \forall X, X', X'' : IO \bullet id_{(X+X')+X''} = \\ (id\ X \mid_a id_{X'+X''}) \centerdot (id_{X+X''} \mid_a id\ X') \end{array} \right)$$

6 Future Work

In future, we hope to address further developments which are not exposed yet in this paper. In this section, we briefly discuss the direction of these developments. For the category **uPro** of processes types, we also develop a category **uPro′** equipped with the following structures:

- Let IO be the set of inputs or outputs. Every element X in IO is called an object of **uPro′**.
- For each pair (X, Y) of objects of **uPro′**, let $agent(X, Y)$ be an object of **uPro**, called the *agent-object* of X and Y.
- For each object X of **uPro′**, let $id(X)$ be a morphism in **uPro** from id_X to $agent(X, X)$, called the *identity morphism* of X. That is $id_X \xrightarrow{id(X)} agent(X, X)$
- For each triple (X, Y, Z) of objects of **uPro′**, let

$$\begin{array}{c} agent(X, Y) \times agent(Y, Z) \\ \Big\downarrow \text{“}\centerdot\text{”} \\ agent(X, Z) \end{array}$$

  be a morphism in **uPro** called the composition morphism of X, Y and Z.

If all $agent(X, Y)$ satisfy three axioms including associativity, left identity and right identity then **uPro′** is a category *enriched* over **uPro**.

7 Conclusions

This paper has dealt with the ubiquitous computing support for information and services integration based on formal agent-oriented approach. Agent-orientation has been featured using categorical structures from which its useful properties

emerge. For ubiquitous agents, our developments have stemmed from formulating type of process as $X \longrightarrow Y$, type-based agent as $a : X \longrightarrow Y$ and category **uPro** of processes types. Then extensional monoidal structure and symmetry monoidal structure of the category of processes types have been constructed in order to consider the significant properties of agents on UCSs.

References

1. Denko, M.K., Yang, L.T., Zhang, Y. (eds.): Autonomic Computing and Networking, 1st edn., 464 pages. Springer, USA (2009)
2. Kwon, O.B.: Multi-agent System Approach to Context-aware Coordinated Web Services under General Market Mechanism. Decision Support Systems 41(2), 380–399 (2006)
3. Mangina, E., Carbo, J., Molina, J.M.: Agent-based Ubiquitous Computing, 1st edn., 216 pages. Atlantis Press (2009)
4. Parashar, M., Hariri, S. (eds.): Autonomic Computing: Concepts, Infrastructure and Applications, 1st edn., 568 pages. CRC Press (2006)
5. Vinh, P.C.: Formal Aspects of Self-* in Autonomic Networked Computing Systems. In: Autonomic Computing and Networking, 1st edn., pp. 383–412. Springer, USA (2009)
6. Vinh, P.C.: Categorical Approaches to Models and Behaviors of Autonomic Agent Systems. The International Journal of Cognitive Informatics and Natural Intelligence (IJCiNi) 3(1), 17–33 (2009)

Prediction of Rainfall Time Series Using Modular RBF Neural Network Model Coupled with SSA and PLS

Jiansheng Wu[1,2]

[1] School of Information Engineering, Wuhan University of Technology Wuhan, 430070, Hubei, China
[2] Department of Mathematics and Computer, Liuzhou Teacher College, Liuzhou, 545004, Guangxi, China
wjsh2002168@163.com

Abstract. In this paper, a new approach using an Modular Radial Basis Function Neural Network (M-RBF-NN) technique is presented to improve rainfall forecasting performance coupled with appropriate data–preprocessing techniques by Singular Spectrum Analysis (SSA) and Partial Least Square (PLS) regression. In the process of modular modeling, SSA is applied for the time series extraction of complex trends and finding structure. In the second stage, the data set is divided into different training sets by used Bagging and Boosting technology. In the third stage, then modular RBF–NN predictors are produced by different kernel function. In the fourth stage, PLS technology is used to choose the appropriate number of neural network ensemble members. In the final stage, least squares support vector regression is used for ensemble of the M-RBF-NN to prediction purpose. The developed RBF-NN model is being applied for real time rainfall forecasting and flood management in Liuzhou, Guangxi. Aimed at providing forecasts in a near real time schedule, different network types were tested with the same input information. Additionally, forecasts by M-RBF-NN model were compared to the convenient approach. Results show that that the predictions using the proposed approach are consistently better than those obtained using the other methods presented in this study in terms of the same measurements. Sensitivity analysis indicated that the proposed M-RBF-NN technique provides a promising alternative to rainfall prediction.

Keywords: Singular Spectrum Analysis, Radial Basis Function Neural Network, Partial Least Square Regression, Rainfall prediction, Least Squares Support Vector Regression.

1 Introduction

Accurate and timely rainfall prediction is essential for the planning and management of water resources, in particular for flood warning systems because it can provide an information of help prevent casualties and damages caused by natural disasters [1]. For example, a flood warning system for fast responding

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 509–518, 2012.

catchments may require a quantitative rainfall forecast to increase the lead time for warning. Similarly, a rainfall forecast provide advance information for many water quality problems [2]. Rainfall prediction is one of the most complex and difficult elements of the hydrology cycle to understand and to model due to the complexity of the atmospheric processes involved and the variability of rainfall in space and time [3], [4].

Recently, the concept of coupling different models has been a very popular research topic in hydrologic forecasting, which has attracted scientists from other fields including Statistics, Machine Learning and so on. They can be broadly categorized into ensemble models and modular (or hybrid) models. The basic idea behind ensemble models is to build several different or similar models for the same process and to integrate them together. Their success largely arises from the fact that they lead to improved accuracy compared to a single classification or regression model. Typically, ensemble methods comprise two phases: a) the production of multiple predictive models, and b) their combination. Recent work, has been the main consideration the reduction of the ensemble size prior to combination [5] [6].

Different from the previous work in this paper, one of the main purposes is to develop a Modular Radial Basis Function Neural Network (MRBF–NN) coupled with appropriate data–preprocessing techniques by Singular Spectrum Analysis (SSA) and Partial Least Squar (PLS) to improve the accuracy of rainfall forecasting. The rainfall data of Liuzhou in Guangxi is predicted as a case study for our proposed method. An actual case of forecasting monthly rainfall is illustrated to show the improvement in predictive accuracy and capability of generalization achieved by our proposed MRBF–NN model.

The rest of this study is organized as follows. Section 2 describes the proposed MRBF–NN, ideas and procedures. For further illustration, this work employs the method set up a prediction model for rainfall forecasting in Section 3. Discussions are presented in Section 4 and conclusions are drawn in the final Section.

2 The Building Process of the MRBF–NN

Firstly, Singular Spectrum Analysis (SSA) is used to reduce noises in original rainfall time series, and to reconstruct the new time series in this section. Secondly, a triple–phase nonlinear modular RBF–NN model is proposed for rainfall forecasting based on the different activation function and training data. Then an appropriate number of RBF–NN predictors are selected from the considerable number of candidate predictors by the Partial Least Square technology. Finally, selected RBF–NN predictors are combined into an aggregated neural predictor in terms of LS–SVR.

2.1 Singular Spectrum Analysis

The Singular Spectrum Analysis (SSA) technique is a novel and powerful technique of time series analysis incorporating the elements of classical time series

analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. method. Broomhead and King [7] was presented SSA because they show that the singular value decomposition (SVD) is effective in reducing noises. The aim of SSA is to make a decomposition of the original series into the sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and a structure less noise. The basic SSA algorithm is described by the related literature [8].

2.2 Radial Basis Function Neural Network

Radial basis function was introduced into the neural network literature by Broomhead and Lowe [9] [10], which was motivated by the presence of many local response neurons in human brain. On the contrary to the other type of NN used for nonlinear regression like back propagation feed forward networks, it learns quickly and has a more compact topology.

The network is generally composed of three layers: an input layer, a single layer of nonlinear processing neuron and output layer. The output of the RBF–NN is calculated according to

$$y_i = f_i(x) = \sum_{k=1}^{N} w_{ik}\phi_k(\|x - c_k\|), i = 1, 2, \cdots, m \tag{1}$$

where $x \in \Re^{n\times 1}$ is an input vector, $\phi_k(\cdot)$ is a function from $\Re^+$ to $\Re$, $\|\cdot\|_2$ denotes the Euclidean norm, w_{ik} are the weights in the output layer, n is the number of neurons in the hidden layer, and $c_k \in \Re^{n\times 1}$ are the centers in the input vector space. The functional form of $\phi_k(\cdot)$ is assumed to have been given, and some typical choices are shown in Table 1.

Table 1. Types of kernel function name and formula

Modual	Functional name	Function formula
A	Linear function	$\phi(x) = x$
B	Cubic approximation	$\phi(x) = x^3$
C	Thin-plate-spline function	$\phi(x) = x^2 lnx$
D	Guassian function	$\phi(x) = exp(-x^2/\sigma^2)$
E	Multi-quadratic function	$\phi(x) = \sqrt{x^2 + \sigma^2}$
F	Inverse multi-quadratic function	$\phi(x) = \frac{1}{\sqrt{x^2+\sigma^2}}$

The training procedure of the RBF networks is a complex process, this procedure requires the training of all parameters including the centers of the hidden layer units $(c_i, i = 1, 2, \cdots, m)$, the widths (σ_i) of the corresponding Gaussian functions, and the weights $(\omega_i, i = 0, 1, \cdots, m)$ between the hidden layer and output layer. In this paper, the the orthogonal least squares algorithm (OLS) is used to training RBF based on the minimizing of SSE. The more detailed about algorithm is described by the related literature [11].

2.3 Selecting Appropriate Ensemble Members

When data were completed the training, each Modular RBF-NN predictor has generated its own result. However, if there are a great number of individual members, we need to select a subset of representatives in order to improve ensemble efficiency. In this paper, the Partial Least Square (PLS) regression technique is adopted to select appropriate ensemble members. Partial least squares (PLS) regression analysis was developed in the late seventies by Herman O. A. Wold [12]. PLS regression is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). Interested readers can be referred to [13] for more details.

2.4 Least Squares Support Vector Regression

Support vector regression (SVR) was derived form support vector machine (SVM) technique. LS-SVM is a least squares modification to the Support Vector Machine [14]. When SVM can be used for spectral regression purpose, it is called least squares support vector regression (LS–SVR) [15]. Where $\{x_i, i = 1, 2, \cdots, N\}$ are the output of linear and nonlinear forecasting predictors, $\{y_i, i = 1, 2, \cdots, N\}$ are the aggregated output and the goal is to estimate a regression function f. Basically we define a N dimensional function space by defining the mappings $\varphi = [\varphi_1, \varphi_2, \cdots, \varphi_N]^T$ according to the measured points. The LS-SVM model is of the form $\hat{f(x)} = \omega^T \varphi(x) + b$ where ω is a weight vector and b is a bias term. The optimization problem is the following:

$$\begin{cases} \min J(\omega, \epsilon) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum\limits_{i=1}^{N}\epsilon_i^2 \\ s.t. \;\; y_i = \omega^T\varphi(x_i) + b + \epsilon_i, i = 1, 2, \cdots, N \end{cases} \tag{2}$$

where the fitting error is denoted by ϵ_i. Hyper-parameter γ controls the trade-off between the smoothness of the function and the accuracy of the fitting. This optimization problem leads to a solution by solving the linear Karush–Kuhn–Tucker (KKT) [16]:

$$\begin{bmatrix} 0 & \mathbf{I_n^T} \\ \mathbf{I_n} & \mathbf{K} + \gamma^{-1}\mathbf{I} \end{bmatrix} \begin{bmatrix} b_0 \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{3}$$

where $\mathbf{I_n}$ is a $[n \times 1]$ vector of ones, $\mathbf{T}$ means transpose of a matrix or vector, γ a weight vector, $\mathbf{b}$ regression vector and b_0 is the model offset. $\mathbf{K}$ is kernel function. A common choice for the kernel function is the Gaussian function:

$$K(x, x_i) = e^{\frac{\|x - x_i\|^2}{2\sigma^2}} \tag{4}$$

2.5 The Establishment of Modular RBF–NN

To summarize, the proposed Modular RBF–NN model consists of five main steps. In the process of modular modeling, firstly, SSA is applied for the time series extraction of complex trends and finding structure. Secondly, the data set is

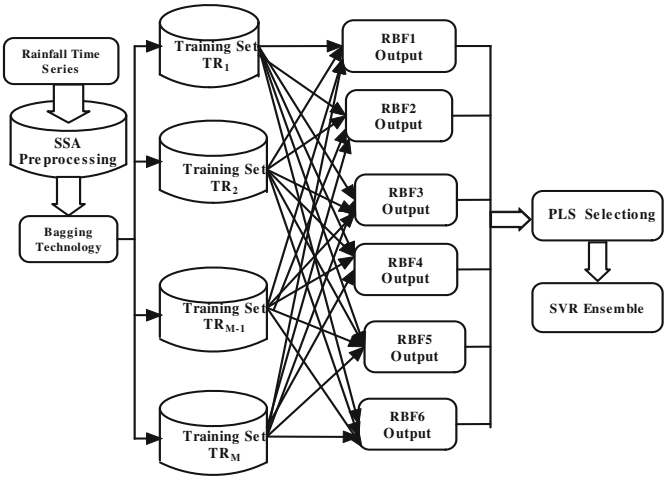


Fig. 1. The Modular RBF–NN architecture

divided into different training sets by used Bagging and Boosting technology. Thirdly, then modular RBF–NN predictors are produced by different kernel function. Fourthly, PLS technology is used to choose the appropriate number of neural network ensemble members. Finally, LS-SVR is used for ensemble of the M-RBF-NN to prediction purpose. The basic flow diagram can be shown in Figure 1.

3 Results and Discussion

3.1 Empirical Data

Liuzhou is one of the highly developing cities in southwest of China, which is the capital and commercial city of Guangxi. Historical monthly rainfall data was collected from 24 stations of the Liuzhou Meteorology Administration rain gauge networks for the period from 1949 to 2010. After analyzed data, the period from January 1949 to December 2006 was selected to train MRBF-NN models, and the data from January 2007 to December 2010 were used as a testing set. Thus the training data set contained 696 data points in time series for MRBF–NN learning, and the other 48 data were used to test sample for MRBF–NN Generalization ability. Fig.2 shows the average monthly rainfall, taken over a period from 1949 to 2010, in Liuzhou. There is one peak of rainfall during a year in August.

3.2 Criteria for Evaluating Model Performance

Three different types of standard statistical performance evaluation criteria were employed to evaluate the performance of various models developed in this

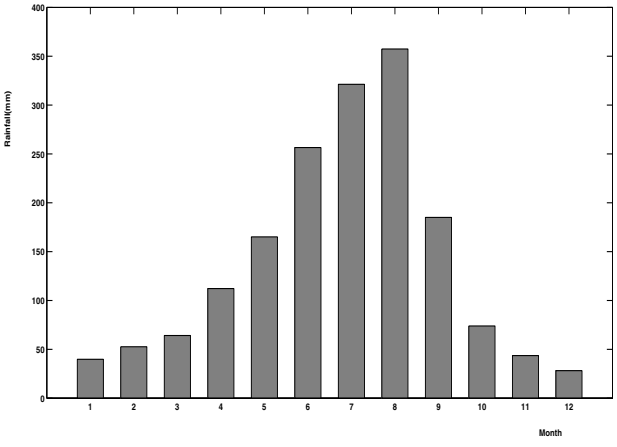


Fig. 2. Average monthly rainfall in Liuzhou

paper. These are average absolute relative error (AARE), root mean square error (RMSE), and the correlation coefficient (CC)which be found in many paper [18].

According to the previous literature, there are a variety of methods for rainfall forecasting model in the past studies. The author used Eviews statistical packages to formulate the ARIMA model. Akaike information criterion (AIC) was used to determine the best model. The model is generated from the data set is AR(5). The equation used is presented in Equation 5.

$$x_t = 1 - 0.30x_{t-1} - 0.02x_{t-2} - 0.11x_{t-3} + 0.91x_{t-4} + 0.05x_{t-5} \tag{5}$$

For the purpose of comparison by the same four input variables, we have also built other three rainfall forecasting models: multi–layer perceptron neural network (MLP–NN) model, single RBF–NN and stacked regression (SR) ensemble [17] method based on RBF–NN.

The standard RBF–NN with Gaussian-type activation functions in hidden layer were trained for each training set, then tested as an ensemble for each method for the testing set. Each network was trained using the neural network toolbox provided by Matlab software package. In addition, the best single RBF neural network using cross–validation method [13] (i.e., select the individual RBF network by minimizing the MSE on cross–validation) is chosen as a benchmark model for comparison.

3.3 Analysis of the Results

Table 2 illustrates the fitting and testing accuracy and efficiency of the model in terms of various evaluation indices for 696 training and 48 testing samples. From the table 2, we can generally see that learning ability of M–RBF–NN outperforms the other four models under the same network input. As a consequence, poor

Table 2. Performance statistics of the five models for rainfall fitting and forecasting

Model	AR(5)	MLP-NN	S-RBF-NN	SR	M-RBF-NN
Index		Training data (from 1949 to 2006)			
AARE	92.90	82.63	83.84	61.74	52.64
RMSE	87.85	72.14	73.14	56.69	44.25
PRC	0.8403	0.8939	0.8901	0.9239	0.9612
		Testing data (from 2007 to 2010)			
AARE	117.09	118.85	107.10	76.25	68.63
RMSE	85.96	67.98	68.93	67.92	48.46
PRC	0.7269	0.7540	0.7518	0.7936	0.8944

performance indices in terms of AARE, RMSE and CC can be observed in AR(5) model than other four model. Table 2 also shows that the performance of M-RBF-NN is the best in case study for training samples.

The more important factor to measure performance of a method is to check its forecasting ability of testing samples in order for actual rainfall application. Table 2 shows the forecasting results of five different models for 48 testing samples, we can see that the forecasting results of M-RBF-NN model are the best in all models, the M-RMF-NN can better capture the mapping relation than using other four model. Table 2 also shows that the forecasting performance of four different models from different perspectives in terms of various evaluation indices.

Figure 3-6 show the forecasting results of five different models for 48 testing samples, we can see that the forecasting results of M-RBF-NN model are best in all models. From the graphs and table, we can generally see that the forecasting results are very promising in the rainfall forecasting under the research where either the measurement of fitting performance is goodness or where the forecasting performance is effectiveness. It also can be seen that there was consistency in the results obtained between the training and testing of these M-RBF-NN model.

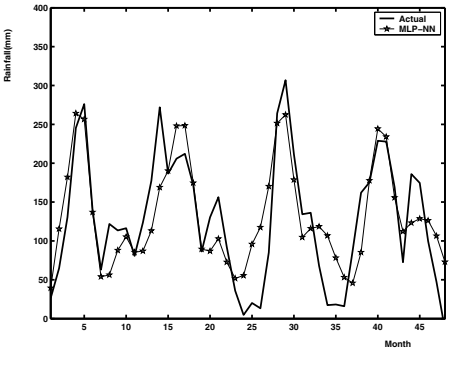


Fig. 3. Forecasting for Model MLP-NN

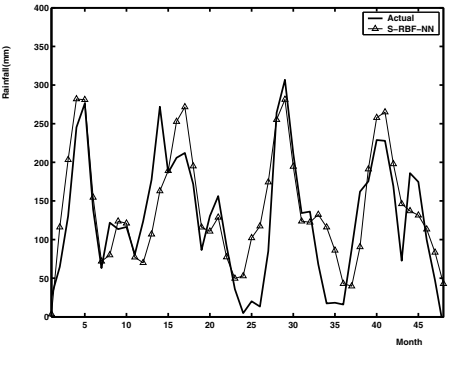


Fig. 4. Forecasting for Model S-RBF-NN

Comparison of Model AR(5) with Model MLP-NN, both of which used the same input data, the Model MLP-NN yielded better results than Model AR(5) for both the training and testing samples. The results show the rainfall system is complex nonlinear system and the traditional statistical model is very difficult for accuracy prediction.

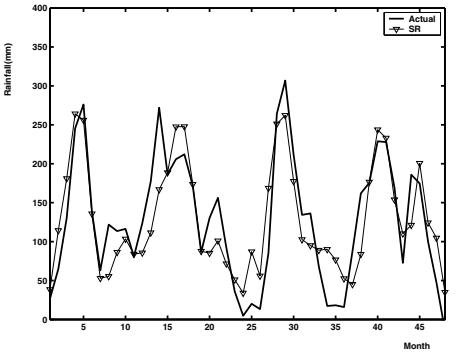


Fig. 5. Forecasting for Model SR

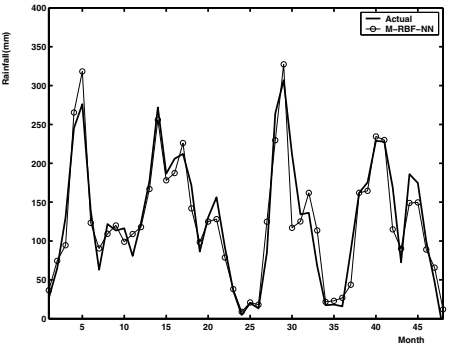


Fig. 6. Forecasting for Model M-RBF-NN

For model MLP-NN and S-RBF-NN, the results of these two models is closer in testing samples. As shown in Table 2, models MLP-NN and S-RBF-NN are based on neural network theory, but those algorithms are different. Those results indicate that model neural network is capable to model without prescribing hydrological processes, catching the complex nonlinear relation of input and output, and solving without the use of differential equations. Model M-RBF–NN, which involved the same input data of rainfall at the Liuzhou, produced the highest performance. For example, the AARE of the M-RBF–NN is 68.63, the RMSE of the M-RBF–NN model is 48.46, and the PRC of the M-RBF–NN model is 0.94. The values of AARE and RMSE is the minimum and the values of PRC is the maximum in all models. The results indicate that the deviations between original values and forecast are very small, is capable to capture the average change tendency of the daily rainfall data.

From the experiments presented in this study we can draw that the M-RBF-NN model is superior to other model about the fitting and testing cases in terms of the different measurement, as can be seen from Table 2. There are three main reasons for this phenomenon. Firstly, the rainfall system contain complex nonlinear pattern, SSA can extract complex trends and find structure in rainfall time series. Using different the kernel function form of an effective nonlinear mapping can establish the effective nonlinear mapping for rainfall forecasting. Secondly, the output of different models has the high correlative relationship, the high noise, nonlinearity and complex factors. If PLS technology don't reduce the dimension of the data and extract the main features, the results of

the model is unstable. At last, LS-SVR is used to combine the selected individual forecasting results into a nonlinear ensemble model, which keeps the flexibility of the nonlinear model. So the proposed nonlinear modular ensemble model can be used as a feasible approach to rainfall forecasting.

4 Conclusion

Accurate rainfall forecasting is crucial for a frequent unanticipated flash flood region to avoid life losing and economic loses. In this study, an modular Radial Basis Function Neural Network model was employed to forecast monthly rainfall for Liuzhou, Guangxi. In terms of the different forecasting models, empirical results show that the developed modular model performs the best for monthly rainfall on the basis of different criteria. Our experimental results demonstrated the successful application of our proposed new model, M–RBF–NN, for the complex forecasting problem. It demonstrated that it increased the rainfall forecasting accuracy more than any other model employed in this study in terms of the same measurements. So the M–RBF–NN ensemble forecasting model can be used as an alternative tool for monthly rainfall forecasting to obtain greater forecasting accuracy and improve the prediction quality further in view of empirical results, and can provide more useful information, avoid invalid information for the future forecasting.

Acknowledgment. The authors would like to express their sincere thanks to the editor and anonymous reviewers comments and suggestions for the improvement of this paper. This work was supported by Program for Excellent Talents in Guangxi Higher Education Institutions, by Natural Science Foundation of Guangxi under Grant No. 2011GXNSFE018006 and by the Natural Science Foundation of China under Grant No.11161029.

References

1. Wu, J., Liu, M.Z., Jin, L.: A Hybrid Support Vector Regression Approach for Rainfall Forecasting Using Particle Swarm Optimization and Projection Pursuit Technology. International Journal of Computational Intelligence and Applications 9(3), 87–104 (2010)
2. Wu, J., Jin, L.: Study on the Meteorological Prediction Model Using the Learning Algorithm of Neural Networks Ensemble Based on PSO agorithm. Journal of Tropical Meteorology 15(1), 83–88 (2009)
3. French, M.N., Krajewski, W.F., Cuykendall, R.R.: Rainfall Forecasting in Space and Time Using Neural Network. Journal of Hydrology 137, 1–31 (1992)
4. Gwangseob, K., Ana, P.B.: Quantitative Flood Forecasting Using Multisensor Data and Neural Networks. Journal of Hydrology 246, 45–62 (2001)
5. Parag, P., Preeti, B., Ajith, A., Prasanna, P., Amol, D.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(3), 227–235 (2011)

6. Partalas, I., Hatzikos, E., Tsoumakas, G., Vlahavas, I.: Ensemble Selection for Water Quality Prediction. In: Proeedings of 10th International Conference on Engineering Applications of Neural Networks, pp. 428–435 (2007)
7. Broomhead, D.S., King, G.P.: Extracting Qualitative Dynamics from Experimental Data. Physica D 20, 217–236 (1986)
8. Alexandrov, T., Bianconcini, S., Dagum, E.B., Maass, P., McElroy, T.S.: A Review of Some Modern Approaches to The Problem of Trend Extraction. Technical report, US Census Bureau RRS2008/03 (2008)
9. Wu, J.: A Semiparametric Regression Ensemble Model for Rainfall Forecasting Based on RBF Neural Network. In: Wang, F.L., Deng, H., Gao, Y., Lei, J. (eds.) AICI 2010, Part II. LNCS (LNAI), vol. 6320, pp. 284–292. Springer, Heidelberg (2010)
10. Moravej, Z., Vishwakarma, D.N., Singh, S.P.: Application of Radial Basis Function Neural Network for Differential Relaying of a Power Transformer. Computers and Electrical Engineering 29, 421–434 (2003)
11. Ham, F.M., Kostanic, I.: Principles of Neurocomputing for Science & Engineering. The McGraw-Hill Companies, New York (2001)
12. Wold, S., Ruhe, A., Wold, H., Dunn, W.J.: The Collinearity Problem in Linear Regression: the Partial Least Squares Approach to Generalized Inverses. Journal on Scientific and Statistical Computing 5(3), 735–743 (1984)
13. Pirouz, D.M.: An Overview of Partial Least Square. Technical report, The Paul Merage School of Business, University of California, Irvine (2006)
14. Suykens, J., Gestel, T., Van, J.: Least Squares Support Vector Machines. The World Scientific Publishing, Singapore (2002)
15. Schökopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge (2002)
16. Wang, H., Li, E., Li, G.Y.: The Least Square Support Vector Regression Coupled with Parallel Sampling Scheme Metamodeling Technique and Application in Sheet Forming Optimization. Materials and Design 30, 1468–1479 (2009)
17. Chang, F.C., Huang, H.C.: A Refactoring Method for Cache-Efficient Swarm Intelligence Algorithms. Information Sciences, doi:10.1016/j.ins.2010.02.025
18. Wu, J.: An Effective Hybrid Semi-Parametric Regression Strategy for Rainfall Forecasting Combining Linear and Nonlinear Regression. International Journal of Applied Evolutionary Computation 2(4), 50–65 (2011)

Heuristic Algorithms for Solving Survivability Problem in the Design of Last Mile Communication Networks

Vo Khanh Trung, Nguyen Thi Minh, and Huynh Thi Thanh Binh

Ha noi University of Science and Technology
trungvokhanh@yahoo.com, ntminh1988@gmail.com,
binhht@soict.hut.edu.vn

Abstract. Given a connected, weighted, undirected graph G = (V, E), a set infrastructure nodes and a set customers C includes two customer types where by customers C1 require a single connection (type-1) and customers C2 need to be redundantly connected (type-2). Survivable Network Design Problem (SNDP) seeks sub-graph of G with smallest weight in which all customers are connected to infrastructure nodes. This problem is NP-hard and has application in the design of the last mile of the real-world communication networks. This paper proposes a new heuristic algorithm for solving SNDP. Results of computational experiments are reported to show the efficiency of proposed algorithm.

Keywords: Survivable Network Design, Steiner Tree Problem, Fiber Optic Network, Heuristic Algorithm, Local Search.

1 Introduction

In the recent years, the increasing of communication demands requires more extended network and the standard quality of service is higher than ever. The customers require not only fast but also reliable connections. The word "reliable" has many meanings, but one of the most important meanings is survivable ability – having at least one back-up connection. It means that if the main connection is down, the network still works fine. Now, due to advantages of optic cable such as large capacity, small size, light weight, security…, it has become popular around the world. In this paper, we only focus optic cable. However, the same as coaxial cable, it also needs survivability to ensure reliability. So all the service providers need to solve the problem: how to connect to those complex clients with the lowest cost while still ensuring the survivable ability? This is one of the problems in SNDP.

In this work, we consider the problem of augmenting an existing network infrastructure by additional links (and switches) in order to connect potential customer nodes. There are two types of customers. In type-1, a standard, single link connection is sufficient, while type-2 customers require more reliable connections, ensuring connectivity even when a single link or routing node fails.

Before the SNDP problem can be formally stated, we need some definitions relating to connected constraints. Given infrastructure nodes represented for the

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 519–528, 2012.

existed network infrastructure and customer node is partitioned into subsets C1 and C2, (node sets C1, C2 $\subseteq$ C and C1 $\cap$ C2 = 0), the following conditions specify how customer nodes are to be connected:

- Simple connection: A customer node k from C_1 is feasibly connected if there exists a path from node k to infrastructure nodes.
- Redundant connection: A customer node k from C_2 is feasibly connected if there exists two nodes (and edges) disjoint paths from node k to infrastructure nodes.

Let G = (V, E) be a connected undirected graph with positive edge weight w (e) and set of customer nodes C $\subset$ V. The SNDP problem can be formulated as follows: among sub-graph of G whose satisfy connected conditions for each customer (|C1| |C2| > 0), find the sub-graph with the minimal cost (sum of the weights on edges of the sub-graph).

More precisely, we can formulate the problem as:

Find a sub- graph G' = (V', E') that minimizes:

$$\text{Cost}(G') = \sum_{e \in E'} w(e)$$

G' is sub-graph of G, connecting all customers C and satisfies all the connection requirements posed by the nodes C1 and C2.

This problem is known to be NP-hard for |C1| > 0 and |C2| > 0 [1].

In this paper, we propose new heuristic algorithms for solving SNDP: Random Node Selection – RNS which is known to often perform based on the idea of greedy algorithm. RNS builds a random Steiner tree, and then constructs k solution for SNDP. Finally, RNS selects the best solution. In order to illustrate the effectiveness of proposed algorithm, we experiment on real-world and random instances with three heuristic algorithms (GRASP- Greedy Randomized Adaptive Search Procedure [3]; MSSP- Multi Source Shortest Path [20]; OSSP – One Source Shortest Path [20]) which are the best in the known heuristic algorithms. The results, obtained from new algorithm, were better than previous works in [2] [3] [8] and [20].

The rest of this paper is organized as follows: In Section 2, we present related works. Section 3 describes the proposed algorithm for solving SNDP. The details of our experiments and the computational and comparative results are given in section 4. The paper concludes with section 5 with some discussions on the future extension of this work.

2 Related Work

A survey on methods for SNDP which can be seen as a more general version of our problem can be found in [1]. Techniques for solving the SNDP problem may be classified into categories: exact methods and heuristic methods.

Exact approaches for solving the SNDP problem are based on mixed linear integer programming. Wagner [2] modeled this problem as an integer linear program (ILP)

by means of an extended multi-commodity network flow (MCF) formulation. With the general purpose ILP-solver CPLEX [6], instances with up to 190 total nodes, 377 edges but only 6 customer nodes could be solved to prove optimality, and instances up to 2804 nodes, 3082 edges and 12 customer nodes could be solved with a final gap of about 7%. This approach is unsuitable for larger instances and/or in particular instances with larger number of customer nodes. In [7], this problem is approached with a different formulation based on directed connectivity constraints. By using a branch-and-cut algorithm, this model could be solved relatively well, and we were able to find proven optimal solutions for instances with up to 190 nodes, 377 edges, and 13 customer nodes. Liubíc et al [9] presented an exact method for the PCSTP (price-collecting Steiner tree problem) based on directed connection cuts. Other successful mathematical programming based approaches include a relax-and-cut by Cunha et al. [10] and a cutting plane method by Lucena et al. [11]. However, being deterministic and exhaustive in nature, these approaches could only be used to solve small problem instances (e.g sparse graphs with number of customers less than 15 customers)

To solve this problem with larger instances, heuristic methods are especially interested in it. In [8], Thomas Bucsics used meta-heuristic approaches such as local search and Simulated Annealing, Variable Neighborhood Descent and Variable Neighborhood Search. This approach has solved and obtained some significant improvements for some problem instances. In [3], Leitner formulated it as an abstract integer linear program and apply Lagrangian Decomposition to obtain relatively tight lower bounds as well as feasible solutions. Furthermore, hybrids of a Variable Neighborhood Search and a GRASP, approach are applicable to larger problem instances. Canuto et al. [12] described an effective multi-start local search approach based on perturbation of the nodes prizes, where path-relining and variable neighborhood search are used to further improve the obtained solutions. Experiments to reduce the number of nodes and edges that need to be considered in an instance of the PCSTP have been described by Uchoa [14]. And Chapovska et al. [13] discuss complexity of and solution methods for several variants of the PCSTP. In [17] Leitner has used Mixed Integer Programming and Hybrid Optimization Methods to solve this problem. Both approaches are able to derive proven optimal solutions or high quality solutions with small optimality gaps for mediums sized instances within reasonable time.

The most recently, we solved the SNDP in the design of last mile communication network [20] using heuristic algorithms and given better results than above works.

Other related problems are the various variants of the SNDP in [15] and see e.g. [1], [16] for relevant surveys.

In the next section, we introduce new heuristic algorithm to solve this problem.

3 Proposed Algorithms

As in the paper [20] which indicated specifically, there are two steps for solving SNDP:

- Construct a tree connected all customer nodes (include C_1 and C_2) to infrastructure nodes J (in infrastructure network).
- Augment some edges to satisfy redundancy connected condition for each type-2 customer node.

In this section, we present new heuristic algorithm based on Steiner tree construction to improve step 1 above and keep step 2 in the algorithm All-Pairs-Shortest-Path-adapting - APSPx of Thomas and Raidl in GRASP[8] for solving SNDP.
From the two our algorithms which are proposed in [20], we realize that the optimal Steiner tree does not always give us the best solution of SNDP. So, we propose a new random algorithm (call Random Node Selection - RNS) to construct Steiner tree as following. Constructing random node V' from V (V is edges set in graph G = (V, E)) and $|V'| = k*|V|$ ($0 < k \leq 1$). Then, we build the minimum path tree from each vertices of V' to terminal node. The rest of RNS is quiet similar to the above algorithms which can be seen detail at [20]. Finally, RNS selects the best solution from k solution of SNDP.

Sketch of the RNS algorithm is presented bellow.

```
1. Procedure RNS-Algorithm
2. Initialization: V' ⊆ V.
3. Handing: construct k solution for SNDP.
4. Gbest ← φ
5. Cost (Gbest) ← Infinity
6. For n:=1 to k do
7.      Gs ← φ
8.      For all v in V' do
9.          //Find shortest path from v to terminal nodes
10.             Tk ← number of terminal nodes
11.             Gs ← φ
12.             For i := 1 to Tk do
13.                     t • ith terminal node
14.                     For all edges e in
shortest_path(from v to t) do
15.                             Add e to Gs
16.                     End For
17.             End For
18.     End For
19.     Augmentation_Process
20.     Cost (Gs) ← Sum all cost of edges in Gs
21.     If (Cost (Gbest) > Cost (Gs)) then
22.             Gbest ← Gs
23.             Cost(Gbest) ← Cost (Gs)
24.     End if
25. Return Gbest
26. End Procedure
```

RNS used the same idea in step 2 (augmentation process) with the previous algorithms in [20]. The different points are approaches in the step 1. In MSSP and OSSP algorithm [20], we try to optimize cost of the Steiner tree which connectivity all customers to infrastructure nodes. In this paper, RNS algorithm constructs the random Steiner. It focuses the end process to improvement solution for SNDP. This is a new approach.

Computational results in section 4 will show the effectiveness of the proposed algorithms.

4 Computational Results

4.1 Problem Instances

The problem instances used in our experiments are the real-world instances and random instances that its detail is given in problem instances in [20].

We created two sets of experiments. In the first set of experiment, we compare the performance of the heuristic algorithms: GRASP [3], MSSP, OSSP [20] which is the best algorithm in the known heuristic algorithms, and RNS on the real world instances. In the second set of experiment, we reinstall GRASP and then compare the performance of our heuristic algorithms: GRASP, MSSP, OSSP and RNS on random instances.

4.2 System Setting

In the experiment, the system was run 20 times for each problem instance. All the programs were run on a machine with Intel Core 2 Duo T6400 2.0GHz, 4GB RAM, and were installed by C++ language.

4.3 Result of Computational Experiment

In figure 1, 2, 3, 4, 5, 6, 7, and figure 9, horizontal axis represents index of instances in data set and vertical axis represents cost of obtained network.

In figure 8, figure 10, horizontal axis represents index of instances in data set and vertical axis represents average running time (unit: second).

The experiment shows that:

- The results in figure 1, 2, 3, 4, 5, 6, 7 shows that on the real-world problem instances, the results of our new proposed algorithms (RNS) are better than GRASP in [3], and MSSP, OSSP in [20]. It means that the solution found by propose algorithm is the best solution in comparison with the other known heuristic algorithm for solving SNDP on all the real-world instances.
- Figure 1 shows that on ClgSExtra instances: the obtained solutions by our heuristic algorithm (RNS) are better than GRASP [3] about 76%, better than MSSP about 84 %, better than OSSP about 80%.

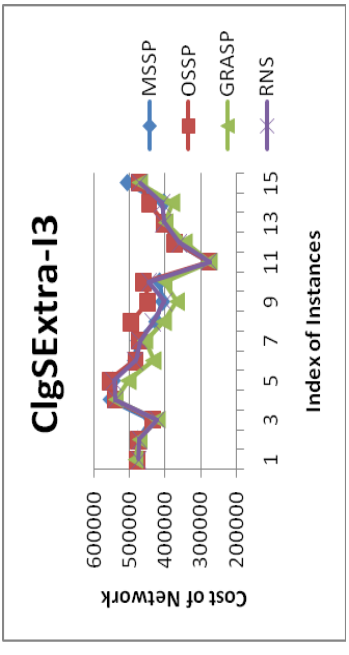


Fig. 3. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgSExtra-I3 instances

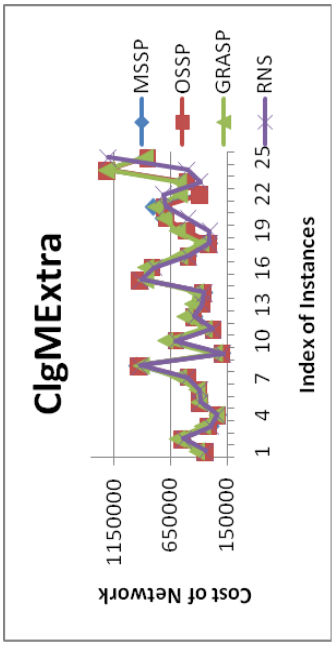


Fig. 4. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgMExtra instances

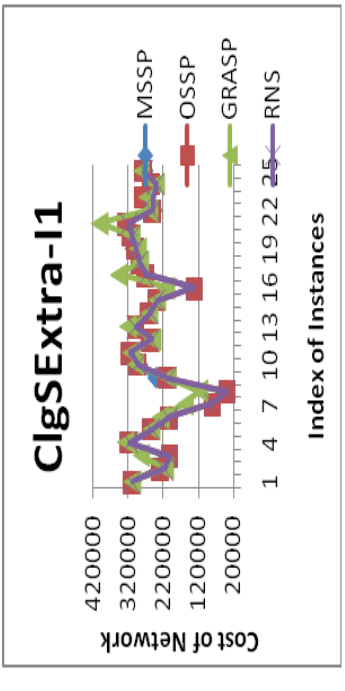


Fig. 1. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgSExtra-I1 instances

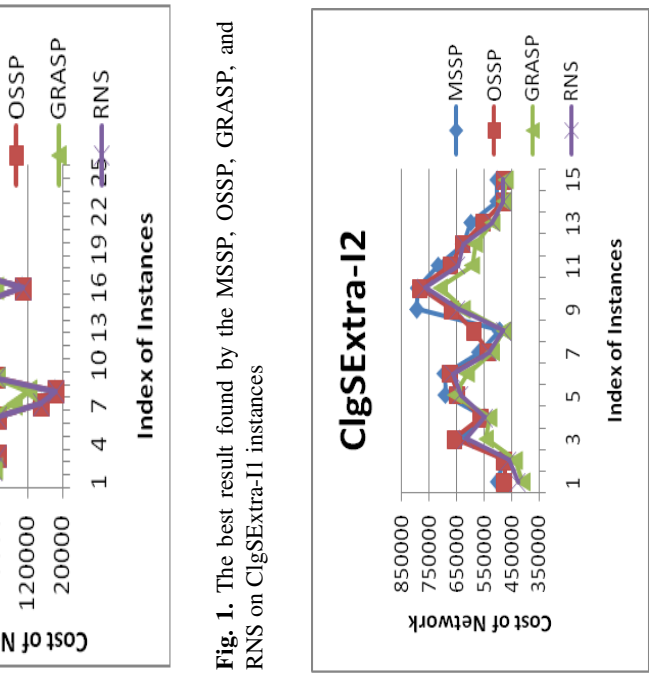


Fig. 2. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgSExtra-I1 instances

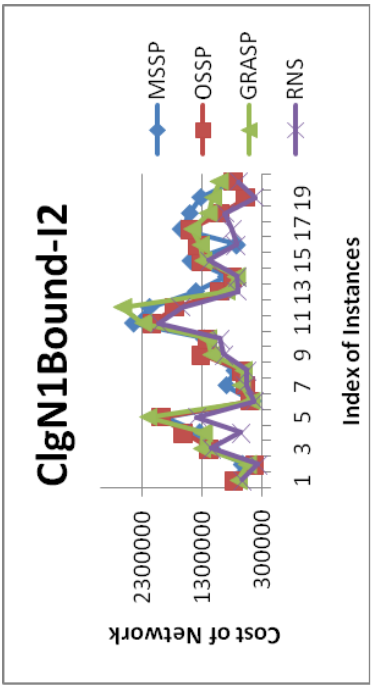


Fig. 7. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgNBound-I2 instances

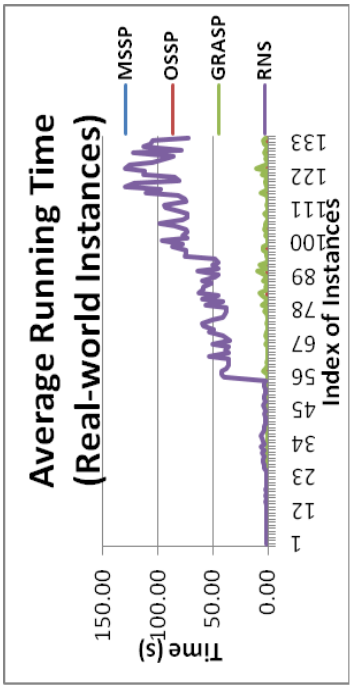


Fig. 8. The average running time of the MSSP, OSSP, GRASP and RNS on the real-world instances

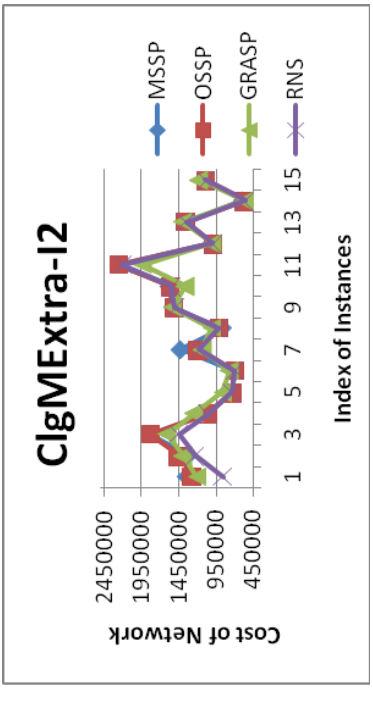


Fig. 5. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgMExtra-I2 instances

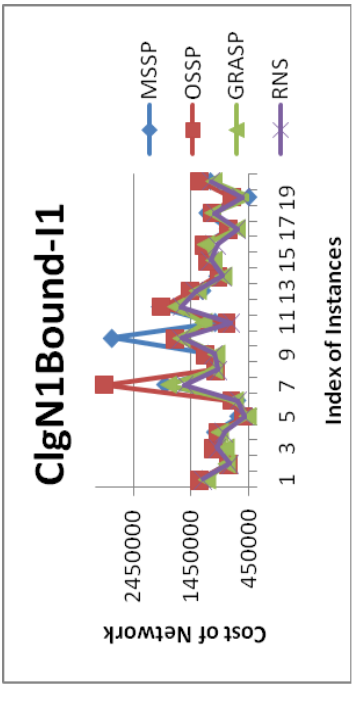


Fig. 6. The best result found by the MSSP, OSSP, GRASP, and RNS on ClgNBound-I1 instances

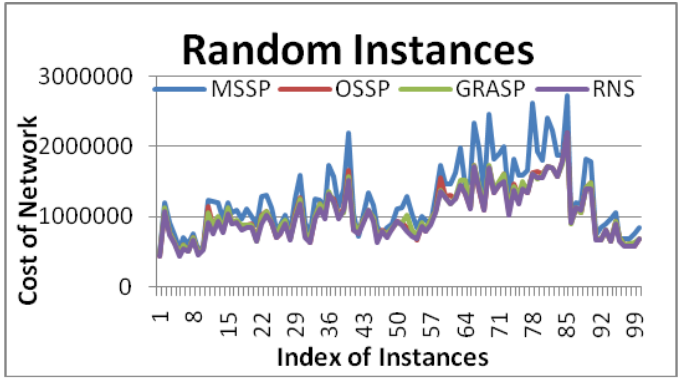


Fig. 9. The best result found by the MSSP, OSSP, GRASP, and RNS on Random instances

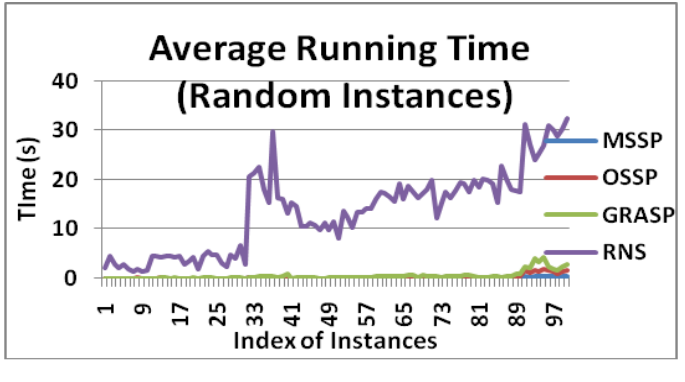


Fig. 10. The average running time of the MSSP, OSSP, GRASP and RNS on the real-world instances

Table 1. The rate of instances in RNS has shown better results of the attained Network than MSSP, OSSP, GRASP on real-world instances

Instances	#	n	m	MSSP	OSSP	GRASP
ClgSExtra	25	190	377	84%	80%	76%
ClgSExtra-I2	15	190	377	100%	100%	13%
ClgSExtra-I3	15	190	377	80%	100%	20%
ClgMExtra	25	1757	3877	68%	88%	84%
ClgMExtra-I2	15	1523	3290	93%	100%	80%
ClgN1BoundI1	20	2804	3082	85%	100%	55%
ClgN1BoundI2	20	2804	3082	95%	95%	100%

Table 2. The rate of instances in RNS has shown better results of the attained Network than MSSP, OSSP, GRASP on random instances

Instances	#	n	m	MSSP	OSSP	GRASP
SNDP.R1xx	30	190	500	100%	90%	96%
SNDP.R2xx	9	2000	5000	100%	88%	77%
SNDP.R3xx	10	1523	3290	80%	70%	70%
SNDP.R4xx	10	1757	3877	100%	90%	100%
SNDP.R5xx	20	2400	5000	100%	60%	90%
SNDP.R6xx	6	2800	3500	100%	16%	0%
SNDP.R7xx	5	3867	8477	100%	40%	60%
SNDP.R8xx	10	3867	46307	100%	60%	80%

- The same in Figure 2, 3, 4, 5, 6, 7 also shows that on random problem instances, the results found by propose algorithm are the best in four algorithms. Concrete results are detailed in Table 1.
- Figure 9 shows that on random problem instances, the results found by MSSP [20] are the worst result and found by RNS are the best.
- Figure 8 and figure 10 shows that, on the real-world instances and random instances, RNS needs using the longest average running time, but within acceptable time.
- Table 1 shows that the results found by RNS algorithm are better than MSSP, OSSP, and GRASP algorithm on real-world instances and the results found by RNS are superior to GRASP (100%).
- Table 2 shows that the results found by RNS are superior to MSSP, OSSP, and GRASP on random instances.

5 Conclusion

In this paper, we propose new heuristic algorithm for solving SNDP called RNS. With the opposite idea to MSSP and OSSP [20], RNS algorithm doesn't try to construct the best Steiner tree that focuses the selection process the best solution in k solutions of SNDP to improvement solution for SNDP. We experimented on 135 real-world instances derived from SNDLib network and 100 random instances. The results show that the proposed approach can be attractive for solving SNDP. On the real-world instances from a German city and random instances, the cost of obtained result from RNS algorithm are better than the other known heuristic algorithm for solving SNDP called GRASP, MSSP, and OSSP. In the future, we are planning to improve the algorithm for solving larger instances. Moreover, a direction for further research could be a study of the other survivable models such as multicast, any-cast models…

Acknowledgement. This work was partially supported by the project "Models for next generation of robust Internet" funded by the Ministry of Science and Technology, Vietnam and the project "Computational intelligence: Evolution, adaptation, and knowledge-based techniques" funded by the National Foundation for Science and Technology Development under grant number 102.01.14.09.

References

1. Kerivin, H., Mahjoub, A.R.: Design of survivable networks: A survey. Networks 46(1), 1–21 (2005)
2. Wagner, D., Raidl, G.R., Pferschy, U., Mutzel, P., Bachhiesl, P.: A multi-commodity flow approach for the design of the last mile in real-world fiber optic networks. In: Waldmann, K.H., Stocker, U.M. (eds.) Operations Research Proceedings 2006, pp. 197–202. Springer, Heidelberg (2007)

3. Leitner, M., Raidl, G.R.: Lagrangian Decomposition, Metaheuristics, and Hybrid Approaches for the Design of the Last Mile in Fiber Optic Networks. In: Blesa, M.J., Blum, C., Cotta, C., Fernández, A.J., Gallardo, J.E., Roli, A., Sampels, M. (eds.) HM 2008. LNCS, vol. 5296, pp. 158–174. Springer, Heidelberg (2008)
4. Leitner, M., Raidl, G.R.: Branch-and-cut and price for capacitated connected facility location. Technical Report TR 186{1{10{01, Vienna University of Technology, Vienna, Austria (2010)
5. Leitner, M., Raidl, G.R.: Strong lower bounds for a survivable network design problem. In: International Symposium on Combinatorial Optimization (ISCO 2010), Hammamet, Tunisia (March 2010)
6. ILOG: CPLEX 10.0 (2006), http://www.ilog.com
7. Wagner, D., Pferschy, U., Mutzel, P., Raidl, G.R., Bachhiesl, P.: A directed cut model for the design of the last mile in real-world fiber optic networks. In: Fortz, B. (ed.) Proceedings of the International Network Optimization Conference 2007, Spa, Belgium, pp. 1–6, 103 (2007)
8. Bucsics, T., Raidl, G.: Metaheuristic Approaches for Designing Survivable Fiber-Optic Networks. Institute for Computer Graphics and Algorithms of the Vienna University of Technology (2007)
9. Ljubic, I., Weiskircher, R., Pferschy, U., Klau, G., Mutzel, P., Fischetti, M.: An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. Mathematical Programming, Series B 105(2-3), 427–449 (2006)
10. da Cunha, A.S., Lucena, A., Maculan, N., Resende, M.G.C.: A relax-and-cut algorithm for the prize-collecting Steiner problem in graph. Discrete Applied Mathematics 157(6), 1198–1217 (2009)
11. Lucena, A., Resende, M.G.C.: Strong lower bounds for the prize collecting Steiner problem in graphs. Discrete Applied Mathematics 141(1-3), 277–294 (2004)
12. Canuto, S.A., Resende, M.G.C., Ribeiro, C.C.: Local search with perturbations for the prize-collecting Steiner tree problem in graphs. Networks 38, 50–58 (2001)
13. Chapovska, O., Punnen, A.P.: Variations of the prize-collecting Steiner tree problem. Networks 47(4), 199–205 (2006)
14. Uchoa, E.: Reduction tests for the prize-collecting Steiner problem. Operations Research Letters 34(4), 437–444 (2006)
15. Fortz, B., Labbe, M.: Polyhedral approaches to the design of survivable networks. In: Resende, M.G.C., Pardolas, P.M. (eds.) Handbook of Optimization in Telecommunications, pp. 367–389. Springer, Heidelberg (2006)
16. Stoer, M.: Design of Survivable Networks. LNCS, vol. 1531. Springer, Heidelberg (1992)
17. Leitner, M.: Solving Two Network Design Problems by Mixed Integer Programming and Hybrid Optimization Methods. PhD thesis, Vienna University of Technology, Institute of Computer Graphics and Algorithms, Vienna, Austria (May 2010); supervised by Raidl, G.R., Pferschy, U.
18. Bachhiesl, P.: The OPT- and the SST-problems for real world access network design basic definitions and test instances. Working Report 01/2005, Carinthia Tech Institue, Department of Telematics and Network Engineering, Klagenfurt, Austria (2005)
19. https://www.ads.tuwien.ac.at/people/mleitner/sndp/sndpinstances.tar.gz
20. Minh, N.T., Trung, V.K., Binh, H.T.T.: Heuristic Algorithms for Solving the Survivable Problem in the Design of Last Mile Communication Networks. In: The 9th IEEE – RIVF International Conference on Computing and Communication Technologies, Ho Chi Minh city, Vietnam, Feburary 27-March 01 (accepted, 2012)

HEp-2 Cell Images Classification Based on Textural and Statistic Features Using Self-Organizing Map

Yi-Chu Huang[1], Tsu-Yi Hsieh[2], Chin-Yuan Chang[1], Wei-Ta Cheng[1], Yu-Chih Lin[1], and Yu-Len Huang[1]

[1] Department of Computer Science, Tunghai University, Taichung, Taiwan
[2] Division of Allergy, Immunology and Rheumatology, Taichung Veterans General Hospital, Taichung, Taiwan
ylhuang@thu.edu.tw

Abstract. Indirect immunofluorescence (IIF) with HEp-2 cells has been used to detect antinuclear auto-antibodies (ANA) for diagnosing systemic autoimmune diseases. The aim of this study is to develop an automatic scheme to identify the fluorescence patterns of HEp-2 cell in IIF images. The self-organizing map (SOM) neural network with 14 textural and statistic features were utilized to classify the fluorescence patterns. This study evaluated 1020 autoantibody fluorescence patterns that were divided into six pattern categories, i.e. diffuse, peripheral, coarse speckled, fine speckled, discrete speckled and nucleolar patterns. Experimental results show that the proposed approach can identify autoantibody fluorescence patterns with a high accuracy and is therefore clinically useful to provide a second opinion for diagnosing systemic autoimmune diseases.

Keywords: Images classification, indirect immunofluorescence, self-organizing map, antinuclear auto-antibodies, autoimmune diseases.

1 Introduction

Autoimmune diseases have been confirmed to be in connection with the occurrence of disease specific autoantibodies, such as systemic autoimmune rheumatic diseases, primary biliary cirrhosis and dermatomyositis [1;2]. Using indirect immunofluorescence (IIF) to identify the antinuclear autoantibodies (ANA) patterns of HEp-2 cell during the serological hallmark, the physician could diagnose the autoimmune diseases. The fluorescence patterns are usually manually identified with physician visually inspecting the slides with the help of a microscope. This procedure still needs highly specialized and experienced physician to make diagnoses due to lacking in satisfied automation of inspection. As ANA testing becomes more widespread used, clinical application of functional automatic inspection system is in great demand.

Perner et al. [3] combined the decision tree model and texture features for Hep-2 cell image classification, and the error rate is 25%. This study changed the classifier, and used the texture and statistic features attempt to reduce the error rate. Soda et al.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 529–538, 2012.

[4] combined expert system and classification technique to completion the computer-aided diagnosis systems. They used wavelet to extract 360 features and select 180 effective features for identify the well cells. The large feature space can provide the better classification result, but the selection and train may increase the system operation time. This study used unsupervised neural network to improve the whole system performance. Kurdthongmee [5] using self-organizing map (SOM) neural network to classification the highly uniform color image of wood boards, and without human intervention the classification result accuracy is 95%. Hence this study hope to use the SOM to classify the homogenous cells, and accomplish automatically identified the six primary autoantibody fluorescence (ANA) patterns.

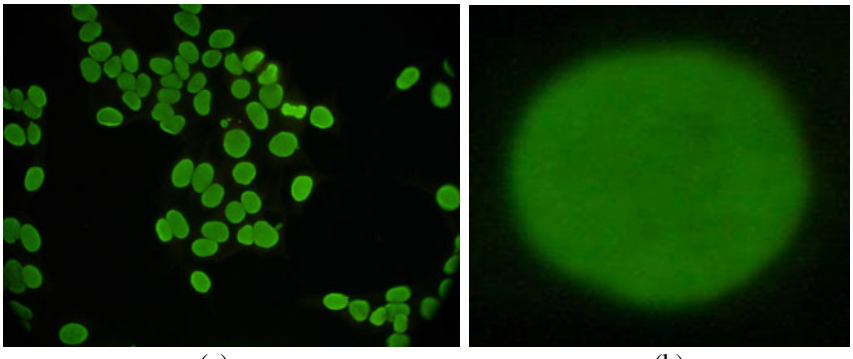
An unsupervised learning classification scheme, the SOM [6] classifier, with four textural features and ten statistics features was performed to identify fluorescence patterns. After image analysis, the extracted features included standard deviation, uniformity, image intensity, autocorrelation value and other cells area features. The proposed method evaluated 1020 cells image with six distinct fluorescence patterns. Computer simulations revealed that the accuracy of each pattern was over 80.0% and total average accuracy was 92.4%. The results show that the proposed SOM model with the textural and statistics features is clinically feasible for fluorescence pattern classification.

This study presented a scheme on classification of indirect immunofluorescence with HEp-2 cells using image analysis and artificial neural network techniques. The most popular method to detect autoimmune disease in ANA testing is IIF, and the diagnosis results depend on fluorescence patterns. In general, physician or technician manually inspects the slides of ANA with the help of a microscope to identify fluorescence patterns [7]. However, valid inspection need highly experienced technician or physician to handle due to a faulty identification of fluorescence patterns might make an incorrect diagnosis. For this reason, automatic identification and classification for fluorescence patterns in IIF image could offer physicians in making correct diagnosis without relevant experience.

2 Materials and Methods

2.1 Data Acquisition

This study used slides of HEp-2 substrate, at a serum dilution of 1:80. A physician takes images of slides with an acquisition unit consisting of the fluorescence microscope coupled with a commonly used fluorescence microscope (Axioskop 2, CarlZeiss, Jena, Germany) at 40-fold magnification. The immunofluorescence images were taken by an operator with a color digital camera (E-330, Olympus, Tokyo, Japan). The digitized images were of 8-bit photometric resolution for each RGB (Red, Green and Blue) color channel with a resolution of 3136 × 2352 pixel. Finally, the images were transferred to a personal computer and stored as *.orf-files (Raw data format) without compression. The samples were collected from January 2007 to May 2009 and test image database containing 1020 cells with manual segmentation form the samples. Figure 1 shows the result of manual segmentation for sample.



(a) (b)

Fig. 1. The example of the image acquisition: (a) original indirect immunofluorescence image (b) after the manual segmentation for samples

2.2 Autoantibody Fluorescence Patterns

After analyzing the patient‘s serum with autoimmune disease by indirect immunofluorescence, the results commonly appear six distinct patterns. The six primary ANA patterns include diffuse pattern, peripheral pattern, coarse speckled pattern, fine speckled pattern, discrete speckled pattern, and nucleolar pattern. In the diffuse pattern, the area of cell inside is always apparent homogeneous. In the peripheral pattern, the higher fluorescence intensity is apparent around the nucleus outer region. The nucleolar pattern shows that large fluorescence speckled staining within the nucleus. Moreover, speckled patterns include three types of patterns, i.e. coarse speckled pattern, fine speckled pattern, and discrete speckled pattern. In the coarse speckled and fine speckled pattern, their nucleuses apply homogeneous and some dark speckled within the nucleus. Discrete speckled pattern shows that small fluorescence speckled staining within the nucleus, and other area has the weaker fluorescence. Figure 2 illustrate the distinct autoantibody fluorescence patterns.

2.3 Preprocessing

Original IIF images were of 8-bit photometric with a resolution of 3136×2352 pixel, but the resolution of manual segmented samples were 256×256. The detail images were to magnify the image noises, speckles and tissue textures. For reduce the noises, and keep the texture. This study used Wiener method to enhance images. The Wiener method is a pixel-wise adaptive low-pass filtering method based on statistics estimated from the local neighborhood of each pixel. Due to the diffused and peripheral pattern appear smooth of the cells inside area, the tiny speckle would be reduced of inside area after Wiener method. Thus the proposed method used the image with preprocessing of Wiener to extract the texture feature can made more accurate.

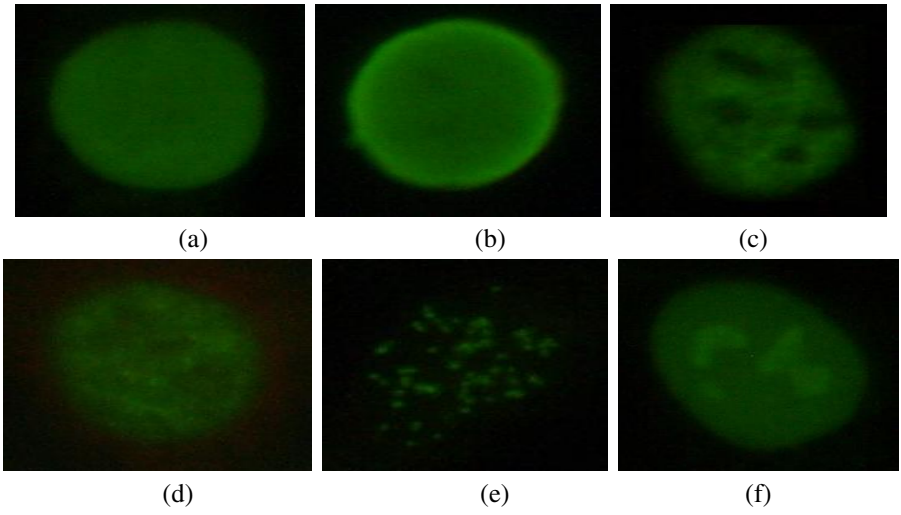


Fig. 2. The six main ANA patterns: (a) diffuse, (b) peripheral, (c) coarse speckled, (d) fine speckled, (e) discrete speckled, and (f) nucleolar patterns

2.4 The Proposed Classification Method

Figure 3 illustrates the overall processes of a proposed method of identified fluorescence patterns used the SOM model. The training set sample images were chose in database excluded test set. This approach can be classified into two main stages. The first stage was feature extraction and the second was identifying the test set sample stage. In the first stage, convert the images channel from RGB to HSI and gray level color spaces, and extract the total of 14 features. The second stage was used the 14 features to created the map of SOM and mapping with test set to evaluated the classification results.

2.5 Feature Extraction

The proposed approach utilized SOM model for classification based on the similarity of cells. Ten statistic features and four textural features from the extracted cells list as follows:

- *Area*: The *area* represents the area of totally cell. Automatic detect the cell area of using Otsu's method.
- *Perimeter*: The *perimeter* represents the boundary of a cell. Due to some features used statistics value of boundary, the area had to mark in a cell.
- A_I (Inside area): The A_I was the extract area except the perimeter of a cell. This feature computes the pixels number in the A_I range.
- A_P (Perimeter area): The A_P was the binary image of detected cell except A_I. The A_P value also computes the pixels number in the A_P range.

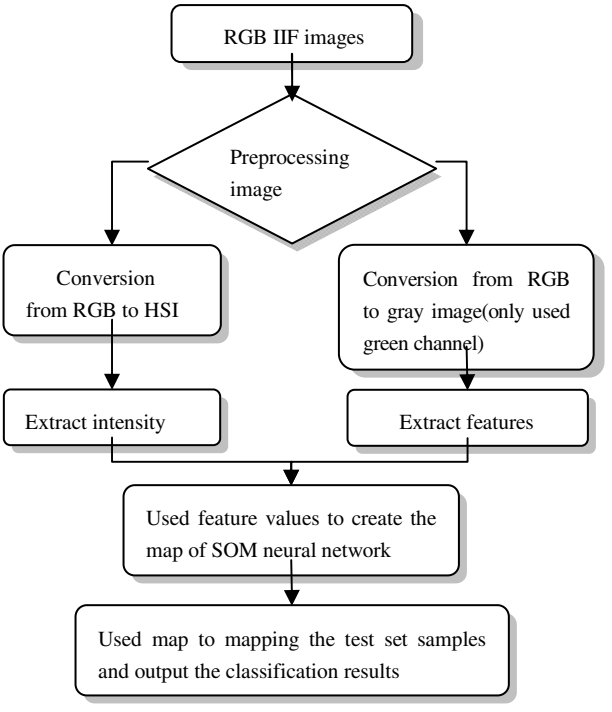


Fig. 3. The flowchart of the proposed method

- *Average intensity* (Average intensity of A_I and A_P): The original image was transformed from RGB to HSI color space, and then measured the average intensity of the A_I and A_P. If a cell had difference high intensity between A_I and A_P, these features can be used to detect them.
- *Higher intensity ratio* (Higher intensity ratio of A_I, A_P and *Area*): The ratio which the area of higher intensity compared with the whole cell, the rate based bright region of the A_I, A_P and *Area* generated as features. These features can detect higher intensity place of a cell, and using these features to classify the cells pattern were intuition.
- *Darker intensity ratio* (Darker intensity ratio of A_I, A_P and *Area*): The way of extract these features were same as *higher intensity ratio.* First step was inverse the *Area*, and used the result to find the darker region of A_I, A_P and *Area*. These features were sensitive with speckle, when the speckle occur the features value were increase.
- *Uniformity*: The uniformity is a useful texture measure based on the histogram of an image, and the smoother fluorescence cell would obtain the larger value. "Uniformity" given by equation (1), where Z is a random variable denoting gray levels and $p(Zi)$, $i = 0, 1, 2, \ldots, L\text{-}1$, is the corresponding histogram, L is the number of distinct gray levels. Because the p's have values in the range

[0, 1] and their sum equals 1, measure U is maximum for an image in which all gray levels are equal (maximum uniform), and decreases from there:

$$U = \sum_{i=0}^{L-1} P^2(\mathrm{Z}_i). \tag{1}$$

- *Standard deviation* A_I (Standard deviation of A_I): The standard deviation of A_I indicated that the distribution of histogram, the rougher inside area would also obtain the larger value.

The auto-coefficients used the 2-D normalized auto-coefficients to measure the features, and the parameter was set as 5 and 10 in this study. Figure 4 illustrates the area of features extraction with diffused pattern.

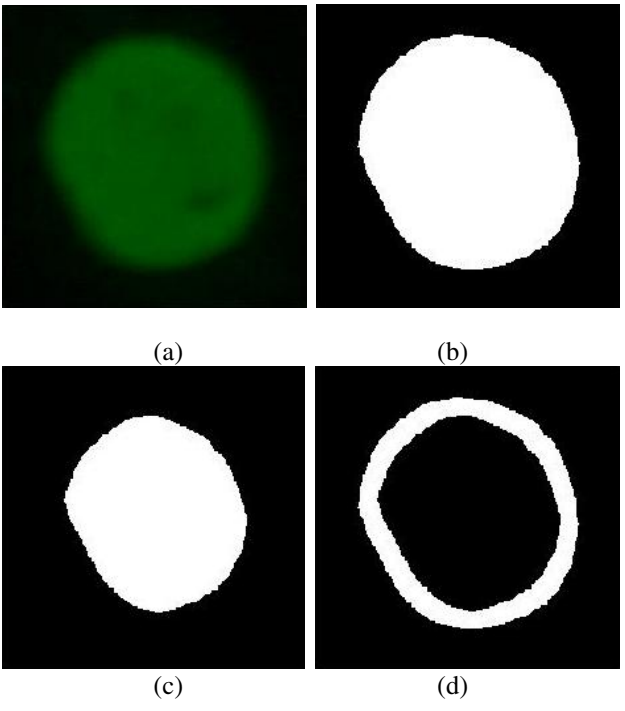


Fig. 4. The area of features extraction: (a) original image with diffused patterns, (b) binary image of detect the cell area, (c) inside features extract area, (d) perimeter features extract area

2.6 Self-Organizing Maps

Self-organizing maps (SOM) an unsupervised learning of artificial neural network technique, has been widely used to classify data in many clusters. The SOM have been applied extensively for classification [8;9], region identify [10;11], and image analysis [11;12]. The proposed method identified fluorescence patterns used the SOM model. In this study, the 14 statistic and textural features from the fluorescence cells were utilized as input of SOM classifier. The output of the SOM was a map address,

the address values corresponded with a fluorescence pattern, and the result was used to observe the classification is correct or not.

This study used 14 characteristic features to create the SOM map. In the training stage, a number of maps were created in order to determine the best fit map, and the best fit map depend on the quantization and topographic error, when the less quantization and topographic error can make map better. The first step was used the features to created SOM map, and the second step was labeling the map. When SOM map was created the test set can used the map to found the similar map address and identify the fluorescence pattern.

2.7 Classification Evaluations

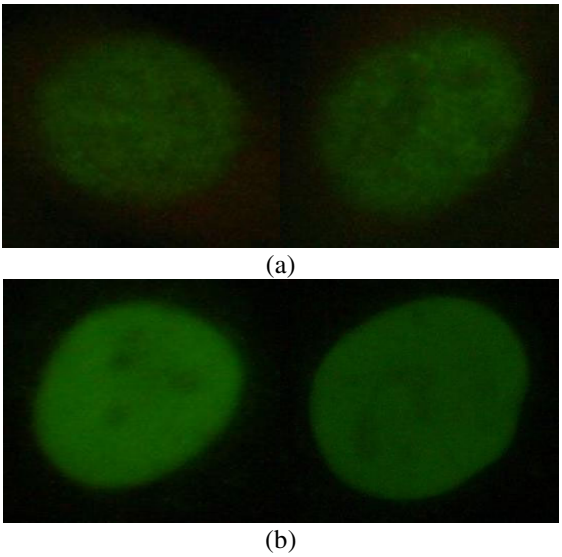
This study used *k*-fold cross-validation method [13;14] to evaluate the classification performance of the proposed SOM system. The *k*-fold method is widely used to evaluation stage. When the data amount is not enough, the *k*-fold method can solve the situation and construction the normal train and test evaluation. Every pattern was divided into *k* groups. The *k*-fold cross-validation method used the *k* groups to create the training set and testing set. The testing set has only one group, which was excluded from the training set. This process was repeated until all *k* groups were used in turn as the group used for testing.

3 Results

The autoantibody fluorescence pattern database contained 1020 cells, and assign equally to *k*-fold (k =10). In experimented stage 9 cases for SOM model training with map size 9×9, 1 case for testing, and run 10 times to find all autoantibody fluorescence cells classification result. These simulations were made on a single CPU Intel(R) Core(TM)2 2.13 GHz personal computer (ASUSTek Computer Inc., Taipei, Taiwan) with Microsoft Windows Vista® operating system. The system was performed by using MATLAB (R2008.a) software (The MathWorks, Inc., Natick, MA).

The ANA pattern database included 130 diffuse patterns, 200 peripheral patterns, 200 coarse speckled patterns, 130 fine speckled patterns, 200 discrete speckled patterns, and 160 nucleolar patterns. Table 1 lists the simulation results of SOM classification. The accuracy of SOM for classifying fluorescence patterns were diffuse pattern 86.1% (112/130), peripheral pattern 100% (200/200), coarse speckled pattern 95.5% (191/200), fine speckled pattern 80% (104/130), discrete speckled pattern 98% (196/200), and nucleolar pattern 87.5% (140/160), and the average accuracy was 92.4% (943/1020). Table 2 lists the classification results of autoantibody fluorescence pattern database, the best case was peripheral pattern, and the worse case was the fine speckled pattern. Figure 5 illustrates the incorrect classification samples of the fine speckled pattern and diffuse pattern, the fine speckled pattern classify to diffuse pattern and the diffuse pattern classify to fine speckled pattern.

This study used the statistic and textural features to classify fluorescence patterns, and texture of the fine speckled samples (as in Fig. 5(a)) was not regular in the normal samples. The diffuse pattern should not appearance speckle within the cells inside, but some samples (like Fig. 5(b)) might contain speckled regions, in the feature extraction step chooses textural features, the cells inside appear speckle or not would effect the feature values. The proposed SOM classifier for the same kind of pattern may produce excessive difference feature values and then obtained an incorrect classification.



(a) (b)

Fig. 5. The samples with incorrect classification: (a) fine speckled patterns and (b) diffuse patterns

Table 1. The simulation results of SOM classification

Fluorescence patterns	Simulation result		
	NT	***NC***	***Accuracy***
Diffuse	130	112	86.1%
Peripheral	200	200	100%
Coarse speckled	200	191	95.5%
Fine speckled	130	104	80.0%
Discrete speckled	200	196	98.0%
Nucleolar	160	140	87.5%
Total	1020	943	92.4%

NT: number of test cells
NC: number of correct classifications
Accuracy: NC / NT × 100%

Table 2. The classification results of fluorescence pattern database

Fluorescence patterns	Diffused	Peripheral	Coarse speckled	Fine speckled	Discrete speckled	Nucleolar
Diffuse	112	0	2	11	0	6
Peripheral	3	200	0	0	0	0
Coarse speckled	0	0	191	8	4	4
Fine speckled	12	0	5	104	0	9
Discrete speckled	0	0	0	0	196	1
Nucleolar	3	0	2	7	0	140
Total	130	200	200	130	200	160

4 Conclusion

This study proposed a classification scheme to identify IIF image fluorescence patterns. The database including 1020 fluorescence cells with manual segmentation for IIF image. The classification based on SOM classifier with 14 features, and used the property of SOM for cluster the feature data. In the six different patterns, the correct rate were 86.1% for diffuse, 100% peripheral for, 95.5% for coarse speckled, 80% for fine speckled, 98% for discrete speckled, 87.5% for nucleolar, and the 92.4% for average accuracy. The property of SOM classifier was cluster the data and used the result to identify test samples. When feature values did not had variance the result of classification was mistake. The large input feature values were to make the train time increase, and the bad feature values may disturb classification result. So the input features had to select carefully. In this study, the 14 features of training SOM were selected. This study used 1020 simples and the map size was 9×9, observed the simulation result the SOM map size was depend on simple number. The size of 9×9 map had 81 cells and the training simples was 918, so every cells always had 10~12 simples, and the used the contest way to decide the pattern of this cell. The SOM classifier rule is Winner-Takes-All, which patterns a sample to the class whose output sample has the biggest value. Hence the map cells can not too many or few. The size of 9×9 map can make better classification results, when the database had 1020 simples. After training, every cell was defining the unique pattern, and test set used the map to find the pattern which represent itself.

From the experimental results, the proposed method for IIF cells image classification using SOM artificial neural network with ten statistic features and four textural features is feasible. As this automated classification system is useful for identifying autoantibody fluorescence patterns and decrease the time of physician diagnosis. Future work should combine the proposed fluorescence pattern classification method with automatic IIF image cells segmentation system, and apply them to computer-aided diagnosis (CAD) system for systemic autoimmune diseases.

Acknowledgments. The authors would like to thank the Taichung Veterans General Hospital (Taiwan) and Tunghai University for financially supporting this research under Contract No. TCVGH-T997802. This research was supported in part by National Science Council of the Republic of China (Taiwan) under Contract No. NSC98-2221-E-029-026.

References

1. Fritzler, M.J.: Challenges to the use of autoantibodies as predictors of disease onset, diagnosis and outcomes. Autoimmunity Reviews 7(8), 616–620 (2008)
2. Worman, H.J., Courvalin, J.C.: Antinuclear antibodies specific for primary biliary cirrhosis. Autoimmunity Reviews 2(4), 211–217 (2003)
3. Perner, P., Perner, H., Muller, B.: Mining knowledge for HEp-2 cell image classification. Artificial Intelligence in Medicine 26(1-2), 161–173 (2002)
4. Soda, P., Iannello, G.: Aggregation of Classifiers for Staining Pattern Recognition in Antinuclear Autoantibodies Analysis. IEEE Transactions on Information Technology in Biomedicine 13(3), 322–329 (2009)
5. Kurdthongmee, W.: Colour classification of rubberwood boards for fingerjoint manufacturing using a SOM neural network and image processing. Computers and Electronics in Agriculture 64(2), 85–92 (2008)
6. Kita, E., Kan, S., Fei, Z.: Investigation of self-organizing map for genetic algorithm. Advances in Engineering Software 41(2), 148–153 (2010)
7. Conrad, K., Schoessler, W., Hiepe, F.: Autoantibodies in systemic autoimmune diseases. Pabst Science Publishers (2002)
8. Varez-Guerra, M., Gonzalez-Pinuela, C., Andres, A., Galan, B., Viguri, J.R.: Assessment of Self-Organizing Map artificial neural networks for the classification of sediment quality. Environment International 34(6), 782–790 (2008)
9. Mohamed, M.: Clustering the ecological footprint of nations using Kohonen's self-organizing maps. Expert Systems with Applications 37(4), 2747–2755 (2010)
10. Lin, G.F., Chen, L.H.: Identification of homogeneous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology 324(1-4), 1–9 (2006)
11. Vijayakumar, C., Damayanti, G., Pant, R., Sreedhar, C.M.: Segmentation and grading of brain tumors on apparent diffusion coefficient images using self-organizing maps. Computerized Medical Imaging and Graphics 31(7), 473–484 (2007)
12. Guler, I., Demirhan, A., Karakis, R.: Interpretation of MR images using self-organizing maps and knowledge-based expert systems. Digital Signal Processing 19(4), 668–677 (2009)
13. Weiss, S.M., Kapouleas, I.: An empirical comparison of pattern recognition neural nets and machine learning classification methods. In: Proc. 11th Int. Joint Conf. Artificial Intelligence, pp. 234–237 (1989)
14. Wiens, T.S., Dale, B.C., Boyce, M.S., Kershaw, G.P.: Three way k-fold cross-validation of resource selection functions. Ecological Modelling 212(3-4), 244–255 (2008)

Identifying Smuggling Vessels with Artificial Neural Network and Logistics Regression in Criminal Intelligence Using Vessels Smuggling Case Data

Chih-Hao Wen[1,2], Ping-Yu Hsu[1], Chung-yung Wang[2], and Tai-Long Wu[2]

[1] Department of Business Administration, National Central University,
300 Jhongda Road, Jhongli, Taiwan, R.O.C.
[2] Department of Logistics Management, National Defense University,
70 Sec. 2 Chung-Yang North Road, Taipei, Taiwan, R.O.C.
{Chih-Hao Wen,chwen}@mgt.ncu.edu.tw

Abstract. In spite of the gradual increase of the academic studies on smuggling crime, they seldom focus on the subject of applying data mining to crime prevention. Artificial Neural Networks and Logistic Regression are used to conduct classification and prediction. This study establishes models for vessels of different tonnage and operation purpose, which can provide the enforcers with clearer judgment criteria. The study results show that the application of Artificial Neural Networks to smuggling fishing vessel can get the average precision as high as 76.49%, the application of Logistic Regression to smuggling fishing vessel can get the average precision as high as 61.58%, both of which are of significantly higher efficiency compared with human inspection. The information technology can greatly help to increase the probabilities of seizing smuggling vessels, what's more, it can make better use of the data in the database to increase the probabilities of seizing smuggling crimes.

Keywords: Artificial intelligent, Crime data mining, Smuggling predicts, Artificial neural networks, Logistics regression, Hybrid model.

1 Introduction

As Taiwan is an island nation, the oceanic area is about three times of the land area. Under this circumstance, the oceanic transportation is so prosperous that the vessel transportation becomes the most commonly-used way for the smugglers to conduct various smugglings. In 2009, the seized smuggling of agricultural and fishery products reached 90.2 tons in total [5]. Furthermore, according to the official data of Taiwan's Ministry of Finance, among the seized smuggling items over the past 20 years, the annual sum reaches over $24 million on average [6], but it is estimated that the value of actual smuggling is between $160.8 million and $184.6 million [1, 6].

Coastal Patrol Directorate General (CPDG) of Coast Guard Administration (CGA) in Taiwan is in charge of all maritime affairs related to smuggling prevention, costal

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 539–548, 2012.

line patrol and inspection of vessels in ports. The inspection of vessels on ports is in the judiciary of compulsory officers and soldiers. Among all the vessels being investigated, only 5.03% are found to be guilty. Thus, CGA needs highly effective information tools to spot smuggling vessels. This study shows that both methods can effectively distinguish smuggling fish boats from normal fishing vessels by learning from the records kept in Coast Guard Information System (CGIS), which is implemented by CGA in 2002. The system has recorded all detail information of the vessels leaving and returning the ports, including whether it is for smuggling, and the item and content of the smuggling.

In this research, the average precision of ANN model is as high as 76.49% (63.2-92.0%), while that of LR model is 61.58% (42.2-77.7%). Comparing to the meager 7.7% precision of human patrolling method, the proposed methods can greatly improve the efficiency of CGA.

2 The Application of Information Technology to Crimes

The ability to predict criminal incidents is vital for all types of law enforcement agencies [22]. In the past, the crime prediction was believed to infeasible [23]. In the practice, it needs to work out the small location range to achieve the good performance for the police patrolling. Currently, the commonly-used method is the statistical sampling analysis, but there are few discussions on the prediction precision of the model. Thus, the model precision has some problems when applied in small location or range. Some are the problems related to the prediction scale in the statistics [24, 25], while it is probably because the data quantity accumulated in the system is too large, so it can't be analyzed and used in real-time. When getting giant data that can't be processed effectively, it may cause the situation of data dump, but data mining can solve this problem [26], especially the crime data mining for the criminal behaviors.

Crime data mining uses traditional data mining techniques such as association analysis, classification and prediction, cluster analysis and outlier analysis to identify the patterns [27]. The pattern can be used to predict the crime trend, and make definition through the related attributed to provide suggestions for decision-making and police deployment.

This study is the first one to apply both methods against marital smuggling. The attributes recorded in CGIS is different from the systems reported by other research. Nonetheless, the research shows that both method are still very effective, comparing to approach replying on human primarily.

3 Methods

3.1 Samples

Technology can help law enforcement agencies identify a gun or a person, arrest someone for a crime, listen in on incriminating conversations, or prosecute and convict persons charged with crimes [28]. Therefore, CGIS that combines with technology can

also help the coastal patrolling agencies to identify whether the fishing vessel is for smuggling.

225 fishing ports in Taiwan all have inspection offices. Before leaving port, all fishing vessels need to register in the inspection offices to record their out-port data. If the owner of fishing vessel has some records of smuggling crime, the inspection human will get aboard for inspection. For other vessels, the on-site inspection human will determine the vessels and numbers for sampling checking according to the previous personal experience and objective observation. If some illegal persons or objects are found during the inspection, it will follow the related regulations to deal with. Other vessels, after getting the port clearance, can drive the vessel out of the port. When the vessels return to the port, they also have to follow the same procedure to inspect and record the inspection course and results. All these data will be uploaded and saved on the hosting server of the coastal patrolling agency. We will use these data to analyze and predict the smuggling vessels, as well as establish the judgment model.

In order to work out the smuggling behavior of the fishing vessel when leaving and returning the port, we use CGA's "CGIS" data. This study has officially submitted application to the administrative agencies, so we directly obtain the original data files stored in the national large hosting server, and can use these data through an information security management window. The data to be used covers all smuggling records of eight years from 2002 to 2009, which has recorded more than 150,000 cases. Moreover, the data about the vessels leaving and returning the ports covers the years of 2008 and 2009, with nearly 3 million record items. The fishing ports include 225 ports in Taiwan region.

Table 1. Vessel type and classification level

Vessel Type	Vessel tonnage or application
CT0	Less than 5 tons
CT1	Greater than 5 tons but less than 10 tons
CT2	Greater than 10 tons but less than 20 tons
CT3	Greater than 20 tons but less than 50 tons
CT4	Greater than 50 tons but less than 100 tons
CT5	Greater than 100 tons but less than 200 tons
CT6	Greater than 200 tons but less than 500 tons
CTR	Raft with powered
CTS	Sampan with powered
PENGHU	The ship use only for recreation in Penghu County
YACHT	Yacht

The study data includes the vessel's basic information, data of leaving and returning port and smuggling record data. The basic information of vessel includes the vessel register data and the personal information of the vessel owner. The data of leaving and returning port includes the vessel name, the information of security inspection unit and the attributes describing the vessel state at that time. The smuggling record data

includes the enforcement unit, vessel number and various attributes describing the smuggling. This study, under the suggestions of coastal patrolling officers, chooses the following 16 attributes as the prediction variables.

This study establishes the model based on different vessel types, which can be classified into 11 types (see Table 1). The yacht type has no smuggling records, so it is removed in advance.

3.2 Logistics Regression (LR)

Logistics Regression (LR) model is a generalized linear model that is used for binomial regression in which the predictor variables can be either numerical or categorical [29, 30]. It is principally used to solve problems caused by automobile insurance and corporate fraud[31], but rarely in the smuggling field.

LR model is one of the commonly-used prediction or classification methods [32]. The greatest difference from the traditional regression is that it will find out the linear relations among the variables after converting the binary variables. LR model has been widely applied in the fields of financial risk [33], fraud detection [31, 34, 35], insurance [36-38] and credit ranking [32, 39], medical and epidemiology [40, 41]. This study however applies LR model to the prediction and application of fishing vessel smuggling. The smuggling is judged by a kind of binary identification string in this study, namely, smuggling or non-smuggling. Thus, LR model is suitable to be applied in the statistic prediction.

3.3 Artificial Neural Network (ANN)

ANN are biologically inspired computational methods, which can be used to capture complex and non-linear relationships between data [42]. The ANN basically consist of several non-linear processing units, called neurons or nodes and they are connected in a massive parallel architecture [43]. ANN simulates the human brain with the intent to collect the empirical evidence during the learning process, and inter-neural connections (synapses) are used to store the knowledge. An important feature of ANN, in addition to the ability of learning, is the ability to generalize the learned knowledge [44]. The neurons are interconnected with connection links which have weights which are multiplied by the signal transmitted in the network [45]. The fundamental architecture of ANN is composed of input layer, hidden layer and output layer. The data is input through input layer, and the weights are repeatedly adjusted by the conversion function in hidden layer. After getting the best learning rate, the prediction results will be output by output layer. Learning occurs in the hidden layer where input data are summed and weighted with statistical functions to generate a predicted value that is then passed on to the output layer [46].

3.4 Performance Measures

The quality of classification is usually measured with classification accuracy given as a proportion of correctly classified objects. In this study, as in Bhattacharyya, et al.,[39],

Sensitivity (or recall), Specificity and Precision, are used to measure the performance. Sensitivity is the positive precision displayed in the inspection results of the smuggling fishing vessel, which is calculated by TP / (TP + FN). Specificity is the precision of the model evaluating the non-smuggling cases, or the negative precision displayed in the inspection results of the non-smuggling fishing vessel, which is calculated by TN / (FP + TN). Precision is the precision of this model to the smuggling detection, which is calculated by TP / (TP + FP).

In the computation process, we specially pay attention to the value of precision, which refers to the accuracy rate of the model in this study to predict whether the vessel is for smuggling. In other words, it means the accuracy of this model detecting a vessel may be for smuggling. The possibility of misjudging a real smuggling vessel is 1 – precision, which is the same as the type I error in the statistics. When a smuggling vessel is misjudged as non-smuggling, it may pay huge public security cost. However, if non-smuggling is misjudged as smuggling, the enforcement agencies just need to put in manpower costs, which is comparatively smaller.

4 Finding

This section presents the experimental results. Several types of smuggling detection models are devised. Training data are fed into ANN and LR methods to create related models. 70 % of 150,000 logs kept in CGIS is used as training data and the remaining 30% of data is reserved as testing data compute the precision and recalls of the models.

For ANN, we used Multilayer perceptions (MLPs). Multilayer perceptions (MLPs) are the most commonly used neural network [46, 47]. MLPs can use two or more hidden layers to make conversion and learning, mapping data in the way of non-linear conversion through S conversion function. Generally, the neural networks have performed as well or better than other methods in prediction accuracy on validation samples [15, 44]. The input layer neurons were 53, hidden layer-1 were 39, hidden layer-2 were 19, and the output layer was one neuron. For LR, we used Backwards Stepwise method. At each step, terms in the model are evaluated, and any terms that can be removed without significantly detracting from the model are removed [48]. Qualitative response models are appropriate when dependent variable is categorical [49]. In this study, our dependent variable “smuggling” is binary, and logistic regression is a widely used technique in such problems [39].

Fig.1. shows the compare performance of ANN and LR model, which includes accuracy, sensitivity (recall), specificity and precision. The accuracy of ANN model is better than LR model, but the difference is insignificant, in which the minimum value 0.833 is the CT4 model of LR. For the part of Sensitivity, the CTR value of ANN model is the lowest 0.052, but this value is still eight times better than that of human inspection.

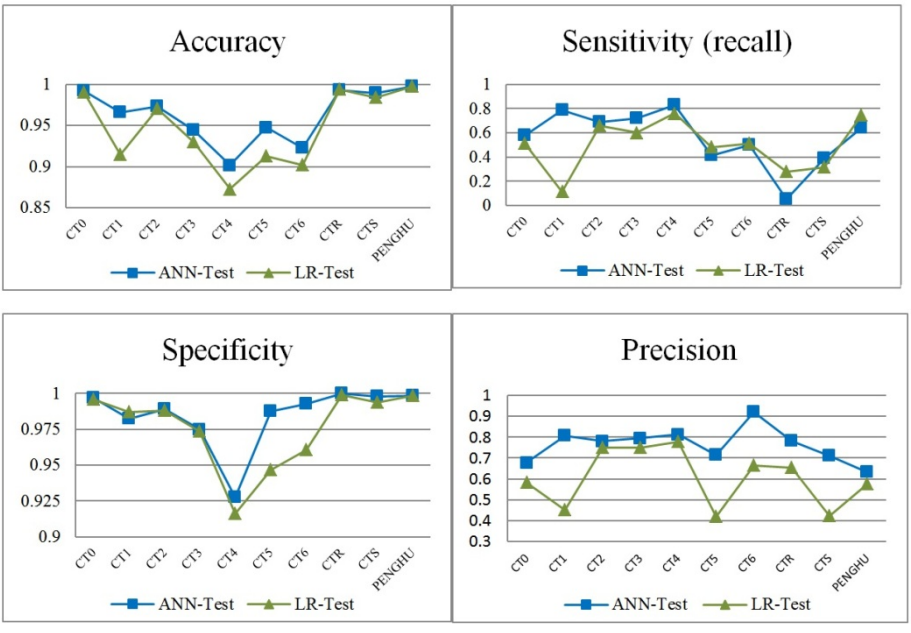


Fig. 1. Performance across different smuggling rates in testing data

Table 2. Performance of precision across SI, LR and ANN model in testing data

Vessel	Human Inspection (HI)	LR	ANN
CT0	0.011	0.583	0.677
CT1	0.085	0.450	0.806
CT2	0.053	0.751	0.780
CT3	0.119	0.751	0.795
CT4	0.274	0.777	0.812
CT5	0.070	0.417	0.713
CT6	0.132	0.664	0.920
CTR	0.007	0.655	0.781
CTS	0.014	0.423	0.712
PENGHU	0.004	0.576	0.632
AVERAGE	0.077	0.605	0.763

The human inspection method demonstrates some marginal better performance in CT3, CT4 and CT6. However, on the whole, the pure personnel dispatching approach can take much manpower resource with very weak performance. However, in ANN model, the poorest performance shows in the type of PENGHU, with only 0.632, but this value can still improve that of the current human inspection, namely, 0.004, which can provide significant improvement.

5 Discussion

The purpose of using information technology is to reduce the on-duty loading and increase the identification rate of the smuggling fishing vessels, finally increase the seizing performance through the strength of information technology. The model describes the patterns and shapes of the previous smuggling fishing vessel in the way of rule set, and includes the possibilities of judging the smuggling vessel, so as to provide clear reference foundation and reduce the vague identification space for the enforcement staff.

Table 2 shows the performances of seizing smuggling fishing vessel by using ANN model, LR model and human inspection. Comparing ANN model with human inspection, it can be found the best precision of ANN is 0.92 in CT6, which improves the performance for same-type HI vessel by 7 times, and the poorest precision of ANN is 0.632 in PENGHU, which also improves the performance for same-type HI vessel by 158 times. Comparing LR model with human inspection, it can be found the best precision of LR is 0.777 in CT4, which improves the performance for same-type HI vessel by 2.8 times, and the poorest sensitivity of LR is 0.417 in CT5, which also improves the performance for same-type HI vessel by 5.9 times.

By using the ANN model, we can get the average identification precision of 0.763, which is higher than that of human inspection 0.077. Besides obtaining the best identification performance, it can also greatly reduce the needed manpower. Taking the seizing performance of human inspection for 10,000 vessels as the basis, if using hybrid model, it just needs to check 953 vessels to get the same figure of smuggling fishing vessels, which can save about 90.47% of the manpower loading (see Fig. 2).

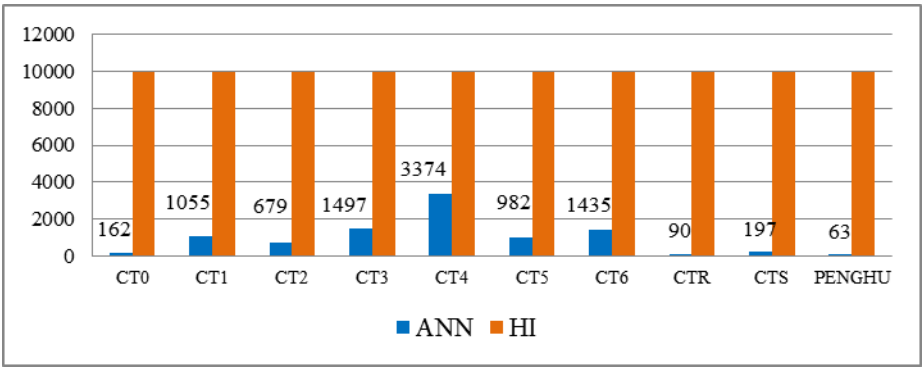


Fig. 2. Compare the performance with sample sizes in testing data (base on human inspection)

6 Conclusions

This study uses ANN model of information technology and LR model of statistic technology to analyze the data of smuggling fishing vessels in Taiwan region, and also establish the identification model. On the whole, the ANN model of information

technology is much better than that of LR model of statistic technology. Their performance is significantly better than that of using human inspection. ANN simulates neurons of the human neural and gets good performance after learning and adjusting over and over again. LR uses the statistic regression method to present the linear relations among the attributes after binary conversion, just needs a shorter time to establish model, but with poorer precision than ANN.

The findings of this research should lead to case-based system which will allow CGA to develop automatic identifying mechanism. In general, they will be useful to improve efficiency of smuggling boats and particularly to identifying more automatically. With the tool in hand, even CGA face the organization change, it still has enough manpower to effectively patrol the ports.

References

1. Farzanegan, M.R.: Illegal trade in the Iranian economy: Evidence from a structural model. European Journal of Political Economy 25, 489–507 (2009)
2. Tseng, T.-Y., Shiue, Y.-R., Ning, K.-C., Lin, S.-W., Cheng, W.-M.: Using new attribute construction to incorporate the expertise of human experts into a smuggling vessels classification system. Expert Systems with Applications 36, 7773–7777 (2009)
3. Niewiarowski, S., Gogbashian, A., Afaq, A., Kantor, R., Win, Z.: Abdominal X-ray signs of intra-intestinal drug smuggling. Journal of Forensic and Legal Medicine 17, 198–202 (2010)
4. Klar, A., Linker, R.: Feasibility study of automated detection of tunnel excavation by Brillouin optical time domain reflectometry. Tunnelling and Underground Space Technology 25, 575–586 (2010)
5. Coast Guard Administration: Coastal Patrol Annual Statistical Report. Coast Guard Administration Executive Yuan R.O.C., Taipei (2010)
6. Ministry of Finance: Yearbook of financial statistics of the Republic of China. Ministry of Finance R.O.C., Taipei (2010)
7. Ministry of National Defense: National Defense Report. Ministry of National Defense R.O.C (2009)
8. Kirby, B.C.: Parole prediction using multiple correlation. American Journal of Sociology 59, 539–550 (1954)
9. Glaser, D.: Prediction Tables as Accounting Devices for Judges and Parole Boards. Crime & Delinquency 8, 239–258 (1962)
10. Wilkins, L.T., Macnaughton-Smith, P.: New Prediction and Classification Methods in Criminology. Journal of Research in Crime and Delinquency 1, 19–32 (1964)
11. van Alstyne, D.J., Gottfredson, M.R.: A Multidimensional Contingency Table Analysis of Parole Outcome New Methods and Old Problems in Criminological Prediction. Journal of Research in Crime and Delinquency 15, 172–193 (1978)
12. Simon, F.H.: Statistical Methods of Making Prediction Instruments. Journal of Research in Crime and Delinquency 9, 46–53 (1972)
13. Gottfredson, S.D., Gottfredson, D.M.: Screening for Risk A Comparison of Methods. Criminal Justice and Behavior 7, 315–330 (1980)
14. Wilbanks, W.L.: Predicting failure on parole. In: Farrington, D.P., Tarling, R. (eds.) Prediction in criminology, State University of New York Press, NY (1985)

15. Caulkins, J., Cohen, J., Gorr, W., Wei, J.: Predicting criminal recidivism: A comparison of neural network models with statistical methods. Journal of Criminal Justice 24, 227–240 (1996)
16. Urbaniok, F., Endrass, J., Rossegger, A., Noll, T., Gallo, W.T., Angst, J.: The prediction of criminal recidivism: The implication of sampling in prognostic models. European Archives of Psychiatry and Clinical Neuroscience 257, 129–134 (2007)
17. Poortinga, E., Lemmen, C., Majeske, K.: A comparison of criminal sexual conduct defendants based on victim age. Journal of Forensic Sciences 52, 1372–1375 (2007)
18. Andresen, M.A., Jenion, G.W.: The unspecified temporal criminal event: What is unknown is known with aoristic analysis and multinomial logistic regression. Western Criminology Review 5, 1–11 (2004)
19. Keyvanpour, M.R., Javideh, M., Ebrahimi, M.R.: Detecting and investigating crime by means of data mining: A general crime matching framework, pp. 872–880 (Year)
20. Wang, P., Mathieu, R., Ke, J., Cai, H.J.: Predicting criminal recidivism with support vector machine (Year)
21. Le Khac, N.A., Kechadi, M.T.: Application of data mining for anti-money laundering detection: A case study, pp. 577–584 (Year)
22. Brown, D.E., Gunderson, L.F.: Using clustering to discover the preferences of computer criminals. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 31, 311–318 (2001)
23. Gorr, W., Harries, R.: Introduction to crime forecasting. International Journal of Forecasting 19, 551–555 (2003)
24. Bunn, D.W., Vassilopoulos, A.I.: Comparison of seasonal estimation methods in multi-item short-term forecasting. International Journal of Forecasting 15, 431–443 (1999)
25. Duncan, G., Gorr, W., Szczypula, J.: Forecasting analogous time series. Kluwer Academic Publishing, Boston (2001)
26. Keim, D.A., Panse, C., Sips, M., North, S.C.: Pixel based visual data mining of geo-spatial data. Computers & Graphics 28, 327–344 (2004)
27. Appavu, S., Rajaram, R., Muthupandian, M., Athiappan, G., Kashmeera, K.S.: Data mining based intelligent analysis of threatening e-mail. Knowledge-Based Systems 22, 392–393 (2009)
28. Nunn, S.: Measuring criminal justice technology outputs: The case of Title III wiretap productivity, 1987-2005. Journal of Criminal Justice 36, 344–353 (2008)
29. Yeh, I.C., Lien, C.-H.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36, 2473–2480 (2009)
30. Spathis, C.T.: Detecting false financial statements using published data: some evidence from Greece. Managerial Auditing Journal 17, 179–191 (2002)
31. Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems 50, 559–569 (2011)
32. Sinha, A.P., Zhao, H.: Incorporating domain knowledge into data mining classifiers: An application in indirect lending. Decision Support Systems 46, 287–299 (2008)
33. Peng, Y., Wang, G., Kou, G., Shi, Y.: An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing 11, 2906–2915 (2011)
34. Ravisankar, P., Ravi, V., Raghava Rao, G., Bose, I.: Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems 50, 491–500 (2011)

35. Jin, Y., Rejesus, R.M., Little, B.B.: Binary choice models for rare events data: A crop insurance fraud application. Applied Economics 37, 841–848 (2005)
36. Artís, M., Ayuso, M., Guillén, M.: Detection of automobile insurance fraud with discrete choice models and misclassified claims. Journal of Risk and Insurance 69, 325–340 (2002)
37. Saliba, B., Ventelou, B.: Complementary health insurance in France Who pays? Why? Who will suffer from public disengagement? Health Policy 81, 166–182 (2007)
38. Beirlant, J., Derveaux, V., De Meyer, A.M., Goovaerts, M.J., Labie, E., Maenhoudt, B.: Statistical risk evaluation applied to (Belgian) car insurance. Insurance: Mathematics and Economics 10, 289–302 (1992)
39. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C.: Data mining for credit card fraud: A comparative study. Decision Support Systems 50, 602–613 (2011)
40. Zarzaur, B.L., Stair, B.R., Magnotti, L.J., Croce, M.A., Fabian, T.C.: Insurance Type is a Determinant of 2-Year Mortality After Non-neurologic Trauma. Journal of Surgical Research 160, 196–201 (2010)
41. Ye, J., Xu, Z., Aladesanmi, O.: Provider recommendation for colorectal cancer screening: Examining the role of patients' socioeconomic status and health insurance. Cancer Epidemiology 33, 207–211 (2009)
42. Scott, D.J., Coveney, P.V., Kilner, J.A., Rossiny, J.C.H., Alford, N.M.N.: Prediction of the functional properties of ceramic materials from composition using artificial neural networks. Journal of the European Ceramic Society 27, 4425–4435 (2007)
43. Patra, J.C., Van Den Bos, A.: Modeling and development of an ANN-based smart pressure sensor in a dynamic environment. Measurement: Journal of the International Measurement Confederation 26, 249–262 (1999)
44. Hájek, P.: Municipal credit rating modelling by neural networks. Decision Support Systems 51, 108–118 (2011)
45. Marcano-Cedeño, A., Quintanilla-Domínguez, J., Andina, D.: WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications 38, 9573–9579 (2011)
46. Somers, M.J., Casal, J.C.: Using Artificial Neural Networks to Model Nonlinearity: The Case of the Job Satisfaction Job Performance Relationship. Organizational Research Methods 12, 403–417 (2009)
47. Swingler, K.: Applying Neural Networks: A Practical Guide. Academic Press, New York (1996)
48. SPSS: Clementine 12.0 Modeling Nodes. Integral Solutions Limited, Chicago (2003)
49. Magder, L.S., Hughes, J.P.: Logistic Regression When the Outcome Is Measured with Uncertainty. American Journal of Epidemiology 146, 195–203 (1997)

Replication Techniques in Data Grid Environments

Noriyani Mohd. Zin, A. Noraziah, Ainul Azila Che Fauzi, and Tutut Herawan

Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia
noriyanimz@gmail.com, {noraziah,tutut}@ump.edu.my,
ainulazila@yahoo.com

Abstract. A millions of data has been produce by cross-organizational research and collaborations must be managed, shared and analyzed. Data grid is a useful technique to solve these tasks that applicable to process the large number of data produced by scientific experiments. It also enables an organization to operate and manage distributed resources over the internet as a secure, robust, and flexible infrastructure. Some problems must be considered in managing data grid such as reliability and availability of the data to the user access, network latency, failures or malicious attacks during execution and etc. The replication strategy is the solution to solve these problems that can minimize the time access to the data by creating many replicas and storing replicas in appropriate locations. In this paper, we present some reviews on the existing dynamic replication replacement strategies due to the limited storage used on data grid and also to improve the management of data grid. It is shown that replication techniques able to improve availability and reliability of data, network latency, bandwidth consumption, fault tolerance and etc in data grid environments.

Keywords: Replica placement strategies, Data grid.

1 Introduction

Nowadays, there is a tendency of storing, retrieving, and managing different types of data such as experimental data that are produced from many projects. These data play a fundamental role in all kinds of cross-organizational researches and collaborations. For example, several scientific applications such as Particle Physics, High Energy Physics [1,2,3,4,5] and Genetics, earthquake engineering [1], climate change modeling [4,5] and astronomy [5], to cite a few, manage and generate an important amount of data which can reach terabytes and even petabytes, which need to be shared and analyzed. A community of hundreds or thousands of researchers distributed worldwide must share these datasets [1,2]. It's difficult, even impossible, to store such amount of data in the same location. Moreover, an application may need data produced by another geographically remote application. For this reason, data grid is suitable for the above situation.

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 549–559, 2012.

Data grid is a very important and useful technique to process the large number of data produced by scientific experiments and simulations [5]. Grid enables an organization to operate and manage distributed resources as a secure, robust, and flexible infrastructure – particularly as this infrastructure grows, shrink and changes in response to our needs [6]. Furthermore, it can efficiently manage and transfer the terabytes or petabytes of resources which are geographically distributed storage resources [1,2,3] located in different places, and enables users to share data and other resources [7,8]. Data grid can become as a single virtual data management system [9]. Therefore, data grid can be treated as a suitable solution for high-performance and data-intensive [9,10] computing applications that generate large data sets [5,11,12,13]. According to the type of resources shared, there are computational grids and data grids. Both of them are not mutually exclusive. In fact, computational grids and data grids play important roles in grid environment and are complementary to each other [14].

Grid computing in general comes from high-performance computing, super computing and later cluster computing where several processors or work stations are connected via a high-speed interconnect in order to compute a mutual program [15]. The grid research field can further be categorized into two large sub-domains: *Computational Grid* and *Data Grid*. Whereas the *Computational Grid* is a natural extension of the former cluster computer which are large computing tasks have to be computed at distributed computing resources, meanwhile a *Data Grid* deals with the efficient management, placement and replication of large amounts of data [15]. Many data grid applications are being developed, such as DoD's Global Information Grid (GIG) for both business and military domains, NASA's Information Power Grids, GMESS Health-Grid for medical services, data grids for Federal Disaster Relief [9], Johnson & Johnson (J&J), China Grid, Amazon, Google, eBay [6], etc. Besides that, it also provides new and more powerful ways of working, such as science portals, distributed computing for large-scale data analysis or collaborative work [16].

Some issues related to data management in grid environment are becoming increasingly important. Handling and managing data in grid is not easy. Some related problems must be considered. In terms of data management, the grid allows keeping a large number of replicas of data objects, possibly with different versions or levels of freshness, to allow for a high degree of availability, reliability and performance so as to best meet the needs of users and applications [16]. Furthermore, the size of data managed by data grid is continuously growing [8].

In the data grid, when a user requests a file, a large amount of bandwidth could be spent to send the file from the server to the client and the delay or response time involved could be high [3,7]. Besides that, maintaining local copies of data on each accessing site are cost prohibitive while storing all data in a centralized manner is impractical due to remote access latency [11,12,13,17]. This may lead the Internet turns to be the bottleneck in accessing the files in the grid. Due to the high latency of the Wide Area Network (WAN), the main issue is to design the strategy for efficient data access and share around the world with considerably low time complexity in data grid research [5]. Furthermore, in order to manage the data grid there are another several problems must be considered such as failures or malicious attacks during

execution, fault tolerance, scalability of data and etc. These problems can be solving by using the replication techniques in [1,2,3,4,5,7,8,11,17,18].

In this paper, we review the replication techniques that have been proposed in handling and managing data grid environment. The management of data grid is important to ensure the availability and reliability of data access to the users, to reduce the network latency and to provide fault tolerance. The rest of this paper is structured as follows. Section 2 presents some related works on data replication, where several different replication techniques are reviewed and compared. Finally, the conclusion of this work is described in section 3.

2 Data Management on Grid

In this section, we review some related works on the replication techniques in data grid.

2.1 Replication on Grid

In data replications architecture, the data will be replicated into several sites. If one of the sites has failed, it will failed independently and not affect to others node. Therefore, the data replication will improve the reliability, availability and performance of data. Replication in distributed environment receives particular attention for providing efficient access to data, fault tolerance [7] and enhance the performance of the system [19,20,21]. The replication strategy can minimize the time access to the file by creating many replicas and storing replicas in appropriate locations, which provide nearby data access. Furthermore, using replication is to reduce bandwidth consumption [1,2,3,17] to achieve efficient and dependable data access in grids, improve access time [3,5] fault tolerance [5,7,17] and load balancing [8]. In managing the replication on data grid, we will discuss about the replica placement strategy due to the limited storage use on data grids. Besides that, we also consider the problems that will face in grid that have been discussed on Section 1. There are three fundamental questions [2,3,4,5,18,22,23] that must be answer in managing replica placement strategy in data grids:

a. When should the replicas be created?
b. Which files should be replicated?
c. Where the replicas should be placed?

Different replication strategies have been developed or design to answer these questions above. The simulation is used to evaluate the performance of the strategy, usually using *OptorSim* to simulate the strategy.

2.2 Replication Techniques in Data Grid

2.2.1 A Weight-Based Dynamic Replica Replacement

In [5] proposed a weight-based dynamic replica replacement strategy in data grids. This strategy calculated the weight of replica based on the access time in the future time window, based on the last access history. After that, calculate the access cost

which embodies the number of replicas and the current bandwidth of the network. The replicas with high weight will be helpful to improve the efficiency of data access, so they should be retained and then the replica with low weight will not make sense to the rise of data access efficiency, and therefore, should be deleted [5]. The access history defines based on the zip-like distribution [5].

2.2.2 Distributed Popularity Based Replica Placement Algorithm

The distributed popularity based replica placement (DPBRP) [17] algorithm was developed for hierarchical data grids. This strategy exploits data access histories to recognize popular files. Furthermore, to determine optimal replication locations in order to improve data access performance by minimizing replication overhead (access and update) assuming a given traffic pattern. The dynamic programming is used to formulate this problem. Figure 1 shows the example of hierarchical data grid.

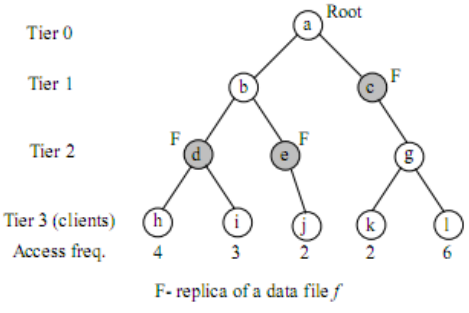


Fig. 1. An example of hierarchical data grid [17]

In DPBRP, the popular data files are determined by analyzing file access histories. The algorithm is invoked at usual intervals to process the access histories in order to decide new replica locations. It is based on file popularity, and replica update frequencies. If the old replicas are still in a newly determined replica set, it is retained. Other replicas from it set are deleted, and new replicas are created as required. The interval chosen is determined by the access request rate. Hence, a short interval result for high arrival rates and this incurs greater overhead but adapts more quickly to changing access patterns [17].

2.2.3 A New Replication Strategy for Dynamic Data Grids

A new replication strategy for dynamic data grids proposed in [3], which take into account the dynamic of sites. This strategy can increase the file availability, improved the response time and can reduce the bandwidth consumption. Moreover, it exploits the replicas placement and file requests in order to converge towards a global balancing of the grid load [3]. This strategy will focus on read-only-access as most grids have very few dynamic updates because they tend to use a "load" rather than "update" strategy [3]. There are three steps [3] provided by this algorithm, which are:

a. Selection of the best candidate files for replication;
   Selected based on requests number and copies number of each files.
b. Determination of the best sites for files placement which are selected in the previous step;
   Selected based on requests number and utility of each site regarding to the grid.
c. Selection of the best replica.
   Taking account the bandwidth and the utility of each site.

2.2.4 Enhance Fast Spread Replication Strategy

Enhance Fast Spread (EPF) [7] is an enhanced version of Fast Spread for replication strategy in the data grid. This strategy was proposed to improve the total of response time and total bandwidth consumption. Its takes into account some criteria such as the number and frequency of requests, the size of the replica, and the last time the replica was requested. EFS strategy and keeps only the important replicas while the other less important replicas are replaced with more important replicas [7]. This is achieved by using a dynamic threshold that determines if the requested replica should be stored at each node along its path to the requester [7]. This strategy takes four factors or criteria that have been stated as above into consideration when calculating the threshold. The number of requests shows the sums or how many times the replica has been requested by its node. The frequency of requests shows how many times the replica has been requested by its node within a specific time interval [7]. The size of replica is also a significant factor in deciding if the replica should be stored. The number and frequency of requests in addition to the last time the replica was requested give a hint of the probability of requesting the replica again [7].

2.2.5 A Value-Based Replication Strategy

In [24] proposed a value-based replication strategy (VBRS) to decrease the network latency and meanwhile to improve the performance of the whole system. In VBRS, threshold was made to decide whether to copy the requested file, and then solve the replica replacement problem. VBRS has two steps: (1) the threshold to decide whether a file should be replicated in the local storage device is introduced according to the access history and the storage capacity, (2) a measure based on the values of the local replicas, is devised to choose the replica that should be replaced [24]. To evaluate the performance of the VBRS, the Grid simulator *OptorSim* is used that can simulate the real data grid environment. At the first steps, the threshold will be calculated to decide whether the requested file should be copied in the local (nearest) storage site. Then at the second stage, the replacement algorithm will be triggered when the requested file needs to be copied at the local (nearest) storage site does not have enough space. It, firstly, calculates the files' values in the local storage site [24]. The files that have the least value will be deleted by the replacement algorithm. The files' value mostly concerns with three factors: network bandwidth, file's size, and the access history [24]. The files with higher value should be retained, and then the files with lower value will be deleted. The replica replacement policy is developed by considering the replica's value which is based on the file's access frequency and

access time [24]. The experiment results show that the effectiveness of VBRS algorithm can reduce network latency.

2.2.6 Agent Based Replica Placement Algorithm

An agent-based replica placement algorithm [1] was proposed to determine the candidate site for the placement of replica. For each site that holding the master copies of the shared data files will deploy an agent. The main objective of an agent is to select a candidate site for the placement of a replica that reduces the access cost, network traffic and aggregated response time for the applications [1]. Furthermore, in creating the replica an agent prioritizes the resources in the grid based on the resource configuration, bandwidth in the network and insists for the replica at their sites and then creates a replica at suitable resource locations. The agent in this approach is autonomous, self-contained software capable of making independent decisions [1].

There are two important issues that considers in this strategy [1], which are:

a. Choosing a replica location is to place a replica at sites that optimize the aggregated response time.
b. Choosing a replica location is to place a replica at sites that optimize the total execution time of the jobs executed in the grid.

Response time is calculated by multiplying the number of requests at site with the transmission time between the nearest replication site to the requester and the sum of the response times for all sites constitutes the aggregated response time [1]. Based on resource factors that influence the data transmission time between the sites is the decision must be made by an agent at each site. The factors include baud-rate between the sites, CPU Rating, CPU Load, Site Storage Capacity and Local Demand of the replicas at each site [1]. In order to evaluate the resource properties and grade with an appropriate rank, the agent uses a Multi-Dimensional Ranking (MDR) function. The agent preferences are represented by a set of factor weightings, which allow resource rank to be tailored to the current resource characteristics [1].

2.2.7 Predictive Hierarchical Fast Spread (PHFS)

In [25] has proposed a new dynamic replication method in a multi-tier data grid environments. This method called as predictive hierarchical fast spread (PHFS) which is an improve version of common fast spread (CFS). The fast spread is a dynamic replication method in the data grid. The multi-tier [25] is a tree-like structure to build data grid. The PHFS tries to forecast future needs and pre-replicates the min hierarchal manner to increase locality in accesses and improve performance that consider spatial locality. This method is able to optimize the usage of storage resources, which not only replicates data objects hierarchically in different layers of the multi-tier data grid for obtaining more localities in accesses. It is a method intended for read intensive data grids [25]. In PHFS, to predict future needs and pre-replicates them hierarchically in different nodes of different tiers in the multi-tier data grid on the path from the source node to the client by using predictive methods. The nodes in upper layers of multi-tier data grid have more storage capacity and computational power than the lower level nodes [25]. And, the bandwidth of the links

among nodes in upper levels is greater than the links in the lower level nodes [25]. So, the replication method in fast spread can replicate more items in upper level nodes. And, otherwise it can decrease the amount of replicated items on the path to lower levels toward the client. In order to optimize the utilization of resources for replication and provide maximum locality with available resources the hierarchal replication is used. In this paper has compared the PHFS with CFS from the perspective of access latency. Although CFS makes some improvements in some metrics of performance like bandwidth consumption, it shows poor results in local accesses patterns [25]. Therefore PHFS is used by predicting user's subsequent file demands and pre-replicate them earlier in hierarchal manner to increase locality in accesses. The PHFS method use priority mechanism and replication configuration change component to adapt the replication configuration dynamically with the obtainable condition [25]. Besides that, it is developed on the basis of the concept that users who work on the same context will request some files with high probability [25]. The results show that PHFS causes lower latency and better performance compared with CFS. Moreover, compared with CFS it showed that the most of the accesses in PHFS occur in lower levels. Besides that, it is more suitable for applications wherein the clients work on a context for a period of time and the requests of clients are not random, like scientific applications that researchers work on a project [25].

3 Conclusion

A review has been done on the dynamic replication techniques that implemented in a data grid in order to preserve the availability and reliability of data. Based on the review, there are several techniques have been proposed to improve the network latency, bandwidth consumption, and fault tolerance in data grid environments. Table 1 shows the summary of the replication techniques discussed in Section 2.

Table 1. Summary of dynamic replication techniques or strategies

Dynamic Replication Strategy	Purpose and Method
A Weight-Based Dynamic Replica Replacement Strategy [5]	- Calculated the weight of replica based on the access time in the future time window, based on the last access history. - Calculate the access cost which embodies the number of replicas and the current bandwidth of the network. - Replicas with high weight will be helpful to improve the efficiency of data access, so they should be retained and then the replica with low weight will not make sense to the rise of data access efficiency, and therefore should be deleted.

Table 1. (*continued*)

Distributed Popularity Based Replica Placement[17]	- Developed for hierarchical data grids. - Exploits data access histories to recognize popular files and determines optimal replication locations to improve data access performance by minimizing replication overhead (access and update) assuming a given traffic pattern. - Using dynamic programming.
A new replication strategy for dynamic data grids [3]	- Consider dynamicity of sites. - Purpose – to increase the file availability, improve the response time and can reduce the bandwidth consumption. - Focus on read-only-access as most grids have very few dynamic updates because they tend to use a "load" rather than "update" strategy. - Three steps: a. Selection of the best candidate files for replication. b. Determination of the best sites for files placement which are selected in the previous step. c. Selection of the best replica.
Enhance Fast Spread Replication Strategy [7]	- Enhanced version of Fast Spread for replication strategy in data grid. - Purpose – to improve the total of response time and total bandwidth consumption. - Consider on some criteria such as the number and frequency of requests, the size of the replica, and the last time the replica was requested. - Keeps only the important replicas while the other less important replicas are replaced with more important replicas by using a dynamic threshold that determines if the requested replica should be stored at each node along its path to the requester.
A Value-Based Replication Strategy (VBRS) [24]	- Purpose – to decrease the network latency and meanwhile to improve the performance of the whole system. - In VBRS, threshold was made to decide whether to copy the requested file, and then solve the replica replacement problem. - Two steps on VBRS: a. The threshold to decide whether a file should be replicated in the local storage device is introduce according to the access history and the storage capacity. b. A measure based on the values of the local replicas, is devised to choose the replica that should be replaced.

Table 1. (*continued*)

Agent Based Replica Placement Strategy [1]	- Proposed to determine the candidate site for the placement of replica. - Each site that holding the master copies of the shared data files will deploy an agent. - Main objective of an agent – to select a candidate site for the placement of a replica to reduces the access cost, network traffic and aggregated response time for the applications. - Consider two important issues: 1) Choosing a replica location is to place a replica at sites that optimize the aggregated response time. 2) Choosing a replica location is to place a replica at sites that optimize the total execution time of the jobs executed in the grid.
Predictive hierarchical fast spread [25]	- PHFS is an improve version of common fast spread (CFS) - Consider spatial locality, the PHFS tries to forecast future needs and pre-replicates the min hierarchal manner to increase locality in accesses and to improve its performance. - Using predictive methods to predict future needs and pre-replicates them hierarchically in different nodes of different tiers in the multi-tier data grid on the path from the source node to the client. - Used the hierarchical replication to optimize the utilization of resources. - The PHFS causes lower latency and better performance compared with CFS.

Acknowledgement. This work is supported by Postgraduate Research Grant Scheme by Universiti Malaysia Pahang.

References

1. Naseera, S., Murthy, K.V.M.: Agent Based Replica Placement in a Data Grid Environement.. In: The Proceeding of the First International Conference on Computational Intelligence, Communication Systems and Networks, CICSYN 2009, pp. 426–430 (2009)
2. Ben Charrada, F., Ounelli, H., Chettaoui, H.: Dynamic Period vs Static Period in Data Grid Replication. In: The Proceeding of 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pp. 565–568 (2010)
3. Ben Charrada, F., Ounelli, H., Chettaoui, H.: An Efficient Replication Strategy for Dynamic Data Grids. In: The Proceeding of 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pp. 50–54 (2010)

4. AL-Mistarihi, H.H.E., Yong, C.H.: On Fairness, Optimizing Replica Selection in Data Grids. IEEE Transactions on Parallel and Distributed Systems 20(8), 1102–1111 (2009)
5. Zhao, W., Xu, X., Xiong, N., Wang, Z.: A Weight-Based Dynamic Replica Replacement Strategy in Data Grids. In: The Proceeding of Asia-Pacific Services Computing Conference, APSCC 2008, pp. 1544–1549. IEEE Press (2008)
6. Mark Linesch, H.P.: Grid – Distributed Computing at Scale An Overview of Grid and the Open Grid Forum. In: Open Grid Forum 2007, August 28 (2007)
7. Bsoul, M., Al-Khasawneh, A., Eddien Abdallah, E., Kilani, Y.: Enhanced Fast Spread Replication strategy for Data Grid. Journal of Network Computer Application 34(2), 575–580 (2011)
8. Pérez, J.M., García-Carballeira, F., Carretero, J., Calderón, A., Fernández, J.: Branch replication scheme: A new model for data replication in large scale data grids. Future Generation Computer System 26(1), 12–20 (2010)
9. Tu, M., Li, P., Yen, I.-L., Thuraisingham, B., Khan, L.: Secure Data Objects Replication in Data Grid. IEEE Transactions On Dependable and Secure Computing 7(1), 1545–5971 (2010)
10. Zhang, J., Lee, B.-S., Tang, X., Yeo, C.K.: A model to predict the optimal performance of the Hierarchical Data Grid. Future Generation Computer Systems 26(1), 1–11 (2010)
11. Shorfuzzaman, M., Graham, P., Eskicioglu, R.: QoS-Aware Distributed Replica Placement in Hierarchical Data Grids. In: The Proceeding of 2011 IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 291–299 (2011)
12. Sashi, K., Thanamani, A.S.: Dynamic Replica Management for Data Grid. International Journal of Engineering and Technology 2(4), 329–333 (2010)
13. Sashi, K., Thanamani, A.S.: A New Replica Creation and Placement Algorithm for Data Grid Environment. In: The Proceeding of 2010 International Conference on Data Storage and Data Engineering (DSDE), pp. 265–269 (2010)
14. Ruay-Shiung, C., Chun-Fu, L., Shih-Chun, H.: Accessing data from many servers simultaneously and adaptively in data grids. Future Generation Computer System 26(1), 63–71 (2010)
15. Stockinger, H.: Distributed Database Management Systems and the Data Grid. In: The Proceeding of Eighteenth IEEE Symposium on Mass Storage Systems and Technologies (MSS 2001), pp. 1–12 (2001)
16. Laura, C.V., Schuldt, H., Breitbart, Y., Schek, H.J.: Replicated data management in the grid: the Re:GRIDiT approach. In: The Proceedings of the 1st ACM workshop on Data grids for eScience (DaGreS 2009), pp. 7–16 (2009)
17. Shorfuzzaman, M., Graham, P., Eskicioglu, R.: Distributed Popularity Based Replica Placement in Data Grid Environments. In: The Proceeding of 2010 International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), pp. 66–77 (2010)
18. Zhao, W., Xu, X., Wang, Z., Zhang, Y., He, H.: A Dynamic Optimal Replication Strategy in Data Grid Environment. In: The Proceeding of International Conference on Internet Technology and Applications, pp. 1–4 (2010)
19. Gao, M.D., Nayate, A., Zheng, J., Iyengar, A.: Improving Availability and Performance with Application-Specific Data Replication. IEEE Transaction on Knowledge and Data Engineering 17(1), 106–200 (2005)
20. Noraziah, A., Mat Deris, M., Norhayati, M.R., Rabiei, M., Shuhadah, W.N.W.: Managing Transaction on Grid-Neighbour Replication in Distributed System. International Journal of Computer Mathematics 86(9), 1–10 (2008)

21. Tang, M., Lee, B.S., Tang, X., Yeo, C.K.: The impact on data replication on Job Scheduling Performance in the Data Grid. Future Generation of Computer Systems 22, 254–268 (2006)
22. Ranganathan, K., Foster, I.: Identifying dynamic replication strategies for a high performance data grid. In: The Proceeding of the Second International Workshop on Grid Computing, pp. 75–86 (2001)
23. AL-Mistarihi, H.E., Yong, C.: Replica management in data grid. International Journal of Computer Science and Network Security 8, 22–32 (2008)
24. Wuqing, Z., Xianbin, X., Zhuowei, W., Yuping, Z., Shuibing, H.: Improve the Performance of Data Grids by Value - Based Replication Strategy. In: 2010 Sixth International Conference on Semantics, Knowledge and Grids, pp. 313–316 (2010)
25. Leyli, M.K., Isazadeh, A., Shishavan, T.N.: PHFS: A dynamic replication method, to decrease access latency in the multi-tier data grid. Future Generation Computer Systems 27(3), 233–244 (2011)

On Cloud Computing Security Issues

Ainul Azila Che Fauzi, A. Noraziah, Tutut Herawan, and Noriyani Mohd. Zin

Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang
Lebuhraya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia
ainulazila@yahoo.com, {noraziah,tutut}@ump.edu.my,
noriyanimz@gmail.com

Abstract. The cloud is a next generation platform that provides dynamic resource pools, virtualization, and high availability. The concept of cloud computing is using a virtual centralization. This means, in one part, we have a full control on data and processes in his computer. On the other part, we have the cloud computing where the service and data maintenance is provided by vendors. The client or customers usually unaware about the place where processes are running or the data is stored. So, logically speaking, the client has no control over it. This is the reason cloud computing facing so many security challenge. In this paper, we presented selection issues in cloud computing and focus on the security issues. There are four cloud computing security issues that will be focused, namely XML signature, browser security, cloud integrity and binding issues and flooding attacks. Data security on the cloud side is not only focused on the process of data transmission, but also the system security and data protection for those data stored on the storages of the cloud side. There are some considerations that need to be focused in order to achieve better safe environment in cloud computing such as storage and system protection and data protection. In order to achieve better performance in security, cloud computing needs to fulfill five goals which are availability, confidentiality, data integrity, control and audit. By implementing these goals, we hope data security in cloud computing will be more secure. We also hope that cloud computing will have a bright future with arise of a large number of enterprises and will bring an enormous change in the Internet since it is a low-cost supercomputing to provide services.

Keywords: Cloud computing, Services, Issues, Security.

1 Introduction

Computer has been changed in its evolution form several times, as learned from its previous events. However, the trend turned from larger and more expensive, to slighter and more affordable commodity PCs and servers which are tired together to construct something called cloud computing system [1]. Furthermore, cloud has advantages in offering more scalable, fault-tolerant services with even higher performance. Cloud computing integrates and provides different types of services

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 560–569, 2012.

such as Software-as-a-Service (SaaS), the applications are delivered as services over the Internet; Platform-as-a- Service (PaaS) systems software made available over the Internet and Infrastructureas-a-Service (IaaS), when the hardware made available for cloud users. The requirements and demands from users for cloud services vary, resulting in complex design and deployment of resources [2]. However, there still exist many problems in cloud computing today, a recent survey shows that data security and privacy risks have become the primary concern for people to shift to cloud computing because the data is stored as well as processing somewhere on to centralized location called data centers. So, the clients have to trust the provider on the availability as well as data security. Even more concerning, though, is the corporations that are jumping to cloud computing while being oblivious to the implications of putting critical applications and data in the cloud. Moving critical applications and sensitive data to a public and shared cloud environment is a major concern for corporations that are moving beyond their data center's network perimeter defense [3]. To alleviate these concerns, a cloud solution provider must ensure that customers can continue to have the same security and privacy controls over their applications and services by providing evidence to these customers that their organization and customers are secure. Besides, they also need to show that they can meet their service-level agreements, and show how they prove compliance to their auditors [4]. This paper will explore the key concepts and ideas surrounding cloud computing and will be focusing at the security challenges within the cloud computing prototype. The rest of this paper is organized as follows: Section 2 discusses the literature; Section 3 focuses on security in cloud computing; Finally, Section 4 is the conclusion from the finding.

2 Related Works

2.1 Cloud Computing

Cloud computing is a technology which using internet and central remote servers in order to maintain data and applications. A simple example of cloud computing is Yahoo email or Gmail. User doesn't need software or a server to use them. [5]. The service is fully managed by the provider which means user only needs personal computer and Internet access [5].

2.2 Cloud Computing Issues

In the last few years, cloud computing has grown from being a promising business concept to one of the fastest growing segments of the IT industry. Now, recession-hit companies are increasingly realizing that simply by tapping into the cloud they can gain fast access to best-of-breed business applications or drastically boost their infrastructure resources, all at negligible cost [6]. So, the concerns are beginning to grow about how safe an environment it is for everyone. There are several issues that will be highlighted in this paper.

2.2.1 Privacy

Cloud computing utilizes the virtual computing technology, users' personal data may be scattered in various virtual data center rather than stay in the same physical location. At this time, data privacy protection will face the controversy of different legal systems even across the national borders. Attackers can analyze the critical task depend on the computing task submitted by the users [7].

2.2.2 Reliability

The cloud servers also experience downtimes and slowdowns but what the difference is that users have a higher dependent on cloud service provider (CSP) in the model of cloud computing. There is a big difference in the CSP's service model. Once you select a particular CSP, you may be locked-in, thus bring a potential business secure risk [6].

2.2.3 Legal Issues

Regardless of efforts to bring into line the lawful situation, as of 2009, supplier such as Amazon Web Services provide to major markets by developing restricted road and rail network and letting users to choose "availability zones" [8]. On the other hand, worries stick with safety measures and confidentiality from individual all the way through legislative levels.

2.2.4 Open Standard

Open standards are serious to the growth of cloud computing. Most cloud providers expose APIs which are typically well-documented and unique to their implementation thus not interoperable. Some vendors have adopted others' APIs [9] and there are a number of open standards under development. The Open Cloud Consortium (OCC) [10] is working to develop consensus on early cloud computing standards and practices.

2.2.5 Compliance

Managing Compliance and Security for Cloud Computing, provides insight on how a top-down view of all IT resources within a cloud-based location can deliver a stronger management and enforcement of compliance policies. In addition to the requirements to which customers are subject, the data centers maintained by cloud providers may also be subject to compliance requirements [10].

2.2.6 Freedom

Cloud computing does not allow users to physically possess the storage of the data, leaving the data storage and control in the hands of cloud providers. Customers will argue that this is pretty fundamental and affords them the ability to retain their own copies of data in a form that retains their freedom of choice [2].

2.2.7 Security

In the cloud, your data will be distributed over these individual computers regardless of where your base repository of data is ultimately stored. Industrious hackers can

invade virtually any server. There are the statistics that show that one-third of breaches result from stolen or lost laptops and other devices. Besides, there also some cases which from employees' accidentally exposing data on the Internet, with nearly 16 percent due to insider theft [6].

3 Security in Cloud Computing

3.1 Goals in Cloud Computing

Traditionally, cloud computing has five goals that need to be fulfilling in order to achieve an adequate security.

3.1.1 Availability

The goal of availability for Cloud Computing systems is to make sure users can use them at any time and place. As we know, Cloud Computing system enables its users to access the system from anywhere as long as they have internet connection. This principle is valid for all the Cloud Computing services. There are two strategies that are mostly used to enhance the availability of cloud computing which are hardening and redundancy.

In addition, current Cloud system vendors such as Amazon and Skytab, offer the ability to block and filter traffic based on IP address and port only to secure their systems. These security control strategies are hardened into to their virtual machine, which in turn enhances availability of the provided infrastructure.

For redundancy, large cloud computing system vendors offer geographic redundancy in their cloud systems so that they can enable the high availability on a single provider. For example, Amazon builds data centers in numerous regions and various availability zones within those regions. Availability zones are distinct locations that are engineered to be insulated from failures in other availability zones and provide inexpensive, low latency network connectivity to other availability zones in the same region. One can protect applications from failure of a single location using instances in separate availability zones. That's to say, Cloud system has capability in providing redundancy to enhance the high availability of the system in nature [1].

3.1.2 Confidentiality

The confidentiality in cloud systems is a big obstacle for users to step into it. Currently, cloud computing system offers services that are basically public to networks.

Consequently, keeping all confidential data of users' secret in the cloud will attract even more users as this is a fundamental requirement. In order to achieve such confidentiality, there are two basic approaches namely physical isolation and cryptography. As discussed, cloud system services are transmitted through public networks. So, no physical isolation could be achieved. Alternatively, Virtual Local Area Networks should be deployed to achieve the virtual physical isolation [11].

Encrypted storage is another choice to enhance the confidentiality. For example, encrypting data before placing it in a Cloud may be even more secure than unencrypted data in a local data center [1].

3.1.3 Data Integrity

Keeping data integrity is a fundamental task as data is the base for providing cloud computing services. Cloud computing system usually provides massive data procession capability. Herein, massive data means many Tera Bytes (TB) data or even Peta Bytes (PB) data in volume. The challenge is the current development of state for hard disk drivers. The capacity increases are not keeping pace with the data growth. As a result, vendors need to increase the population of hard drives to scale up the data storage in the Cloud Computing systems. Consequently, this may increase high probability of node failure, disk failure, data corruption or even data loss. Second challenge is disk drives are getting bigger and bigger in terms of their capacity, but not getting much faster in terms of data access [1].

3.2 Cloud Computing Security Issues

There are four cloud computing security issues that will be focus in this paper namely XML signature, browser security, cloud integrity and binding issues and flooding attacks.

3.2.1 An XML Signature

XML Signature Element Wrapping is a well known type of attacks on protocols using XML Signature for authentication or integrity protection [12]. This of course applies to Web Services and therefore also for Cloud Computing.

Figures 2 and 3 show a simple example for a wrapping attack to illustrate the concept of this attack. The first figure presents a SOAP message sent by a legitimate client. The SOAP body contains a request for the file “me.jpg” and was signed by the sender. The signature is enclosed in the SOAP header and refers to the signed

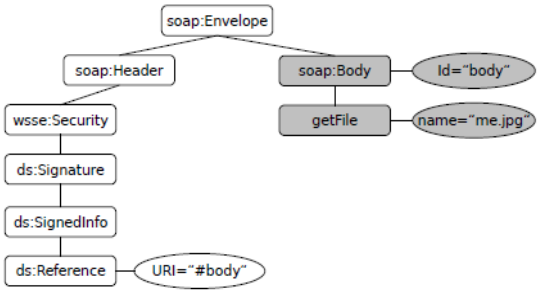


Fig. 1. Example SOAP message with signed SOAP body

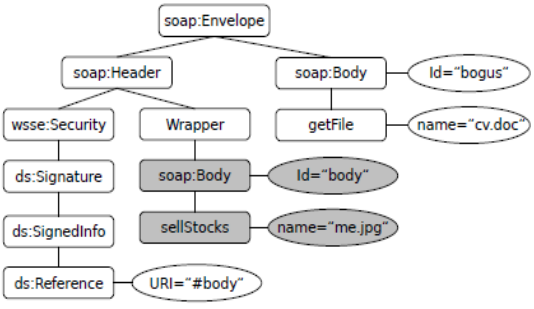


Fig. 2. Example SOAP message after attack

message fragment using an XPointer to the Id attribute with the value "body". If an attacker eavesdrops such a message, he can perform the following attack. The original body is moved to a newly inserted wrapping element (giving the attack its name) inside the SOAP header, and a new body is created. This body contains the operation the attacker wants to perform with the original sender's authorization, here the request for the file "cv.doc". The resulting message still contains a valid signature of a legitimate user, thus the service executes the modified request [13].

3.2.2 Browser Security

In a Cloud, computation is done on remote servers. The client PC is used for I/O and for authentication and authorization of commands to the Cloud. Therefore, it does not make sense to develop client software but have to use a universal platform independent tool for I/O which is a Web browser.

Modern Web browsers with their AJAX techniques are ideally suited for I/O but not really suited for security. Web browsers can not directly make use of XML Signature or XML Encryption. Data can only be encrypted through Transport Layer Security (TLS), and signatures are only used within the TLS handshake. TLS has been introduced, under its more common name "Secure Sockets Layer (SSL)", by Netscape in 1996. It consists of two main parts: The Record Layer encrypts/decrypts TCP data streams using the algorithms and keys negotiated in the TLS Handshake, which is also used to authenticate the server and optionally the client. Today it is the most important cryptographic protocol worldwide, since it is implemented in every web browser [13].

3.2.3 Cloud Integrity and Binding Issues

Maintaining and coordinating instances of virtual machines (IaaS) or explicit service implementation modules (PaaS) are the responsibility of a cloud computing system. Cloud system is responsible for determining and eventually instantiating a free-to-use instance of the requested service implementation type on request of any user. Then, the address for accessing that new instance is to be communicated back to the requesting user. Generally, this task requires some metadata on the service implementation modules, at least for identification purposes. For the specific PaaS

case of Web Services provided via the Cloud, this metadata may also cover all Web Service description documents related to the specific service implementation. For instance, the Web Service description document itself should not only be present within the service implementation instance, but also be provided by the cloud system in order to deliver it to its users on demand. Most of these metadata descriptions are usually required by any user prior to service invocation in order to determine the appropriateness of a service for a specific purpose. Thus, these metadata should be stored outside of the Cloud system, resulting in a necessity to maintain the correct association of metadata and service implementation instances [13].

3.2.4 Flooding Issues

A major aspect of Cloud Computing consists in outsourcing basic operational tasks to a cloud system provider. Among these basic tasks, one of the most important ones is server hardware maintenance. Thus, instead of operating an own, internal data center, the paradigm of cloud computing enables companies which is users to rent server hardware on demand (IaaS). This approach provides valuable economic benefits when it comes to dynamics in server load, as for instance day-and-night cycles can be attenuated by having the data traffic of different time zones operated by the same servers. Hence, instead of buying sufficient server hardware for the high workload times, cloud computing enables a dynamic adaptation of hardware requirements to the actual workload occurring. Technically, this achievement can be realized by using virtual machines deployed on arbitrary data center servers of the cloud system. If a company's demand on computational power rises, it simply is provided with more instances of virtual machines for its services.

Unfortunately, under security concerns, this architecture has a serious drawback. Though the feature of providing more computational power on demand is appreciated in the case of valid users, it poses severe troubles in the presence of an attacker. The corresponding threat is that of flooding attacks, which basically consist in an attacker sending a huge amount of nonsense requests to a certain service. In the specific case of cloud computing systems, the impact of such a flooding attack is expected to be amplified drastically. This is due to the different kinds of impact, which are discussed next [13].

3.3 Security Approach in Cloud Computing Security Issues

3.3.1 Security Issue of Enterprises Adopting the Application of Cloud Computing

Data security on the cloud side is not only focused on the process of data transmission, but also the system security and data protection for those data stored on the storages of the cloud side. The service provider has to pay attention to find out the possible occurred problems and possess the capability of perfect database and file management especially when there are a lot of users on the client side of the cloud computing accessing the same folders or even the same files on the cloud side. There are some considerations that need to be focused in order to achieve better safe environment in cloud computing which are storage and system protection and data protection

The service provider of cloud computing must present the necessary documents for the third party that plays the role of supervisory for auditing as a process defined in the specification of physical security. The user of cloud computing should also adopt the specification of physical security to watch over the physical retrieval procedure of the storage on the cloud side. Record the whole abolishing process including the abolishing location, verification procedure, and the process of demagnetization and recycling at the resource recycling station by video-taping as a reference or evidence. Besides, an enterprise of the user of cloud computing should also pay attention to the data security on the storage. Confidential file or sensitive data should be encrypted by the enterprise before uploading. After then, those encrypted data could be uploading to the storage designated and provided by service provider of the cloud computing through secure channel as shown in Figure 4 [14].

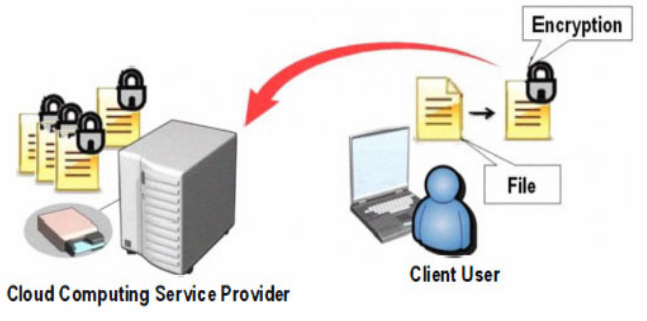


Fig. 3. Diagram of data encryption scheme before uploading

3.3.2 Digital Signature with RSA Encryption Algorithm to Enhance the Data Security of Cloud in Cloud Computing

This scheme is proposed to ensure the security of data in cloud. Till now, it is the only asymmetric algorithm used for private/public key generation and encryption. Both digital signature scheme and public key cryptography are included to enhance the security of cloud computing [15].

Assuming there are two enterprises A and B. An enterprise A has a public cloud with data, software's and applications. Company B wants a secure data from A's Cloud. Here, we will send a secure data to B by using Digital signature with RSA algorithm.

a. Step 1: A takes a document that B needs from cloud.
b. Step 2: The document will be crunched into few lines by using some Hash function the hash value is referred as message digest (refer Figure 4).
c. Step 3: A's software then encrypts the message digest with its private key. The result
d. is the digital signature (refer Figure 5)
e. Step 4: Using RSA Algorithm, A will encrypt digitally signed signature with B's

f. public key and B will decrypt the cipher text to plain text with its private key and A
g. public key for verification of signature (refer Figure 6).

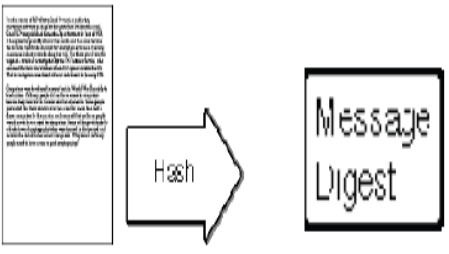


Fig. 4. Document crunched into message digest

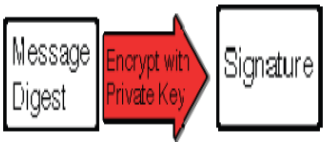


Fig. 5. Encryption of message digest into Signature

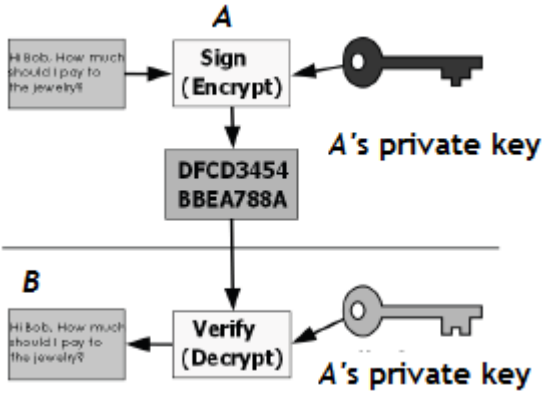


Fig. 6. Encryption of Digital Signature into Cipher text browser

4 Conclusion

In this paper, we have presented a selection of issues in cloud computing and focusing more on security issue. This is because security issues indicate problems which might arise from time to time. It is very important to take security and privacy into account when designing and using cloud services. We also discussed the approaches in order to enhance the cloud computing security issues such as fulfilling the five goals which are availability, confidentiality, data integrity, control and audit in cloud computing. From our observations, a first good starting point for improving cloud computing security consists of strengthening the security capabilities of both Web browsers and Web Service frameworks. In addition, cloud computing will bring a revolutionary change in the Internet since it has announced a low-cost supercomputing to provide services. Whereas there are a large number of enterprises behind, there are no doubt that cloud computing has a bright future.

Acknowledgement. This work is supported by Postgraduate Research Grant Scheme by Universiti Malaysia Pahang.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, D., Stoica I., Zaharia, M.: Above the clouds: A Berkeley view of cloud computing. Technical report UC Berkeley Reliable Adaptive Distributed Systems Laboratory (2009)
2. Minqi, Z., Rong, Z., Wei, X., Weining, Q., Aoying, Z.: Security and Privacy in Cloud Computing: A Survey. In: The Proceeding of Sixth International Conference of Semantics Knowledge and Grid (SKG 2010), pp. 105–112 (2010)
3. Chang-Lung, T., Uei-Chin, L., Chang, A.Y., Chun-Jung, C.: Information security issue of enterprises adopting the application of cloud computing. In: The Proceeding of Sixth International Conference Networked Computing and Advanced Information Management (NCM 2010), pp. 645–649 (2010)
4. Tharam, D., Chen, W., Elizabeth, C.: Cloud computing: issues and challenges. In: The Proceeding of 24th IEEE International Conference on Advanced Information Networking and Applications AINA, pp. 27–33 (2010)
5. http://searchcloudcomputing.techtarget.com/definition/cloud-computing
6. Pooja, T.: Demystifying Cloud Computing, http://www.technology-digital.com/blogs/editor/demystifying-cloud-computing (retrieved on January13, 2011)
7. Jianfeng, Y., Zhibin, C.: Cloud Computing Research and Security Issues. In: The Proceeding of International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1–3 (2010)
8. Brockmeier, J.: Eucalyptus Completes Amazon Web Services Specs with Latest Release, http://ostatic.com/blog/eucalyptus-completesamazon-web-services-specs-with-latest-release (retrieved on January 17, 2011)
9. Open Cloud Consortium, http://opencloudconsortium.org/home/ (retrieved on January 17, 2011)
10. Bardin, J.: Security Guidance for Critical Areas of Focus in Cloud Computing, http://www.cloudsecurityalliance.org/guidance/csaguide.pdf (retrieved on January 13, 2011)
11. Brodkin, J.: Gartner: Seven cloud-computing security risks, http://www.infoworld.com/d/security-central/gartner-seven-cloud-computingsecurity-risks-853 (retrieved on January 11, 2011)
12. Gantz, J.F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., Toncheva, A.: The diverse and exploding digital universe, IDC Future Report, 1–12 (2008)
13. Jensen, M., Schwenk, J., Gruschka, N., Iacono, L.L.: On Technical Security Issues in Cloud Computing Cloud Computing. In: The Proceeding of IEEE International Conference on Cloud Computing CLOUD 2009, pp. 109–116 (2009)
14. Somani, U., Lakhani, K., Mundra, M.: Implementing digital signature with RSA encryption algorithm to enhance the Data Security of cloud in Cloud Computing. In: The Proceeding of 1st International Conference on Parallel Distributed and Grid Computing (PDGC 2010), pp. 211–216 (2010)
15. Wood, K., Pereira, E.: An investigation into cloud configuration and security. In: The Proceeding of International Conference Internet Technology and Secured Transactions (ICITST), pp. 1–6 (2010)

Author Index