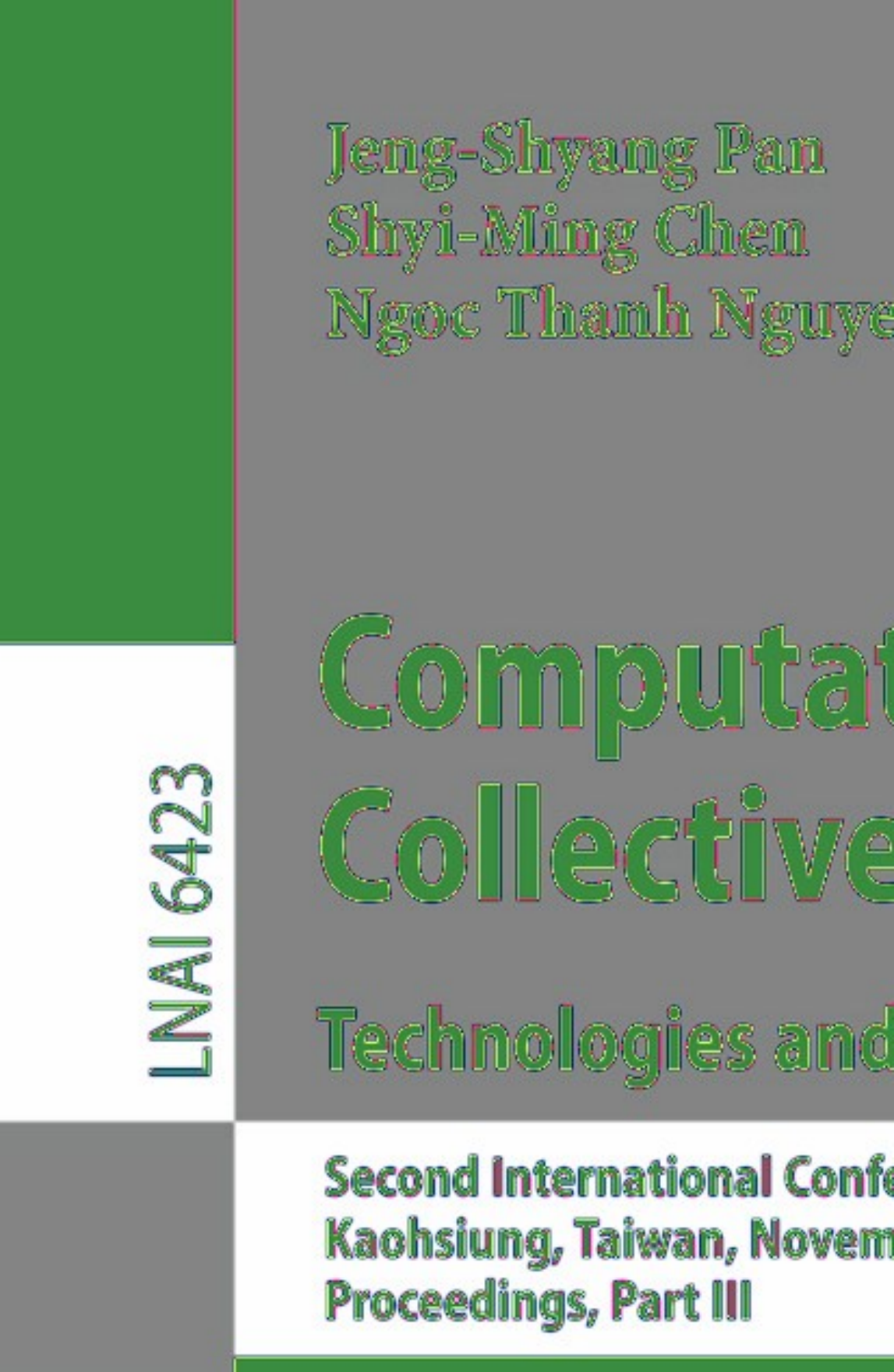

Jeng-Shyang Pan
Shyi-Ming Chen
Ngoc Thanh Nguye
Computat
Collective
Technologies and
LNAI 6423
Second International Confe
Kaohsiung, Taiwan, Novem
Proceedings, Part III

Lecture Notes in Artificial Intelligence 6423

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

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

Computational Collective Intelligence

Technologies and Applications

Second International Conference, ICCCI 2010
Kaohsiung, Taiwan, November 10-12, 2010
Proceedings, Part III

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Jeng-Shyang Pan
National Kaohsiung University of Applied Sciences
Department of Electronic Engineering
415 Chien-Kung Road, Kaohsiung 807, Taiwan
E-mail: jspan@cc.kuas.edu.tw

Shyi-Ming Chen
National Taiwan University of Science and Technology
Department of Computer Science and Information Engineering #43, Sec.4
Keelung Rd., Taipei, 106,Taiwan
E-mail: smchen@mail.ntust.edu.tw

Ngoc Thanh Nguyen
Wroclaw University of Technology, Institute of Informatics
Str. Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
E-mail: ngoc-thanh.nguyen@pwr.wroc.pl

Library of Congress Control Number: 2010937276

CR Subject Classification (1998): I.2, I.2.11, H.3-4, C.2, D, H.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-642-16695-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16695-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

This volume composes the proceedings of the Second International Conference on Computational Collective Intelligence—Technologies and Applications (ICCCI 2010), which was hosted by National Kaohsiung University of Applied Sciences and Wroclaw University of Technology, and was held in Kaohsiung City on November 10-12, 2010. ICCCI 2010 was technically co-sponsored by Shenzhen Graduate School of Harbin Institute of Technology, the Tainan Chapter of the IEEE Signal Processing Society, the Taiwan Association for Web Intelligence Consortium and the Taiwanese Association for Consumer Electronics. It aimed to bring together researchers, engineers and policymakers to discuss the related techniques, to exchange research ideas, and to make friends. ICCCI 2010 focused on the following themes:

- Agent Theory and Application
- Cognitive Modeling of Agent Systems
- Computational Collective Intelligence
- Computer Vision
- Computational Intelligence
- Hybrid Systems
- Intelligent Image Processing
- Information Hiding
- Machine Learning
- Social Networks
- Web Intelligence and Interaction

Around 500 papers were submitted to ICCCI 2010 and each paper was reviewed by at least two referees. The referees were from universities and industrial organizations. 155 papers were accepted for the final technical program. Four plenary talks were kindly offered by: Gary G. Yen (Oklahoma State University, USA), on "Population Control in Evolutionary Multi-objective Optimization Algorithm," Chin-Chen Chang (Feng Chia University, Taiwan), on "Applying De-clustering Concept to Information Hiding," Qinyu Zhang (Harbin Institute of Technology, China), on "Cognitive Radio Networks and Its Applications," and Lakhmi C. Jain (University of South Australia, Australia), on "Intelligent System Design in Security."

We would like to thank the authors for their tremendous contributions. We would also express our sincere appreciation to the reviewers, Program Committee members and the Local Committee members for making this conference successful. Finally,

we would like to express special thanks for the financial support from the National Kaohsiung University of Applied Sciences, Kaohsiung City Government, National Science Council and Education Ministry, Taiwan, in making ICCCI 2010 possible.

Novermber 2010

Ngoc Thanh Nguyen
Jeng-Shyang Pan
Shyi-Ming Chen
Ryszard Kowalczyk

ICCCI 2010 Conference Organization

Honorary Chair

Chun-Hsiung Fang	National Kaohsiung University of Applied Sciences, Taiwan
Jui-Chang Kung	Cheng Shiu University, Taiwan

General Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Program Committee Chair

Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Ryszard Kowalczyk	Swinburne University of Technology, Australia

Special Session Chairs

Bao-Rong Chang	National University of Kaohsiung, Taiwan
Chang-Shing Lee	National University of Tainan, Taiwan
Radoslaw Katarzyniak	Wroclaw University of Technology, Poland

International Advisory Chair

Bin-Yih Liao	National Kaohsiung University of Applied Sciences, Taiwan

International Publication Chair

Chin-Shin Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Bing-Hong Liu	National Kaohsiung University of Applied Sciences, Taiwan

Local Organizing Committee Chair

Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan

ICCCI 2010 Steering Committee

Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Co-chair

Ryszard Kowalczyk	Swinburne University of Technology, Australia
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Adam Grzech	Wroclaw University of Technology, Poland
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, South Korea
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Ryszard Tadeusiewicz	AGH-UST, Poland
Toyoaki Nishida	Kyoto University, Japan

ICCCI 2010 Technical Program Committee

Jeng Albert B.	Jinwen University of Science and Technology, Taiwan
Gomez-Skarmeta Antonio F.	Murcia University, Spain
Shih An-Zen	Jinwen University of Science and Technology, Taiwan
Andres Cesar	Universidad Complutense de Madrid, Spain
Hsieh Cheng-Hsiung	Chaoyang University of Technology, Taiwan
Lee Chin-Feng	Chaoyang University of Technology, Taiwan
Badica Costin	University of Craiova, Romania
Godoy Daniela	Unicen University, Argentina
Barbucha Dariusz	Gdynia Maritime University, Poland
Greenwood Dominic	Whitestein Technologies, Switzerland
CAPKOVIC Frantisek	Slovak Academy of Sciences, Slovakia
Yang Fuw-Yi	Chaoyang University of Technology, Taiwan
Huang Hsiang-Cheh	National University of Kaohsiung, Taiwan
Chang Hsuan-Ting	National Yunlin University of Science and Technology, Taiwan
Lee Huey-Ming	Chinese Culture University, Taiwan
Deng Hui-Fang	South China University of Technology, China
Czarnowski Ireneusz	Gdynia Maritime University, Poland
Lu James J.	Emory University, USA
Kacprzyk Janusz	Polish Academy of Sciences, Poland

Marecki Janusz	IBM T.J. Watson Research, USA
Sobecki Janusz	Wroclaw University of Technology, Poland
Jung Jason J.	Yeungnam University, South Korea
Nebel Jean-Christophe	Kingston University, USA
Dang Jiangbo	Siemens Corporate Research, USA
Huang Jingshan	University of South Alabama, USA
Chang Jui-fang	National Kaohsiung University of Applied Sciences, Taiwan
Nunez Manuel	Universidad Complutense de Madrid, Spain
Gaspari Mauro	University of Bologna, Italy
Khurram Khan Muhammad	King Saud University, Saudi Arabia
Sheng Quan Z.	University of Adelaide, Australia
Katarzyniak Radoslaw	Wroclaw University of Technology, Poland
Unland Rainer	University of Duisburg-Essen, Germany
Ching Chen Rung	Chaoyang University of Technology, Taiwan
Shen Rung-Lin	National Taipei University, Taiwan
Yang Sheng-Yuan	St. John's University, Taiwan
Yen Shu-Chin	Wenzao Ursuline College of Languages, Taiwan
Chen Shyi-Ming	National Taiwan University of Science and Technology, Taiwan
Zadrozny Slawomir	Polish Academy of Sciences, Poland
Hammoudi Slimane	ESEO, France
Hong Tzung-Pei	National University of Kaohsiung, Taiwan
Hsu Wen-Lian	Academia Sinica, Taiwan
Pedrycz Witold	University of Alberta, Canada
Baghdadi Youcef	Sultan Qaboos University, Oman
Lo Yu-lung	Chaoyang University of Technology, Taiwan
Cheng Yuh Ming	Shu-Te University, Taiwan
Huang Yung-Fa	Chaoyang University of Technology, Taiwan
Ye Yunming	Harbin Institute of Technology, China

Keynote Speakers

Gary G. Yen	Oklahoma State University, USA
Lakhmi C. Jain	University of South Australia, Australia
Chin-Chen Chang	Feng Chia University, Taiwan
Qinyu Zhang	Harbin Institute of Technology Shenzhen Graduate School, China

Program Committee of Special Sessions

Dariusz Barbucha	Gdynia Maritime University, Poland
Bao-Rong Chang	National University of Kaohsiung, Taiwan
Hsuan-Ting Chang	National Yunlin University of Science and Technology, Taiwan

Chuan-Yu Chang	National Yunlin University of Science and Technology, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Kazimierz Choroś	Wrocław University of Technology, Poland
Mohamed Hassoun	ENSSIB Villeurbanne, France
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Chien-Chang Hsu	Fu-Jen Catholic University, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Chien-Feng Huang	National University of Kaohsiung, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Huey-Ming Lee	Chinese Culture University, Taiwan
Che-Hung Lin	Cheng Shiu University, Taiwan
Lily Lin	China University of Technology, Taiwan
Piotr Jędrzejowicz	Gdynia Maritime University, Poland
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Chia-Nan Wang	National Kaohsiung University of Applied Sciences, Taiwan

Table of Contents – Part III

Intelligent Computing for Image Analysis (I)

Intelligent Digital Watermarking and Pattern Recognition

Advanced Knowledgement Management (II)

Intelligent Computing for Image Analysis (II)

Innovative Information System and Application

Intelligent Computing for Networks

Soft Computing to Industrial Management and Applications

Innovations in Pervasive Computing

Biological Computing

Feature Analysis and Classification of Lymph Nodes

Chuan-Yu Chang[1], Shu-Han Chang[1], and Shao-Jer Chen[2]

[1] Department of Computer Science and Information Engineering,
National Yunlin University of Science & Technology. Douliou, Yunlin, Taiwan
{chuanyu,g9717703}@yuntech.edu.tw
[2] Department of Medical Imaging, Buddhist Tzu Chi General Hospital,
Dalin, Chia-Yi, Taiwan
shaojer.chen@msa.hinet.net

Abstract. Pathological changes in lymph nodes (LN) can be diagnosed using biopsy, which is a time consuming process. Compared to biopsy, sonography is a better material for detecting pathology in the LN. However, there is lack of consistency between different ultrasound systems, which tend to produce images with different properties. To overcome this problem, a method was proposed in this paper to identify and select universal imaging features to standardize the classification of LN for different ultrasound imaging systems. This will help in the diagnosis of various pathological conditions. The support vector machine (SVM), which combines correlation and performance analysis for the selection of proper imaging features, was adopted for this classification system. Experimental results demonstrated that each selected feature set could be used to classify respective pathological conditions in the LN for images acquired from different ultrasound imaging machines.

Keywords: Lymph nodes, Ultrasound image, Feature selection, Classification, Support Vector Machine, Correlation analysis.

1 Introduction

Lymph nodes (LN) exist in the body as small, discrete nodal masses formed by the collection of lymphocytic cells and tissue as a component of the immunological system. Various conditions, ranging from self-limited throat infections to life-threatening cancers, may cause inflammatory and reactive changes in the LN. Pathological examples of such changes include lymphoid hyperplasia, lymphoma, and metastasis. Clinical manifestations of these pathological changes are important in prompting clinicians to a diagnosis of malignancy. Biopsy allows clinicians to diagnose and examine the LN, a process that can take several days before results can be obtained. As such, during this time clinicians may also simultaneously acquire an ultrasound image to aid diagnosis. However, ultrasound images are limited by poor resolution and varied echo patterns, which may often confuse and deter diagnosis. Another prominent issue with sonography is the lack of consistency between images obtained from different imaging systems. This limitation could be overcome with the universal implementation of a standardized computer-aided diagnostic system. In this

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 1–10, 2010.

paper, a method was proposed by which the significant morphological features of certain pathologies were identified using different imaging technologies and standardized to their respective pathological conditions.

There are six pathological conditions identified by an experienced radiologist and confirmed using biopsy, where the green rectangles highlight the regions of interest (ROIs). Apart from the normal state (NL), pathological changes can be categorized into the following disease states, namely, tuberculosis (TB), squamous cell carcinoma (SCC), metastatic changes (Meta), lymphadenitis, and hyperplasia [1]. It is difficult for radiologist to identify the correct pathological condition from the images.

The feature extraction process requires significant computation time. The process of feature extraction might also be inefficient when a large number of features have to be extracted. A significant/specific feature set can help reduce the extraction time for other features and also improve the accuracy of classification. Feature selection essentially involves the sifting of significant/specific features from all extracted features. Various feature selection methods have been proposed such as the sequential forward floating selection (SFFS) algorithm [2] and the ReliefF algorithm [3]. In addition, the F-score is a statistical criterion that was introduced to rank or score features [4]. However, the SFFS, F-score and ReliefF algorithms do not take into consideration the data from different sources.

In this study, significant features were selected based on the correlation between features obtained from different imaging systems as well as the F-score. The correlation analysis is basically a statistical analysis [5], which evaluates the statistical relationship between two or more random variables or observed data values. The correlation ratio was a value ranging from [0, 1] [6], and a calculation of different imaging systems data values was used to select the coefficient of the correlation ratio. In this paper, a correlation analysis was applied to evaluate the correlation of features from images obtained from different imaging systems. The use of a good classifier increased the accuracy of classification. The support vector machine (SVM) [7] was been employed to test and verify the selected feature sets. The SVM can be used for the purposes of classification and regression, which enables the construction of a hyper plane in N-dimensional spaces.

The paper is organized as follows: In Section 2, the process of feature extraction in the ROIs of LN ultrasound images is described. Section 3 provides a description of the methods used to select features from images acquired using different ultrasound imaging systems. The structure of a multi-class SVM is explained in section 4. Section 5 discusses the experimental results obtained using the proposed method while Section 6 summarizes the findings of this paper.

2 Feature Extraction

136 features were extracted from each ROI, for which the pathological conditions were identified by a radiologist and confirmed using biopsy. In this paper, the gray level range of an image was [0, L], and $I(x,y)$ denotes the gray level of the position (x,y) in the image. The features are described as follows:

(1) *Spatial Gray Level Co-occurrence Texture Matrix:* The spatial gray level co-occurrence textural matrix (SGLCM) is a statistical distribution in the spatial

domain [8]. Statistical results were recorded in this two dimensional array, which has a size of $(L+1) \times (L+1)$. In this paper, the distance was chosen so as one pixel to eliminate the effect of attenuation in the ultrasound image, and four different angles (0°, 45°, 90°, 135°) were selected to preserve spatial details. Fifty-two features were obtained using the SGLCM, namely: (**F1-F4**) correlation, (**F5-F8**) difference of entropy, (**F9-F12**) difference of variance, (**F13-F16**) sum of average, (**F17-F20**) sum of entropy, (**F21-F24**) sum of squares, (**F25-F28**) sum of variance, (**F29-F32**) contrast, (**F33-F36**) energy, (**F37-F40**) entropy, (**F41-F44**) local homogeneity, (**F45-F48**) cluster shade and (**F49-F52**) cluster prominence.

(2) *Statistical Feature Matrix:* The statistical feature matrix (SFM) can be used to describe the characteristic distance between image pixels [9]. We set the distances Δx and Δy to be equal to one. The (**F53**) dissimilarity in the ROI was calculated using the SFM.

(3) *Gray Level Run-Length Texture Matrix:* Similar to the SGLCM, the gray level run-length textural matrix (GLRLM) is also constructed using a certain distance and angles [10]. Twenty features were extracted from the GLRLM using different angles (0°, 45°, 90°, 135°), namely: (**F54-F57**) short run emphasis, (**F58-F61**) long run emphasis, (**F62-F65**) gray level uniformity, (**F66-F69**) run length uniformity, and (**F70-F73**) run percentage.

(4) *Laws' Texture Energy Matrix:* Laws' texture energy matrix measures the amount of variation occurring within a fixed window size[11]. Statistical results can be obtained by convolving the five Laws' 5×5 mask with the images. Ten features were calculated from Laws' texture energy matrix, namely: the (**F74**) *LE* mean, (**F75**) *EL* mean, (**F76**) *SL* mean, (**F77**) *EE* mean, (**F78**) *LS* mean, (**F79**) *LE* variance, (**F80**) *EL* variance, (**F81**) *SL* variance, (**F82**) *EE* variance, and (**F83**) *LS* variance.

(5) *Neighboring Gray Level Dependence Texture Matrix:* Neighboring Gray Level Dependence Textural Matrix (NGLDTM) is a two-dimensional matrix constructed using the gray level relationship between every pixel and its neighbors in an image [1]. The following features were obtained using the NGLDTM: (**F84**) small number emphasis, (**F85**) large number emphasis, (**F86**) number non-uniformity, (**F87**) second moment and (**F88**) entropy.

(6) *Wavelet feature Neighboring:* The image was decomposed for the double lower (*LL*) frequency sub-bands using low-pass filters in the horizontal direction. The coefficients of the low-pass filter were set using Antonini *et al.*'s method [12]. Wavelet features were calculated using the standard deviation and Laws' features of the *LL* sub-band. They are (**F89**) the mean, (**F90**) standard deviation, and (**F91-F100**) Laws' features of the *LL* sub-band.

(7) *Fourier Features Based on Local Fourier Coefficients:* The Fourier transform was used to obtain the local Fourier coefficients image map [13]. Each coefficient consists of two parameters: the intensity and the phase angle. (**F101-F108**) represent the means of eight magnitudes, (**F109-F116**) represent the means of eight phase angles, (**F117-F124**) are the standard deviations of the

eight magnitudes and (**F125-F132**) are the standard deviations of the eight phase angles.

(8) *Histogram:* The (**F133**) histogram feature was calculated to obtain an image with its pixels occupying the entire range of possible gray levels. The histogram is a basis for numerous spatial domain processing techniques and provides useful image statistics.

(9) *The Discrete Cosine Transform:* The Discrete Cosine Transform (DCT) was used in the frequency domain to calculate the frequency variation of an image. The Direct Current (DC) component refers to the image energy found at the center of the low frequency range. The value extracted from the DC component of the DCT gives (**F134**) the intensity at position (0,0) in the DCT image map.

(10) *The Inverse Probability Block Difference:* This feature (**F135**) was used to evaluate the difference between the number of pixels in a block and the ratio of the sum of the pixel intensities in the block to the maximum intensity in the block [9].

(11) *Normalized Multi-scale Intensity Difference:* The Normalized Multi-scale Intensity Difference (NMSID) feature (**F136**) characterizes a pair of pixels using four directions, namely the horizontal, vertical, diagonal and the asymmetric-diagonal. The average of the values from the four directions is then used to determine the nature of surface roughness [9].

3 Feature Selection

The 136 extracted features can be directly used to classify LN diseases. However, the direct approach would be time-consuming and does not guarantee a high level of accuracy. In addition, the extraction of irrelevant features further complicates the classification process. In this study, a correlation analysis was combined with the F-score to only select certain significant features among images acquired from different imaging systems. This initially involved feature correlation analysis, which was used to evaluate the correlation between features acquired from different imaging systems. Subsequently, the F-score method was applied to evaluate the significance of these features. Following this, the desired feature set was selected from different evaluation combinations.

3.1 Correlation Image Feature Analysis for Different Imaging Systems

In this subsection, a feature evaluation method is proposed. The features extracted from one system are considered to be good if they are similar to those extracted from another imaging system. The correlation method was adopted to evaluate the same type of extracted features acquired from different imaging systems. In this subsection, the problem of finding and evaluating the available feature has been addressed. The feature essentially symbolizes a correlation standard between different imaging systems. Each feature was evaluated using correlation analysis. Firstly, the data set with the *j*th feature for the *i*th pathological condition was defined as $\mathbf{X}_{\mathrm{F}_j}^{i}$, and its subset was defined to contain the *k*th ultrasound imaging system data. The objective

here was to evaluate the correlation η_i of the *j*th feature, between the data sets of two ultrasound imaging systems, for the *i*th pathological condition:

$$\eta_i\left(\mathbf{X}_{\mathrm{F}_j}^{i,1},\mathbf{X}_{\mathrm{F}_j}^{i,2}\right)=\sqrt{\frac{\sum_{k=1}^{2}\left|\mathbf{X}_{\mathrm{F}_j}^{i,k}\right|\left(\mathrm{E}\left(\mathbf{X}_{\mathrm{F}_j}^{i,k}\right)-\mathrm{E}\left(\mathbf{X}_{\mathrm{F}_j}^{i}\right)\right)^2}{\sum_{\forall x}\left(x-\mathrm{E}\left(\mathbf{X}_{\mathrm{F}_j}^{i}\right)\right)^2}} \text{ where } \forall x\in\mathbf{X}_{\mathrm{F}_j}^{i} \tag{1}$$

where the symbol $|\cdot|$ represents the number of elements present in the input data set, $\mathrm{E}(\cdot)$ is the average value of the elements in the input data set, and x is the value of the element from a single data set. The value of η_i lies between [0, 1]. If $\mathbf{X}_{\mathrm{F}_j}^{i,1}$ and $\mathbf{X}_{\mathrm{F}_j}^{i,2}$ are completely correlated, η_i has a value of one; if $\mathbf{X}_{\mathrm{F}_j}^{i,1}$ and $\mathbf{X}_{\mathrm{F}_j}^{i,2}$ are totally independent, then η_i is zero. If the data sets of the two systems exhibit a purely linear relationship, the result will be equal to the square of the Pearson product-moment correlation coefficient. In other cases, the correlation ratio will be greater in magnitude, and can be used for judging the properties of non-linear relationships. Based on the above equation, feature correlation between different ultrasound imaging systems was evaluated, for a specific pathological condition.

3.2 Discriminate Feature Analysis of Different Imaging Systems

Fisher's criterion was used to evaluate the discrimination of every feature extracted from different imaging systems. Fisher's criterion for feature discrimination is defined as follows:

$$D_i\left(\mathbf{Y}_{\mathrm{F}_j}^{k}\right)=\frac{\left(\mathrm{E}\left(\mathbf{Y}_{\mathrm{F}_j}^{i,k}\right)-\mathrm{E}\left(\mathbf{Y}_{\mathrm{F}_j}^{k}\right)\right)^2+\left(\mathrm{E}\left(\mathbf{Y}_{\mathrm{F}_j}^{k}-\mathbf{Y}_{\mathrm{F}_j}^{i,k}\right)-\mathrm{E}\left(\mathbf{Y}_{\mathrm{F}_j}^{k}\right)\right)^2}{\left(\sigma\left(\mathbf{Y}_{\mathrm{F}_j}^{i,k}\right)\right)^2+\left(\sigma\left(\mathbf{Y}_{\mathrm{F}_j}^{k}-\mathbf{Y}_{\mathrm{F}_j}^{i,k}\right)\right)^2} \tag{2}$$

where $\mathbf{Y}_{\mathrm{F}_j}^{k}$ represents the data set describing the *k*th ultrasound imaging system with the *j*th feature, $\mathbf{Y}_{\mathrm{F}_j}^{i,k}$ is a subset of $\mathbf{Y}_{\mathrm{F}_j}^{k}$, containing the data for the *i*th pathological condition, and $\sigma(\cdot)$ is the standard error of the input data set. By evaluating the above equation, it was possible to discriminate the *j*th feature of the *k*th ultrasound imaging system, for the *i*th pathological condition. A high value of D indicates a strong discrimination ability and vice versa.

3.3 A Combination of Correlation and Discrimination Analysis for Proper Feature Selection

In this subsection, the correlation and discrimination analysis results obtained in Section 3.2 and Section 3.3, respectively, have been combined to select features for each pathological condition. The equation is defined by:

$$Q_i\left(\mathrm{F}_j\right)=\left(\sum_{k=1}^{2}D_i\left(\mathbf{Y}_{\mathrm{F}_j}^{k}\right)\right)/\left(1-\eta_i\left(\mathbf{X}_{\mathrm{F}_j}^{i,1},\mathbf{X}_{\mathrm{F}_j}^{i,2}\right)\right) \tag{3}$$

where $Q_i(\cdot)$ represents the specificity and certainty analysis of the jth feature F_j with respect to the ith pathological condition. A high evaluation value indicates an optimally selected feature. Lastly, the set of features required to describe the ith pathological condition were selected using the following equations.

$$\begin{aligned} &\mathbf{R}_i^{t+1} = \mathbf{R}_i^t \cup \arg\max_{\mathrm{F}_j} \left(Q_i\left(\mathrm{F}_j\right)\right), \\ &\text{if } \sum_{\forall \mathrm{F}_w} Q_i\left(\mathrm{F}_w\right) \le T_Q \text{ where } \forall \mathrm{F}_j \notin \mathbf{R}_i^t \text{ and } \forall \mathrm{F}_w \in \mathbf{R}_i^t \end{aligned} \tag{4}$$

where $\mathbf{R}_i$ denotes the feature set required for the ith pathological condition, t is the tth iteration, F_w is the wth feature, and T_Q is a real number greater than or equal to zero. The value of T_Q is user-defined, and it represents the degree of specificity and certainty of the required feature for the ith pathological condition; A higher value enables more features to be selected into the set of eligible features, and vice versa.

4 Multi-class Support Vector Machine (SVM)

In this paper, the multi-class SVM was implemented using the LIBSVM [14], and the radial basis function (RBF) was selected as the kernel function. The SVM has primarily been designed for the purposes of binary classification. It can be extended to solve multi-class problems [15]. The proposed multi-class SVM consists of several two-class SVMs. Fig. 1 shows the structure of the proposed multi-class SVM. It is used to test and verify the results of feature selection to classify the diseases of the LN in different imaging systems. Fifteen SVMs were trained with the data from each class. Each classification result contributed to the tally of a particular class. Results were verified by making a comparison between the voting result and class of the test data. The multi-class SVM can categorize the ROIs representing the pathological condition of the LN into C1) LN Metastasis (Meta), C2) LN Hyperplasia, C3) normal LN C4) Lymphadenitis, C5) Tuberculosis (TB) of the LN, and C6) Squamous cell carcinoma (SCC). The three-fold cross-validation method was used to evaluate the multi-class SVM of the RBF kernel parameters [14].

5 Experimental Results

Ultrasound images of LNs were obtained using different commercial sonographic imaging systems. These included the GE LOGIQ 700 ultrasound system (system1) and the ATL HDI 3000 ultrasound machine (system2). LN ultrasound images were taken from 358 patients during 2005 and 2009. The images were obtained using a B-mode linear array with an operating frequency in the 5 to 10 MHz range. Parameters affecting image acquisition, such as the time-gain compensation and focal zones, were kept constant. System1 had seven focal zones within a depth range of 2 cm and system2 had 3 focal zones within a depth range of 2 cm. The other system1 parameters were set as follows: the dynamic range was set to 78 dB; the gain was set to 34; the edge enhancement was set as E2; the gray map was set as MC; and the

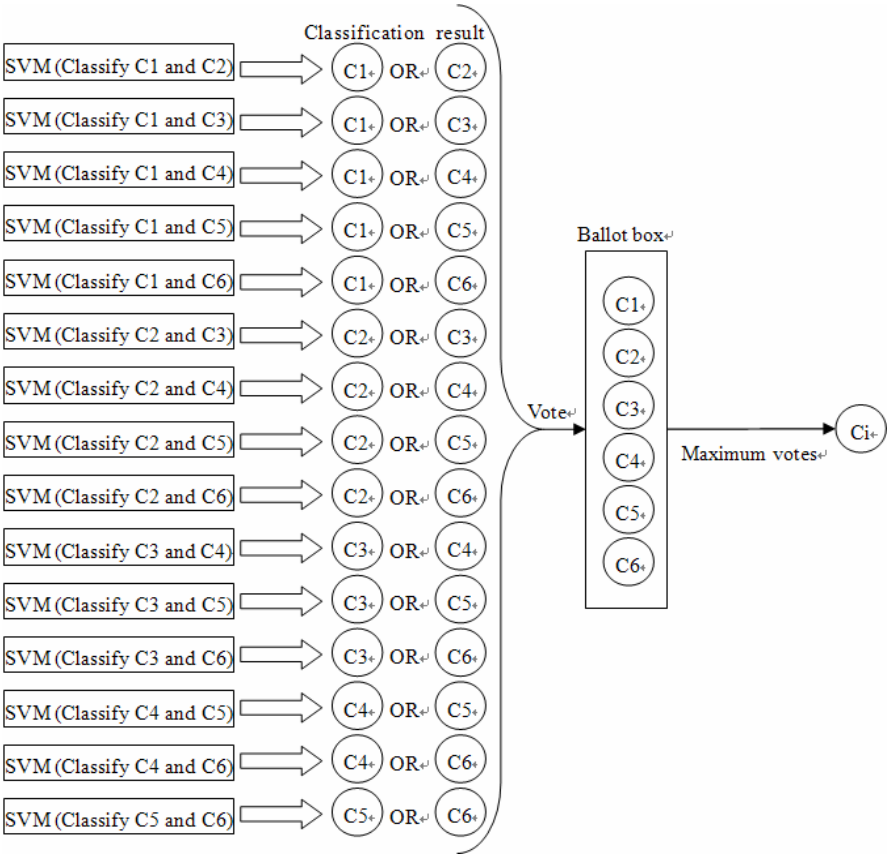


Fig. 1. The multi-class SVM structure

frame average settings were set as A2 [16]. The remaining parameters were set as follows: the dynamic range was set to 55dB, the 2D gray maps were assigned a value of eight, the persistent rate was set in the high state, and the 2D frame rate was assigned a medium setting [17]. A total of 2674 ROIs were obtained from LN ultrasound images, of which 335 ROIs indicated C1) LN Metastasis; 691 ROIs indicated a C2) LN Hyperplasia; 478 ROIs were found to be C3) normal; 555 ROIs pointed towards C4) Lymphadenitis; 250 ROIs showed C5) TB of the LN; and 365 ROIs indicated a C6) SCC section of the LN, as shown in Table 1. The table summarizes the number of patients involved in the study, and training and testing data statistics for each class and imaging system.

To test a common discriminative feature subset for each class between systems 1 and 2, the feature vectors obtained from the two systems were set as the training and testing data sets of the SVMs. The accuracy of the proposed method is shown in Table2, where the number of features was fixed by setting T_Q to zero. Table 2 shows that each class and category of every system can be correctly classified by selecting

Table 1. Number of patients involved in the study and data statistics for different imaging systems

Class	System	Patients	ROI	Training samples	Testing samples
1	1	17	164	82	82
1	2	18	171	86	85
2	1	41	303	152	151
2	2	56	388	194	194
3	1	44	126	63	63
3	2	59	352	176	176
4	1	32	242	121	121
4	2	32	313	156	157
5	1	16	160	80	80
5	2	11	90	45	45
6	1	20	188	94	94
6	2	18	177	89	88

the best features of the proposed method. The proposed method was compared with other feature selection method, such as SFFS and was shown to have superior performance in Tables 3. Tables 2 and 3 clearly demonstrate that the accuracy of the proposed method was higher than the SFFS method.

Table 2. Results of feature selection and feature classification using the proposed method in different imaging systems (T_Q=0)

Class	Feature Set	System 1 Accuracy	System 2 Accuracy	Average Accuracy
1	F1	97.6%	98.8%	95.2%
2	F9	99.0%	90.1%	92.5%
3	F88	93.2%	90.5%	92.1%
4	F15	100%	98.3%	97.1%
5	F62	100%	100%	100%
6	F39	96.6%	97.9%	97.8%

Table 3. Results of feature selection and classification for features selected by the SFFS method

Class	Feature Set	System 1 Accuracy	System 2 Accuracy	Average Accuracy
1	F77	92.9%	96.3%	86.2%
2	F37	99%	94%	93.6%
3	F32	88.1%	90.5%	84.1%
4	F52	99.4%	86.8%	92.1%
5	F134	100%	100%	100%
6	F101	98.9%	95.7%	97.3%

Tables 2, 4 and 5 show the accuracy for different T_Q values. The T_Q values in Table 2, 4, and 5 were set to 0, 1, and 2, respectively, among which the highest accuracy was obtained when T_Q is equal to 1.

Table 4. Results of feature selection and classification for features selected by the proposed method in different imaging systems (T_Q=1)

Class	Feature Set	System 1 Accuracy	System 2 Accuracy	Average Accuracy
1	F1	100%	98.8%	98.8%
2	F9,F37,F124	98.5%	95.4%	97.7%
3	F88,F65	96.6%	96.8%	94.1%
4	F15	100%	98.3%	98.2%
5	F62	100%	100%	99.2%
6	F39	100%	97.9%	97.8%

Table 5. Results of feature selection and classification for features selected by the proposed method in different imaging systems (T_Q=2)

Class	Feature Set	System 1 Accuracy (AC)	System 2 AC	Average AC
1	F1,F131,F127	97.6%	97.6%	96.8%
2	F9,F37,F124,F118,F63,F64	100%	93.4%	95.9%
3	F88,F65,F63	97.2%	95.2%	95.4%
4	F15,F16	98.7%	98.3%	99.3%
5	F62	100%	100%	100%
6	F39,F38	98.9%	97.9%	97.8%

6 Conclusion

A universal feature set for each pathological condition helps in the establishment of a standardized method of computer aided diagnosis, which can be applied to images acquired from different imaging systems. A feature selection method was proposed in this paper to select significant features for each pathological condition, which were extracted from images acquired using different imaging systems. The multi-class SVM was used to test these selected image features. Experimental results demonstrated that the proposed feature selection method, which combined correlation evaluation and the F-score algorithm, was effective in selecting universal features from different imaging systems. The accuracy of each imaging system was higher than 90%. This demonstrates that the proposed method is able to detect pathological changes in the LN images of different imaging systems.

Acknowledgement. This work was supported by the National Science Council, Taiwan, under grant NSC 96-2218-E- 224-007.

References

1. Chang, C.Y., Lai, C.T., Chen, S.J.: Applying the Particle Swarm Optimization and Boltzmann Function for Feature Selection and Classification of Lymph Node in Ultrasound Images. In: 2008 Eighth International Conference on Intelligent Systems Design and Applications, pp. 55–60. IEEE Press, Taiwan (2008)
2. Chen, X.Y., Zheng, S.J., Tao, T.: Framework for Efficient Letter Selection in Genetic Algorithm Based Data Mining. In: 2008 International Symposium on Distributed Computing and Applications, pp. 334–338. ISTP Press, China (2008)

3. Chang, C.C., Lin, T.Y.: Linear feature extraction by integrating pairwise and global discriminatory information via sequential forward floating selection and kernel QR factorization with column pivoting. Pattern Recognition 41, 1373–1383 (2008)
4. Chen, S., Yu, S., Tzeng, J., Chen, Y., Chang, K., Cheng, K., Hsiao, F., Wei, C.: Characterization of the major histopathological components of thyroid nodules using sonographic textural features for clinical diagnosis and management. Ultrasound in Medicine & Biology 35, 201–208 (2009)
5. Chang, C.Y., Chen, S.J., Tsai, M.F.: Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. Pattern Recognition 43, 3494–3506 (2010)
6. Lewandowski, D., Cooke, R., Tebbens, R.: Sample-based estimation of correlation ratio with polynomial approximation. ACM Transactions on Modeling and Computer Simulation 18, 1–17 (2007)
7. Milko, S., Melvar, E., Samset, E., Kadir, T.: Evaluation of bivariate correlation ratio similarity metric for rigid registration of US/MR images of the liver. International Journal of Computer Assisted Radiology and Surgery 4, 147–155 (2009)
8. Coelho, S.T., Ynoguti, C.A.: A Histogram Based Method for Multiclass Classification Using SVMs. Advances in Experimental Medicine and Biology 657, 233–242 (2010)
9. Chang, C.Y., Huang, H.C., Chen, S.J.: Automatic Thyroid Nodule Segmentation and Component Analysis in Ultrasound Images. Biomedical Engineering: Applications, Basis and Communications 22, 81–89 (2010)
10. Chen, S.J., Cheng, K.S., Chen, Y.T., Dai, Y.C., Sun, Y.N., Chang, K.Y., Yu, S.N.: Quantitative correlation between sonographic textural feature and histopathological components for breast cancer: preliminary results. Clinical Imaging 32, 93–102 (2008)
11. Sun, X., Chuang, S., Li, J., McKenzie, F.: Automatic diagnosis for prostate cancer using run-length matrix method. Progress in Biomedical Optics and Imaging 7260, 1–8 (2009)
12. Chang, C.Y., Wu, Y.L., Tsai, Y.S.: Integrating the Validation Incremental Neural Network and Radial-Basis Function Neural Network for Segmenting Prostate in Ultrasound Images. In: 2009 Ninth International Conference on Hybrid Intelligent Systems, pp. 198–203. IEEE Press, China (2009)
13. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. IEEE Transactions on Image Processing 1, 205–220 (1992)
14. Chen, D.R., Chang, R.F., Chen, C.J., Ho, M.F., Kuo, S.J., Chen, S.T., Hung, S.J., Moon, W.K.: Classification of breast ultrasound images using fractal feature. Clinical Imaging 29, 235–245 (2005)
15. LIBSVM: a library for support vector machines, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
16. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13, 415–425 (2002)
17. Paste, A.: LOGIQ 700 Expert Series: Conformance Statement for DICOM V3.0. Operating Instructions, UK (1997)

Automatic and Accurate Image Matting

Wu-Chih Hu[1], Deng-Yuan Huang[2], Ching-Yu Yang[1],
Jia-Jie Jhu[1], and Cheng-Pin Lin[1]

[1] Department of Computer Science and Information Engineering,
National Penghu University of Science and Technology, Penghu, Taiwan
{wchu,chingyu,d1095405044,d1095405020}@npu.edu.tw
[2] Department of Electrical Engineering,
Dayeh University, Changhua, Taiwan
kevin@mail.dyu.edu.tw

Abstract. This paper presents a modified spectral matting to obtain automatic and accurate image matting. Spectral matting is the state-of-the-art image matting and also a milestone in theoretic matting research. However, using spectral matting without user guides, the accuracy is usually low. The proposed modified spectral matting effectively raises the accuracy. In the proposed modified spectral matting, the palette-based component classification is proposed to obtain the reliable foreground and background components. In contrast to the spectral matting, based on these reliable foreground and background components, the accuracy of obtained alpha matte is greatly increased. Experimental results show that the proposed method has better performance than the spectral matting. Therefore, the proposed image matting is very suitable for image and video editing.

Keywords: spectral matting, image matting, alpha matte.

1 Introduction

Image matting plays an important role in image and video editing. Image matting is the process of extracting a foreground object from an image along with an opacity estimate for each pixel covered by the object. Therefore, image matting belongs to image segmentation, but it is not a binary segmentation.

Image matting was first mathematically established by Porter and Duff [1]. Image matting takes the problem of estimating the partial opacity of each pixel in a given image. The given image is assumed to be a composite of a foreground image and a background image using the compositing equation, where the color of the given image is assumed to be a convex combination of corresponding foreground and background colors with the associated opacity value (alpha value). For each pixel in color images, compositing equation gives us 3 equations (RGB channels) in 7 unknowns. Consequently, image matting is a highly under-constrained problem. 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).

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 11–20, 2010.

Although image matting has been studied for more than two decades, image matting has received increasing attention in the last decade and many methods have been proposed for image matting [2], especially in the last five years, such as robust matting [3], easy matting [4], closed-form matting [5], and flash matting [6]. Generally, image matting can be roughly classified into three types [2, 3]: sampling-based methods, propagation-based methods, and matting with extra information.

Sampling-based methods assume that an unknown pixel can be estimated as the foreground and background colors based on examining nearby pixels that have been specified by the user as foreground or background. Next, these color samples are used to directly estimate the alpha value. Propagation-based methods assume that foreground and background colors are locally smooth. Next, the foreground and background colors are systematically eliminated from the optimization process to obtain the alpha matte. Matting with extra information is designed to provide additional information or constraints to matting algorithms in order to obtain better results on natural images.

In contrast to matting with extra information, sampling-based methods and propagation-based methods are suitable for video matting. However, almost sampling-based methods and propagation-based methods are the supervised matting (non-automatic matting). Therefore, video matting based on these supervised methods is a time-consuming and troublesome process for the user. Unsupervised (automatic) image matting is thus an important and interesting issue. It is a big challenging task for unsupervised image matting to solve the alpha matte from the highly under-constrained problem without any user input.

The unsupervised image matting (spectral matting) was first proposed by Levin et al. [7], and spectral matting is the only unsupervised method in the current proposed image matting algorithms. Spectral matting extends the ideas of spectral segmentation [8, 9]. In spectral matting, a set of fundamental fuzzy matting components are automatically extracted from an input image based on analyzing the smallest eigenvectors of a suitably defined Laplacian matrix (matting Laplacian [5]). The smallest eigenvectors of the matting Laplacian span the individual matting components of the image, thus recovering the matting components of the image is equivalent to finding a linear transformation of the eigenvectors. These matting components are then combined to form the complete alpha matte. However, using spectral matting without user guides, the accuracy is usually low. Therefore, it is a challenging task for unsupervised image matting to obtain an accurate alpha matte.

In this paper, the modified spectral matting is proposed to obtain automatic and accurate alpha matte. The proposed method extends the idea of spectral matting. In the modified spectral matting, the palette-based component classification is proposed to find the components of foreground, background and unknown regions. Next, the corresponding matting components of foreground components, background components, and components of unknown region are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian matrix. Finally, only matting components of foreground and unknown regions are combined to form the complete alpha matte based on minimizing the matte cost. Therefore, the accuracy of obtained alpha matte is greatly increased. Experimental results show that the proposed method can obtain the high-quality alpha matte for natural images without any user input.

The rest of this paper is organized as follows. The proposed modified spectral matting is presented in Section 2. Image composition with consistency of color temperature is described in Section 3. Section 4 presents experimental results and their evaluations. Finally, the conclusion is given in Section 5.

2 Modified Spectral Matting

Image matting typically assume that each pixel I_i in an input image is a linear combination of a foreground color F_i and a background color B_i, as defined in Eq. (1), where α_i is the pixel's foreground opacity. Eq. (1) is also called as the compositing equation.

$$I_i = \alpha_i F_i + (1-\alpha_i) B_i \tag{1}$$

In spectral matting, the compositing equation 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) [7], where α_i^k must satisfy the conditions in Eq. (3). The K vectors α^k are the matting components of the 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 input image consists of K distinct components $C_1 \sim C_K$ such that $C_i \bigcap C_j = \phi$ for $i \neq j$. Compute the eigenvectors of $N \times N$ Laplacian matrix L (matting Laplacian [5]) as $E = [e^1, ..., e^M]$, so 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 [10] 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:

$$A_q(i,j) = \begin{cases} \delta_{ij} - \dfrac{1}{|w_q|}\left(1 + (I_i - \mu_q)^T \left(\sum_q + \dfrac{\varepsilon}{|w_q|} I_{3\times 3}\right)^{-1} (I_j - \mu_q)\right), (i,j) \in w_q \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{, otherwise} \end{cases} \tag{4}$$

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 $I_{3\times 3}$ is the 3×3 identity matrix.

Fig. 1 is the result of distinct component detection using spectral segmentation with the *k*-means algorithm, where Fig. 1(a) is the input image, Fig. 1(b) is the clustering result, and Fig. 1(c) is the result of distinct components.

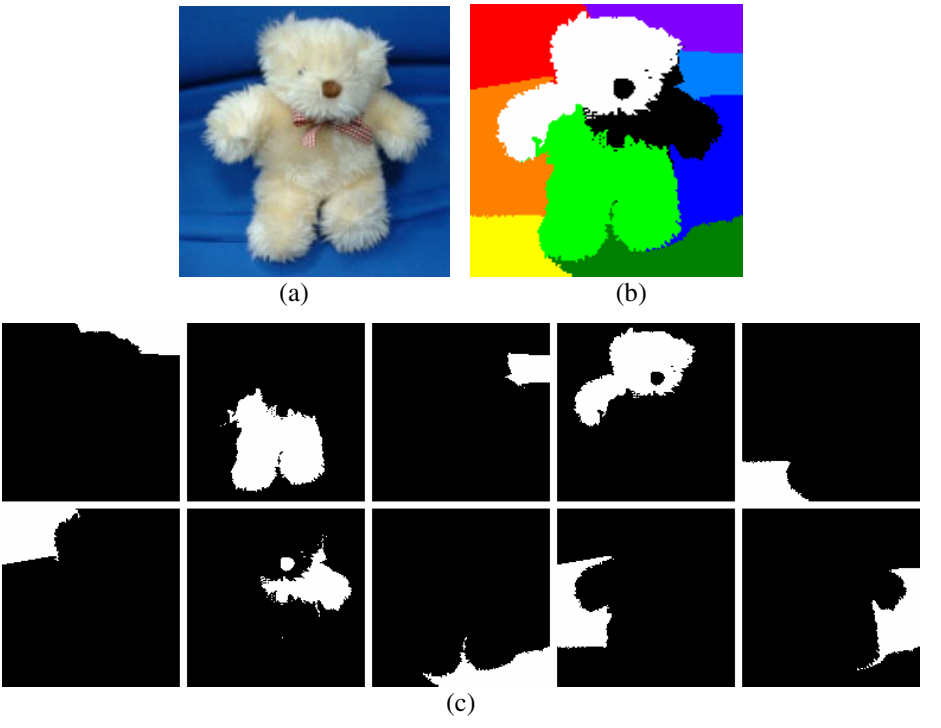


Fig. 1. Distinct component detection. (a) Input image; (b) clustering result; (c) distinct components.

In order to greatly increase the accuracy of alpha matte, the palette-based component classification is proposed to obtain the reliable foreground and background components. The palette-based component classification is described as follows. First, the boundary pixels (with a width of one pixel) of the input image are extracted to analysis components of RGB. If the R component is the maximum, then suppose the background is the red palette. Green and blue palettes are with the same assumption. Next, the RGB color space of distinct component is transformed to the HSV color space with Eq. (5), where $Max = \max(R,G,B)$, $Min = \min(R,G,B)$, and HSV color space is shown in Fig. 2.

$$H = \begin{cases} (G-B)*60/(Max-Min) & , \text{ if } R = Max \\ 180+(B-R)*60/(Max-Min) & , \text{ if } G = Max \\ 240+(R-G)*60/(Max-Min) & , \text{ if } B = Max \end{cases} \tag{5}$$

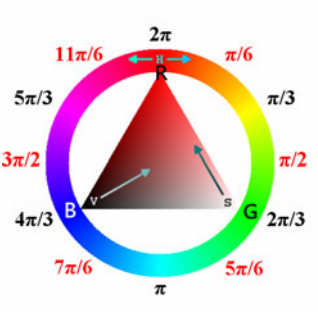


Fig. 2. HSV color space

Calculate the average hue $\tilde{H}$ of each distinct component. The candidate components of foreground and background are obtained using Eq. (6) and Eq. (7), respectively, where $\theta = \pi/6$ is set from experience.

$$\begin{cases} \pi-\theta \le \tilde{H} \le \pi+\theta & \text{, if pallette} = R \\ (5\pi/3)-\theta \le \tilde{H} \le (5\pi/3)+\theta & \text{, if pallette} = G \\ (\pi/3)-\theta \le \tilde{H} \le (\pi/3)+\theta & \text{, if pallette} = B \end{cases} \tag{6}$$

$$\begin{cases} 2\pi-\theta \le \tilde{H} \le 2\pi+\theta & \text{, if pallette} = R \\ (2\pi/3)-\theta \le \tilde{H} \le (2\pi/3)+\theta & \text{, if pallette} = G \\ (4\pi/3)-\theta \le \tilde{H} \le (4\pi/3)+\theta & \text{, if pallette} = B \end{cases} \tag{7}$$

These candidate components of foreground and background are checked again using the average value of the palette to obtain the final components of foreground and background. If the average value of the palette of the candidate component of the foreground is smaller than the average value of the palette of the input image, then it is the final component of the foreground. If the average value of the palette of the candidate component of the background is bigger than the average value of the palette of the input image, then it is the final component of the background. Using the proposed palette-based component classification, the distinct components are classified as the components of foreground, background and unknown regions. Fig. 3 is the result of component classification using the proposed palette-based component classification, where the components of the foreground and background are framed in red and blue, respectively, and the other are the components of unknown regions.

The corresponding matting components of foreground components, background components, and components of unknown region are obtained via 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. (8), where m^C denotes the indicator vector of the component *C* as defined in Eq. (9).

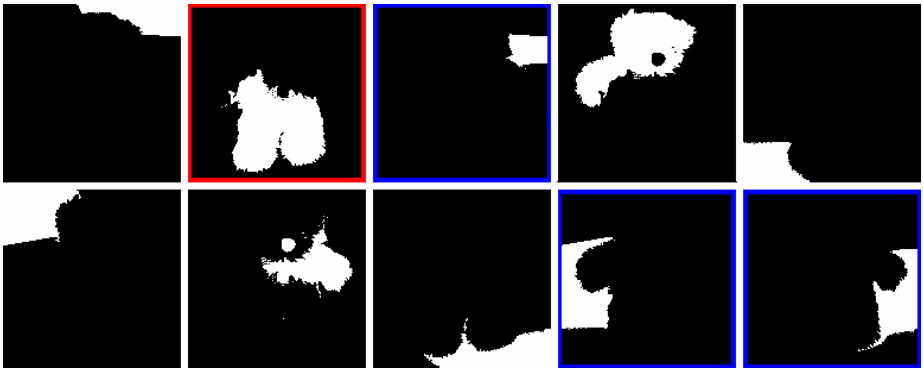


Fig. 3. Component classification

$$\alpha^k = \tilde{E}\tilde{E}^T m^{C^k} \tag{8}$$

$$m_i^C = \begin{cases} 1 & , i \in C \\ 0 & , i \notin C \end{cases} \tag{9}$$

Compute matting components by minimizing an energy function defined as Eq. (10) subject to $\sum_k \alpha_i^k = 1$. This goal is to find a set of K linear combination vectors y^k. The above energy function is minimized optimally using Newton's method, where γ is chosen to be 0.9 for a robust measure. Fig. 4 is the matting components of the resulting components.

$$\sum_{i,k} \left|\alpha_i^k\right|^\gamma + \left|1-\alpha_i^k\right|^\gamma, \text{ where } \alpha^k = \tilde{E} y^k \tag{10}$$

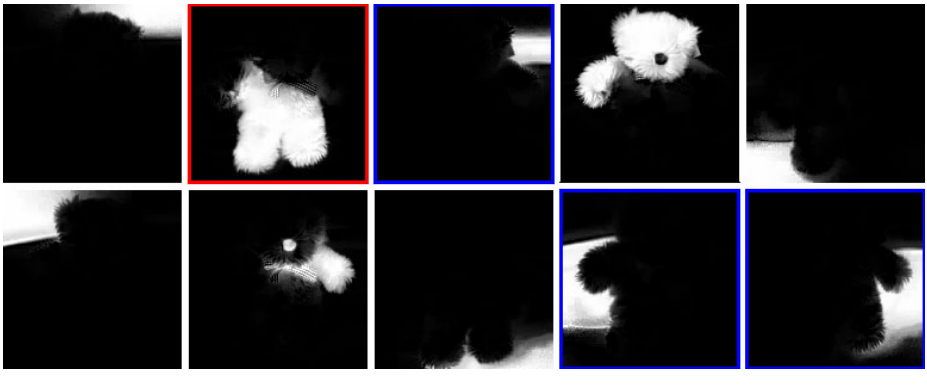


Fig. 4. Matting components of the resulting components

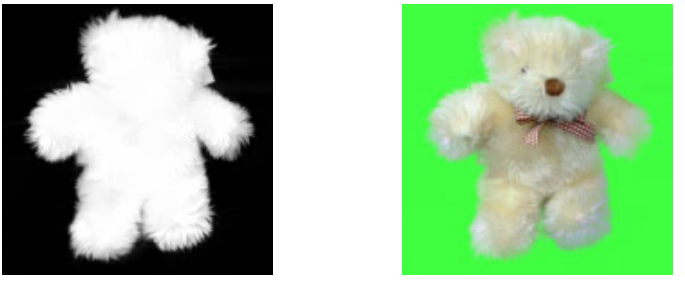
Finally, only matting components of the foreground and unknown region are combined to form the complete alpha matte based on minimizing the matte cost, as

defined in Eq. (11). In order to do this more efficiently, 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. (12). Then, the matting cost is computed by Eq. (13), where b is a K-dimensional binary vector indicating the selected matting components. Fig. 5 is the result of unsupervised image matting, where Fig. 5(a) is the alpha matte and Fig. 5(b) is foreground extraction with a constant-color background.

$$J(\alpha) = \alpha^T L \alpha \tag{11}$$

$$\Phi(k,l) = \alpha^{k^T} L \alpha^l \tag{12}$$

$$J(\alpha) = b^T \Phi b \tag{13}$$



(a) (b)

Fig. 5. Result of unsupervised image matting. (a) Alpha matte; (b) extracted foreground with a constant-color background.

It is worth mentioning that the matting component of the foreground region is included in the matte cost and that the matting component of the background region is removed from the matting cost, which greatly increases the accuracy of obtained alpha matte.

3 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.

In order to analyze the performance of alpha matte detection using spectral matting and the proposed method, the best five alpha mattes were obtained, with the best framed in red in the figures below. The ground truth of the alpha matte was obtained by manually selecting the matting components of the foreground.

Three tested images were used to evaluate the performance of the spectral matting and proposed image matting. Figs. 6-8 are a face image, a wolf image, and a fox image, respectively. Figs. 6(b)-8(b) are the clustering results, Figs. 6(c)-8(c) are the alpha matte detection using the spectral matting, and Figs. 6(d)-8(d) are the alpha matte detection using the proposed method.

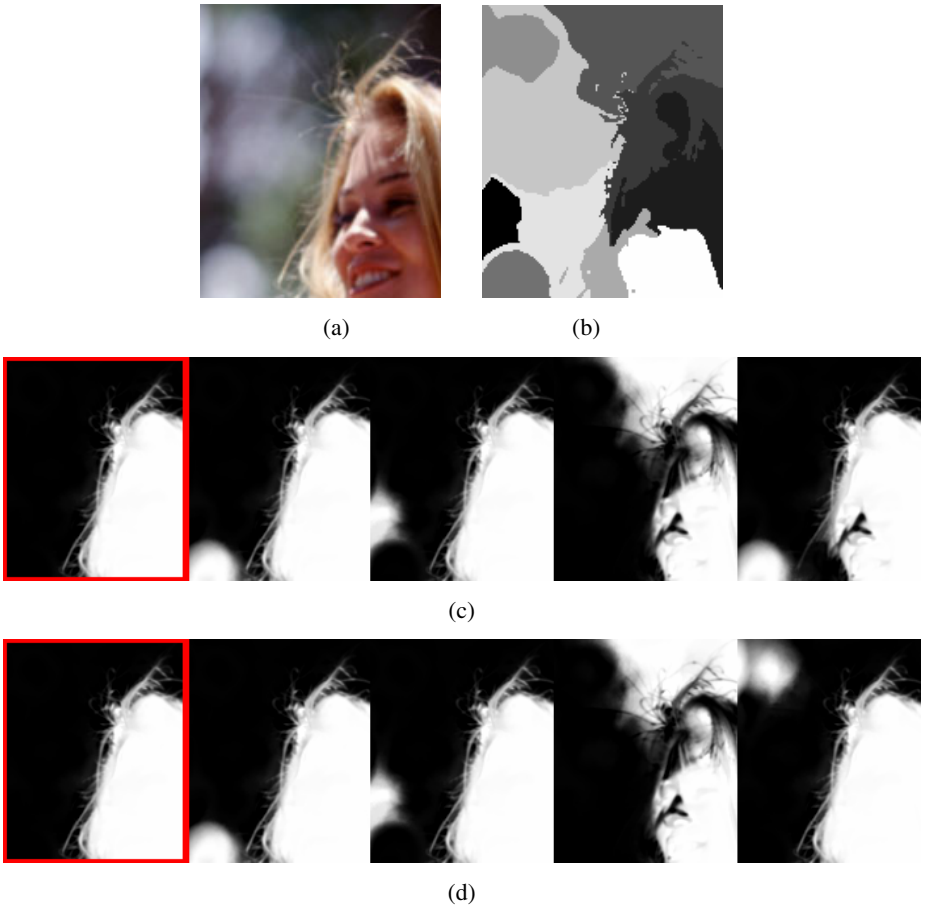


Fig. 6. Image matting of a face image. (a) Original image; (b) clustering result; (c) alpha matte detection using the spectral matting; (d) alpha matte detection using the proposed method.

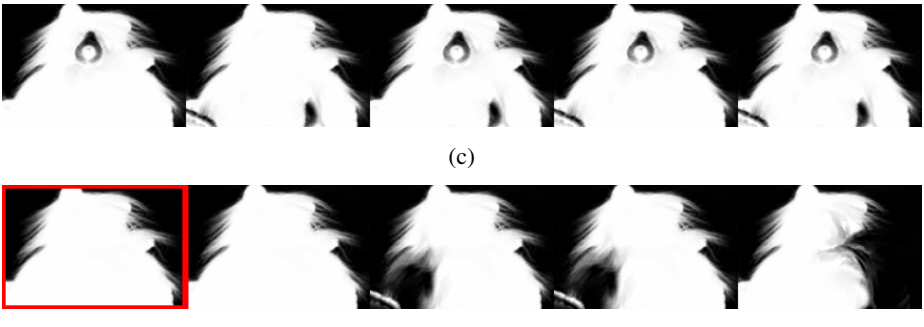
Fig. 7. Image matting of a wolf image. (a) Original image; (b) clustering result; (c) alpha matte detection using the spectral matting; (d) alpha matte detection using the proposed method.

(c)

(d)

Fig. 7. (*continued*)

(a) (b)



(c)

(d)

Fig. 8. Image matting of a fox image. (a) Original image; (b) clustering result; (c) alpha matte detection using the spectral matting; (d) alpha matte detection using the proposed method.

4 Conclusion

The modified spectral matting was proposed to obtain automatic and accurate image matting. In the proposed method, the palette-based component classification is proposed to obtain the reliable foreground and background components. Using the proposed palette-based component classification, the distinct components are further classified as the components of foreground, background and unknown regions. Next, the corresponding matting components of foreground components, background components, and components of unknown region are obtained via a linear transformation of the smallest eigenvectors of the matting Laplacian matrix. Finally, only matting components of foreground and unknown regions are combined to form the complete alpha matte based on minimizing the matte cost. Therefore, the accuracy of the alpha matte is greatly increased. Experimental results show that the proposed method can obtain the high-quality alpha matte for natural images without any user input, and the proposed method has better performance than the spectral matting.

Acknowledgments. This paper has been supported by the National Science Council, Taiwan, under grant no. NSC99-2221-E-346-007.

References

1. Porter, T., Duff, T.: Compositing Digital Images. In: Proceedings of ACM SIGGRAPH, vol. 18(3), pp. 253–259 (1984)
2. Wang, J., Cohen, M.F.: Image and Video Matting: A Survey. Foundations and Trends in Computer Graphics and Vision 3(2), 1–78 (2007)
3. Wang, J., Cohen, M.F.: Optimized Color Sampling for Robust Matting. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 1–8 (2007)
4. 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)
5. Levin, A., Lischinski, D., Weiss, Y.: A Closed-Form Solution to Natural Image Matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), 1–15 (2008)
6. Sun, J., Li, Y., Kang, S.B., Shum, H.Y.: Flash Matting. In: Proceedings of ACM SIGGRAPH, vol. 25(3), pp. 772–778 (2006)
7. Levin, A., Rav-Acha, A., Lischinski, D.: Spectral Matting. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10), 1–14 (2008)
8. Weiss, Y.: Segmentation using Eigenvectors: A Unifying View. In: Proceedings of International Conference on Computer Vision, pp. 975–982 (1999)
9. Yu, S.X., Shi, J.: Multiclass Spectral Clustering. In: Proceedings of International Conference on Computer Vision, pp. 313–319 (2003)
10. Ng, A., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm. In: Proceedings of Advances in Neural Information Processing System (2001)

Improved DCT-Based Watermarking through Particle Swarm Optimization

Chih-Chin Lai, Wei-Shun Wang, and Ci-Fong Jhan

Department of Electrical Engineering
National University of Kaohsiung
Kaohsiung, Taiwan 81148
cclai@nuk.edu.tw

Abstract. With the widespread use of computers and Internet, users have more chances to use multimedia data and digital contents. As a result, illegal reproduction of digital information started to pose a real problem. Digital watermarking has been regarded as an effective solution to protect various kinds of digital contents against illegal use. In this paper, a watermarking technique which applies the discrete cosine transformation and particle swarm optimization is presented. Experimental results demonstrate that the proposed technique is able to withstand a variety of attacks.

Keywords: watermarking; discrete cosine transformation; particle swarm optimization.

1 Introduction

The rapid expansion of the Internet and digital technologies in the past years have sharply increased the ease of the production and distribution of digital media. The phenomenon has lead to a matched ease in the illegal and unauthorized manipulation of multimedia products. Protecting the ownership of digital products while allowing a full utilization of the Internet resources becomes an urgent issue. One technical solution to make law enforcement and copyright protection for digital products possible and practical is digital watermarking. Digital watermarking is the process of embedding digital information called watermark into multimedia element, where the watermark remains detectable as long as the quality of the content itself is not rendered useless. The most important properties of watermarking include imperceptibility, robustness, and security.

A basic watermarking algorithm, an image for example, consists of a cover image, a watermark structure, an embedding algorithm, and an extraction or detection algorithm. There are numerous digital watermarking algorithms in the literature [5,7,10]. In general, there are two ways to embed the watermark based on the domain in which the watermark is inserted: spatial-domain and frequency-domain methods. Embedding the watermark into the spatial-domain component of the original image is the straightforward method. It has the advantages of low complexity and easy implementation. However, the spatial domain watermarking

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 21–28, 2010.

algorithms are generally fragile to image processing operations or other attacks. On the other hand, the representative frequency-domain techniques embed the watermark by modulating the magnitude of coefficients in a transform domain, such as discrete cosine transform (DCT), discrete Fourier transform (DFT), and discrete wavelet transform (DWT) [2,3]. Although frequency-domain methods can yield more information embedding and more robustness against many common attacks, the computational cost is higher than spatial-domain watermarking methods.

In the literature, several watermarking techniques based on particle swarm optimization have been proposed. Wang et al. [11] presented a blind watermark extracting scheme using the DWT and PSO algorithm. The watermark is embedded to the discrete multiwavelet transform coefficients larger than some threshold values, and watermark extraction is efficiently performed via particle swarm optimization algorithm. Aslantas et al. [1] proposed a fragile watermarking scheme based on DCT using PSO algorithm. In their method, embedding watermarks in frequency domain can usually be achieved by modifying the least significant bits of the transformation coefficients. After embedding process is completed, a number of rounding errors appear due to different domain transformation and PSO is used to correct these rounding errors. Lee et al. [9] presented a hybrid watermarking technique, in which the parameters of perceptual lossless ratio (PLR) for two complementary watermark modulations are first derived. Furthermore, a hybrid algorithm based on genetic algorithm (GA) and PSO is simultaneously performed to find the optimal values of PLR instead of heuristics. Zhu and Liu [13] proposed a watermarking scheme based on adaptive quantization index modulation and singular value decomposition in the hybrid DWT and DCT. The secret watermark bits are embedded on the singular values vector of blocks within low frequency sub-band in host image hybrid DWT-DCT domain. To embed watermark imperceptibly and robustly, they model the adaptive quantization steps by utilizing human visual system (HVS) characteristics and PSO algorithm. In this paper, we present a DCT-based watermarking technique and the PSO is also considered to improve the visual quality of the watermarked image and the robustness of the watermark. Experimental results show that the proposed approach has good performance against several attacks.

The rest of this paper is organized as follows. Section 2 reviews the basic concepts of DCT and PSO. Next, the proposed watermarking scheme is introduced in Section 3. Simulations of our technique with respect to attacks are conducted in Section 4. Finally, conclusions are given in Section 5.

2 Background Review

2.1 Discrete Cosine Transform

Discrete Cosine Transform is used to convert the time domain signal into the frequency domain. It is the basis for many image and video compression algorithms, especially the baseline JPEG and MPEG standards for compression of still and video images respectively. In all frequency domain watermarking schemes, there

is a conflict between robustness and transparency. If the watermark is inserted in the low-frequency components, the scheme tends to be robust to image processing attacks. On the contrary, if the watermark is inserted in the high-frequency components, it is easier to achieve high visual quality of the watermarked image. Therefore, applying the DCT to the image watermarking in the literature, an image is split up into pseudo frequency bands, and then the watermark is inserted into the middle band frequencies.

The two-dimensional DCT and inverse DCT of a block of $N \times N$ samples of a two-dimensional signal $f(x, y)$ are formulated as

$$C(u,v) = \frac{1}{\sqrt{2N}}\alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y) \times cos\left[\frac{(2x+1)u\pi}{2N}\right] cos\left[\frac{(2y+1)v\pi}{2N}\right], \quad (1)$$

$$f(x,y) = \frac{1}{\sqrt{2N}}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} \alpha(u)\alpha(v)C(u,v) \times cos\left[\frac{(2x+1)u\pi}{2N}\right] cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (2)$$

which $u, v = 0, 1, \ldots, N-1$, and $\alpha(k)$ are defined as

$$\alpha(k) = \begin{cases} \sqrt{\frac{1}{2}}, \text{for} \quad u = 0, \\ 1, \quad \text{otherwise.} \end{cases} \quad (3)$$

2.2 Particle Swarm Optimization

Particle swarm optimization (PSO) [4,8] algorithm is a new branch of evolutionary computation, which include stochastic search algorithms inspired by the mechanics of natural selection and genetics to emulate evolutionary behaviors in biological systems. The PSO is a population-based algorithm introduced by Kennedy and Eberhart, which is based on the simulation of simplified social models such as bird flocking, fish schooling and the swarming theory.

In general, the PSO contains a fixed-size population particles over the search space. Each particle represents a candidate solution to the considered optimization problem. The i-th particle has three attributes: (1) the current position in an N-dimensional search space $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N})$, (2) the current velocity $V_i = (v_{i,1}, v_{i,2}, \ldots, v_{i,N})$, and (3) each particle has its own best previous position $P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,N})$. In additional to individual information for each particle, the sharing information among conspecifics also plays a key role in searching solution. This can be achieved by employing the publicized knowledge P_g (the global best particle), which represents the best position found so far among all the particles in the swarm at generation t.

Initially, a population of particles is randomly created and then optimum is searched by increasing generations. In each generation, a particle profits from

the discoveries and previous experience of other particles during the exploration. Therefore, a new population is created based on a preceding one and the particles are updated by the following equations:

$$v_{i,j}(t+1) = wv_{i,j}(t) + c_1 r_1(p_{i,j} - x_{i,j}(t)) + c_2 r_2(p_{g,j} - x_{i,j}(t)), j = 1, 2, \ldots, N, \quad (4)$$

$$x_{i,j}(t+1) = x_{i,j}(t) + v_{i,j}(t+1), j = 1, 2, \ldots, N, \quad (5)$$

where w is called the inertia factor, for which value is typically setup to vary linearly from 1 to 0 during the iterated processing. Acceleration coefficients c_1 and c_2 are used to control how far a particle will move in a single iteration. Typically, these are both set to a value of 2.0 [6]. Two independent random numbers r_1 and r_2 are uniformly distributed in the range of $[0, 1]$. Velocity values must be within a range defined by two parameters v_{min} and v_{max}.

3 The Proposed Watermarking Scheme

3.1 The Watermark Embedding Procedure

The proposed watermark embedding is formulated as follows.

Step 1. The cover image was first partitioned into blocks with 8×8 pixels.
Step 2. Apply DCT technique to each block and then obtain DCT domain frequency bands.
Step 3. Apply z-score transformation to the watermark.
Step 4. Apply the PSO algorithm to determine the proper scaling factors in order to optimize watermark embedding process.
Step 5. Apply inverse DCT technique to produce the watermarked image.

Determining the proper values of multiple scaling factors is a difficult problem. Therefore, an efficient and powerful algorithm is required for this purpose. Here we use the PSO algorithm to automatically determine these values without making any assumption.

- Solution representation: In the PSO algorithm, a solution representation is needed to describe the population of interest. In our approach, the possible values of scaling factors are encoded as a particle. The multiple scaling factors should be in the form of diagonal matrix whose elements are only nonzero on the diagonal.
- Initial population: Usually, a collection of potential solutions are initialized in the beginning of the PSO process. Here, we randomly generate the particles in the initial population.
- Fitness function: Within the PSO-based methods, a fitness function is the survival arbiter for particles. The fitness function used in the proposed approach is defined as

$$f = \text{PSNR} + \sum_{i=1}^{K} \lambda_i \cdot Q_i, \quad (6)$$

where PSNR (Peak Signal-to-Noise Ratio) is defined in the following formula.

$$\text{PSNR} = 10 \cdot log_{10} \frac{255^2}{MSE}, \tag{7}$$

$$MSE = \frac{1}{WH} \sum_{i}^{W} \sum_{j}^{H} (x_{ij} - x'_{ij})^2. \tag{8}$$

Here, λ_i is the weighting factor for the Q value, the notations W and H represent the width and height of an image, x_{ij} is the pixel value of coordinate (i, j) in an original image, and x'_{ij} is the pixel value after the watermark embedding procedure. The symbol Q in the fitness function indicates a universal objective image quality index [12], and the definition is

$$Q = \frac{4\sigma_{xy}\overline{x}\,\overline{y}}{(\sigma_x^2 + \sigma_y^2)[(\overline{x})^2 + (\overline{y})^2]} \tag{9}$$

where

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \quad \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

$$\sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2,$$

$$\sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})^2,$$

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}).$$

Here x_i and y_i represent the original and processed image signal.

3.2 The Watermark Extracting Procedure

The watermark extraction sequence consists of the following steps:

Step 1. The watermarked image was first partitioned into blocks with 8×8 pixels.
Step 2. Apply DCT to the spatial domain pixels of each block to obtain DCT domain frequency bands.
Step 3. Extract the z-score value from each DCT transformed block.
Step 4. Apply inverse z-score transformation to produce the extracted watermark.

4 Experimental Results

In this section, some experiments are carried out to evaluate the performance of the proposed scheme. The image Bridge with size 256×256 and a 32×32

(a) (b) (c)

Fig. 1. (a) The cover image, (b) watermark, and (c) watermarked image ($PSNR = 48.82$)

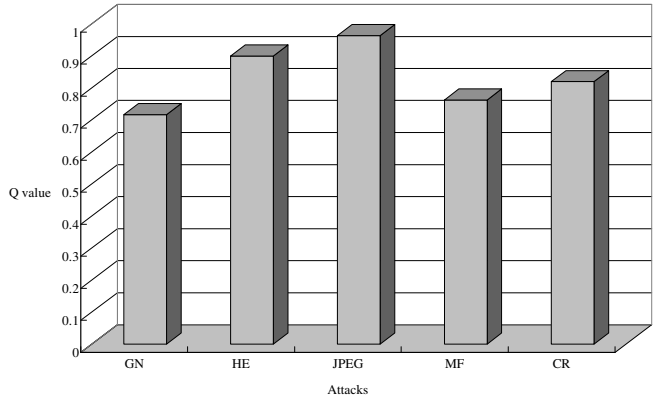


Fig. 2. Q values under different attacks

gray image Cameraman that are illustrated in Figure 1(a) and (b) were used as cover image and watermark in the experiments. Five common image processing attacks are applied to the proposed approach: Gaussian noise (GN), histogram equalization (HE), JPEG compression (QF = 75), median filtering (MF), and cropping (CR) on the border.

To evaluate the image quality of the watermarked and original images, the PSNR (Peak Signal-to-Noise Ratio) which is defined in Eq.(7) is used. From the Figure 1(c), we can find that the proposed approach can achieve high perceptual quality of the watermarked images.

Furthermore, the robustness is another requirement of the watermarking technique. The universal objective image quality index which is defined as Eq.(9) is used to measure the image quality of the extracted watermark. The value of Q

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

Fig. 3. Extracted watermarks under different signal processing attacks. (a) Gaussian noise, (b) histogram equalization, (c) JPEG compression, (d) median filtering, and (e) cropping.

is in the interval $[-1, 1]$. The value 1 means that the two images are exactly the same, and -1 means totally unrelated. This quality index models any distortion as a combination of three different factors: loss of correlation, luminance distortion, and contrast distortion [12]. The quality measurement results are listed in Figure 2 and the corresponding extracted watermarks are also shown in Figure 3. From the results, we found that in the cases of experiments, the extracted watermarks are visually recognizable and thus indicates that the proposed method is robust against such types of signal manipulation.

5 Conclusion

In this paper, a transformation-based image watermarking technique considering the DCT and the PSO algorithm is presented. Instead of directly embedding the watermark to the cover image, our approach is that applying z-score transformation to the watermark, and then insert the obtained z-score into the transform coefficients of the cover image. In order to efficiently find the proper values of scaling factors to control the strength of the embedded watermark, we used the PSO to achieve this goal. Experimental results show that both the significant improvement in imperceptibility and the robustness under attacks. Further work of considering the human visual system characteristics into our approach and other image quality metrics are in progress.

Acknowledgments. This paper is supported by National Science Council, Taiwan, under grant NSC 98-2221-E-390-027.

References

1. Aslantas, V., Ozer, S., Ozturk, S.: A Novel Fragile Watermarking Based on Particle Swarm Optimization. In: 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, pp. 269–272 (2008)
2. Barni, M., Bartolini, F., De Rosa, A., Piva, A.: Optimal Decoding and Detection of Multiplicative Watermarks. IEEE Trans. Signal Processing 51, 1118–1123 (2003)
3. Briassouli, A., Strintzis, M.G.: Locally Optimum Nonlinearities for DCT Watermark Detection. IEEE Trans. Image Processing 13, 1604–1617 (2004)
4. Cui, Z., Zeng, J., Sun, G.: A Fast Particle Swarm Optimization. International Journal of Innovative Computing, Informatin and Control 2, 1365–1380 (2006)

5. Cox, I.J., Kilian, J., Leighton, F.T., Shamoon, T.: Secure Spread Spectrum Watermarking for Multimedia. IEEE Trans. Image Processing 6, 1673–1687 (1997)
6. Eberhart, R.C., Simpson, P., Dobbins, R.: Computational Intelligence PC Tools. Academic, MA (1996)
7. Hartung, F., Kutter, M.: Multimedia Watermarking Techniques. Proceedings of the IEEE 87, 1079–1107 (1999)
8. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: IEEE International Conference on Neural Network, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
9. Lee, Z.-J., Lin, S.-W., Su, S.-F., Lin, C.-Y.: A Hybrid Watermarking Technique Applied to Digital Images. Applied Soft Computing 8, 798–808 (2008)
10. Nguyen, T.V., Patra, J.C.: A Simple ICA-Based Digital Image Watermarking Scheme. Digital Signal Processing 18, 762–776 (2008)
11. Wang, Z., Sun, X., Zhang, D.: A Novel Watermarking Scheme Based on PSO Algorithm. In: Li, K., Fei, M., Irwin, G.W., Ma, S. (eds.) LSMS 2007. LNCS, vol. 4688, pp. 307–317. Springer, Heidelberg (2007)
12. Wang, Z., Bovik, A.C.: A Universal Image Quality Index. IEEE Signal Processing Letters 9, 81–84 (2002)
13. Zhu, S., Liu, J.: A Novel Adaptive Watermarking Scheme Based on Human Visual System and Particle Swarm Optimization. In: Bao, F., Li, H., Wang, G. (eds.) ISPEC 2009. LNCS, vol. 5451, pp. 136–146. Springer, Heidelberg (2009)

A New Adaptive B-spline VFC Snake for Object Contour Extraction

Hoang-Nam Nguyen and An-Chen Lee

Mechanical Engineering Department
National Chiao Tung University
1001 University Road, Hsinchu, Taiwan 300
Namnguyen822003@yahoo.com, aclee@mail.nctu.edu.tw

Abstract. We propose a new adaptive B-spline VFC Snake model for object contour extraction. Bing Li et al. proposed vector field convolution (VFC) snake which has the advantages of superior noise robustness, reducing computational cost, and large capture range. However, it suffers from slow convergence speed due to large number of control points, as well as from difficulties in determining the weight factors associated to the internal energies constraining the smoothness of the curve. There is also no relevant criterion to determine the number of control points in VFC snake method. Our alternative approach expresses the curve as a non-uniform B-spline, in which fewer parameters are required and most importantly, internal energy calculation is eliminated because the smoothness is implicitly built into the model. A novel formulation of control points' movement estimation was established based on the least square fitting of non-uniform B-spline curve and VFC external force for the snake evolution process. A novel strategy of adding control points quickly matches the snake to desired complex shapes. Experimental results demonstrate the capability of adaptive shape description with high convergence speed of the proposed model.

Keywords: Computer vision, active contour, contour extraction, B-spline, B-snake.

1 Introduction

Snakes, or active contours, have been widely used for many active research areas in image analysis, computer vision, and medical imaging. Snake methods usually suffer from limited capture range, noise sensitivity and poor convergence speed due to large number of coefficients to be optimized. Many techniques have been proposed as an alternative method for original Snakes. An external force for snakes, called vector force convolution (VFC) [1] was introduced to overcome two key difficulties of snakes, capture range and the ability to capture concavities by diffusing the gradient vectors of an edge map generated from the image. Other techniques proposed alternative methods for presenting a curve: B-spline [3, 4], auto-regressive model [5], HMM models [6], and Wavelets [7], etc. The curve expressed as a parametric B-spline has the properties of built-in smoothness and fewer parameters. Furthermore,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 29–36, 2010.

the B-spline snake approach also naturally permits the local control of the curve by adjusting individual control points.

In this paper, we present an adaptive B-spline VFC snake model with a novel formulation of control points' movement estimation of a non-uniform B-Spine curve instead of dynamic programming approach. We also introduce a new strategy of adding control points to adaptively capture the complex object shape from the small number of control points in the initialized step. These novel properties are demonstrated by experimental results in section 4.

2 B-spline VFC Snake

2.1 Non-uniform B-spline VFC Snake

A traditional snake [8] is represented by a parametric curve $V(s) = [\, x(s) \;\; y(s) \,]$, $s \in [0,1]$; that deforms through the image to minimize the energy functional.

$$E_{snake} = \int_{s=0}^{s=1} \left[\frac{1}{2} \alpha (\left| V'(s) \right|)^2 + \beta (\left| V'' \right|)^2 + E_{ext} \right] ds$$

where α and β are weighting parameters representing the degree of the smoothness and tautness of the contour, respectively. $V'(s)$ and $V''(s)$ are the first derivative and second derivative with respect to s. E_{ext} denotes the external energy.

At the minima of above equation, the contour must satisfy the Euler–Lagrange equation:

$$\alpha V''(s) - \beta V'''' - \nabla E_{ext}(V(s)) = 0$$

The solution of above equation is obtained when the steady state solution of the following gradient descent equation occurs:

$$\frac{\partial V(s,t)}{\partial t} = \alpha V''(s,t) - \beta V''''(s,t) - \nabla E_{ext}(V(s,t))$$

The first term $F_{\text{int}}(V(s)) = \alpha V''(s) - \beta V''''$ is internal force and the second term $F_{ext}(V(s)) = -\nabla E_{ext}(V(s))$ is external force to attract the snake towards the desired FOI (feature of interest).

In the VFC snake method, Bing Li et al. proposed to change the standard external force by the VFC field, calculated by convolving the edge map generated from the image with the vector field kernel, as

$$\frac{\partial V(s,t)}{\partial t} = \alpha V''(s,t) - \beta V''''(s,t) - F_{vfc}(V(s))$$

Patrich Brigger et al. [9] proposed a B-spline snake, expressed as a B-spline curve $S(x)$, in which the smoothness is implicitly built into the model. A variable knot spacing between knot points is introduced to eliminate the need for internal energy. In other words, they control the elasticity of the spline implicitly by varying the spacing between the spline knots. The B-spline snake then deforms to minimize a cost function, which is the summation of external forces over the path of the curve sampled at M consecutive points:

$$\xi(c(k)) = \sum_{i=0}^{M-1} F_{ext}(S(i))$$

where $c(k)$ are B-spline coefficients (or control points) and $F_{ext}(S(i))$ is external force applied on the point $S(i)$. The variable knot spacing h is an integer, defined as:

$$h = \frac{M}{N}$$

where N is the number of control points.

Adopting from VFC and B-spline snake methods, we propose a new non-uniform B-spline VFC snake, expressed as a non-uniform B-spline $C(u)$, that deforms to minimize a cost function, defined as the summation of VFC external forces over the path of the snake sampled at M consecutive points

$$V(P) = \sum_{i=0}^{M-1} F_{vfc}(C(u_i)) \tag{1}$$

where P is the set of control points, $F_{vfc}(C(u_i))$ is VFC force applied on the sampled point $C(u_i)$. The smoothness of proposed snake is constrained by the variable knot spacing h. Increasing the number of control points will reduce the knot spacing, hence reducing the smoothing effect on the curve [9]. We will typically select M, h sufficiently large so that the curve points are connected, and modify M when adding control point to prevent the decrease of snake smoothness.

2.2 Non-uniform B-spline

For a given set of $n+1$ control points, $\{ P_i = [\, x_i \;\; y_i \,],\ i = 0,1,..,n \}$, the non-uniform B-spline curve: $[0,1] \rightarrow \Omega$ (2D image domain) of degree p is defined as [10]

$$C(u) = \sum_{i=0}^{n} N_{i,p}(u) P_i \tag{2}$$

where $N_{i,p}(u)$ is the i^{th} B-spline basis function of degree p (order $k = p+1$) recursively defined by

$$N_{i,p}(u) = \begin{cases} 1, & if \quad u \in [u_i, u_{i+1}) \\ 0, & otherwise \end{cases}$$

and

$$N_{i,p}(u) = \frac{(u-u_i)N_{i,p-1}(u)}{u_{i+p}-u_i} + \frac{(u_{i+p+1}-u)N_{i+1,p-1}(u)}{u_{i+p+1}-u_{i+1}}$$

where u_i , $i = 0,1,...,n+p+1$ are real numbers referred to as knots. The knot vector of the non-uniform B-spline curve $C(u)$ is defined as [10]

$$U = \{ u_0, u_1, ..., u_p = 0, u_{p+1}, ..., u_n \ \ u_{n+1} = 1, ..., u_{n+p+1} \}$$

2.3 Estimating the Control Points' Movement

To deform the proposed snake to the object boundaries, the cost function has to be minimized. The following equation shows the deformation of snake under the influence of VFC field:

$$\hat{C} = C^t(u) + \lambda F_{vfc}$$

where λ is a step-size constant, $C^t(u)$ is the deformable curve at iteration t and F_{vfc} is a set of VFC external forces applied on the M consecutive points sampled from the curve. Assume that $\hat{C}$ has $M = m+1$ discrete points: $Q = \{ Q_i(x_i, y_i) \mid i = 0,1,..., m \}$. Let t_i be the parameter of point Q_i, the chord-length method is used to parameterize the data points. We can fit the set of control points P^{t+1} to the discrete data set $\hat{C}$ using least square error method [10] as:

$$P^{t+1} = (N^T N)^{-1} N^T \hat{C}$$

where

$$N = \begin{bmatrix} N_0(t_0).....N_n(t_0) \\ ... \\ ... \\ N_0(t_m).....N_n(t_m) \end{bmatrix}$$

So we can estimate the movement of control polygon as

$$\Delta P = P^{t+1} - P^t = \lambda (N^T N)^{-1} N^T F_{vfc} \tag{3}$$

3 Complete Object Contour Extraction Algorithm

In order to extract the desired object contour, an iterative procedure is adopted. This algorithm guarantees the convergence at least to a local minimum of cost function $V(P)$.

(1) Calculate the edge image of original image by using Canny edge detector.
(2) Calculate the VFC field of the edge image as the external force of B-spline VFC snake.
(3) Initialize the set of control points P^0, M, h build the B-spline curve C^0 from P^0.
(4) Compute ΔP from equation (3); compute the change of cost function as:

$$\Delta V(P^t) = \left| \sum_{i=0}^{M-1} F_{vfc}(C^{t+1}(u_i)) - \sum_{i=0}^{M-1} F_{vfc}(C^t(u_i)) \right| \tag{4}$$

(5) If $\frac{1}{M}\Delta V(P^t) \geq T1$ ($T1$ is a predefined threshold), then:

- Set $P^{t+1} = P^t + \Delta P$
- Build C^{t+1} from P^{t+1}
- Go to step (4)

If $\frac{1}{M}\Delta V(P^t) <$ go to step (6)

(6) Check the following condition for all B-spline segments $i = 0,1,...,n$:

$$\frac{1}{N_i} \sum_{u=u_{p+i}}^{u=u_{p+i+1}} \overline{F}_{vfc}(C(u)) > T2$$

where $\overline{F}_{vfc}(C(u_i))$: magnitude of VFC force applied on the point $C(u_i)$, N_i is the number of forces in the i_{th} segment. $T2$ is a predefined threshold.

If the above condition is satisfied then

- Add control point for this segment at $\tilde{u}_i$, where the normal component of VFC force is the maximum of i^{th} segment.
- Build new C^t from new P^t
- Go to step (4)

If no condition is satisfied, then stop the algorithm, last P^t is regarded as the final result.

There is no exact rule to set the thresholds. They depend on the image resolution, convergence speed, and the acceptable error. According to our experience, for image of size 64x64 and $\lambda = 0.5$ T1 and T2 are set to 0.1 and 0.5, respectively. It is the best tradeoff between computational time and matching accuracy.

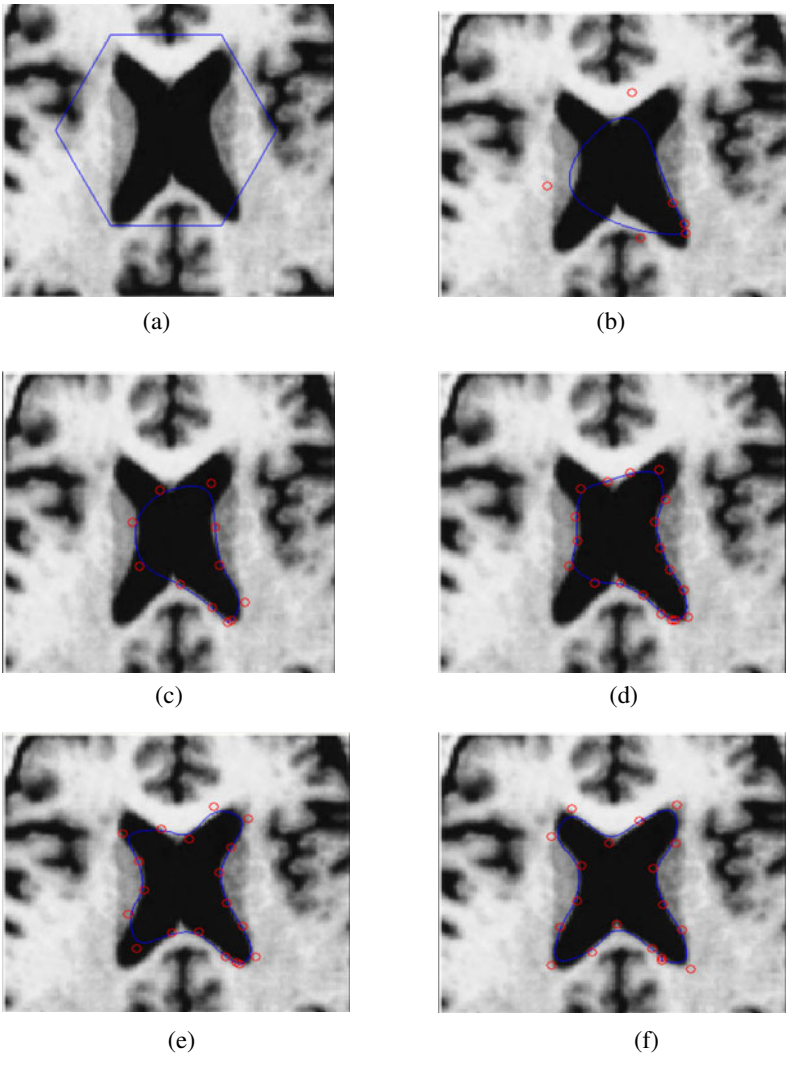


Fig. 2. Results of brain ventricle extraction

4 Results

The proposed snake model was simulated in MATLAB 7 environment using an Intel Core 2 Duo, 3.0 GHz processor with 2 GB of memory, and under Microsoft Windows XP operating system.

The B-spline VFC Snake model was tested on MR medical image of size 256x256 for human brain ventricle contour extraction and a maple leaf to demonstrate the capability of adaptive shape description and object contour extraction. All the images were initialized with 6 control points. The proposed extraction algorithm adaptively increases the number of control points to match the desired contour exactly in the extracting process. The overall processing times without any code optimization for Fig. 2 and Fig. 3 are 0.54, 0.31 second, respectively. The RMSE of Fig.2 and Fig.3 are 1.1 and 0.4 pixel.

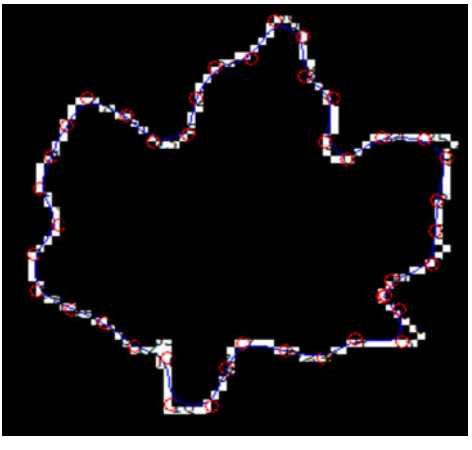


Fig. 3. Extracting maple leaf contour

5 Conclusions

We presented a new adaptive B-spline VFC Snake for object contour extraction. Movement of control points is determined based on the least square error approach and Vector Field Convolution forces.ïVFC external forces does not only pull the B-spline curve toward the contour but also play the role as an effective measure for adding controls algorithm. Furthermore, the proposed algorithm can adaptively add multiple control points at each iterative step during the deformation procedure to increase the convergence speed. The proposed strategy for curve evolution and the low computational cost, improve the overall performance effectively. Future work focuses on the use of non-uniform rational B-spline to facilitate the local control of the snake shape without increasing the number of control points.

References

1. Li, B., Acton, S.T.: Active Contour External Force Using Vector Field Convolution For Image Segmentation. IEEE Transactions on Image Processing 16, 2096–2106 (2007)
2. Meegama, R., Rajapakse, J.: Nurbs Snakes. J. Image Vis. Comput. 21, 551–562 (2003)
3. Huang, Z., Cohen, F.S.: Affine-Invariant B-spline Moments For Curve Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 5, 1473–1480 (1996)
4. Wang, Y., Teoh, E.K., Shen, D.: Dynamic B-snake model for complex objects segmentation. J. Image Vis. Comput. 23, 1029–1040 (2005)
5. Paulik, M.J., Das, M., Loh, N.K.: Nonstationary Auto-regressive Modeling of Object Contours. IEEE Transactions on Signal Process 40, 660–675 (1992)
6. Fred, A.L.N., Marques, J.S., Jorge, P.M.: Hidden Markov Models vs Syntactic Modeling in Object Recognition. In: International Conference on Image Processing (ICIP 1997), Washington DC, USA, pp. 893–896 (1997)
7. Tueng, Q.M., Boles, W.W.: Wavelet-based Affine Invariant Representation: A Tool for Recognizing Planar Objects in 3D Space. IEEE Transactions on PAMI 19, 846–857 (1997)
8. Kass, M., Witkin, A., Terzopoulos, D.: Snakes-Active Contour Models. Int. J. Comput. Vis. 1, 321–331 (1987)
9. Brigger, P., Hoeg, J., Unser, M.: B-spline Snakes: A Flexible Tool for Parametric Contour Detection. IEEE Transactions on Image Processing 9, 1484–1496 (2000)
10. Piegl, L., Tiller, W.: The NURBS book. Springer, New York (1997)

Algorithm for Computational Measure of Color Constancy

S.J. Jerome Teng

Department of Computer Science and Information Engineering
Kainan University, Taoyuan, Taiwan, ROC

Abstract. Color constancy (CC) is the ability to perceive or retrieve constant image colors despite changes in illumination. CC has long been a research subject in color and machine vision. This paper presents a computational algorithm that offers an optimized solution for CC. The proposed CC algorithm exploits the strategy of RGB channel gain variation and the suppressing mechanism of grayscale pixel maximization (GPM). For most natural scenes under a single illuminant, the CC offset gain location can be revealed with a distinct GPM peak. The optimization scheme provides 2D and 3D graphical illustrations for user-friendly visualization and analysis. Operating mechanism and experimental results of this algorithm are clearly illustrated.

Keywords: color constancy, white balance, computer vision, grayscale pixel maximization.

1 Introduction

The human visual system has the capability of seeing roughly constant colors from object under different illuminations, while video cameras and image sensors do not. This inherent ability of color adaptation by adjusting the spectral response and discounting the illuminant changes is known as color constancy (CC) [1, 2]. Color constancy is linked to chromatic adaptation in which our visual system adjusts its sensitivity according to the context of the illuminated scene. The mechanism tends to discount the effect of the illuminant. However, human color constancy is possible only under a limited range of illuminants [3]. For example, human CC fails in scenes under unnatural illumination such as a sodium-vapor light at night. Although human color constancy has been studied extensively, its mechanisms are still not yet well-understood [4].

Machine vision has no such inherent capability of chromatic adaptation. Sensor captured images can vary substantially or have an undesirable color cast due to changes of illuminant. As a result, color constancy plays an important role in color and machine vision. Many computer vision tasks, such as object recognition, tracking, and surveillance, require CC. The goal is to acquire an illuminant invariant image by estimating and correcting the external illuminant. The methods to correct the color inconstancy are also called color correction, color balancing, or white balancing (WB). For example, WB is a key function to remove undesirable color cast in digital

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 37–46, 2010.

cameras. Such a WB process needs image analysis, cast evaluation, and RGB channel-gain adjustment.

There are different theories and experiments that have tried to interpret the human color constancy [1-5]. Likewise, many algorithms and techniques for solving the machine CC problem have been proposed [6-11]. Machine CC is often considered as a complex, under-constrained, and ill-posed problem, because the object intrinsic color and illuminant color can not be uniquely separated. The spectral sensitivities of the RGB detector channels add additional complexity to these already nonlinear relations. Thus, an accurate computational CC algorithm is considered difficult because of the nonlinear complexities involved.

Most color constancy approaches rely on some assumptions [8]. Existing CC algorithms can be roughly classified into two categories, namely, the global algorithms and the local algorithms. The global algorithm imposes constraints on the scene and/or the illuminant, and uses all pixels in an image for illuminant estimation. A good example is the widely used gray-world (GW) assumption [11, 8], which is also the basis of some other algorithms [12]. The GW algorithm assumes the average color or reflectance within a scene is achromatic. Subsequent equalization adjustment of the image RGB channel gains obtains color constancy. The GW approach is convenient to use, and comparatively accurate if the external color cast is not high, but it fails when dominant intrinsic colors exist in the scene.

On the other hand, the local algorithm uses statistical data and/or those pixels which satisfy some specified conditions [8, 9]. Most machine CC approaches do not always achieve good performance, and they cannot handle situations with more than one illumination present in the scene [10]. Theoretically, given a natural scene and a single illuminant, there should be only one optimized solution for CC. It is desirable to find a CC algorithm that is robust enough to reveal the likely illuminant in a scene and complete the subsequent CC process. Thus, a good CC algorithm, in the first place, should have the robustness and precision of becoming a CC measure. Without such a CC measure, the target of CC solutions becomes an illusive one and comparisons among them are less than meaningful.

This paper presents a cast evaluation method and CC measure which has the characteristics of a numerical analysis and optimization scheme. The proposed algorithm exploits the scanning strategy of RGB channel gain adjustment and the suppressing mechanism of grayscale pixel maximization (GPM). The human color vision is trichromatic similar to that of the RGB model. Human vision perceives color in a 3D fashion, owing to the presence of three types of color receptors. This algorithm computes, compares, and summarizes the status of all the pixels in terms of a saturation threshold in HSV color space. Using the GPM algorithm, the amount of RGB channel-gain adjustment required to achieve CC is graphically indicated.

The GPM method uses a 3D graphical scheme established on a 2D RGB channel-gain variation platform. The method provides 2D and 3D graphical illustrations for user-friendly visualization and analysis. The accumulated numbers of screened grayscale pixels as a function of channel gain variation are graphically displayed. The gain location of the external color cast appears as a prominent spike in the 2D gain variation platform. As a CC measure, the proposed method can evaluate natural images for the precise gain location of the external color cast. The measured amount of color cast in image can be so small that our naked eye is difficult to tell the

difference after adjustment. In this paper, the operating mechanism of this algorithm is described in Section 2. Section 3 shows the experimental results and discussion. Section 4 is the conclusion.

2 Proposed Algorithm

Human color vision is trichromatic and a set of 3D tristimulus value can specify a color signal. The RGB color model corresponds most closely to the nature of the human vision. The three photoreceptors in human eye are tuned approximately to red, green, and blue light (λ = 564, 533, 437nm). The detected RGB components as a function of illuminant wavelength can be expressed as the following

$$R,\ G,\ B \ \propto \ \int_{400}^{700} I(\lambda)\, O(\lambda)\, S_{R,G,B}(\lambda)\, d\lambda \tag{1}$$

where λ is wavelength, I(λ) is the illuminant spectral density. O(λ) is the object spectral density. $S_{R,G,B}(\lambda)$ is the spectral sensitivity of the RGB detector channel.

After an image is captured by a RGB sensor, all the above parameters are factored in and difficult to decorrelate. The above sensor equations are extremely nonlinear. They do not have a significant role in the CC analysis. Human vision system adaptively adjusts the cone sensitivity of each channel to achieve chromatic adaptation. Similarly, for an RGB image, the color correction of an overall color cast is done by adjusting the RGB filters or channel gains. This model is similar to the von Kries hypothesis or coefficient law, first proposed in 1878, that the chromatic adaptation is an independent gain regulation of the three cone signals L, M, S, through three different gain coefficients [13, 11].

These indicate that, at least in human vision, the CC is far from a complicated mathematical issue. For a single illuminant, machine CC can be achieved none other than by adjusting the RGB channel gains of the image. Thus, a robust cast evaluation method needs to incorporate the scheme of the RGB gain variation. For visualization purpose, the gain variation is preferred in a scanning fashion. This is the first feature of the proposed CC algorithm. Equations 2 and 3 describe the relations of channel gain adjustment between two RGB images I_o and I_1, where α, β, and γ are the RGB channel gain scaling coefficients, respectively.

$$\left(I_1\right) = \left(\left[R_1\right], \left[G_1\right], \left[B_1\right] \right) \tag{2}$$

$$\left(I_o\right) = \left(\alpha \cdot \left[R_1\right], \beta \cdot \left[G_1\right], \gamma \cdot \left[B_1\right] \right) \tag{3}$$

Using RGB adjustment, a large variation in image lightness may create distortions of image colors. It is desirable to have normalized RGB gain adjustments so that the lightness levels of images I_o and I_1 remain unchanged. Equations 2 and 3 can be rewritten as Equations 4 to 6, where r, g, b, and K_{rgb} are the gain adjustment ratios and the normalization factors, respectively. In the actual operation of gain adjustment, for example, the value of 'b' can remain as one or constant, only one 2D gain variation platform is required. Notice that no expression of matrix transform, as used in linear algebra or color space transform, is required in Equ. 4 to 7.

$$[R_o] = r \cdot K_{rgb} \cdot [R_1] \tag{4}$$

$$[G_o] = g \cdot K_{rgb} \cdot [G_1] \tag{5}$$

$$[B_o] = b \cdot K_{rgb} \cdot [B_1] \tag{6}$$

$$[I_o] = ([R_o], [G_o], [B_o]) \tag{7}$$

To make this algorithm an optimization scheme, a numerical analysis designed to create meaningful local maxima or minima is required. Natural scenes usually have, more or less, some grayscale or near grayscale pixels in them. These pixels do not have to be whitish pixels or pixels having high lightness values. Inside a RGB color cube, these pixels are located near the diagonal grayscale line. In terms of HSV color model, these pixels have small saturation values. It can be easily observed that a CC image tends to have more such grayish pixels than an image that has a color cast [6, 8, 10].

The second feature of this algorithm is to utilize the maximization effect of grayscale pixels in natural images when a canonical illuminant is applied. By the same token, the number of grayscale pixels is reduced drastically as a color cast is applied to an image. An external color cast shifts the image color and suppresses the number of grayscale pixels. RGB gain adjustment exerts a similar effect on these pixels. A CC process is essentially adjusting the RGB channel gain to compensate for the external cast. The key is to utilize this effect in a systematic fashion. Utilizing these features and the RGB channel-gain scanning, a basic structure of an effective cast evaluation method is thus created.

Addition features of this algorithm include RGB-HSV color space transformation and graphical schemes to illustrate and analyze the gain-response relation. The offset gain required for CC can be easily visualized using the GPM diagrams. The distribution of grayscale pixels on a 2D RG gain platform reveals the location of likely illuminant. For a color image of size (M, N), the proposed algorithm can be mathematically expressed in the following.

$$f(r,g) = \frac{100\,\%}{M \times N} \left\{ \sum_{S<S_{TH}}^{M} \sum_{S<S_{TH}}^{N} S[m, n; r, g] \right\} \tag{8}$$

$$S[m, n; r, g] = \begin{cases} 1 & , \; if \;\; S < S_{TH} \\ 0 & , \; if \;\; S \geq S_{TH} \end{cases} \tag{9}$$

where f(r,g) is the normalized 3D distribution of grayscale pixels (%), as a function of gain adjustment ratios ‘r’ (red) and ‘g’ (green). The ‘b’ (blue) components remain constant. The saturation values ‘S’ are calculated from RGB-HSV color space transform, where the RGB values are firstly obtained after the normalized r-g gain adjustment. The S values are then screened using a preset threshold value ‘S_{TH}’.

Highly-saturated or vivid color pixels in an image usually are not desirable for CC algorithms based on the GW assumption. However, these pixels tend to have

negligible contribution to the GPM. This algorithm has no difficulty in dealing with such a situation, as long as some grayscale pixels exist in the scene. Another feature of this proposed algorithm is the capability of revealing multiple illuminants in different regions of an image. This is because the superposition principle works in this algorithm. Evaluations showing multiple GPM peaks require further gain-adjusted analysis.

3 Experimental Results and Discussion

To demonstrate the performance of the GPM operation and CC measure, two natural images and their experimental results are reported. One image had a slight color cast, while the other had an obvious color cast. Firstly, as shown in Fig. 1, the publicly available color image 'Barbara' was used. Fig. 2 shows the white pixels in the image which are below a HSV saturation threshold of 0.05. The presence of these grayscale pixels in Fig. 2 indicates this image is at or near CC. The 3D contour plot of calculated grayscale pixels (%) vs. r-g gain variation is shown in Fig. 3. The MATLAB software was used. The r-g gain ratios were in the 0.5-1.5 range, or -/+ 50%. The saturation threshold S_{TH} (0.05) is a variable that determines the sensitivity of the grayscale screening. A smaller S_{TH} makes the peak lower but the spike sharper. Fig. 3 shows a single dominant GPM spike. The CC offset gain location is very likely at this point. As shown in Fig.3, the numbers of grayscale pixels in the remaining r-g gain platform were drastically suppressed.

For better viewing of the CC gain location, Fig. 4 shows the 2D contour plot or top view of Fig. 3. The peak is slightly below the central or unit gain point. The CC gain location is still not clearly indicated in Fig. 4. For even better viewing and analysis, two cross-sectional r- and g-gain response profiles at the GPM peak are used, as shown in Fig. 5 and 6. The circles were calculated data and cubic spline interpolations were used to complete the curves. The central vertical lines in Fig.5 and 6 correspond to the unit gain points.

The horizontal shift from unit gain point measures the r- and g-gain adjustments required for CC. In this case, they are about (-0.08, -0.01), or (-8%, -1%). A slight reddish cast is thus quantitatively measured in this image. In other words, to achieve CC, the r- and g-gains in Fig. 1 are to reduce by 8% and 1%, respectively. Secondary or multiple peaks may occur in the GPM plot. These situations can be further analyzed by correlating the gain locations with their corresponding image pixels. For example in Fig. 4, a nearby secondary peak occurs at about (-0.22,-0.03). The pixels of this secondary peak are mainly located at the rug area between the table's leg and Barbara's arm. It may be caused by a different illuminant, or other factors.

The obtained image after the (-8%, -1%) r-g gain adjustment is shown in Fig. 7. For a small channel gain adjustment, such as 2-3%, the changes in image may not be noticeable. In this case, Fig. 7 looks better white balanced and a very slight reddish cast is removed. Fig. 8 confirms that there are more white pixels in Fig. 7, which are below the saturation threshold S_{ST}=0.05, compared to that of Fig. 2. Again, applying the GPM analysis on Fig. 7, the contour plot and evaluation profiles will show that the CC gain location is now at the unit gain. The important aspect of this result is that the computational CC measure and the specific adjustment are achieved using numerical analysis and optimization, not by trial and error, nor algorithms of vague generality.

Fig. 1. Color Image 'Barbara'

Fig. 2. White pixels are below S_{TH}.

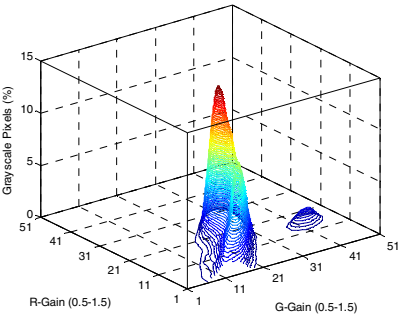


Fig. 3. Contour plot of GPM

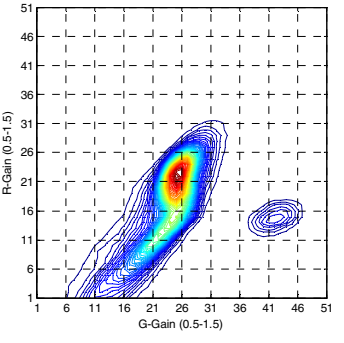


Fig. 4. 2D Contour plot of GPM (Barbara)

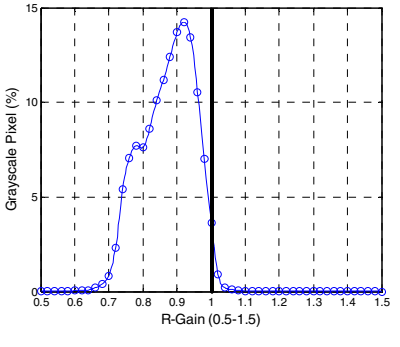


Fig. 5. The R-gain response diagram

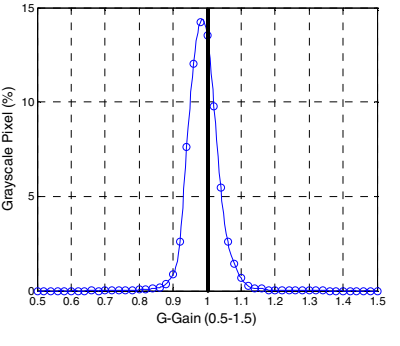


Fig. 6. The G-gain response diagram at GPM

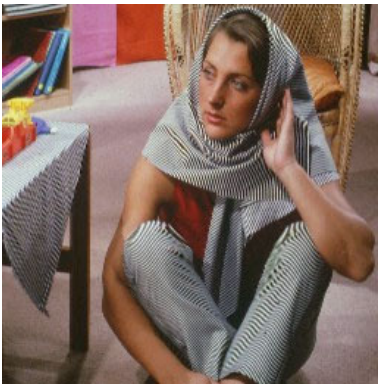


Fig. 7. White balanced image

Fig. 8. White pixels are below S_{TH}

The second set of experimental results is shown in Fig. 9 to Fig. 18. The color image 'peppers' (Fig. 9) was used for evaluation. The image contains vivid colors and looks having a reddish color cast. Its pixel distribution in 2D YCbCr space is shown in Fig. 10. The elliptical circle plotted in Fig. 10 was delineated by the mean values of Cb and Cr components and their respective standard deviations. The neutral point of the CbCr plot is located outside the elliptical circle. This indicates that the image is likely having a color cast. Similar CC criteria have been reported [7, 9]. Due to the vivid colors, the area of pixel distribution still covers the neutral point, but very few of them are grayscale pixels. The image's 3D GPM contour plot shows a very sharp spike, as shown in Fig. 11. Fig. 12 shows the 2D top view of Fig. 11. Having only one sharp GPM spike, the image's CC gain location is clearly indicated. The two r-g gain response profiles at the GPM are shown in Fig. 13 and 14. The measured r-g gain adjustment required for CC is about (-0.30, -0.16). In other words, in order to achieve CC, the red/green channel gains of this image should be reduced by 30% and 16 %, respectively.

The obtained CC image is shown in Fig. 15 for comparison. The pixel distribution in the CbCr plot is shown in Fig. 16. Notice that the neutral point of the CbCr plot is still located outside of the elliptical circle, due to the image's intrinsic color cast. The GW algorithm does not work for this image. Fig. 17 and 18 display the 3D and 2D contour plots of the CC image. As predicted by the GPM algorithm, the CC gain location of the new image is now aligned with the unit gain point, as shown in Fig. 18. An accurate CC measure for up to 30% channel gain adjustment is thus demonstrated. It is observed that the base area of the spike in Fig. 18 becomes larger, as compared to that of Fig. 12. This change reflects the amount of distortion due to the application of von Kries coefficient law. For channel gain adjustment such as less than 30-40%, the chromatic adaptation and CC gain prediction using the GPM method generally works well. For images having a stronger color cast and needing a channel gain adjustment larger than such as 50%, the prediction of CC gain location using the GPM method becomes less accurate or erratic. The color distortion also increases and may become obvious. The von Kries coefficient law becomes insufficient for CC operation.

Fig. 9. Original color image ‘peppers’

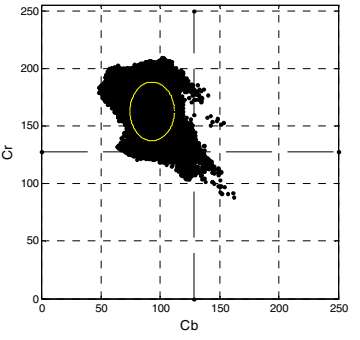


Fig. 10. Image pixel distribution in YCbCr space

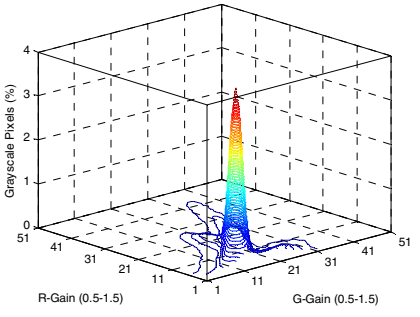


Fig. 11. 3D Contour plot of GPM (‘peppers’)

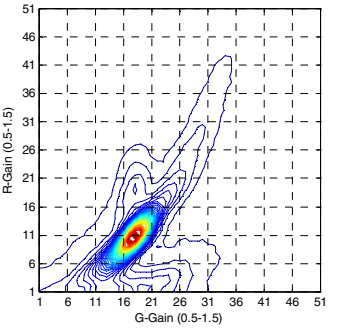


Fig. 12. 2D Contour plot of GPM

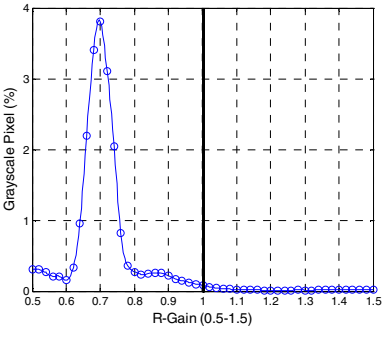


Fig. 13. The R-gain response at GPM

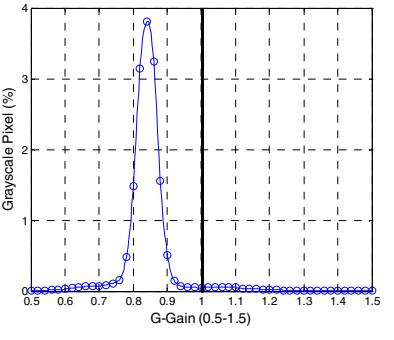


Fig. 14. The G-gain response at GPM

More than 120 natural images have been evaluated using this algorithm. Many images were selected because they contained vivid colors. Such images usually do not satisfy the gray world assumption and thus more difficult for CC algorithms to work

Fig. 15. White balanced image 'peppers'

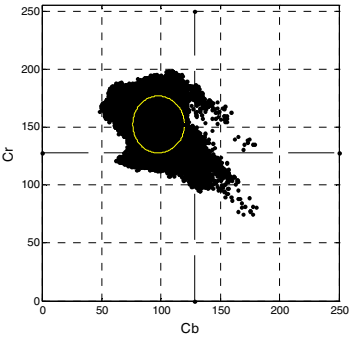


Fig. 16. Image pixel distribution in YCbCr space

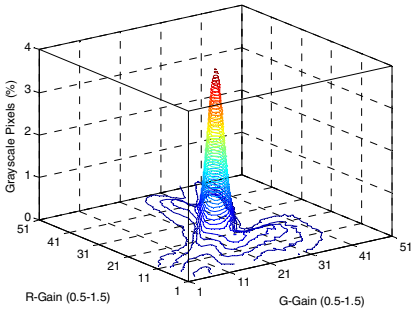


Fig. 17. 3D Contour plot of GPM

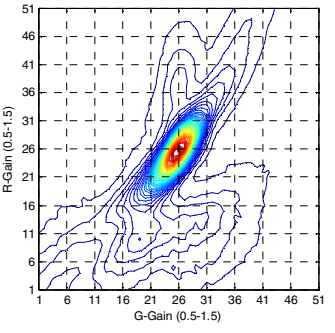


Fig. 18. 2D Contour plot of GPM.(CC)

well. The requirement for the proposed CC measure to work is the presence of some grayscale pixels in the scene. For most images, a dominant GPM peak usually exists in the diagram. Some images may have multiple peaks in the GPM diagram. Depending on the circumstance, further gain-vs-pixel analysis may be required to find out the true CC gain location. In some uncommon cases, the CC gain location may appear as a smaller GPM peak. False peaks or heightened response areas are usually created by large areas of lightly colored pixels, such as light skin color or the like. Secondary peaks may also be created by a different and separate illuminant in the scene. This is a subject beyond simply using the von Kries coefficient law on the whole image.

For most natural scenes, a certain amount of grayscale pixels is always present. The grayscale pixels in the scene need not be whitish pixels or patches for the algorithm to work. The pixel screening is carried out in the HSV color space. Pixels are screened based on a saturation threshold value. This algorithm has no difficulty in dealing with images having vivid or dominant intrinsic colors, as long as some grayscale pixels exist in the scene. Vivid color pixels have high saturation component in HSV space. They would least affect the GPM operation, because they are less likely to become grayscale pixels during the channel gain variation. An effective and accurate CC measure based on an acquired image is thus demonstrated.

4 Conclusion

An algorithm for illuminant estimation and color constancy (CC) is presented. The proposed CC measure exploits the scanning strategy of RGB channel gain adjustment and the suppressing mechanism of grayscale pixel maximization (GPM). It turns the CC algorithm to become a numerical optimization scheme where local maxima are used to find the CC gain solution. The algorithm provides 2D and 3D graphical illustrations for user-friendly visualization and analysis. It is demonstrated as an effective computational measure of color constancy. For most natural scenes having a single illuminant, especially those not having a very large color cast, the CC gain location can be accurately revealed and predicted. Operating mechanism and experimental results of this algorithm have been demonstrated.

References

1. D'Zmura, M., Lennie, P.: Mechanisms of Color Constancy. J. Opt. Soc. Am. A 3, 1662–1672 (1986)
2. Brainard, D.H.: Color Constancy. In: Chalupa, L., Werner, J. (eds.) The Visual Neurosciences, vol. 1, pp. 948–961. MIT Press, Cambridge (2004)
3. Kraft, J.M., Brainard, D.H.: Mechanisms of Color Constancy under Nearly Natural Viewing. Proc. Natl Acad. Sci. USA 96, 307–312 (1999)
4. Kuriki, I., Uchikawa, K.: Limitations of surface-color and apparent-color constancy. J. Opt. Soc. Am. A 13, 1622–1636 (1996)
5. Smithson, H.E., Zaidi, Q.: Colour Constancy in Context: Roles for Local Adaptation and Levels of Reference. J. of Vision 4(9), 693–710 (2004)
6. Liu, Y.C., Chan, W.H., Chen, Y.Q.: Automatic White Balance for Digital Still Camera. IEEE Trans. on Consumer Electronics 41(3), 460–466 (1995)
7. Cooper, T., Tastl, I., Tao, B.: A Novel Approach to Color Cast Detection and Removal in Digital Images. Proc. of SPIE, vol. 3963, pp. 167–177 (2000)
8. Barnard, K., Cardei, V., Funt, B.V.: A Comparison of Computational Color Constancy Algorithms - part I: Methodology and Experiments with Synthesized Data. IEEE Transactions on Image Processing 11, 972–984 (2002)
9. Gasparini, F., Schettini, R.: Color Balancing of Digital Photos using Simple Image Statistics. Pattern Recognition 37(6), 1201–1217 (2004)
10. Agarwal, V., Abidi, B., Koschan, A., Abidi, M.: An Overview of Color Constancy Algorithms. J. of Pattern Recog. Research 1(1), 42–54 (2006)
11. Buchsbaum, G.: A Spatial Processor Model for Object Colour Perception. J. of the Franklin Institute 310, 1–26 (1980)
12. van de Weijer, J., Gevers, T., Gijsenij, A.: Edge-based Color Constancy. IEEE Trans. on Image Processing 16(9), 2207–2214 (2007)
13. Land, E.H.: The Retinex Theory of Colour Vision. Proc. R. Instn. Gr. Br. 47, 23–58 (1974)

GOP-Flexible Distributed Multiview Video Coding with Adaptive Side Information

Lili Meng[1], Yao Zhao[1], Jeng-Shyang Pan[2], Huihui Bai[1], and Anhong Wang[3]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, P.R. China
[2] Department of Electronic Engineering, Kaohsiung University of Applied Sciences, Kaohsiung, 807, Taiwan
[3] Taiyuan University of Science and Technology, Taiyuan, 030024, P.R. China
{07112076,yzhao,hhbai}@bjtu.edu.cn, jspan@cc.kuas.edu.tw, wah_ty@163.com

Abstract. A GOP-flexible multiview video coding (MVC) scheme based on Wyner-Ziv (WZ) coding with adaptive side information (SI) is proposed in this paper. In this scheme, each view is WZ encoded independently at the encoder side, while the views are jointly decoded at the decoder side. Therefore, the communications of different cameras can be avoided at the encoder side. In WZ coding, SPIHT and LDPC are applied to improve compression efficiency. Meanwhile, a flexible MVC structure which can satisfy different bit rate requirements and an adaptive SI selection mechanism which better utilizes temporal and interview correlations to obtain more accurate SI are proposed. The experimental results show better rate-distortion performance of the proposed scheme than other tested schemes.

Keywords: Wyner-Ziv coding, multiview video coding, side information.

1 Introduction

With the rapid development of 3D technology, video displaying technology has developed from 2D to 3D and multiview video has attracted considerable attention [1]. Multiview video consists of video sequences captured by an array of cameras, which can be from different angles and locations. Therefore, multiview video can make scenes more vivid and accurate. At present, some organizations have constituted practical multiview video systems, such as, the multicamera array system designed in Stanford University [2] and real-time multiview video system developed in MSRA [3]. Recently, some new techniques have been proposed by the standard organization, MPEG 3DAV group [4] [5]. However, the data of multiview video is too large to be stored and transported conveniently, so it is very necessary to compress the multiview video efficiently.

With the development of video coding [6] [7] [8], in multiview video, there are not only temporal correlations within each view sequence but also interview correlations among different view sequences. Such correlations can be utilized in MVC. So far, various MVC technologies have been developed. For instance, a matrix of pictures

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 47–55, 2010.

with N view sequences and each with K temporal successive pictures is defined [9], which utilized histogram matching to compensate for interview intensity variations. A sprite generation algorithm for improving coding efficiency is proposed in [10]. In such MVC technologies, both the temporal correlations and interview correlations are exploited at the encoder side. Although the coding efficiency is improved to some extent, there still are some limitations in practical application. For example, the cameras of different views must communicate freely to make the interview correlations accessible. However, the transmission between two cameras is usually unavailable in practice and the computing complexity is also a burden for the camera arrays. Consequently, the existing MVC systems are hard to be applied in practice.

Distributed source coding (DSC) can solve such a problem. DSC is separately encoded and jointly decoded. Slepian-Wolf theory shows that even if correlated sources are encoded independently, coding efficiency can be as good as dependent encoding [11]. Later, Wyner and Ziv proposed lossy source coding with SI [12]. Recently, distributed video coding (DVC) is proposed utilizing temporal correlations at the decoder. Lately, MVC which is based on WZ coding with fixed frame structure and flexible prediction side information is proposed in [13]. In this paper, we will propose flexible frame structure and simpler adaptive selection mode to get more accurate SI.

In this paper, a new flexible MVC based on DVC with adaptive SI is proposed. The frame structure is flexible to satisfy different bit rate requirements. Our previous technologies of DVC [14] are used in this paper. Each frame of multiview video is encoded either as traditional Intra-frame (I frame) or as Wyner-Ziv frame (WZ frame). The DVC exploited in this paper exploits SPIHT and LDPC to improve coding efficiency. Experimental results indicate that the adaptive selection mode can achieve high prediction accuracy due to the better employment of temporal correlations and interview correlations.

This paper is organized as follows. Section 2 describes the proposed flexible MVC scheme based on WZ coding with adaptive SI. Experimental results are reported in Section 3, followed by the conclusion drawn in Section 4.

2 GOP-Flexible MVC Based on WZ Coding with Adaptive SI

2.1 GOP-Flexible Structure of the Proposed MVC Scheme

Structure of the proposed flexible MVC scheme is shown in Fig. 1. Structure of the proposed flexible MVC scheme is shown in Fig. 1. In the Fig. 1, I denotes the frame encoded by H.264 intra frame encoder and W is the frame encoded by WZ encoder. In the proposed scheme, one I frame and $m-1$ W frames are in one group. Length of group is m, which is a flexible integer and greater than two. We can adjust the group length according to the bit rate requirement and the trait of the sequences. For example, if we need low bit rate and the quality of the decoded video can be accepted, the m can bigger; if the motion of the sequences is fast or the distance of cameras is big, and we need high quality decoded video, the m should small. At the encoder side, the frames are be encoded independently while the W frames are jointly decoded with the help of side information which can be generated from the decoded I frames using adaptive selection mode.

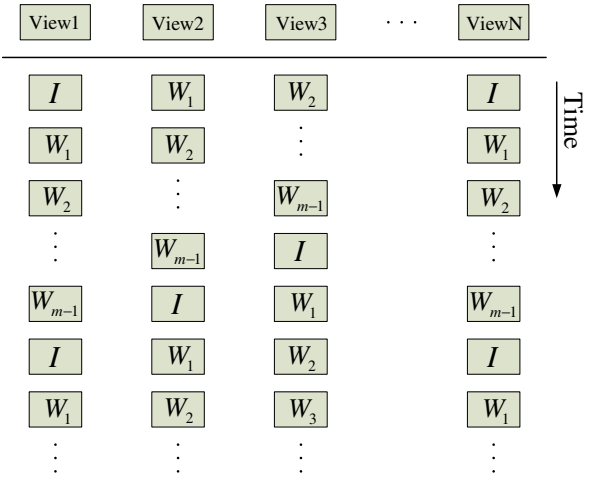


Fig. 1. The structure of proposed MVC

In the proposed MVC scheme, communication between different cameras can be avoided and the selection of decoded view is more flexible, which may satisfy the practical applications.

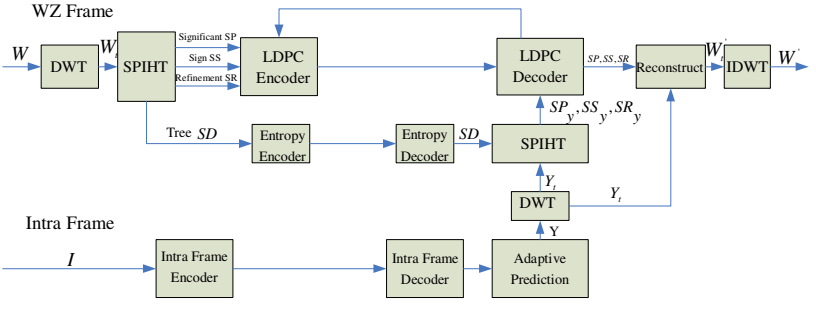


Fig. 2. WZ video coding scheme

In this proposed scheme, WZ video coding which exploits DWT, SPIHT and LDPC is the key part. Fig. 2 illustrates the WZ video coding system. Intra frames are encoded by H.264 intra frame encoder and then decoded by H.264 intra frame decoder. For WZ frames, after being processed by DWT and SPIHT, significant information SP, sign information SS, refinement information SR and tree information SD can be obtained. The tree information SD is encoded by arithmetic coding and sent to the decoder. While SP, SS and SR are encoded by LDPC and only the parity bits are transmitted to the decoder. At the decoder, the adaptive selection mode is implemented to gain the more accurate side information Y. With the help of decompressed SD and side information Y, we can obtain SP_y, SS_y

and SR_y. Then, the main information SP, SS and SR are obtained with the help of SP_y, SS_y, SR_y and received parity bits. Next, SP, SS, SR and Y_t are used to recover the reconstruction frame $W_t^{'}$. At last, the restoration $W^{'}$ is obtained after IDWT processing.

In Fig. 2, the intra frames which include all the intra frames around the current WZ frame are exploited to gain side information Y by the proposed adaptive selection mode described in next part.

2.2 Adaptive Selection Scheme of SI

In DVC, the more relation exists between side information Y and current frame, the more efficiency of the coding is. So the side information Y is much important for improving the coding efficiency. Temporal and interview correlations can be exploited to generate side information Y in the MVC based on WZ coding.

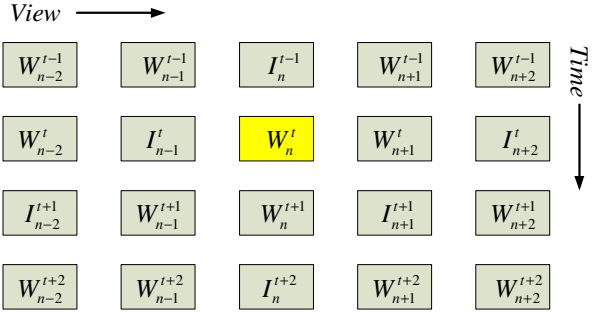
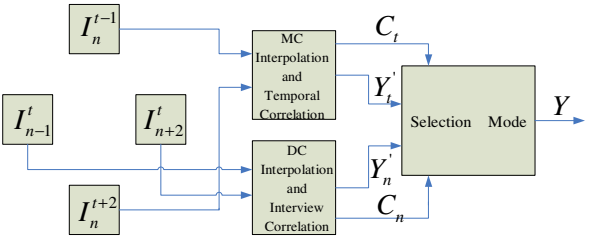


Fig. 3. The structure of frames

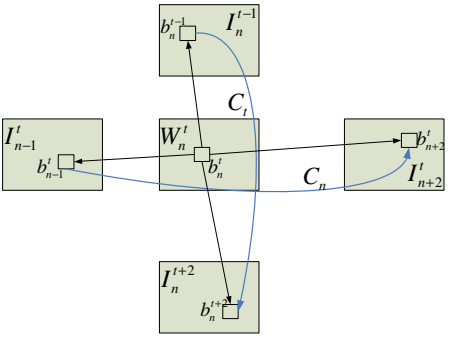
Here, for simplicity we select three frames of each group, one I frame and two W frames. The above Fig. 3 shows the structure of frames. The current frame is W_n^t, which denotes the frame at the n-th view and at the t-th time. The I frames around current frame W_n^t contain temporal frames I_n^{t-1}, I_n^{t+2} and view frames I_{n-1}^t, I_{n+2}^t. In the adaptive selection mode of side information Y, all the I frames around the current W frame- W_n^t are used to generate the SI of current W frame. All the I frames which are exploited in the adaptive selection mode are decoded frames by Intra frame decoder. The structure of proposed adaptive selection mode is showed in Fig. 4(a) in detail. From the Fig. 4(a), we can observe that motion compensation (MC) interpolation and disparity compensation (DC) interpolation are implemented at the decoder. In this proposed mode, firstly, temporal side information $Y_t^{'}$ and disparity side information $Y_n^{'}$ are generated by MC interpolation and DC interpolation respectively; temporal correlation C_t and interview correlation C_n of current block are generated at the same time. Fig. 4(b) shows the generation mode of temporal correlation C_t and interview correlation C_n. The temporal correlation C_t is computed as:

$$C_t = 10\lg \frac{255^2 \times mb^2}{\sum_{i=1}^{mb}\sum_{j=1}^{mb}(b_n^{t+2}(i,j) - b_n^{t-1}(i,j))^2} \tag{1}$$

Where mb is the block size, b_n^{t+2} and b_n^{t-1} are denoted in Fig. 4(b). Interview correlation C_n can be obtained in the same way. The bigger the temporal correlation C_t or interview correlation C_n, the more relation between the temporal frames or view frames. Then the temporal side information $Y_t^{'}$, temporal correlation C_t, disparity side information $Y_n^{'}$ and interview correlation C_n are transmitted to selection mode. The selection mode determines the generation mode of the every block of side information Y, which is gained by MC interpolation or DC interpolation. If temporal correlation C_t is larger than interview correlation C_n, this means the relation of temporal is greater than interview. Current block of side information Y is achieved by MC interpolation; otherwise the DC interpolation is performed to get current block of side information Y.



(a)



(b)

Fig. 4. Adaptive selection mode of side information

3 Experimental Results

A multiview sequence named Rena (640×480) is used to test the proposed scheme. We choose five views: rena_p00039, rena_p00040, rena_p00041, rena_p00042, rena_p00043, and 100 frames of every view are as test frames. The length of group is three and the structure is *IWWIWW* . Visualized side information frames are showed firstly.

(a) MC interpolation (b) DC interpolation

(c) Adaptive selection mode

Fig. 5. Side information frame

Visualized results are showed in Fig. 5: (a) shows side information frame generated from temporal prediction-MC interpolation mode; (b) shows side information frame generated from interview prediction-DC interpolation mode; (c) shows side information frame generated from adaptive selection mode. We can obviously see that MC interpolation mode can achieve good performance in static areas, but this mode is not good in high motion areas. DC interpolation mode can not get good performance in static areas. The proposed adaptive selection mode can achieve good performance not only in high motion areas but also in static areas.

Objective evaluation of the side information is showed in Fig. 6 and Table 1. Fig. 6 shows the performance of every frame of side information. It is obvious that the proposed adaptive selection mode is better than MC interpolation and DC interpolation. From Table 1, we can see the proposed adaptive selection mode outperforms the MC interpolation mode 3.69 dB and outperforms the DC interpolation mode 1.93 dB. Better side information can efficiently improve the coding efficiency of MVC.

The rate-distortion (RD) performance curve of Rena is shown in Fig. 7. We compare three MVC methods according to the prediction mode of side information Y. The curve of "MC Interpolation" indicates the coding performance whose side information is achieved by temporal prediction-MC interpolation; the curve of "DC Interpolation" indicates the coding performance whose side information is achieved by interview prediction-DC interpolation; the curve of "Adaptive Selection Mode" indicates the coding performance whose side information is achieved by the adaptive selection mode which is proposed in this paper; the curve of "Reference Current Frame" indicates that the side information is obtained by comparing with the current frame. From Fig. 7, we can obviously observe the coding efficiency of MVC based on adaptive selection mode is better than the MC interpolation and DC interpolation especially at low bit rate.

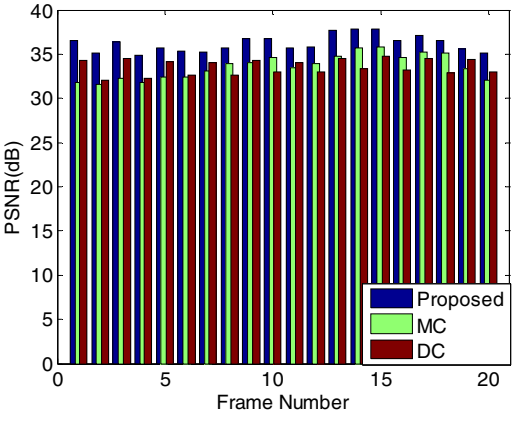


Fig. 6. Performance of side information

Table 1. Comparison of side information

MODE	MC interpolation	DC interpolation	Adaptive mode
Average PSNR(dB)	31.62	33.38	35.31

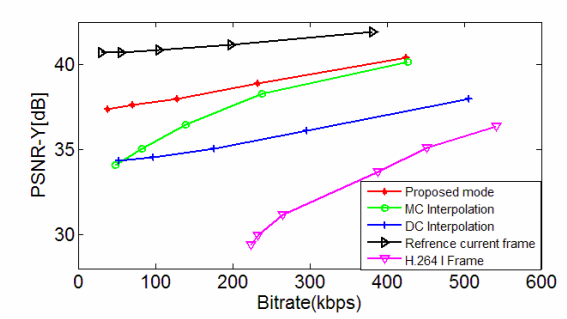


Fig. 7. R-D curves of Rena

4 Conclusion

In this paper, a flexible MVC scheme based on WZ coding with novel adaptive side information is proposed. We can adjust the frame structure according to the bit rate requirement and the character of views. In this scheme, every view is encoded independently, temporal and interview correlations are exploited at the decoder side. Therefore, the communication of cameras is avoided, while the correlations can be also exploited well at the decoder. Side information is very important for WZ coding. This paper proposed a novel adaptive selection mode which determines current block is obtained by MC interpolation or DC interpolation. With the adaptive selection mode, we can get more accurate side information and better coding efficiency of multiview video.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No.60776794, No.60903066, No.60972085), Sino-Singapore JRP (No. 2010DFA11010), Beijing Natural Science Foundation (No.4102049), New teacher Foundation of State Education Ministry (No. 20090009120006).

References

1. Naemura, T., Kaneko, M., Harashima, H.: Compression and Representation of 3-D images. IEICE Trans. INF. & SYST. E82-D(3), 558–565 (1999)
2. Wilburn, B., Joshi, N., Vaish, V., Levoy, M., Horowitz, M.: High speed video using a dense camera array. In: Proc. CVPR, pp. 161–170 (2004)
3. Lou, J., Cai, H., Li, J.: A real time interactive multiview video system. Presented at the 13th ACM Int. Conf. Multimedia, Singapore, November 6-11 (2005)
4. Kimata, H., Kitahara, M.: Multi-view video coding based on scalable video coding for free-viewpoint video. ISO/IEC JTCI/SC29/WG11 M11571, Hong Kong (January 2005)
5. Ho, Y., Yoon, S., Kim, S.: A framework for multi-view video coding using layered depth image. ISO/IEC JTCI/SC29/WG11 M11582, Hong Kong (January 2005)

6. Wang, H., Liang, J., Kuo, J.: Overview of robust video streaming with network coding. Journal of Information Hiding and Multimedia Signal Processing 1(1), 36–50 (2010)
7. Ito, A., Makino, S.: Designing side information of multiple description coding. Journal of Information Hiding and Multimedia Signal Processing 1(1), 10–19 (2010)
8. Lu, Z., Li, Y.: Image compression based on mean value predictive vector quantization. Journal of Information Hiding and Multimedia Signal Processing 1(3), 172–178 (2010)
9. Flierl, M., Mavankar, A., Girod, B.: Motion and disparity compensated coding for multiview video. IEEE Trans. On Circuits and Systems for Video Technology 17(11), 1474–1484 (2007)
10. Grammalidis, N., Strintzis, M.G.: Disparity and Occlusion estimation in multiocular Systems and their coding for the communication of multiview image sequences. IEEE Trans. On Circuits and Systems for Video Technology 8, 328–344 (1998)
11. Slepian, D., Wolf, J.: Noiseless coding of correlated information sources. IEEE Trans. On Information Theory 19, 471–480 (1973)
12. Wyner, A.D., Ziv, J.: The rate-distortion function for source coding with side information at the decoder. IEEE Trans. On Information Theory 22, 1–10 (1976)
13. Guo, X., Lu, Y., Wu, F., Zhao, D., Gao, W.: Wyner-Ziv-based multiview video coding. IEEE Trans. On Circuits and Systems for Video Technology 18(6), 713–724 (2008)
14. Wang, A., Zhao, Y., Pan, J.: Residual distributed video coding based on LQR-hash. Chinese Journal of Electronics 18(1), 109–112 (2009)

A Novel Embedded Coding Algorithm Based on the Reconstructed DCT Coefficients

Lin-Lin Tang[1], Jeng-Shyang Pan[2], and Zhe-Ming Lu[3]

[1] Department of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen 518055, China
[2] Department of Electronic Engineering National Kaohsiung University of Applied Sciences
Kaohsiung, Taiwan
[3] School of Aeronautics and Aeronautics, Zhejiang University,
Hangzhou 310027, P.R. China
Linlintang2009@gmail.com, jspan@cc.kuas.edu.tw,
zheminglu@zju.edu.cn

Abstract. As an efficient tool for image compression, wavelet has been widely used in all kinds of image processing areas. Based on the different encoding effects, wavelet compression algorithms can be probably classified into two categories. They are the embedded wavelet coding algorithms and the non-embedded wavelet coding algorithms. For the convenience of producing the anytime cut coding stream and the progressing reconstruction results, the embedded wavelet coding algorithms have been paid more attention in practice. Such as the embedded wavelet coding algorithms, EZW and SPIHT are the outstanding representatives. The only drawback for this wavelet based embedded coding algorithms is the choice of the different wavelet transform base. We propose a novel embedded coding algorithm based on the reconstructed DCT coefficient to avoid the difficulties brought by the choice of wavelet transform base in this paper. The new algorithm's efficiency can be seen from the experimental results.

Keywords: image compression, embedded wavelet coding algorithm, DCT.

1 Introduction

Fast development of communication has brought a great push to the compression coding algorithms research in this area. Many different kinds of compression algorithms have been proposed based on wavelet. And the wavelet transform has been introduced in all kinds of compression standards, such as the JPEG2000 standard [1]. Most algorithms based on wavelet have made great use of multiresolution analysis property introduced by the wavelet transform. The traditional tree structure which is formed by wavelet transform has also been used in all embedded wavelet coding algorithms. The EZW [2] and SPIHT [3] algorithms are the most famous application among these typical algorithms.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 56–62, 2010.

Although the wavelet based compression coding algorithms have achieved lots of excellent results [4], the complexity brought by selecting the proper wavelet filter and the following boundary extension problems are still there. So, successfully avoiding these difficult problems is a long-term subject to face to. In fact, as an efficient compression tool for image, the DCT transform has more efficient energy concentration property and easier implementation. For these properties, it has already been used in all kinds image processing areas, for example the watermarking orientation [5]. The only short coming for this transform is the lack of multiresolution analysis and the following tree structure decomposition. And these are also the root for not to apply the embedded coding process on the DCT as the wavelet. Actually, the energy distribution for the DCT image also has some regularity. If we can make full use of them, the embedded coding stream also can be generated. A novel embedded compression coding algorithm based on DCT decomposition image is proposed in this paper. After applying the DCT on original image, we redistribute the transform coefficients to make them have some property like multiresolution analysis. According to the reconstructed transform image, we introduce a new definition of directional tree and form a novel embedded coding algorithm based on it. Experimental results based on this new proposed algorithm show its efficiency.

2 Basic Knowledge

2.1 Zerotree and Spatial Orientation Tree

Among all the embedded coding algorithms, zero trees and spatial orientation trees are the two usually used structures. For the proposed new algorithm uses the similar spatial orientation trees in the SPIHT algorithm, we give the detailed definition of it as follows.

After wavelet decomposition, image coefficients arranged by a special way with multiresolution property and the following frequency subband distribution structure. Coefficients lie in the same location in different subband have some natural relation in their values. The following figure 1 gives a clear explanation of the spatial orientation tree. Here, we take the three-level wavelet decomposition image for example.

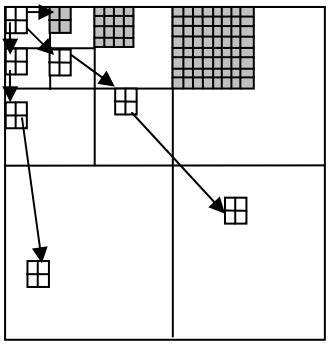


Fig. 1. Spatial Orientation Tree

As we can see from figure 1, every coefficient lies in the decomposition image has four direct offspring besides the ones in the lowest and the highest frequency subbands. The offspring lies in the adjacent same direction higher frequency subband as the father nod. The coefficients lie in the lowest frequency subband has different offspring rule from others, the coefficients are grouped by 2×2 block and the three coefficients but the one in the left corner each has four direct offspring. And the relationship of father and son is shown as in the above figure 1. The embedded wavelet coding algorithms based on this kind of orientation trees make use of this relationship between father and son and record the whole tree information by the root nod. We will not give the detailed description of the algorithm. The author can get some idea from the reference [3].

2.2 DCT and the Redistribution Method

As one the most efficient compression tool, the Discrete Cosine Transform has better energy concentration than wavelet transform, and has been widely used in various compression standards. The most usually used DCT is the 8×8 block transform shown in the following formula.

$$T_{i,j} = \left(0.5K\cos\left[i\pi\left(j+0.5\right)/8\right]\right), \qquad i,j = 0,1,\ldots,7$$
$$\begin{cases} K = 1/\sqrt{2}, & i = 0 \\ K = 1, & i = 1,2,\ldots,7 \end{cases} \tag{1}$$

We take the 512×512 Lena image for example to see the transform result of DCT. The reduced formula is shown below and the result of transform of the whole image is shown in the following figure 2. Here, C and B represent the transformed and the original 8×8 block respectively.

$$C = TBT' \tag{2}$$

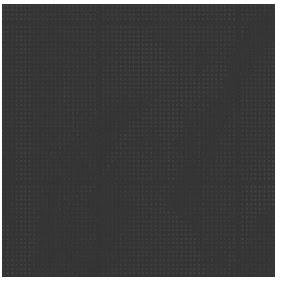


Fig. 2. DCT transform result of 512×512 Lena image

As we can see from figure 2, most of the transform energy is concentrated in the top left corner. So, the whole transform image is shown as 64×64 energy concentration points. If we want to make use of the wavelet multiresolution analysis, the coefficients redistribution operation must be applied on the transform image

which is achieved through DCT. The concrete step is described as the following formula. Though similar idea has already been introduced before, there is no further research after the redistribution application.

$$\begin{aligned} &D\left(\left((m-1)\times(R/8)+i\right),\left((n-1)(C/8)+j\right)\right) \\ &=B\left(\left((i-1)\times 8\right)+m\right),\left(\left((j-1)\times 8+n\right)\right) \end{aligned} \tag{3}$$

Where, D and B represent the reshaped image and the original transform-domain image respectively. The coordinate (i, j) and (m, n) represent the block location in the whole 64×64 transform-domain blocks and the location of each coefficient in the 8×8 transform-domain block. After the reshape step, the original transformed image can be redistributed into a new one with 64 subimages with each size of 64×64. Every subimage among them looks like a reduced image of the original one. And the energy of them is arranged like the wavelet transform image. It means that the reshaped image also has the law looks like the multiresolution analysis. We call this new law pseudo-multiresolution analysis. In fact, based on such a pseudo-multiresolution the traditional embedded wavelet coding algorithm can be applied on. The traditional embedded wavelet coding algorithms make use of the traditional tree structure. Though the redistributed image has the similar energy distribution to the traditional one, the difference in structure will not bring good result just like the case based on the true tree structure coefficients. A new embedded tree structural is introduced in our algorithm to solve the problem.

3 Our Proposed Algorithm

After redistribution of the transform coefficients, the whole image looks like a combination of the subimages of the original one. The size of each subimage is 64×64 and the one with maximum energy lies in the top left corner just like the lowest frequency decomposition block in wavelet decomposition. So, the pixels which lie in the same location belongs to different subimages have different father-son relationship from the traditional wavelet decomposition. The adjacent same location pixel distance is measured by the side length of the subimages. It means that 64 is used as the horizontal and vertical measure unit and $64\sqrt{2}$ is for the diagonal direction. The basic idea of SPIHT algorithm is used to form the new embedded coding algorithm. The concrete step of this new method is described as following.

Firstly, we divide the 64 subimage composed redistributed image into three levels just like the three-level wavelet decomposition image. In the highest frequency part, each direction contains sixteen subimages with each size is 64×64. For the lowest frequency part, each direction contains one subimage with the same size.

Secondly, the new wavelet tree is organized following the new rule in which the same orientation tree is the set of pixels lie in the same space of each subimage. It means that each pixel in the lower frequency subband has four son pixels in the higher frequency subband just as the wavelet tree structure.

Finally, based on the new tree structure, we make use of the principle of classical wavelet embedded coding algorithm SPIHT to realize the coding process. The whole image can be record by the root of each new tree and its location in the whole image. The concrete steps of such a coding process are described as following.

1) $T(i,j)$ is defined as the set of all descendants of pixel (i,j) and the pixel (i,j) itself. It refers to space orientation tree.
2) $D(i,j)$ is defined as the set of all descendants of pixel (i,j) without the pixel (i,j).
3) $O(i,j)$ is defined as the set of direct son of pixel (i,j).
4) $L(i,j)$ is defined as the set of descendants of pixel (i,j) except for those direct sons.

It means that these sets satisfy the relationship below.

$$T(i,j) = c(i,j) + D(i,j) \tag{4}$$

$$D(i,j) = O(i,j) + L(i,j) \tag{5}$$

$$L(i,j) = \sum D(k,l) \qquad (k,l) \in O(i,j) \tag{6}$$

In the ordering process, verification of the important and unimportant pixels will be finished by the multiple times space orientation tree split. Following encoding process is the same as the classical embedded coding algorithms SPIHT. Though the redistributed image has the similar energy distribution to the wavelet image, the energy distribution character is different from it. So, when we choose the proper threshold to finish the coding process, the special value different from the original coding process should be used. The initial threshold value used here is $T_0 = \lfloor \log_2 (\max |c(i,j)|) \rfloor + 1$ and it is the same value with the original algorithm. Different threshold values are used in the following coding steps. The interval used to define the important and unimportant coefficients has been changed from $[T_i, 2T_i]$ to $[T_i, 3T_i]$. This proposed algorithm can be called DCT-SPIHT algorithm.

4 Experiment Results

The 512×512 Lena gray image is used in our experiment here. Table 1 shows the comparison results of the classical comparison algorithm and DCT-SPIHT algorithm under different coding bit rates. All the results based on wavelet are achieved under the three-level decomposition.

As we can see from table 1, our propose method DCT-SPIHT performs well in the comparison. Comparison under the same condition between classical algorithms and proposed DCT-SPIHT algorithm is shown in the following figure 3. The first two is the SPIHT algorithm and ours. The results are achieved under the bit rates 0.25bpp. The following two images are the classical DCT compression algorithm and ours. And the results are achieved under bit rates 1.25bpp.

Table 1. Comparison results between different compression algorithms

Coding algorithms	*PSNR/dB*		
	0.25bpp	0.5bpp	1.0bpp
EZW	33.17	36.28	39.55
SPIHT	34.11	37.21	40.44
DCT-SPIHT	33.09	36.20	39.08

(a) PSNR=34.11 (b) PSNR=33.09

(c) PNSR=32.08 (d) PSNR=39.38

Fig. 3. Compression between the classical algorithm and ours

As we can see from the above experimental results, though there is some distance between our proposed DCT-SPIHT algorithm and the classical wavelet based embedded coding algorithms, the performance is much better than the classical DCT transform based compression algorithm.

5 Conclusion

A new image compression algorithm based on the classical DCT and SPIHT algorithm is proposed in this paper and the novel algorithm is called DCT-SPIHT algorithm. It makes full use of the energy concentration property and the high efficiency of the embedded compression algorithm SPIHT. Good results are shown in the experiments. To find more energy distribution of such a reshaped DCT image and so to find more efficient coding algorithm cording to it is our future work.

References

1. ISO/IEC JCT1/SC29 WG11/N1646, Information technology-JPEG 2000 image coding system: Core coding system (March 2000)
2. Huang, M.H., Zhong, C.X.: A Unified Scheme for Fast EZW Coding Algorithm. In: Proceedings of International Symposium on Computer Science and Computational Technology, vol. 2, pp. 622–626 (2008)
3. Said, A., Pearlman, W.A.: A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees. IEEE Transactions on Circuits Syst. & Video Technol. 6(3), 243–249 (1996)
4. Ersoy, O.K., Nouira, A.: Image coding with the discrete cosine-III transform. IEEE Journal on Selected Areas in Communications 10(5), 884–891 (1992)
5. Lin, C.-C., Shiu, P.-F.: High Capacity Data Hiding Scheme for DCT-based Images. Journal of Information Hiding and Multimedia Signal Processing 1(3), 220–240 (2010)
6. Chan, Y.T., Ho, K.C.: Multiresolution analysis, its link to the discrete parameter wavelet transform, and its initialization. IEEE Transactions on Signal Processing 44(4), 1001–1006 (1996)

A Vehicle License Plate Recognition System Based on Spatial/Frequency Domain Filtering and Neural Networks

Mu-Liang Wang[1], Yi-Hua Liu[1], Bin-Yih Liao[2], Yi-Sin Lin[2], and Mong-Fong Horng[2]

[1] Department of Computer Science and Information Engineering,
Shu-Te University, Kaohsiung, Taiwan
[2] Department of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
mulwang@stu.edu.tw, s2nkcaz.wen@msa.hinet.net, jeltsine@gmail.com
{byliao,mfhorng}@cc.kuas.edu.tw

Abstract. In this paper, we develop an intelligent application based neural networks and image processing to recognize license plate for car management. Through the license recognition, the car number composed of English alphabets and digitals is readable for computers. Recognition of license is processed in two stages including feature extraction and recognition. The feature extraction contains the image locating, segmentation of the region of interest (ROI). Then the extracted ROIs are fed to a trained neural network for recognition. The neural network is a three-layer feed-forward neural network. Test images are produced from real parking lots. There are 500 images of car plates with tile, zooming and various lighting conditions, for verification. The experiment results show that the ratio of successful locating of license plate is around 96.8%, and the ratio of successful segmentation is 91.1%. The overall successful recognition ratio is 87.5%. Therefore, the experimental result shows that the proposed method works effectively, and simultaneously to improve the accuracy for the recognition. This system improves the performance of automatic license plate recognition for future ITS applications.

Keywords: Neural Network, Plate Recognition, Wavelet Transform, Spatial/ Frequency analysis.

1 Introduction

Intelligent Transportation System (ITS) attracted lots of interest from industrials and academics worldwide in the past decade. The ITS investment in globe keeps growing fast. For examples, according to the report from CCID, ITS development in China becomes the focus of urban transportation improvement. The investment over ITS construction was totaled at CNY19.5 billion in 2008, rising by 39.3% of 2006. In United State, since 1992, Department of Transportation (DOT) has invested $1.5 billion to develop an ITS on limited access highways and local roadways across America. Most of this investment was to construct broadband infrastructure. In 2009,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 63–70, 2010.

there are at least 22 states in U. S. sought American Recovery and Reinvestment Act funds to invest in intelligent transportation system technologies including traffic cameras, express toll lanes, and improved traffic signals or accident alert systems. In fields of information and communication technology (ICT), data processing and communication are two pillars to develop feasible applications to improve the functionality and quality of the transportation systems. In other words, ITS is expected to offer more automotive services for drivers, riders and walkers on roads. The first and most significant issue of ITS is to identify vehicles in movement. Thus the technique of car license recognition has been key to successful ITS applications. How to identify the car license in a fast, reliable and accurate way has been acknowledged as a critical issue to be explored [1-7]. In this paper, we develop a license plate recognition system with some image processing technologies to automatically detect if there is any car to be recognized in the screen and rapidly filter out suspicious license plates in these images to recognize characters on the license plate.

The rest of this paper is organized as follows. In Section 2, the previous work related to car identification is reviewed to illustrate the state of art. The proposed Vehicle License Plate Recognition System (VLPRS) is presented to describe the architecture and operations for the license recognition in Section 3. The performance of the developed system is evaluated and analyzed in Section 4. Finally, we conclude this work in Section 5.

2 Related Works

There were various approaches proposed to identify moving vehicles. In Taiwan, Electronic Toll Collection (ETC) was installed in highways for toll services for years. The identification of vehicles is realized by an infrared communication between the transceivers installed on vehicles and on the toll stations. Although this approach is reliable and fast enough for moving vehicles, the pre-installation of transceivers is an obstacle of deployment. Radio-Frequency technology is another approach to identify vehicles in recent years due to its low cost of deployment. However, worse reliability of identification degrades the feasibility of the RFID-based approach. Image-based approach is another attractive solution due to the well-deployed cameras on roads and streets. However, how to interpret the image content and to identify the ownership of vehicles is the key issue to be investigated.

Feature extraction is the first step to identify the license plates of vehicle images. In the past, there were some researches focusing on this topic. Mello *et. al*. proposed a system based on color alternation for acquisition images and fuzzy logic for segmentation of digit images on license plates. Although the reliability and accuracy of that proposal are recognized, the exception conditions in real environments such as lighting dynamic and image tilt are not explored. Lee [2] presented an approach based on neural networks to recognize image patterns and content mining. His neural network simulator is efficient for image data. However, the applications on license identification are rare to evaluate the performance of the simulator. Yu *et. al.* [5] proposed a vertical edge matching algorithm to detect the edge of characters on license plates. Morphology [6] is also another effective scheme to remove the noise introduced during image

captures. Chun *et. al.*[7] proved that a requirement of vehicle identification system on real-time operation is essential to ITS applications. Lin *et. al.*[8] overcame the problem of low-resolution printed digit for character recognition. In this paper, we present a VLPRS implementation based on neural networks and dual-domain signal filtering. The developed system features its ability to learn, train and recognize the license plates in various conditions of tilt, zooming, and dim-lighting. Thus the robustness is better than ever.

3 A Vehicle License Plate Recognition System Based on Spatial/Frequency Domain Filtering and Neural Networks

The architecture of the license plate recognition system is shown in Fig. 1, including preprocesses like license plate locating, license plate image character segmentation, and character recognition. In the stage of license plate locating, after inputting of license plate image, we use Wavelet edge detector and apply the morphology technique to quickly locate candidate areas that might be license plates. Projection and plate number format judgment are used to analyze if the located area is a license plate. If it is a license plate, we extract characters from the license plate. In the stage of plate number recognition, we apply the neural learning method to recognize characters on license plates. Here we introduce each stage of entire system below.

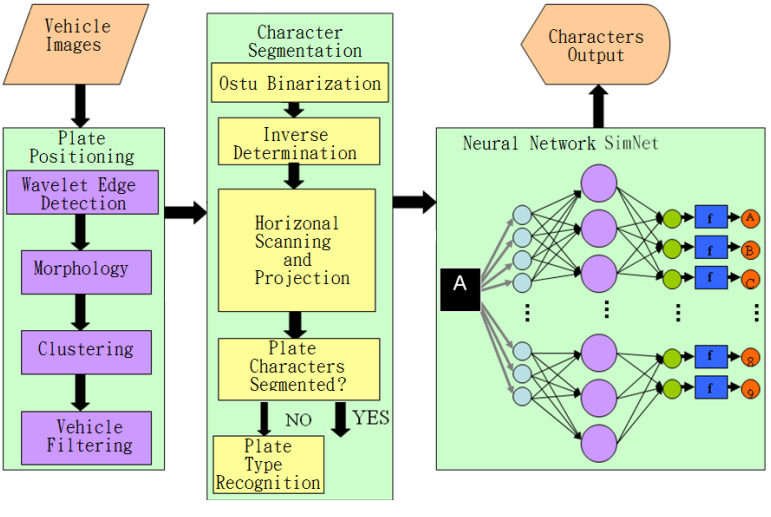


Fig. 1. Architecture of Vehicle License Plate Recognition System

In this paper, we present a rapid recognition of license plate, to find suspicious location of license plate and filter out with license plate formats. The function of wavelet transform is to effectively segment the signals by frequencies. In this paper we use the wavelet transform to process the edge detection of the license plate system. If one image had been processed by the wavelet transform, it would be divided into two

parts: the high-frequency one and the low frequency one. The high frequency part of an image includes most of the edges of objects that the grayscale changes intensively. Ku presented an approach in Wavelet operators for multi-scale edge and corner detection [4], the approach based on 4-tap Daubechies wavelet transform coefficients and operands combines 2-dimension Discrete Periodic Wavelet Transform to derive the operand mask of wavelet edge detection. The computation is relatively complicated. The original image and the edge detected by wavelet are shown in Fig. 2-3; they allow a wider range of the threshold. If the threshold goes down to 50, the wavelet edge detection gets better result in both the edge of the license plate and the complex street view in background, and it also has a better immunity of noise than the Sobel edge detection.

Fig. 2. Original Image

Fig. 3. Edges detected by Wavelet

After the wavelet edge detection, we use the morphology in image process to analyze the shape and the structure of the image to strengthen the structure to locate the license plate. Figure 7-10 show the object we find in contrast with the original image after the morphology process. Closing and opening from morphology can rapidly erase noise and places that do not match the aspect ratio of a license plate. After the morphology process we precede the 8-connected component algorithm. In order to accurately cut out the candidate places for the plates, we set the aspect ratio and the area ratio of object pixels as conditions to filter out things not familiar with a license plate. The figure below shows locations that we get after the license-plate-like condition filtering. Though one of them is not exactly a license plate, but there are still many features to be examined, we put these two objects to the follow-up license plate character procedure.

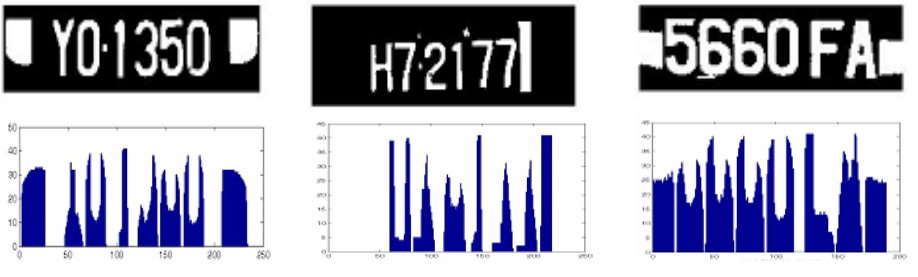
Fig. 4. Detected edge fed to morphological processing

Fig. 5. Result after morphological processing

The second part of the license plate recognition system is character extraction, its main purpose is to extract characters from a license plate object and find the top, right, bottom and left border of each character. Binary threshold is to determine the best threshold. Fig. 6 and 7 below show that the result of the Threshold Selection Method from Gray-Level [9] suffers from the effect of shadow, and fails to divide the characters from background. Before we proceed the character segmentation on the license plate, we must convert the license plate image into objects with single alphabets and digits. The conversion is manipulated by vertical projection of the segmented images. By projection, we detect the positions of characters and the borders of plate. Examples of vertical projections are shown in Fig. 8. Then we use the vertical projection on the horizontal axis in the histogram. We use the vertical-projected aspect ratio as the basis of character and noise prediction. We determine the position of the characters on the license plate by projection.

Fig. 6. Original Image

Fig. 7. Binarized image by Otsu[9]



(a) (b) (c)

Fig. 8. Histograms of Vertical projects for various license images

The segmented ROI image blocks are fed to a feed-forward neural network to identify. In the developed neural network, there are two stages, training and testing, for the presented system. In a training stage, a supervised strategy is used to train the network by adapting the connection weights between neurons till the convergence. As depicted in Fig.1, there are three layers to compose the neural network. The input pattern is given as follows

$$X_{k\times n}=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}=\begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{k,1} & X_{k,2} & \cdots & X_{k,n} \end{bmatrix} \tag{1}$$

where *k, n* are the pattern number and pattern length, respectively. Beside the output of the neural network are given as

$$T_{k\times m}=\begin{bmatrix}T_1\\T_2\\\vdots\\T_k\end{bmatrix}=\begin{bmatrix}T_{1,1} & T_{1,2} & \cdots & T_{1,m}\\T_{2,1} & T_{2,2} & \cdots & T_{2,m}\\\vdots & \vdots & \ddots & \vdots\\T_{k,1} & T_{k,2} & \cdots & T_{k,m}\end{bmatrix} \tag{2}$$

where m is the number of alphabets and digits to recognize. The elements of $T_{k,\,m}$ indicate the recognition results generated by the neural network. When $T_{k,\,m}$ is set to 1, the *k-th* input pattern is recognized as the *m-th* alphabets or digits. The initial weight matrix are depicted as

$$X_{k\times n}=X_{k\times n}/norm(X_{k\times n}) \tag{3}$$

$$W_{n\times m}=X^{T}_{k\times n}\cdot T_{k\times m} \tag{4}$$

During the training stage, the weight matrix is updated according to Eq. 5-8

$$Y_{k\times m}=X_{k\times n}\cdot W_{n\times m} \tag{5}$$

$$dY_{k\times m}=T_{k\times m}-Y_{k\times m} \tag{6}$$

$$dW_{n\times m}=X^{T}_{k\times n}\cdot dY_{k\times m} \tag{7}$$

$$W_{n\times m}=W_{n\times m}+dW_{n\times m} \tag{8}$$

where Y, dY and dW are the actual output, the output error and the adjustments of the connection weights. And the actual output of the neural network is shown in Eq. (9)

$$Y_{k\times m}=f\left(X*W\right) \tag{9}$$

where the activation function of this neural network are indicated as follows,:

$$f(x)=\frac{1-\exp(-\sigma x)}{1+\exp(-\sigma x)} \tag{10}$$

Clearly, the activation function depicted in Eq. 10 produces a bipolar output *Y*, ranging between [-1, 1]. In a stabilized network, the neuron outputs converge to either 1 or -1. If Y_i converges to 1, the *i-th* alphabet is identified, otherwise the *i-th* alphabet is denied. Thus, after the training stage, the trained neural network is used to recognize the license plate through the manipulation of the extracted features. This process is denoted as testing stage. In the following section, we will use 500 vehicle images from real environments to verify the performance of the developed system.

4 Experiment Results

There are four test cases for various conditions such as tilt, zooming, dim-lighting and integrated recognition. In a real application, tilt images are encountered unavoidably.

This image tilt usually is caused by relative positions of camera and vehicles. The width-to-height ratio (WHR) of license plate image is the index to measure the tilt. The images of with no tilt have the WHR around 2.9. The tested image with the WHR of 3.9 is as shown in Fig. 9. The segmentation result is shown in Fig. 10. The second test is to verify the recognition in the conditions of zoomed images. The original images as shown in the first row, Fig. 11. These two images with significant difference are captured in various distances. The second row in Fig. 11 shows the extracted context of license plates. Clearly, the developed system is able to successfully extract the ROI of alphabets and digits.

Fig. 9. Tilted image

Fig. 10. Segmentation results

The third test is to verify the performance of the developed system in a dim-lighting condition. In Fig. 12 (a), the captured original image is shown. The detected edges are depicted in Fig. 12 (b). The extracted ROI of plate image is shown in Fig. 12 (c).

(a)

(b)

Fig. 11. Zoomed images and the extractions

(a)

(b)

(c)

Fig. 12. Performance evaluation in the dim-lighting condition

The accuracy and reliability are two performance indexes to be verified in this test. There are totally 500 vehicle medium-resolution images in the testing of the license plate recognition system. The identification procedure is divided into two parts including the internal testing and the external testing. In internal testing, we use 250 same images to train the neural network and to test the training performance. In an external test, the extra 250 vehicle images are fed to the system to verify the identification accuracy. These images are taken from various environments including day, night, distances between the camera and the car. In this paper we take characters in the 250 license plates as the learning sample of neural network. The recognition rate of the internal testing is 100%. And we use another 250 images which are taken in a worse situation for the external testing.

5 Conclusion

In this paper, we apply wavelet edge detection to extract the edges in the image during the preprocess stage. Wavelet edge detection owns the characteristic of high noise-immunity, and it allows a wider range in threshold we set. Then we apply morphology, to record the top, right, bottom and left border of each object. With projection, we can detect the position of plates and record the height and the top and bottom border of characters in the process of character segmentation. Afterwards, we check the number of characters and determine whether the aspect ratio of characters meets the format of plate numbers. Finally we normalize the character images and apply SimNet to recognize characters.

References

1. Chang, S.L., Chen, L.S., Chung, Y.C., Chen, S.W.: Automatic License Plate Recognition. IEEE Transactions on Intelligent Transportation Systems 5(1), 42–53 (2004)
2. Mello, C.A.B., Costa, D.C.: A Complete System for Vehicle License Plate Recognition. In: 16th International Conference on Signals and Image Processing (IWSSIP 2009), pp. 1–4 (2009)
3. Lee, H.C.: SimNet: A neural network architecture for pattern recognition and data mining, University of Missouri-Rolla (2003)
4. Ku, C.T.: Wavelet operators for multi-scale edge and corner detection. Department of Electrical Engineering, I-Shou University, Taiwan (1998)
5. Yu, M., Kim, Y.D.: An Approach to Korean License Plate Recognition Based on Vertical Edge Matching. In: IEEE Conference on Systems, Man, and Cybernetics, pp. 2975–2980. IEEE Press, New York (2000)
6. Hsieh, J.W., Yu, S.H., Chen, Y.S.: Morphology-based license plate detection from complex scenes. In: 16th International Conference on Pattern Recognition, pp. 176–179. IEEE Press, New York (2002)
7. Chun, B.T., Soh, Y.S., Yoon, H.S.: Design of Real Time Vehicle Identification System. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2147–2152. IEEE Press, New York (1994)
8. Lin, H.H., Chen, C.Y., Chuang, J.H.: Recognition of Printed Digits of Low Resolution. Pattern Recognition and Image Analysis 10(2), 265–272 (2000)
9. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on System, Man, and Cybernetics SMC-9, 62–66 (1979)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, Hoboken (2000)

Reversible Watermarking Based on Invariant Relation of Three Pixels

Shaowei Weng[1], Shu-Chuan Chu[2],
Jeng-Shyang Pan[3], and Lakhmi C. Jain[2]

[1] Information Engineering College,
Guangdong University of Technology, P.R. China
wswweiwei@126.com
[2] School of Electrical and Information Engineering,
University of South Australia
scchu@bit.kuas.edu.tw
[3] Harbin Institute of Technology
Shenzhen Graduate School,P.R. China
jengshyangpan@gmail.com

Abstract. In Lin's method [1], images are divided into non-overlapping three-pixel blocks. Each pixel block contains two pairs composed of two neighboring pixels. Absolute differences between pairs are calculated. And meanwhile, the absolute difference having the highest occurrence rate is obtained. 1-bit watermark is embedded into a pixel pair whose absolute difference is equal to this obtained value. Since each pixel block contains two differences, Lin's method achieves embedding rate of at most $\frac{2}{3}$ bpp (bit per pixel) for a single embedding process. With the aim of further increasing embedding rate, we modify the embedding process in Lin's method to keep the third pixel of a pixel block unaltered in the proposed method. Then, this unchanged pixel is used again to reform into a new pixel block with its two neighboring pixels. As a result, the embedding rate can reach to 1 bpp for a single embedding process. ...

1 Introduction

For some critical applications such as the fields of the law enforcement, medical and military image system, it is crucial to restore the original image without any distortions. The watermarking techniques satisfying these requirements are referred to as the reversible watermarking. The reversible watermarking is designed so that it can be removed to completely restore the original image.

The concept of reversible watermark firstly appeared in the patent owned by Eastman Kodak [2]. Several researchers had developed reversible watermarking [1, 3–10]. Tian [8] expanded the differences between two neighboring pixel values to embed a bit into each pixel pair without causing overflows/underflows after expansion. Alatter [9] used the difference expansion of quads to embed a large amount of data into the grayscale images. His algorithm allowed three bits to be embedded into every quad. Yang *et al.* [10] proposed a simple lossless data

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 71–80, 2010.

hiding method based on the coefficient-bias algorithm by embedding bits in both spatial domain and frequency domain.

In Lin's method [1], images are divided into non-overlapping three-pixel blocks. Each pixel block contains two pairs composed of two neighboring pixels. Absolute differences between pairs are calculated. And meanwhile, the absolute difference having the highest occurrence rate is obtained. 1-bit watermark is embedded into a pixel pair whose absolute difference is equal to this obtained value. Since each pixel block contains two differences, Lin's method achieves embedding rate of at most $\frac{2}{3}$ bpp (bit per pixel) for a single embedding process. With the aim of further increasing embedding rate, the embedding process in Lin's method is changed to keep the third pixel of a pixel block unaltered in the proposed method. Then, this unchanged pixel is used again to form into a new pixel block with its two neighboring pixels. As a result, the embedding rate can reach to 1 bpp for a single embedding process.

The remains of the paper are organized as follows. In Section 2, the proposed method is introduced. Watermark embedding and data extracting and image restoration are presented in Section 2.1 and 2.3, respectively. The performance analysis is discussed in Section 2.2. The experimental results are shown in Section 3 and finally we conclude the paper in Section 4.

2 The Proposed Method

Let $p = (p_1, p_2, p_3)$ denote a pixel block containing three neighboring pixels. Each p contains two absolute differences d_1, d_2, i.e., $d_1 = |p_1 - p_2|$, $d_2 = |p_2 - p_3|$. For a $\mathtt{W} \times \mathtt{H}$-size image, it is partitioned into non-overlapping three-pixel blocks in Lin's method, thus the number of differences occupy only $\frac{2}{3}$ of image size. In the proposed method, with the aim of further increasing the number of differences, p_3 for any three-pixel block p will be kept unchanged in the embedding process, in such case, p_3 can be used again and be reformed into a new three-pixel block with its two right neighboring pixels. Since p_3 is used twice, the number of differences approaches to image size.

Let $g(d)$ denote the number of differences whose absolute values equal d, where $0 \leq d \leq 253$. M_d and m_d are respectively used to represent the absolute values of differences having the largest and the small occurrence, i.e., $g(M_d) \geq g(M^{'})$ and $g(m_d) \leq g(m^{'})$ for $0 \leq M^{'}, m^{'} \leq 253$. In Lin's method, 1-bit watermark is embedded into difference whose absolute value is equal to M_d. For a pixel block p, if $d_1 = M_d$ and $d_2 = M_d$, then this p is capable of carrying two-bit watermark information. In the proposed method, in order to increase the number of differences used for embedding, we modify Lin's method to keep p_3 unaltered. For p, there exists five relations (i.e., $p_1 > p_2 > p_3$, $p_1 < p_2 < p_3$, $p_1 = p_2 = p_3$, $p_1 < p_2 > p_3$ and $p_1 > p_2 < p_3$) among p_1, p_2 and p_3. Take $p_1 > p_2 > p_3$ for example, since $d_1 = M_d$ and $d_2 = M_d$, after watermark embedding, $p_1^{'} = p_1 + w_1 + w_2$, $p_2^{'} = p_2 + w_2$ and $p_3^{'} = p_3$. Note that no change to relation among $p_1^{'}$, $p_2^{'}$ and $p_3^{'}$ occurs before and after watermark embedding, i.e., $p_1^{'} > p_2^{'} > p_3^{'}$. $d_1^{'}$ and $d_2^{'}$ are respectively used to denote the watermarked differences of d_1

and d_2. After watermark embedding, $d_1^{'} = p_1^{'} - p_2^{'} = p_1 - p_2 + w_1 = d_1 + w_1$, $d_2^{'} = p_2^{'} - p_3^{'} = p_2 - p_3 + w_2 = d_2 + w_2$. Hence, in the extraction process, by comparing $d_1^{'}$ (or $d_2^{'}$) with M_d and m_d, 2-bit watermark can be correctly extracted. Detailed modification is illustrated in Procedure 1.

2.1 Watermark Embedding

In the watermark embedding process, 1-bit watermark is embedded into differences whose absolute value equals M_d. Since each p have two absolute differences d_1 and d_2. d_1 is divided into three intervals (i.e., $d_1 = M_d$, $m_d \geq d_1 > M_d$ and $d_1 < M_d$ `or` $d_1 > m_d$) according to its value. Similarly, d_2 is also divided into three intervals (i.e., $d_2 = M_d$, $m_d \geq d_2 > M_d$ and $d_2 < M_d$ `or` $d_2 > m_d$). d_1 and d_2 are combined to generate nine combinations. Nine combinations respectively correspond to nine different embedding strategy detailedly shown in the following embedding procedure.

```
if d1 == Md
{
    if d2 == Md
        call embed_2_bits;
    else
        {
        if Md < d2 ≤ md
            call embed_1_bit_and_increase_difference;
        else
            call embed_1_bit_and_leave_unchanged;
        }
    }
elseif Md < d1 ≤ md
{
    if d2 == Md
        call increase_difference_and_embed_1_bit;
    else
    {
        if Md < d2 < md
            call increase_2_difference;
        else
            increase_difference_and_leave_unchanged;
        }
    }
else
{
    if d2 == Md
        call leave_unchanged_and_embed_1_bit;
    else
```

```
    {
    if M_d < d_2 < m_d
        call leave_unchanged_and_increase_difference;
    else
        Do nothing;
    }
}
```

When $m_d < d_1$ `or` $d_1 < M_d$ and $m_d < d_2$ `or` $d_2 < M_d$, we do nothing. Each of the remaining eight strategy will call a function to realize the embedding process. Eight functions are respectively illustrated in Procedure 1 to Procedure 8. With the help of Table 1 to Table 8, p can be correctly retrieved in the extraction process by comparing $d_1^{'}$ (or $d_2^{'}$) with M_d and m_d.

Procedure 1. Embedding_two_bits

$p_1 > p_2 > p_3$	$p_1^{'} = p_1 + w_1 + w_2, p_2^{'} = p_2 + w_2$
$p_1 < p_2 < p_3$	$p_1^{'} = p_1 - w_1 - w_2, p_2^{'} = p_2 - w_2$
$p_1 = p_2 = p_3$	if $w_1 == 1$ and $w_2 == 1$ $p_2^{'} = p_2 + 1$ elseif $w_1 == 0$ and $w_2 == 1$ $p_1^{'} = p_1 - 1, p_2^{'} = p_2 - 1$ else $p_1^{'} = p_1 - w_1, p_2^{'} = p_2 - w_2$
$p_1 < p_2 > p_3$	if $w_1 == 1$ and $w_2 == 1$ $p_2^{'} = p_2 + 1$ elseif $w_1 == 0$ and $w_2 == 1$ $p_1^{'} = p_1 + 1, p_2^{'} = p_2 + 1$ else $p_1^{'} = p_1 - w_1, p_2^{'} = p_2 + w_2$
$p_1 > p_2 < p_3$	if $w_1 == 1$ and $w_2 == 1$ $p_2^{'} = p_2 - 1$ elseif $w_1 == 0$ and $w_2 == 1$ $p_1^{'} = p_1 - 1, p_2^{'} = p_2 - 1$ else $p_1^{'} = p_1 + w_1, p_2^{'} = p_2 - w_2$

Table 1. Changes to Procedure 1 after watermark embedding

$p_1 > p_2 > p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$	$d_1^{'} \in \{M_d, M_d + 1\}$	$d_2^{'} \in \{M_d, M_d + 1\}$
$p_1 < p_2 < p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$		
$p_1 = p_2 = p_3$	$p_1 \neq p_2^{'} \neq p_3^{'}$		
$p_1 < p_2 > p_3$	$p_1^{'} < p_2^{'} > p_3^{'}$		
$p_1 > p_2 < p_3$	$p_1^{'} > p_2^{'} < p_3^{'}$		

Procedure 2. Embed_1_bit_and_increase_difference

$p_1 > p_2 > p_3$	$p_1^{'} = p_1 + w_1 + 1, p_2^{'} = p_2 + 1$
$p_1 < p_2 < p_3$	$p_1^{'} = p_1 - w_1 - 1, p_2^{'} = p_2 - 1$
$p_1 < p_2 > p_3$	if $w_1 == 1$ $p_2^{'} = p_2 + 1$ else $p_1^{'} = p_1 + 1, p_2^{'} = p_2 + 1$
$p_1 > p_2 < p_3$	if $w_1 == 1$ $p_2^{'} = p_2 - 1$ else $p_1^{'} = p_1 - 1, p_2^{'} = p_2 - 1$

Table 2. Changes to Procedure 2 after watermark embedding

$p_1 > p_2 > p_3$	$p_1^{'} \geq p_2^{'} > p_3^{'}$	$d_1^{'} \in \{M_d, M_d + 1\}$	$M_d + 1 < d_2^{'} < m_d + 1$
$p_1 < p_2 < p_3$	$p_1^{'} \leq p_2^{'} > p_3^{'}$		
$p_1 < p_2 > p_3$	$p_1^{'} < p_2^{'} > p_3^{'}$		
$p_1 > p_2 < p_3$	$p_1^{'} > p_2^{'} < p_3^{'}$		

Procedure 3. Embed_1_bit_and_leave_unchanged

$p_1 > p_2$	$p_1^{'} = p_1 + w_1$
$p_1 \leq p_2$	$p_1^{'} = p_1 - w_1$

Table 3. Changes to Procedure 3 after watermark embedding

$p_1 > p_2$	$p_1^{'} > p_2^{'}$	$d_1^{'} \in \{M_d, M_d + 1\}$	$d_2^{'} < M_d$ or $d_2^{'} > m_d$
$p_1 < p_2$	$p_1^{'} < p_2^{'}$		

Procedure 4. Increase_difference_and_embed_1_bit

$p_1 > p_2 > p_3$	$p_1^{'} = p_1 + 1 + w_2, p_2^{'} = p_2 + w_2$
$p_1 < p_2 < p_3$	$p_1^{'} = p_1 - 1 - w_2, p_2^{'} = p_2 - w_2$
$p_1 < p_2 > p_3$	if $w_2 == 1$ $p_2^{'} = p_2 + 1$ else $p_1^{'} = p_1 - 1$
$p_1 > p_2 < p_3$	if $w_2 == 1$ $p_2^{'} = p_2 - 1$ else $p_1^{'} = p_1 + 1,$

Table 4. Changes to Procedure 4 after watermark embedding

$p_1 > p_2 > p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$	$M_d + 1 < d_1^{'} < m_d + 1$	$d_2^{'} \in \{M_d, M_d + 1\}$
$p_1 < p_2 < p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$		
$p_1 < p_2 > p_3$	$p_1^{'} < p_2^{'} > p_3^{'}$		
$p_1 > p_2 < p_3$	$p_1^{'} > p_2^{'} < p_3^{'}$		

Procedure 5. Increase_2_differences

$p_1 > p_2 > p_3$	$p_1^{'} = p_1 + 2, p_2^{'} = p_2 + 1$
$p_1 < p_2 < p_3$	$p_1^{'} = p_1 - 2, p_2^{'} = p_2 - 1$
$p_1 < p_2 > p_3$	$p_2^{'} = p_2 + 1$
$p_1 > p_2 < p_3$	$p_2^{'} = p_2 - 1$

Table 5. Changes to Procedure 5 after watermark embedding

$p_1 > p_2 > p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$	$M_d + 1 < d_1^{'} < m_d + 1$	$M_d + 1 < d_2^{'} < m_d + 1$
$p_1 < p_2 < p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$		
$p_1 < p_2 > p_3$	$p_1^{'} < p_2^{'} > p_3^{'}$		
$p_1 > p_2 < p_3$	$p_1^{'} > p_2^{'} < p_3^{'}$		

Procedure 6. increase_difference_and_leave_unchanged

$p_1 > p_2$	$p_1^{'} = p_1 + 1$
$p_1 \leq p_2$	$p_1^{'} = p_1 - 1$

Table 6. Changes to Procedure 6 after watermark embedding

$p_1 > p_2$	$p_1^{'} > p_2^{'}$	$M_d + 1 < d_1^{'} < m_d + 1$	$d_2^{'} < M_d$ or $d_2^{'} > m_d$
$p_1 < p_2$	$p_1^{'} < p_2^{'}$		

Procedure 7. Leave_unchanged_and_embed_1_bit

$p_2 > p_3$	$p_1^{'} = p_1 + w_2, p_2^{'} = p_2 + w_2$
$p_2 \leq p_3$	$p_1^{'} = p_1 - w_2, p_2^{'} = p_2 - w_2$

Table 7. Changes to Procedure 7 after watermark embedding

$p_2 > p_3$	$p_2^{'} > p_3^{'}$	$d_1^{'} < M_d$ or $d_1^{'} > m_d$	$d_2^{'} \in \{M_d, M_d + 1\}$
$p_2 < p_3$	$p_2^{'} < p_3^{'}$		

Procedure 8. Leave_unchanged_and_increase_difference

$p_2 > p_3$	$p_1^{'} = p_1 + 1, p_2^{'} = p_2 + 1$
$p_2 < p_3$	$p_1^{'} = p_1 - 1, p_2^{'} = p_2 - 1$

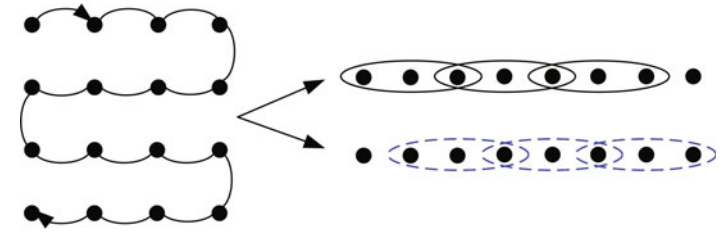
Table 8. Changes to Procedure 8 after watermark embedding

$p_2 > p_3$	$p_2^{'} > p_3^{'}$	$d_1^{'} < M_d$ or $d_1^{'} > m_d$	$M_d + 1 < d_2^{'} < m_d + 1$
$p_2 < p_3$	$p_2^{'} < p_3^{'}$		

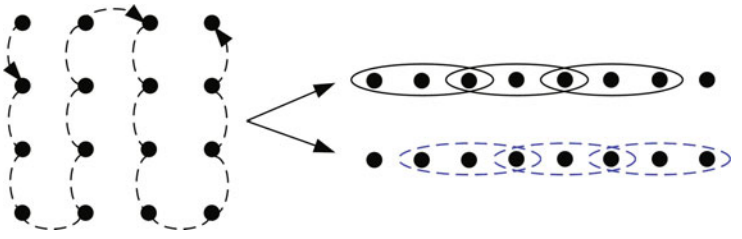
I is converted into a one-dimension pixel list respectively according to the arrow directions shown in Fig. 1(a) and Fig. 1(b). Each list has two different ways (respectively marked by black and blue) to partition all pixels into overlapped three pixel blocks. For each way, we find the number of absolute differences having the maximum occurrence. Hence, two lists will generate four absolute differences with the maximum occurrence, respectively use M_{d1}, M_{d2}, M_{d3} and M_{d4} to denote them. Find the maximum value from M_{di}, $1 \leq i \leq 4$, and $i \in \mathbb{Z}$ and use M_d to denote this value. I is converted into a one-dimension pixel list I_{D1} according to the selected order.

Three consecutive pixels p_1, p_2 and p_3 of I_{D1} is grouped into a pixel block (p_1, p_2, p_3), where $0 \leq p_1 \leq 255$, $0 \leq p_2 \leq 255$ and $0 \leq p_3 \leq 255$. For each p, calculate its two differences d_1 and d_2, and then select the corresponding procedure from Procedure 1 to Procedure 8 according to the combination of d_1 and d_2, finally obtain $p_1^{'}$, $p_2^{'}$ and $p_3^{'}$ based on the relation among p_1,p_2 and p_3.

A location map L_M is generated and denoted by a bit sequence of size $W \times H$. For a pixel block p, if $0 \leq p_1^{'} \leq 255$, $0 \leq p_2^{'} \leq 255$ and $0 \leq p_3^{'} \leq 255$ after performing embedding procedure, this p is marked by '1' in the map L_M. otherwise by '0'. The location map is compressed losslessly by an arithmetic



(a) Partition way in horizontal direction



(b) Partition way in vertical direction

Fig. 1. Partition way for a 4×4 images block

encoder and the resulting bitstream is denoted by $\mathcal{L}$. L_S is the bit length of $\mathcal{L}$. Embedding procedure is performed if p is marked by '1' in L_M. Otherwise, p is kept unchanged.

After the first L_S pixels have been processed, their LSBs are replaced by $\mathcal{L}$, and then the L_S LSBs to be replaced are appended to $\mathcal{P}$. Finally, the rest of $\mathcal{P}$ and the L_S appended bits are embedded into the remaining unprocessed blocks. After all the blocks are processed, a new watermarked image I_W is obtained.

2.2 Performance Analysis

The embedding distortion caused by the embedding procedure is illustrated in this subsection. Following that is the definition of data-hiding capacity.

Modifications to Pixels. Take p with $d_1 = M_d$ and $d_2 = M_d$ for example. Since $d_1 = M_d$ and $d_2 = M_d$, then two-bit watermark information is embedded into p. If $p_1 > p_2 > p_3$, after watermark embedding, $p_1^{'} = p_1 + w_1 + w_2$, $p_2^{'} = p_2 + w_2$ and $p_3^{'} = p_3$. For pixel p_1, the difference d_{p1} between $p_1^{'}$ and p_1 equals $p_1^{'} - p_1 = w_1 + w_2$. Similarly, $d_{p2} = p_2^{'} - p_2 = w_2$, $d_{p3} = 0$. However, in Lin's method, for $p_1 > p_2 > p_3$, $d_{p1} = w_1$, $d_{p2} = 0$ and $d_{p3} = w_2$. Through the comparison of the proposed method and Lin's method, the modification to p_1 caused by watermark embedding is increased. To conclude from the above example, we increase the number of differences at the cost of enhancing the embedding distortions.

Hiding Capacity. The maximum hiding capacity D is given by:

$$D = g(M_d) - L_S \tag{1}$$

From Eq. (1), a part of the available capacity is consumed by $\mathcal{L}$.

2.3 Data Extraction and Image Restoration

The watermarked image I_W is converted into a one-dimensional pixel list in the same way as was done in embedding.

For I_W, LSBs of the first L_S watermarked pixels are collected into a bitstream $\mathcal{B}$. $\mathcal{B}$ are decompressed by an arithmetic decoder to retrieve the location map. By identifying the EOS symbol in $\mathcal{B}$, the bits from the start until EOS are decompressed by an arithmetic decoder to retrieve the location map. Data extraction and pixel restoration is carried out in inverse order as in embedding. M_d and m_d are transmitted to the receiving side. With help of M_d and m_d, the watermark sequence can be correctly extracted. For each pair $p^{'} = (p_1^{'}, p_2^{'}, p_3^{'})$, if $p^{'}$'s location is associated with '0' in the location map, then it is ignored. Otherwise, p can be retrieved in virtue of Table 1 to Table 8.

3 Experimental Results

The proposed method is implemented and tested on various standard test images using MATLAB. The performance for four most frequently used 512×512 grayscale standard test image is presented in Table 9, where we use number of bits and PSNR values for measurement for capacity and distortion.

We modify Lin's method by keeping p_3 of any pixel block p. As a result, the proposed method increases the number of differences whose absolute value equals M_d at the cost of enhancing the modifications to p_1 and p_2. Take Procedure 1 for example, for the situation of $p_1 > p_2 > p_3$, when $w_1 = 1$ and $w_2 = 1$, $d_1^{'} = 2$, $d_1^{'} = 1$. Therefore, decreasing the modifications to p_1 is the key to increasing the payload while reducing the distortions. We will be devoted to research in increasing the performance of the proposed method in the future work.

Table 9. Performance of the proposed method

Image	Embedding capacity (bits)	PSNR value (dB)
Baboon	19265	46.7062
Boat	50407	47.1568
Lena	54069	47.6528
Barbara	37454	46.7380

4 Conclusions

In Lin's method, images are divided into non-overlapping three-pixel blocks. Each pixel block contains two pairs composed of two neighboring pixels. Absolute differences between pairs are calculated. And meanwhile, the absolute difference having the highest occurrence rate is obtained. 1-bit watermark is embedded into a pixel pair whose absolute difference is equal to this obtained value. Since each pixel block contains two differences, Lin's method achieves embedding rate of at most $\frac{2}{3}$ bpp (bit per pixel) for a single embedding process. With the aim of further increasing embedding rate, we modify the embedding process in Lin's method to keep the third pixel of a pixel block unaltered in the proposed method. Then, this unchanged pixel is used again to reform into a new pixel block with its two neighboring pixels. As a result, the embedding rate can reach to 1 bpp for a single embedding process.

References

1. Lin, C.-C., Hsueh, N.-L.: A lossless data hiding scheme based on three-pixel block differences. Pattern Recognition 41(4), 1415–1425 (2008)
2. Honsinger, C.W., Jones, P., Rabbani, M., Stoffel, J.C.: Lossless recovery of an original image containing embedded data. US patent, No 77102/E-D (1999)
3. Goljan, M., Fridrich, J., Du, R.: Distortion-free data embedding for images. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 27–41. Springer, Heidelberg (2001)
4. Celik, M.U., Sharma, G., Tekalp, A.M., Saber, E.: Reversible data hiding. In: Proceedings of IEEE 2002 International Conference on Image Processing, vol. 2, pp. 157–160 (2002)
5. Xuan, G.R., Yang, C.Y., Zhen, Y.Z., Shi, Y.Q.: Reversible Data Hiding Using Integer Wavelet Transform and Companding Technique. In: Cox, I., Kalker, T., Lee, H.-K. (eds.) IWDW 2004. LNCS, vol. 3304, pp. 115–124. Springer, Heidelberg (2004)
6. Celik, M.U., Sharma, G., Tekalp, A.M.: Lossless watermarking for image authentication a new framework and an implementation. IEEE Trans. on Image Processing, 1042–1049 (2006)
7. Thodi, M., Rodrguez, J.J.: Prediction-error based reversible watermarking. In: Proc. of ICIP, Genova, vol. 3, pp. 1549–1552 (October 2004)
8. Tian, J.: Reversible data embedding using a difference expansion. IEEE Trans. Circuits and Systems for Video Tech. 13(8), 890–896 (2003)
9. Alattar, A.M.: Reversible watermark using difference expansion of quads. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 3, pp. 377–380 (2004)
10. Yang, C.-Y., Hu, W.-C., Lin, C.-H.: Reversible data hiding by coefficient-bias algorithm. Journal of Information Hiding and Multimedia Signal Processing 1(2), 91–100 (2010)

Alphanumeric Shape Recognition of Fingertip Writing Trajectory

Ming-Fang Wu, Jen-Hsing Li, Ping-Tsung Wang, and Ruei-Tang Lin

Department of Electrical Engineering, Kun Shan University,
No. 949, Dawan Rd., Yongkang City, Tainan County 710, Taiwan (R.O.C.)
wumifa@mail.ksu.edu.tw

Abstract. The purpose of this thesis is to develop a recognition system of handwriting letter. This system uses web camera as the input device, and discriminates the alphanumeric shape after detecting out the trajectory of handwriting. This research presents a new method of fingertip identification. In the method, the fingertip is found out by the characteristic that the position of fingertip is usually on the external frame of hand block, and the position of fingertip can be determined according to the established conditions in this thesis. The identification system of this article divides into two stages. The first stage is the fingertip trajectory record. In this stage, after video frames input the system, the image processing of each frame and the identification of hand position are proceeded, and then the position of fingertip is determined so that we can track and record the trajectory of fingertip. The second stage is the recognition of alphanumeric shape. In this stage, we do the front disposal of trajectory and then extract features of the trajectory for the recognition of alphanumeric shape. The outcome of the experiment shows that the system can track on the fingertip smoothly, and it also has high success rate on the identification of alphanumeric shape.

Keywords: fingertip, tracking, trajectory, alphanumeric shape, recognition.

1 Introduction

The handwriting recognition technology based on gesture can be divided into two major kinds. One kind is to take down the information of hand based on apparatus such as the glove [1], another kind is based on image [2]. The research of glove-based methods must use the data glove to make the detailed information of the hand, for instance the crooked degree and the interval of finger. The advantage of glove-based methods is that it can obtain the comparatively accurate data, and it is free from interruption by external environment. But its disadvantage involves: Glove has restriction of the scope of activity, the cost is comparatively high, and the glove life-span is relatively short. In the hand tracking technique that performed based on images, it retrieves information about hands through the picture shooting performed by one or multiple sets of cameras. There are two kinds of ways to recognize the information about hands. One is based on models [3]. In this method, a picture of the hand is obtained by a camera; then, the hand position is figured out via computer and

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 81–91, 2010.

matched with the target sample. However, such a method needs more tremendous operand. The other one is based on hand contour [4][5]. In this method, the image of a hand is retrieved by a camera. After the hand position is identified, the necessary information calculation is performed via hand fringe, ex. palm position or fingertip tracking. This method has lesser operand, more suitable for real-time applications.

After the fingertip trajectory is acquired, the word recognition will be performed, to recognize what kind of word trajectory the written trajectory is. There are three most common ways to perform the recognition: Hidden Markov Model [6][7], Dynamic Time Warping[8], and Neural Network. Hidden Markov Model is called "HMM" in abbreviation, which is developed from statistics and is often used in Acoustic Model Recognition. The feature of Dynamic Time Warping is that it can change the lengths of the matched data, often used in researches on Speech Recognition originally, similar to Hidden Markov Model. Neural Network performs various kinds of recognition by means of imitating biological neuron systems, which has extensive applications.

In order to achieve the purpose of recognizing fingertip handwritten words, the screening of skin color areas is performed through the techniques of color system conversion and morphological image processing in the paper first. Next, the hand and the fingertips are extracted, and the handwritten trajectory is tracked. Then, the recognition of the recorded trajectory is performed. Finally, actual images of handwriting are used for the implementation of the experiments in every procedure and stage. Therefore, this paper is divided into five sections. Chapter 1 is the Introduction, mainly describing the research motivation and the procedures for the system of fingertip handwriting recognition. Section 2 introduces the methods for the screening of skin color areas and the determination and tracking of fingertip positions. Section 3 describes the methods for fingertip trajectory pre-processing and handwritten word recognition. The experiment results are discussed in Section 4. The final conclusions are described in Section 5.

2 The Determination of Skin Color Areas and the Tracking of Fingertip Trajectory

In order to reduce the influence of illuminant intensity on colors, the HSV color space [9] is adopted in this study to sift out skin color areas. In this study, all the pixel points conforming to the skin color conditions in the images are found out, and these pixel points are binaryzated to be grey scale value 255, while other pixel points not conforming to the skin color conditions were grey scale value 0. For example, Fig. 1(b) shows the results of Fig. 1(a) after skin color testing.

After the skin color areas are acquired and the noise is removed, the fingertip trajectory tracking is performed. In the process, the determination of the hand areas is undertaken, to calculate the fingertip positions. Then, the coordinates of fingertip positions are recorded one by one to complete the steps for trajectory tracking, namely that the process for fingertip trajectory tracking is divided into the procedures such as the determination of the hand areas, the determination of the fingertip positions, and the record of fingertip trajectory coordinates, which are discussed respectively in Section 2.1, 2.2 and 2.3.

(a) Original Image (b) Results of skin color screening

Fig. 1. HSV skin color screening

2.1 The Determination of the Hand Areas

Under a general operating environment, the skin color picture retrieved by a camera is not limited to the hands. Usually, the facial part will be included in the picture as well; or, the hands are not within the picture. Therefore, we should ponder on how to mark the hand areas correctly, so as to correctly find out where the fingertips are. Such a procedure is achieved mainly through the two steps—the marking of the skin color areas and the removal of facial areas.

(1) Marking of the skin color areas

In this section, the connected component labeling [11] is mainly utilized in this step. The connected pixels belonging to the same object in an image are found out, so as to mark every independent object in an image. In the binarized image whose skin color areas have been marked completely, the facial parts may be also included in the picture at the same time, in addition to the hands. Besides, the square measure of the hands is usually smaller than that of the face. Thus, after the top two areas conforming to the threshold value (1200 pixels) have been selected and framed, the removal of the facial part should be undertaken. Shown as Fig. 2, the red rectangle frame shows the results of the determination of the top two skin color areas in an image.

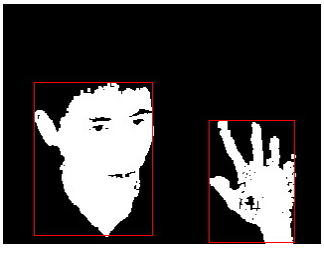


Fig. 2. Skin color areas marking

(2) Removal of facial areas

Generally speaking, the skin texture of palm is different from the back of hand and the face. Usually, it is whiter and pinker than the face. Also, such a kind of skin

color accounts for larger proportion in the hand area than the facial area does. Therefore, in this paper, the hand is determined based on the proportion of this skin color [12]. For calculating the proportion of this skin color, a threshold value should be set necessarily. Then, the statistics of the pixels of white and pink colors in the candidate skin color areas are made, and the proportion of the total pixels of the pixels and skin color areas conforming to this threshold value is calculated. Such a proportion is called "the proportion of red skin color". If both the proportions of red skin color in the two areas are greater than certain value, the one with the greater proportion is the hand. If there is only one area greater than the threshold of the proportion of red skin color, this area is the hand. If neither of the two conforms to the threshold, the picture is determined to have no hand areas. For example, Fig. 3(b) shows the results of hand area determination processed based on Fig. 3(a). The white area in the figure is the hand, while the gray one is the removed facial area.

(a) Original Image (b)Results of hand screening

Fig. 3. Hand screening based on the proportion of red skin color

2.2 Determination of the Fingertip Position

In order to speed up the processing time, this study proposes three-stage methods for completing the determination of fingertip positions: (1) First, calculate the outer frame of the hand area; (2) Then, search for the coordinates of the sharp points on the fringe from the outer frame; (3) Finally, determine the fingertip position from the coordinates of the sharp points on the fringe. The three stages are explained respectively as follows:

(1) Calculate the outer frame of the hand area
The way to calculate the outer frame of the hand area is: The statistics of the coordinate position of every point is processed, regarding the pixels of the hand area from up to down and left to right, to figure out the maximum and minimum of the coordinate X and Y; then, based on this value, the rectangle outer frame is marked, namely the outer frame of the hand image.

(2) Calculate the sharp points on the fringe
After the outer frame of the hand position is figured out, then, the possible fingertip position is calculated. Because the finger will be on the outer frame of the hand position, therefore, in this study, the fingertip position is searched, from the four edges of the outer frame and the fringe of the hand position. As for the

way to search, the four edges of the outer frame of the hand position are divided into four parts first, to search clockwise orderly, based on the four corners as the starting points. When any pixel of the hand fringe is encountered on the outer frame, this point is set to be the fingertip midpoint. Then, based on the fingertip midpoint as the starting point, the origin is set to be 0, after the neighboring coordinate is found; also, this coordinate is set to be the center for next search. After the first search is completed according to this way, the recursion is performed for 30 times to the front and to the back respectively, so as to search the front point and the back point of the fingertip. After the search for the four edges of the outer frame is completed, all the sharp points on the outer frame of the image on the hand fringe, as well as the coordinates located at the 30th point in the front and back directions of the sharp point, can be acquired. The results are shown as Fig. 4. In the figure, the yellow triangle is drawn, based on the connection lines between the sharp point and the front and back points at a distance of 30 pixels.

Fig. 4. Searching result of hand cusp

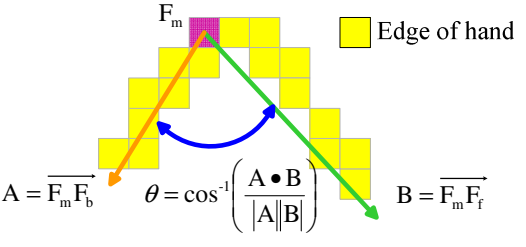


Fig. 5. Angular calculation of hand cusp

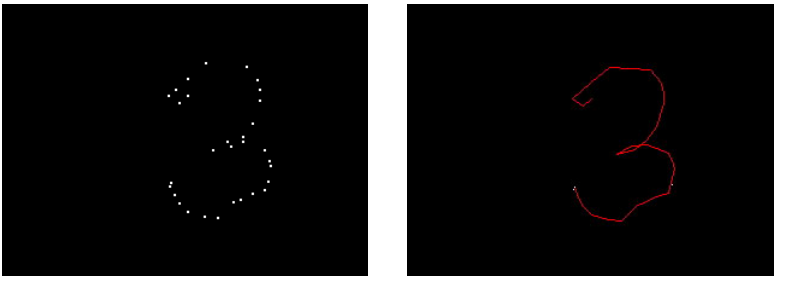
(3) Determine the fingertip position

After the sharp point positions on the four edges of the image's outer frame are acquired, then, the screening for fingertip locations is processed among these four sharp points, in order to select the real fingertip position. Shown as Fig. 5, in this step, the two vectors $\overrightarrow{F_m F_f}$ and $\overrightarrow{F_m F_b}$ are figured out first, based on the sharp point F_m and the front and back points F_f and F_b. Then, the included angle is figured out, based on the two vectors; then, the included angle between the sharp point coordinate and the coordinates of the 30^{th} pixel on the front and back points can be acquired.

On the hand image, because the fingertips on the sharp points on the four edges of the outer frame form the smallest included angle, therefore, the fingertip will be acquired by means of contrasting the sharp point with the smallest included angle on the four edges. However, there is an exception: When no fingers are extended from the hand, it means that it is not in a writing situation and the fingertips do not show up. Thus, the maximum of the included angle between the fingertips should be stipulated. When this angle is exceeded, it means that this included angle is not the fingertip position. From the actual experience in this study, the maximum of the included angle of the fingertip is 45°.

2.3 Record of Fingertip Trajectory

After the fingertip coordinate is determined, the record of fingertip trajectory can be processed. As for the way of recording in this paper, the fingertip recognition is processed on every picture. The record will be taken down, whenever there is any fingertip showing up. The record will not be taken down, if no conditions of the fingertip are conformed. Shown as the white coordinate points in Fig. 6(a), when the record is completed, the trajectory pattern will be a sequence of coordinates. The connection of the coordinate points in Fig. 6(a) will be Fig. 6(b). Such a sequence of coordinates can be used for fingertip trajectory recognition.



(a) Points of the handwritten trajectory (b) Connection of the points

Fig. 6. Handwritten trajectory

3 Word Trajectory Recognition

After the fingertip trajectory tracking is completed, we can go on to process the stage of word trajectory recognition. The complete procedure of this stage can be divided into two parts. The first comprises real-time fingertip trajectory tracking and word trajectory recognition. In this part, the trajectory pre-processing is undertaken first. The steps for retrieving the characteristic values of the trajectory and for comparing the word trajectory are undertaken. After the matching is made, the results can be output. The other part comprises the establishment of the trajectory sample database. In this part, the trajectory samples are collected first to undergo the preprocessing, for retrieving the characteristic values, and then the establishment of a matching database is undertaken. The systems able to be determined by the sample categories in the database can be used for recognizing word categories. In this paper, the words are mainly based on the ones that can be written in a single stroke.

3.1 Trajectory Processing

In the process of word handwriting, unnecessary encumbrances will occur on the trajectory due to many factors. For example, the noise misjudged during the trajectory recording, the length of the word stroke, the different writing habits, etc. All of these will make the trajectory more complicated, increasing difficulty in recognition. Therefore, the smoothing processing and the screening for trajectory samples [14]

should be performed first, under the situation in which the trajectory will not affect the characteristics, which was explained in detail as follows:

(1) Screening for trajectory sample points
The trajectory screening is divided into two parts. In the first part, the noise caused by the erroneous judgment of trajectory is removed. The distance between the noise and the front and back trajectory points is usually larger than that between correct trajectory points, which can be judged according to the size of the distance between every point in the trajectory. In the second part, the excessively-near distance is eliminated, and the quantity of trajectory points is adjusted. The way to eliminate is to orderly check the distance of the neighboring points in the trajectory. If the distance is excessively-near, the point will be eliminated. The check will be processed, until the necessary quantity of trajectory points is achieved.

(2) Trajectory smoothing
As for the smoothing method in this study, every trajectory point in the trajectory sequence is calculated orderly. The area formed by every trajectory point and its front and back trajectory points is viewed as a triangle, to calculate the barycenter. Based on this triangle barycenter as the new trajectory point, the new trajectory undergoing the smoothing can be acquired, after the calculation of the trajectory sequence is processed orderly. Shown as Fig. 7, the red line shows the trajectory of the original curve, and the yellow line shows the result of smoothing.

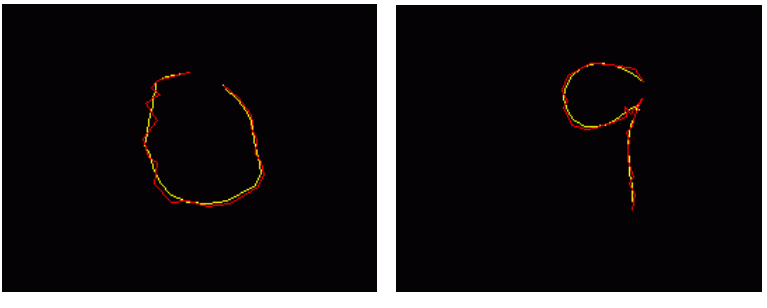


Fig. 7. Result of trajectory smoothing

3.2 Characteristic Value of the Trajectory

In this study, the matching is undertaken, based on the two kinds of characteristics in the word trajectory. The first one is eight-directional code for the trajectory points, and the second one is the head-tail position of the trajectory, respectively explained in detail as follows:

(1) Eight-directional code
In this paper, the 360° will be divided into 8 portions. The ± 25.5° from 0° is defined to be No.1. The 8 numbers are defined counterclockwise. Then, the angles between every point in the trajectory sequence are matched, and the codes are processed according to the partition it belongs to.

(2) Head-tail position of the trajectory
In this study, the way to recognize similar word trajectory is processed mainly according to the 16-lattices within the range of the word trajectory. Each lattice is given a number. Also, the number regarding the position of the starting point and the final point of the word trajectory is calculated further, and then the head-tail position of the trajectory can be acquired.

3.3 Word Trajectory Matching

In this study, the method for comparing word trajectory comprises two parts. The first part is based on Markov Models[6][7], to construct the state transition probability through every change in word trajectory, for performing the calculation aimed at the input word trajectory based on this state transition probability. Then, the trajectory matching aimed at the probability and the sum is performed**.** In the second part, the matching database is constructed based on the characteristics of the starting point and the final point of the word trajectory. Then, the matching aimed at the input word trajectory in this database is performed, while the values are given. Then, the matching values of the two parts are added. The highest value is the matching result.

Because the matching database is divided into the two parts mentioned above, the undertaking of the matching method is also divided into two parts. First, the recognition of trajectory direction is performed. The sequence of the input eight-directional code of the trajectory as well as the transition data is matched in terms of probability, and then multiplied by a parameter. After all the models of word font are matched completely, the matching data of this part can be acquired. In the second part, the matching of the starting point and the final point in the trajectory is performed. The positions of the input starting points and the final points of the word trajectory are matched with the database. The numeral values are given, according to the conformation degree. Then, the matching data of this part can be acquired. Next, the sums of the corresponding matching data of the two parts are calculated. The higher numeral values are the matching results.

4 Experiment Results

In real situations, a human face along with the hands tend to be included in the picture, and the range of human facial skin color may be more obvious than that of the hands. Thus, the hand tacking will be interfered severely. Therefore, before the hand tracking is processed, the human facial part should be removed. In order to verify the way to remove facial areas proposed by this study, the human face together with the hands are actually included in the picture simultaneously in the experiment in this section, to perform hand detection and tracking. Fig. 8 shows the picture of the results of hand tracking. The framed areas in the figure show the detected hand positions. The color pictures on the upper row show the results of the actual hand tracking. The black-and-white pictures on the lower row show the results of facial area removal, in which the white areas represent the hand areas, while the gray scale areas represent the removed facial areas.

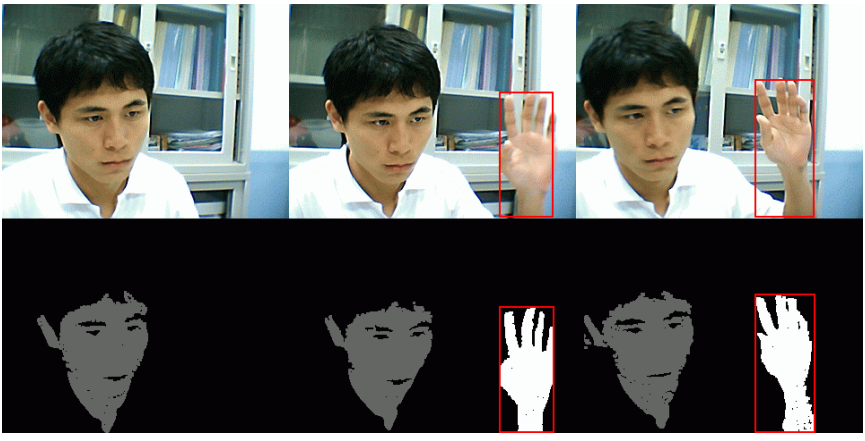


Fig. 8. Results of hand tracking

After the hand area has been detected and tracked, then, this study performs the tracking based on the fingertip tracking method proposed in Section 2.2. The results are shown as Fig. 9. In the figure, the yellow circles represent the fingertip positions. Because the tracking method proposed in this study will not be affected by the fingertip angle, the results show that recognition and tracking can be performed successfully aimed at the fingertips in different people and different angles.

Fig. 9. Picture of fingertip tracking

After the fingertip has been tracked, the word writing is undertaken immediately, while its trajectory is recorded. The trajectory recorded is a series of coordinates. If the coordinates are listed directly, it will be harder to observe the results via naked eyes directly. Therefore, in this paper, the sequence of coordinates is transformed into a picture, and the result is shown as Fig. 10, in which the white light-spots are namely the trajectory points of fingertip writing.

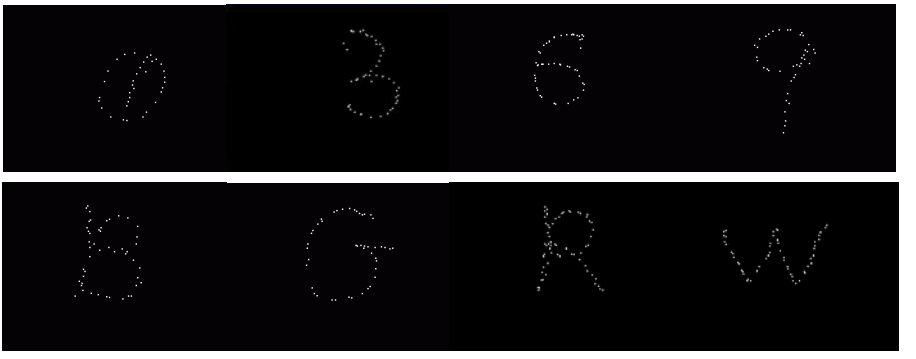


Fig. 10. Picture of trajectory recording

5 Conclusions

In this study, based on image processing, the word recognition technique of handwriting input is developed, which has achieved the purpose of recognizing handwritten numerals and English alphabets via the input image through a web camera. In the process, the handwritten input images orderly undergo pre-processing, hand area recognition, fingertip tracking and trajectory recording. Also, after the recorded word trajectory have undergone the screening for sample points, the processing of trajectory smoothing, and the retrieval of characteristic values, the matching is performed, so as to obtain the results of word recognition.

According to the experiment results in the last section, more than 93% of success rate can be achieved in the stages of hand determination, fingertip recognition and tracking, which has verified that the method proposed in this study can perform tracking aimed at the positions of hands and fingertips smoothly. As for the aspect of word recognition, although the conditions for word recognition vary, the total average of the success rate can achieve 92.84%, which has verified the accuracy and feasibility of the word recognition method in this study.

References

1. Hung, C.H.: A Trajectory-Based Approach to Gesture Recognition. Master Thesis, Institute of Computer Science & Information Engineering, National Central University (2006)
2. Oka, K., Sato, Y., Koike, H.: Real-time Fingertip Tracking and Gesture Recognition. IEEE Computer Graphics and Applications 22(6), 64–71 (2002)
3. Huang, T.S., Wu, Y., Lin, J.: 3D Model-based Visual Hand Tracking. In: IEEE Conf. on Multimedia and Expo., vol. 1, pp. 905–908 (2002)
4. Davis, J., Shah, M.: Visual Gesture Recognition. In: Proc. of IEE Image and Signal Processing, vol. 141, pp. 101–106 (1994)
5. Oka, K., Sato, Y., Koike, H.: Real-time Tracking of Multiple Fingertips and Gesture Recognition for Augmented Desk Interface Systems. IEEE Automatic Face and Gesture Recognition, 429–434 (May 2002)

6. Rabiner, L.R., Juang, B.: An Introduction to Hidden Markov Models. ASSP Magazine of the IEEE 3(1), 4–16 (1986)
7. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
8. Martens, R., Claesen, L.: Dynamic Programming Optimization for on-line Signature Verification. Document Analysis and Recognition 2, 653–656 (1997)
9. Gonzalez, R.C., Woods, R.E.: Digital Image Processing 3/E. Prentice-Hall, New Jersey (2007)
10. Lin, C.F.: The Application of Color Vision Technique on Face-Shape Decision and Skin-Character Analysis. Master Thesis, Department of Electrical Engineering, Kun Shan University (2005)
11. Shapiro, L.G., Stockman, G.C.: Computer Vision, pp. 65–68. Prentice-Hall, New Jersey (2001)
12. Lin, Y.C.: Design and Implementation of A Vision-Based Fingertip Writing Interface. Master Thesis, Department of Computer Science and Information Engineering, National Dong Hwa University (2005)
13. Yeh, T.H.: Application of Web Camera in Specific Fingertip Trajectory Pattern Recognition. Master Thesis, Department of Computer Science and Information Engineering, Southern Taiwan University (2006)
14. Liu, N., Lovell, B.C., Kootsookos, P.J.: Evaluation of HMM Training Algorithms for Letter Hand Gesture Recognition. In: IEEE International Symposium on Signal Processing and Information Technology, Darmstadt, December 14-17, vol. 1(1), pp. WA4–7 (2003)
15. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Roy. Stat. Soc. 39(1), 1–38 (1977)
16. Ni, H.W.: Automatic Bird-Species Recognition System Based on Syllable-Type HMM. Department of Computer Science and Information Engineering, Chung-Hwa University (2005)

Recognition of Tire Tread Patterns Based on Gabor Wavelets and Support Vector Machine

Deng-Yuan Huang[1], Wu-Chih Hu[2], Ying-Wei Wang[3], Ching-I Chen[1], and Chih-Hsiang Cheng[1]

[1] Department of Electrical Engineering, Dayeh University, 515 Changhua, Taiwan
{kevin,r9703029}@mail.dyu.edu.tw
[2] Department of Computer Science and Information Engineering, National Penghu University of Science and Technology, 800 Penghu, Taiwan
wchu@npu.edu.tw
[3] Department of Marketing and Logistics Management, National Penghu University of Science and Technology, 800 Penghu, Taiwan
ywwang@npu.edu.tw

Abstract. In this paper, we propose a novel algorithm based on Gabor wavelets and support vector machine (SVM) for recognition of tire tread patterns. Input tire images are first preprocessed by morphological opening to enhance the features (or textures) on tire surface. The grooves in tire surface are salient important features for a tire matching system. We detect the tire tread patterns of being grooved or wavy and use this feature to train various SVM classifiers. The features of tire tread patterns are then represented by Gabor wavelets, and feature extraction is further carried out by principal component analysis (PCA). Finally, the matching processes are achieved by the classifiers of SVM, Euclidean distance and cosine distance. Result shows that the recognition rate of 60% for tire images can be obtained by the SVM classifier when 15 tire tread patterns are used.

Keywords: Tire tread patterns, Gabor wavelets, Principal component analysis (PCA), Support vector machine (SVM), Pattern recognition.

1 Introduction

Tire tread patterns are broadly used in the investigation of traffic accidents for recognizing the responsibility of car drivers. They are quite useful especially for a hit-and-run accident by identifying a given tread pattern to match existing tires to further reduce the scope of investigation. The tread patterns and tire marks at the scene of accident are commonly used in analysis of the behavior of car drivers, collision locations, vehicle direction, and post-collision trajectories. The above information is of importance in forensic evidence for law enforcement agencies [1]. However, most pattern matching processes are manually operated, which are labor-intensive and require visual identifications of extensive database of tire tread patterns. In this work, we propose to automate the matching process of tire tread patterns by creating

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 92–101, 2010.

effective features representation, extraction and classification, and then to locate candidate matches from the database of existing tread pattern images.

The first step for classifying a given tire tread pattern to existing models is to perfectly reconstruct the imprints of tires. Two popular methods that are based on 3D photogrammetry [2,3] and 2D image processing [4,7] are commonly used. Thali et al. [2] used a CAD-supported photogrammetry for 3D reconstruction of the face of the injury run over by a vehicle tire. They serially photographed the objects (injury and surface of tires). These photo sequences were then evaluated with the RolleiMetric multi-image evaluation system. Their system measured the spatial location of points in the photo sequences, and created 3D data models of the objects. The results show the perfect 3D match between the tire tread and the facial injury, revealing how greatly 3D methods surpass the classic 2D image processing method. Another similar study [3] is to use high resolution 3D optical surface scanner to capture footwear impressions and tire tracks in snow, which is also very satisfactory in different meteorological conditions. The only disadvantage for 3D methods is very computationally expensive.

To identify tire tread patterns by 2D image processing methods, feature representations for tire textures are commonly achieved by image transformations such as 2D fast Fourier transform (FFT) [4] and Gabor wavelets [5]. The curse of dimensionalities is the problem that arises from many more features than training images, which is our case. High dimensional data of tire images are difficult to tackle with; therefore, PCA [6] is usually adopted for dimension reduction by projecting the image data into the most expressive feature axes that are also called principal components. This process is known as feature extraction. The matching process of features are then accomplished by some popular statistical methods such as Fuzzy C-mean, K-nearest neighbor (KNN), neural network, SVM, Euclidean distance, or cosine distance that is used to measure the cross correlation of two variables.

Colbry et al. [4] developed a tire matching system to automate the process of tire pattern recognition that uses 2D FFT coefficients to represent the most informative features of tire imprints. To reduce the complexities of high dimensional image data, both PCA and PSA (power spectrum analysis) are used in the processes of features extraction. A KNN classifier is then used to find the best matches in the database for test samples of tire. However, the evaluations on tire matching rates in their work are absent. Jung et al. [7] proposed a hierarchical fuzzy pattern matching classifier for recognizing tire tread patterns. They constructed a binary decision tree by using fuzzy C-means (FCM) algorithm in their design. The results show the superiority of their proposed method.

In this paper, we propose a tire matching system that is based on Gabor wavelets and SVM classifier. The Gabor wavelets are first used to represent the features of tire tread patterns with different scales and orientations. Several SVM classifiers are then trained according to the type of tire patterns to be grooved or wavy. The matching processes are carried out by the classifiers of SVM, Euclidean distance (ED) and cosine distance (CD). Results show that the tire recognition rate of 60% is obtained by the SVM classifier when 15 tire patterns are tested.

The remainder of the paper is organized as follows: Section 2 gives a detailed description of the proposed method for recognizing tire tread patterns of vehicles. Section 3 describes the preparation of training and testing database for tire images. The experimental results are discussed in Section 4, and Section 5 contains the concluding remarks of this work.

2 The Proposed Method for Tire Pattern Recognition

The flowchart of the proposed method for tire patterns recognition is shown in Fig.1. The tire matching system consists of the following major components including (1) image preprocessing, (2) detection of groove number, (3) transformation of Gabor wavelets, (4) dimensionality reduction of features, and (5) feature matching. First, the image preprocessing is to enhance the features of tire tread patterns to be more salient. Second, we detected the tire patterns of being grooved or wavy and then used this feature to train various SVM classifiers. Third, the features (or textures) of tire images are represented by the Gabor wavelets with 5 scales and 8 orientations. Finally, PCA is used to reduce the high dimensionalities of Gabor features that are further classified by the methods of SVM, ED, and CD, respectively.

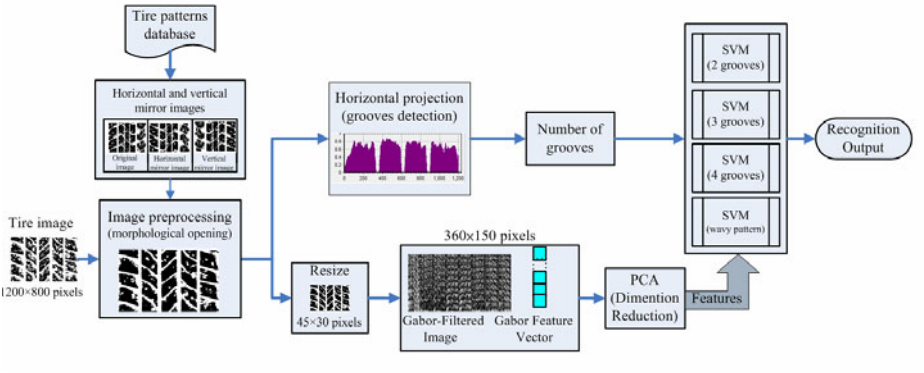


Fig. 1. The flowchart of the proposed method for recognizing tire tread patterns

2.1 Image Preprocessing

Since the orientation of tire patterns significantly affects the performance of a tire matching system; therefore, we manually corrected the tire patterns to be frontal and upright. Subsequently, the tire images are processed by morphological opening to eliminate some fine holes in the image. As shown in Fig. 2(a), the original image used is of size 1200×800 pixels. The features of tire patterns becomes more salient after image preprocessing (see Fig. 2(b)) due to the elimination of small holes.

Fig. 2. (a) Original tire image, (b) Result after image preprocessing procedures

2.2 Detection of Groove Number

The detection of the groove number of tires was performed on the binary images. As shown in Fig. 2(b), the rib patterns that are oriented primarily in the rolling direction of tire are the grooves with white pixels. In this work, we projected all the pixels of tire image onto the horizontal axis (i.e., *x*-axis) and then accumulated the black pixels point by point along this axis. Suppose the binarized tire image is of size with *n* rows and *m* columns. The probability of occurrence of black pixels $p[x]$ can be given as

$$p[x] = \sum_{y=0}^{n-1} \frac{G_{xy}}{n},\ 0 \le x \le m-1$$

$$\textit{where} \tag{1}$$

$$G_{xy} = \begin{cases} 1 & \textit{if } f(x,y) = 0 \\ 0 & \textit{if } f(x,y) = 255 \end{cases}$$

where $f(x,y)$ represents the gray level of pixel at location (x, y) on tire image. If $f(x,y)=0$, it represents a black pixel; otherwise, it is white. Therefore, the number of grooves on tire surface can be determined by the constraints of both $p[x] \le t_1$ and $\Delta x \ge t_2$, where t_1 is a threshold set to 0.1, and t_2 is the total number of continuous points with $p[x] \le t_1$ along the horizontal axis, which is set to 10. The result of 3 grooves in tire image (see Fig. 2(b)) detected by the proposed method is shown in Fig. 3.

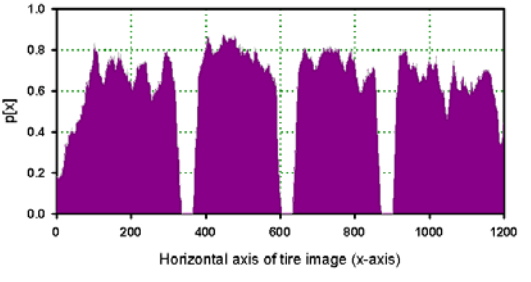


Fig. 3. Three grooves in Fig. 2(b) detected by the method of horizontal projection

2.3 Principle of Gabor Wavelets

Gabor wavelets (or called filters, kernels) can capture the most salient visual properties such as spatial locality, orientation selectivity, and spatial frequency characteristics. Prior to capturing the features of tire tread patterns by Gabor wavelets, the input tire images are resized from 1200×800 pixels to 45×30 pixels by a bilinear interpolation method for greatly reducing computational complexities.

Mathematically, a 2D isotropic Gabor wavelet is the product of a 2D Gaussian and a complex exponential function. The general expression can be expressed as

$$g_{\theta,\gamma,\sigma}(x,y) = \exp\left(-\frac{x^2+y^2}{\sigma^2}\right)\exp\left(\frac{j\pi}{\lambda}(x\cos\theta + y\sin\theta)\right) \tag{2}$$

The parameter θ represents the orientation, λ is the wavelength, and σ indicates the scale at orthogonal direction. However, with this set of parameters the Gabor wavelet does not scale uniformly as the parameter σ changes. It is better to use a parameter γ=λ/σ to replace λ so that a change in σ corresponds to a true scale change in the Gabor wavelets. Moreover, it is convenient to use a 90° counterclockwise rotation to Eq. (2), such that θ expresses the normal direction to the Gabor wavelet edges. The Gabor wavelets can be rewritten as

$$g_{\theta,\gamma,\sigma}(x,y)=\exp\left(-\frac{x^2+y^2}{\sigma^2}\right)\exp\left(\frac{j\pi}{\gamma\sigma}(x\sin\theta-y\cos\theta)\right) \tag{3}$$

In this work, a set of parameters of the Gabor wavelets is used as γ=0.785, θ={0°, 90°, 72°, 45°, 36°, -72°, -45°, -36°}, and σ={1, 2, 3, 4, 5} shown in Fig. 4. Thus, 40 Gabor responses for each tire image can be obtained. Each Gabor filter is then concatenated into a pattern vector with 54,000 (=360×150) elements for a tire image of size 45×30 pixels. The responses of Gabor wavelets for 3-groove and wavy patterns of tire images are shown in Fig. 5 and Fig. 6, respectively.

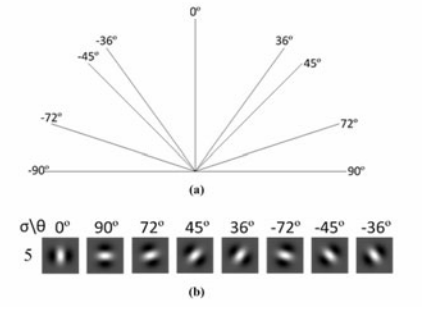


Fig. 4. (a) Schematic diagram of orientations for Gabor wavelets, (b) The response of Gabor wavelets for different orientations at scale σ=5

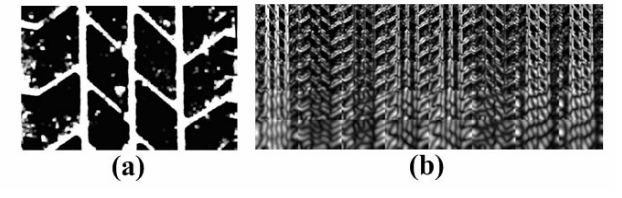


Fig. 5. (a) Tire image with 3-groove pattern, (b) Gabor response of image (a)

2.4 Principal Component Analysis

The method of PCA [6] is a popular dimensionality reduction technique with the purpose to find a set of orthonormal vectors in the data space, which maximize the data's variance and map the data onto a lower dimensional subspace spanned by these vectors.

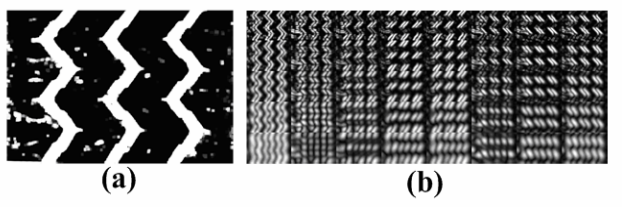


Fig. 6. (a) Tire image with wavy pattern, (b) Gabor response of image (a)

Consider a dataset with M images, $x_i \in \Re^N (i = 1,..., M)$, belonging to C_T classes, and N is the total number of pixels in an image. The global mean image of the training set is set to μ, and suppose $A = [x_1 - \mu,..., x_M - \mu] \in \Re^{N \times M}$. The total scatter matrix, $S_T \in \Re^{N \times N}$, is defined as

$$S_T = \sum_{i=1}^{M} (x_i - \mu)(x_i - \mu)^T = AA^T \tag{4}$$

Direct solving the eigenproblem of S_T is computationally expensive due to its high dimensionalities. Instead of direct finding the eigenvector W_{PCA} of S_T, we solved the eigenvalue problem, $RV_{PCA} = V_{PCA}\Lambda$, to obtain the eigenvectors, $V_{PCA} \in \Re^{M \times P}$, and the eigenvalues, $\Lambda = diag[\lambda_1,...,\lambda_P] \in \Re^{P \times P}$, with decreasing order $\lambda_1 \geq ... \geq \lambda_P > 0$, where λ_i is the nonzero eigenvalue of the matrix $R = A^T A \in \Re^{M \times M}$ $(M \ll N)$. Consequently, the PCA subspace W_{PCA} can be formed by multiplying the matrix A with the eigenvectors V_{PCA}, that is, $W_{PCA} = AV_{PCA} \in \Re^{N \times P}$. Therefore, the feature vector $\boldsymbol{y}$ of an image $\boldsymbol{x}$ can be obtained by projecting $\boldsymbol{x}$ onto the subspace spanned by W_{PCA}, that is

$$y = W_{PCA}^T (x - \mu) \in \Re^P \tag{5}$$

2.5 Support Vector Machine

In principle, one SVM classifier searches for an optimal hyperplane that maximizes the margins of their decision boundaries to ensure that their worst-case generalization errors are minimized, which is known as structural risk minimization (SRM).

To perform the classification between two classes, a nonlinear SVM classifier is commonly used by mapping the input data (x_i, y_i) into a higher dimensional feature space using a nonlinear operator $\Phi(x)$, where $x_i \in \Re^d$ and $y_i \in \{+1, -1\}$. Therefore, the optimal hyperplane can be estimated as a decision surface. That is,

$$f(x) = \mathrm{sgn}\left(\sum_i y_i \alpha_i K(x_i, x) + b \right) \tag{6}$$

where sgn(•) represents the sign function, and $K(x_i, x) = \Phi(x_i)^T \Phi(x)$ is the predefined kernel function that satisfies Mercer's condition [8]. In our work, radial basis function (RBF) is used, and it is defined as follows

$$K(x_i, x) = \exp\left(-\gamma \left\| x_i - x \right\|^2\right), \ \gamma > 0 \tag{7}$$

where γ=0.25. The coefficients α_i and b in Eq. (6) can be determined by the following quadratic programming (QP) problem

$$\begin{aligned} &\max\left[\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)\right] \\ &s.t. \sum_i \alpha_i y_i = 0 \\ &0 < \alpha_i < C, \ \forall i \end{aligned} \tag{8}$$

The parameter C is a penalty that represents the tradeoff between minimizing the training set error and maximizing the margin, where C=*8* is determined empirically. Since SVM is a binary classifier, it should be extended for a C_T-class problem in hand gesture recognition. We used the so called one against one approach, which is a pairwise method and needs to train $C_T(C_T - 1)/2$ SVM classifiers.

3 Database of Tire Images

Since the image datasets of tire tread patterns are hard to collect, the training of SVM classifiers becomes more difficult due to the deficiency of tire samples. To overcome such dilemmas, vertical and horizontal mirror images are created for each training sample. Thus, the total training samples are greatly increased to three times of original training set. The typical vertical and horizontal mirror images of the original tire image (see Fig. 7(a)) are shown in Fig. 7(b) and Fig. 7(c), respectively.

15 tire tread patterns with 7 different manufactures in our dataset, as shown in Fig. 8, are collected. For each tire pattern, the training samples are increased from 2 to 6 by creating vertical and horizontal mirror images. To evaluate the performance of the proposed tire matching system, 2 testing samples per class are used. Table 1 lists the tire tread patterns in our dataset and their corresponding class number in Fig. 8.

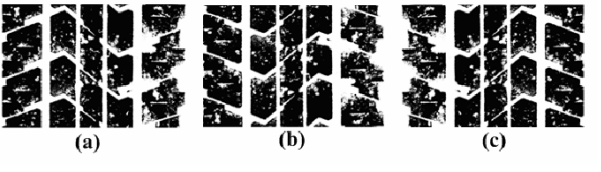


Fig. 7. (a) Original tire image, (b) Vertical mirror image, (c) Horizontal mirror image

No.	Tire pattern	Image	No.	Tire pattern	Image
1	BRIDGESTONE 175 70R13 82S		9	GOODYEAR 185 65R14 86H	
2	BRIDGESTONE 185 70R13 86H		10	GOODYEAR 185 R14C 102 100N	
3	BRIDGESTONE 185 R14C 8PR		11	MAXXIS 185 65R14 86H	
4	BRIDGESTONE 185 R14C 8PRN		12	MAXXIS 205 55R16 91V	
5	BRIDGESTONE 195 60R15 88V		13	SONAR 185 65R14 86H	
6	DUNLOP 185 65R14 86H		14	SONAR 195 60R14 86H	
7	FEDERAL 185 60R14 82H		15	STEEL RADEAL 175 R13C UE 168	
8	FEDERAL 185 65R14 86H				

Fig. 8. 15 different tire tread patterns with related manufactures in our dataset

Table 1. Tire patterns and their corresponding class number in training and testing sets

Tire tread patterns (15 total classes)	Training set (6 samples per class)	Testing set (2 samples per class)
wavy pattern	2 classes (No. 10 and 15)	
2-groove pattern	3 classes (No. 4, 7 and 8)	
3-groove pattern	4 classes (No. 1, 3, 6 and 9)	
≥ 4 grooves pattern	6 classes (No. 2, 5, 11, 12, 13, and 14)	
Total samples	15 classes × 6 sample/class = 90 samples	15 classes × 2 sample/class = 30 samples

4 Results and Discussion

The experimental results for 15 tire tread patterns with 30 total testing samples by the classification methods of Euclidean distance (ED), cosine distance (CD), and support vector machine (SVM) are listed in Table 2, where the data format xx(yy) means that the number of matching images is xx from the total testing samples yy. The total recognition rates of pattern vectors of tire images for ED, CD, and SVM are 40%, 56.7%, and 60%, respectively. Clearly, the recognition performance of SVM is higher than those of ED and CD methods.

As indicated in Table 2, the recognition rates of the class with ≥4 grooves for all classifiers are obviously lower than those of other classes. This result may arise from the closer texture patterns between the tire images in this class such that they are difficult to be recognized by all classifiers. Another possible reason may cause by the capturing quality of tire images by CCD sensors due to noise to blur or deteriorate acquired images.

The recognition rates for the classes of 2-groove and ≥4-groove tire patterns are 50% and 42%, respectively, for the SVM classifiers. This result may be improved by increasing the training samples of tire images. In contrast to those two classes, the recognition rates of SVM are 75% for wavy patterns, and 87.5% for 3-groove patterns. This result illustrates the superiority of support vector machine for recognizing tire tread patterns compared to the other two measures of Euclidean distance and cosine distance.

Table 2. Recognition performance of 15 tire patterns for the proposed tire matching system

Tire tread patterns	Euclidean distance (ED)	Cosine distance (CD)	Support vector Machine (SVM)
wavy pattern	2(4)	2(4)	3(4)
2-groove pattern	2(6)	3(6)	3(6)
3-groove pattern	6(8)	6(8)	7(8)
≥4 grooves pattern	2(12)	6(12)	5(12)
Correct matching number	12(30)	17(30)	18(30)
Total recognition rates	12/30*100%= 40%	17/30*100%= 56.7%	18/30*100%= 60%

5 Conclusion and Future Work

Three methods of ED, CD, and SVM are used to classify the pattern vectors of Gabor wavelets of tire images in our tire matching system. The recognition performance of SVM is higher than those of the other two methods, indicating the superiority of classification capability for SVM. Since the training samples of tire tread patterns are quite limited, the work of collecting more samples is still continuing. To improve the recognition performance, the grooves on tire surface can be further categorized into fine grooves and coarse grooves to increase the discriminative power of tire patterns. More methods for feature representation and extraction of tire pattern are continuously explored to improve the recognition capability of our tire matching system.

Acknowledgments. This research was fully supported by a grant from National Science Council, Taiwan, under contract NSC-97-2221-E-346-003-.

References

1. Wang, Y.W., Lin, C.N.: A Line-Based Skid Mark Segmentation System Using Image-Processing Methods. Transportation Research Part C 16, 390–409 (2008)
2. Thali, M.J., Braun, M., Brüschweiler, W., Dirnhofer, R.: Matching Tire Tracks on the Head Using Forensic Photogrammetry. Forensic Science International 113, 281–287 (2000)

3. Buck, U., Albertini, N., Naether, S., Thali, M.J.: 3D Documentation of Footwear Impressions and Tyre Tracks in Snow with High Resolution Optical Surface Scanning. Forensic Science International 171, 157–164 (2007)
4. Colbry, D., Cherba, D., Luchini, J.: Pattern Recognition for Classification and Matching of Car Tires. Tire Science and Technology 33, 2–17 (2005)
5. Moreno, P., Bernardino, A., Victor, J.S.: Gabor Parameter Selection for Local Feature Detection. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3522, pp. 11–19. Springer, Heidelberg (2005)
6. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
7. Jung, S.W., Bae, S.W., Park, G.T.: A Design Scheme for a Hierarchical Fuzzy Pattern Matching Classifier and Its Application to the Tire Tread. Fuzzy Sets and Systems 65, 311–322 (1994)
8. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, New York (1998)

Terminological and Assertional Queries in KQL Knowledge Access Language

Krzysztof Goczyła, Piotr Piotrowski, Aleksander Waloszek,
Wojciech Waloszek, and Teresa Zawadzka

Gdańsk University of Technology, Department of Software Engineering,
Narutowicza 11/12, 80-233 Gdańsk, Poland
{kris,piopio,alwal,wowal,tegra}@eti.pg.gda.pl

Abstract. One of the directions of development of information systems in recent years is the evolution of data-based systems into the knowledge-based systems. As a part of this process there is ongoing work on a whole range of languages for accessing knowledge bases. They can be used in a variety of applications, however their main drawback is the lack of clearly defined algebra representing a theoretical basis for them. For instance, such a basis is the relational algebra for SQL. The paper presents a new language for accessing knowledge bases that is built on a solid, theoretical foundation – the algebra of ontological modules.

Keywords: query language, knowledge base, ontology, modularization.

1 Introduction

The recent years in the information science brought more and more discussion about the knowledge and knowledge-based systems. Although there is no clear definition of what the knowledge is (in fact, there are many different definitions), we all agree that the knowledge, unlike the data, requires *understanding* and *reasoning*. The problem of building knowledge-based systems is not a new one, but with the growth of the idea of Semantic Web it gained a new meaning. There appeared a very real need for providing the "engineering tools" that would allow for expressing certain facts and querying systems for conclusions arising from these facts. In practice, this need translates into a number of initiatives aimed at developing a new language to access the knowledge base [1], [2], [3]. Unfortunately, those languages, though useful in various applications, do not have a solid mathematical theoretical basis. In our opinion the development of a universal language for accessing knowledge bases must be based on assumptions similar to those adopted and practically proven in the case of SQL – beside the others, theoretical mathematical basis, language closure and availability of commands for creating, manipulating and controlling the knowledge. In [4] there the algebra of ontological modules (in other words s-modules) was presented that constitutes a theoretical basis for the KQL language presented in this paper. Section 2 briefly presents the basis of algebra of ontological modules, Section 3 describes the formal

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 102–111, 2010.

model of s-modular knowledge base, in Section 4 we present assumptions and rules for creating terminological and assertional queries in KQL while in Section 5 we provide a set of examples presenting the usage of the language.

2 The Algebra of s-modules

Recent years have brought increased interest in modularization of knowledge bases. This stems primarily from the need to provide users with ability to create and process modular knowledge bases, but also ability to execute more advanced operations such as extracting a part of a module, connecting this fragment to a fragment of another module, etc. Another goal is using methods of automatic (and invisible to the user) decomposition of a knowledge base in order to increase performance of inference process. As a response for such requirements the s-modular knowledge bases have been proposed. The s-modular knowledge base is a set of ontological modules and a set of inter-modular Horn rules. An s-module is a structure that represents all models (in terms of first-order logic) of a fragment of knowledge, regardless of how the part is formulated (e.g. regardless of the description language). One of the most important, in the cognitive as well as practical point of view, assumptions that were adopted is the fact that the user is interested in conclusions derived from a given fragment of knowledge, and not in the form of sentences that were used to create the module. As a consequence of this assumption the term of an s-module has been defined strictly semantically focusing solely on its first-order logic interpretation.

Definition 1 (s-module). A s-module $M = (S,W)$ is a pair of a signature S (called a signature of the s-module) and a class W of interpretations. The two parts of M are denoted respectively as $S(M)$ and $W(M)$. Each interpretation from W is called a model of M. A signature S (the set of used names) $S = C \uplus R \uplus I$ consists of concept names (C), role names (R), and individual names (I) ($\uplus$ denotes disjoint union). ■

Due to the fact that with the notion of s-module defined in a such way we are losing the sentence structure (e.g. there is no typical division into a terminological part TBox and an assertional part ABox [5]), the formalism was called algebra of ontological modules [6] [7]. This, of course, does not mean that the sentences are not used to define s-modules. The definition of all basic s-modules (i.e. not created on the basis of other ones) clearly is derived from a defined set of statements. The statements are formulated using (possibly complex) concepts, roles and individuals with operators of a specific dialect of description logic $\mathcal{L}$ (using operators of $\mathcal{L}$ and names from the signature $\mathbf{S}$, denoted respectively by $\mathcal{L}_C(\mathbf{S})$, $\mathcal{L}_R(\mathbf{S})$, $\mathcal{L}_I(\mathbf{S})$). In addition, algebra of ontology modules introduces a set of algebraic operations (similar to those defined in the Codd algebra [8]) that allow for creation of new s-modules from existing ones (note that, however, these new s-modules may not be expressible as a set of statements).

3 Formal Model of an s-modular Knowledge Base

Below, there is presented a framework of a formal (expressed in a description logic) model of an s-modular knowledge base (further denoted as a metamodel). It should be noted that the present metamodel provides expressiveness of language for defining s-modules that corresponds to expressiveness of OWL DL version 1.0 [11] and expressiveness of language for defining rules that corresponds to expressiveness of SWRL [12] with a limited set of predicates. The metamodel does not define ability of building s-modules with the use of s-module algebra. It was assumed that s-modules are created as independent objects, and how they are created is not reflected in the model.

In the s-modular knowledge base there exist two types of objects: named (`NamedObject`) and unnamed. Each named object takes exactly one value for the `hasName` attribute (axioms 1 and 2 in Table 1 below). To the named objects belong: basic s-modules (`NamedConglomeration`), named concepts (`NamedConcept`), named roles (`NamedRoleDef`), individuals (`Individual`) and rules (`Rule`) (axioms 3, 4, 5, 6, 7). Moreover, concepts (`Concept`) and properties (`PropertyDef`) are defined within named s-module (`NamedConglomeration`) (axioms 8 and 9), while rules use (`usesTermsFrom`) terms defined within s-modules (axioms 10 and 11).

Concepts has been divided into abstract ones (`AbstractConcept`) and concrete ones (`ConcreteConcept`) (those related to concrete domains) (they are not included in the presented framework due to space limitations). Within abstract concepts there have been defined named concepts (`NamedConcept`) and complex concepts (`ComplexConcept`).

Next, the formal model defines properties, including (as an analogy to OWL) attributes (similarly to concrete concepts, they are not included in the presented framework due to space limitations) and roles (axioms 18 and 19.) The model defines functional roles (axiom 20), transitive roles (axiom 21) and complex roles (axiom 22).

Complex concepts and roles are modeled with the use of `operator` and `operand` roles, that define respectively the operators used to create the given concept or role and components that can be subject of operator's action (axiom 23).

Among the complex roles there are defined complex roles in which the order of operands is important (for instance, such a role is a chain of roles) (axiom 24). For such roles an additional operand denoted by the role `operand1` must be defined (axiom 25).

Moreover, the axioms 26 to 28 define the inheritance hierarchy of roles and concepts. This hierarchy is defined with the roles `subsumer` and `subsumee`.

The last element of the model are instances of properties and concepts. For this reason there have been defined three concepts that represent instances of concepts (`Individual`), instances of roles (`PairOfIndividuals`) and instances of attributes (`PairIndividualValue`) (axioms 29 and 30). The concept `PairOfIndividual` consists of two individuals, while the `PairIndvidualValue` concept consists of a single individual and a single value (axioms 31 – 33).

Table 1. Metamodel framework

NamedObject ⊑ ∃hasName.*string* (1)	TransitiveRole ⊑ RoleDef (21)
NamedObject ⊑ =1hasName.*string* (2)	ComplexRoleDef ⊑ RoleDef (22)
NamedConglomeration ⊑ NamedObject (3)	ComplexConcept ⊔ ComplexRoleDef ⊑ ∃operator.⊤ ⊓ ∃operand.⊤ (23)
NamedConcept ⊑ NamedObject (4)	ComplexRoleOrderDef ⊑ ComplexRoleDef (24)
NamedRoleDef ⊑ NamedObject (5)	ComplexRoleOrderDef ⊑ =1operand1.RoleDef ⊓ =1operand.RoleDef (25)
Individual ⊑ NamedObject (6)	∃subsumer.⊤ ⊑ AbstractConcept ⊔ Role (26)
Rule ⊑ NamedObject (7)	⊤ ⊑ ∀subsumer.AbstractConcept ⊔ Role (27)
∃isDefinedIn.⊤ ⊑ Concept ⊔ PropertyDef (8)	subsumee ≡ subsumer⁻ (28)
⊤ ⊑ ∀isDefinedIn.NamedConglomeration (9)	PairOfIndividual ⊑ =2consistsOf.Individual (29)
∃usesTermsFrom.⊤ ⊑ Rule (10)	PairIndividualValue ⊑ =1consistsOf.Individual ⊓ =1value (30)
∀usesTermsFrom.NamedConglomeration ⊑ ⊤ (11)	Concept ⊑ ∀hasInstance.Individual (31)
AbstractConcept ⊑ Concept (12)	RoleDef ⊑ ∀hasInstance.PairOfIndividuals (32)
ConcreteConcept ⊑ Concept (13)	AttributeDef ⊑ ∀hasInstance.PairOfIndividuals (33)
NamedConcept ⊑ AbstractConcept (14)	
ComplexConcept ⊑ AbstractConcept (15)	
RoleDef ⊑ PropertyDef (18)	
AttributeDef ⊑ PropertyDef (19)	
FunctionalRole ⊑ RoleDef (20)	

4 Basic Assumptions for KQL

Development of new knowledge representation and development of new kwnowledge access language were guided by the same objective: to create „engineering" tools that could be used for creation knowledge based systems. Therefore it has been proposed a set of requirements (characteristics) that should be meet by a uniform language for accessing an s-modular knowledge base. The essential characteristics are: (1) existence of statements for creating knowledge bases, manipulating them and controlling access to them, (2) closure (query statements should return the same structures they operate upon), (3) ability of query nesting, (4) uniformity of definition and modification language (results of queries can be used in queries operating on the knowledge base and creating knowledge base), (5) ability to pose terminological queries, (6) support for ontology modularization.

These characteristics have been proposed in an analogy to characteristics of SQL. KQL, in analogy to SQL, provides statements for creating knowledge bases, manipulating them and controlling access to them.

Statements for creating a knowledge base include creation of: knowledge bases (ontologies); s-modules as conceptual parts of a knowledge base; and rules and mappings to external data sources.

The main statement for manipulating an s-module is the SELECT statement of the following structure:

```
SELECT concept_list, role_list, attributes_list
  ADD axiom_list
  FROM conglomeration_expression
  WHERE concept_expression
  HAVING concept_expression
```

With respect to the s-module algebra, the SELECT clause corresponds to the projection—the choice of terms for a newly created s-module, the ADD clause corresponds to the selection—the contraction of the set of allowed interpretations, the WHERE and HAVING clauses correspond to the projection with respect to individual names. The difference between the WHERE and HAVING clauses lies in the fact that the WHERE clause selects those individuals that belong to a specified concept of the original s-module (as defined in the FROM clause), while the HAVING clause selects those individuals that belong to a specified concept of the target s-module. The query is conceptually executed in the following steps: (1) Determining the basic s-module on the basis of the FROM clause. (2) Reducing the individual names on the basis of the WHERE clause. (3) Extending the alphabet with new concepts / roles / attributes / individuals that occur in the statements contained in the ADD clause (selection). (4) "Adding" (i.e. extending the ontology) to the s-module the statements from the ADD clause. (5) Projection of the alphabet only to the concepts / roles / attributes contained in the SELECT clause. (6) Reducing the individuals names basing on the HAVING clause.

The least matured commands in the current version of the KQL are commands of knowledge access control. The language, however, is designed in a way allowing for its further development also in this, very important from a practical point of view, direction.

KQL is **a closed language**. Utilizing the s-module algebra as a theoretical basis for the language allowed for developing a language that operates upon and returns s-modules. Results of queries that operate on s-modules are also s-modules. A direct consequence of KQL closure is ability of query nesting. Every time an s-module is needed, one can use named s-modules defined in knowledge base or s-modules created on the fly by KQL expressions with the use of s-module algebra operators (such as INTERSECT, JOIN, UJOIN) or the SELECT clause. As a consequence of KQL closure and the ability of **query nesting,** the **uniformity of modification and definition language** has been achieved: one can use query nesting also in s-module creation statements (in KQL, an s-module is called “conglomeration”):

```
CREATE CONGLOMERATION conglomeration_name
…
FROM conglomeration_expression
```

In KQL it is assumed that **terminological queries** are issued against a metaontology or a meta s-module. A metaontology is nothing but a single s-modular knowledge base that stores information about s-modules and rules. A meta s-module is also a single s-modular knowledge base that stores information about exactly one s-module. A metaontology and meta s-modules are built in accordance to the following assumptions:

- Any s-module, rule, concept, role, attribute, and individual of an s-modular knowledge base is reified to individuals that are instances of relevant concepts of the model;
- Description of world is infinite, because every concept and role are generating an infinite number of individuals: each concept and complex rule corresponds to a single individual, and every concept and rule may be defined in countless ways (e.g. A, A ⊔ A, A ⊔ A ⊔ ... ⊔ A). All named terms defined in a domain s-modular knowledge base are reified to individuals of names of terms of the domain knowledge base.

As a consequence of using the metaontology it is possible to issue terminological queries in a similar way how the assertional queries are issued; the only difference is that a query is directed to the metaontology rather than to the knowledge base itself. This follows from the fact that s-modules, concepts, roles, attributes, and individuals from a knowledge base are reified to individuals in the metaontology and the relationships between them are reified to the corresponding binary relationships between individuals.

4 Examples of KQL Usage

The examples below (1-3) have already been defined in terms of the s-module algebra in [7]. Here the same examples have been reformulated in KQL. Example 4 shows sample terminological queries.

Example 1 (simple import). We consider two s-modules: M_1 describes human resources, and M_2 describes a structure of a hospital.

```
CREATE CONGLOMERATION M1
   ADD CONCEPT HTBusinessUnit
   ADD CONCEPT Expert
   ADD CONCEPT Employee
   ADD ROLE isManagerIn

CONSTRAIN M1
   EXIST isManagerIn
     HTBusinessUnit ISSUB Expert
   Expert ISSUB Employee
```

```
CREATE CONGLOMERATION M2
  ADD CONCEPT Department
  ADD ROLE leadsDepartment
  ADD INDIVIDUAL johnSmith
  ADD INDIVIDUAL neurosurgery

CONSTRAIN M2
  (johnSmith, neurosurgery) IS
     leadsDepartment
  neurosurgery IS Department
```

In KQL each s-module is created in two steps. In the first step the s-module signature is created (`CREATE CONGLOMERATION` statement). In the second step the constraints (sets of statements that must be satisfied) are added (`CONSTRAIN` statement). In the example above, the first s-module defines an employee and an expert. We assume that each expert is an employee, and anyone who manages at least one business unit is an expert. The second s-module describes John Smith who leads the department of neurosurgery. We want to ask whether `johnSmith` is the expert.

```
SELECT Expert
  ADD leadsDepartment ISSUB isManagerIn
  ADD Department ISSUB HTBusisnessUnit
  FROM M1 INTERSECT M2
  WHERE {johnSmith}
```

To merge the information from the two s-modules in order to check whether `johnSmith` is an expert we first create an intersection of the s-modules (FROM statement), and then restrict the set of model by introducing additional "bridge" axioms (ADD statement). The result of the query is a new s-module whose signature consists of two terms: `johnSmith` and `Expert`. In our example `johnSmith` is an instance of `Expert` concept.

Example 2 (basic ontology alignment)
In the above example we did not encountered any name confilicts. In general, such a conflict may easily occur. In this example we show how to align two s-modules in which the same set of terms has been used to express different mappings. Let us assume that there exists two s-modules that define three concepts: `HSRoom`, `ASRoom` i `LSRoom` that denote rooms of high, average and low standard, respectively. The first s-module M_o defines the standard of rooms rented for a single night, while the s-module M_l defines those rented for more than one night. Our objective is to create a new s-module *M* that would define a different room standard in accordance to the rent time. Moreover, while creating the new s-module it has been assumed that rooms with bathrooms are of high standard in case of renting them for a single night, and rooms with no bathroom are of low standard if rented for more than one night.

```
CREATE CONGLOMERATION Mo              CREATE CONGLOMERATION Ml
       ADD CONCEPT HSRoom                    ADD CONCEPT HSRoom
       ADD CONCEPT ASRoom                    ADD CONCEPT ASRoom
       ADD CONCEPT LSRoom                    ADD CONCEPT LSRoom

            CREATE CONGLOMERATION M
            FROM(
               SELECT *
                 ADD RoomsWithBathroom ISSUB HSRoomON
                 ADD (NOT RoomsWithBathroom) ISSUB LSRoom
                 FROM(
                   SELECT HSRoomON, ASRoomON, LSRoomON
                     ADD HSRoomON EQ HSRoom
                     ADD ASRoomON EQ ASRoom
                     ADD LSRoomON EQ LSRoom
                     FROM (Mo INTERSECT Ml)
                   )
            )
```

To create *M* module firstly we gather information from both modules (intersection of M_1 module and M_o module with renamed concepts). In that way we create a temporary, unnamed module that defines six different classes of rooms. Secondly we have to establish a translation between “one night” and “longer” concepts. We extend the temporary module by the criteria which were used for the assessment in both cases (the two sentences about rooms with bathrooms).

Example 3 (different versions and what-if problems)
This example illustrates use of union and negation in KQL. Let us consider an s-module *M*:

```
CREATE CONGLOMERATION M          CONSTRAIN CONGLOMERATION M
   ADD CONCEPT TrustedWitness       Top ISSUB <= 1 INV (murdered)
   ADD CONCEPT CrimeScene               {victim}
   ADD ROLE murdered                EXIST INV(accuses) TrustedWitness
   ADD ROLE accuses                 ISSUB EXIST INV(murdered) {victim}
   ADD ROLE presentAt               TrustedWitness ISSUB EXIST
   ADD INDIVIDUAL victim                          presentAt CrimeScene
```

The *M* s-module describes a world that assumes the only one murderer who is accused by a trusted witness (who has been present at the crime scene). We consider two (mutually exclusive) versions of facts (e.g. collected by two investigating agents:

John Shady and Henry Brilliant). To achieve that we define two new s-modules (one for each agent).

```
CREATE CONGLOMERATION M1
FROM(
   SELECT *
   ADD johnShady IS
     TrustedWitness
   ADD (johnShady, tedInnocent)
     IS accuses
   FROM M
)
```

```
CREATE CONGLOMERATION M2
FROM(
   SELECT *
   ADD henryBrillant IS TrustedWitness
   ADD (henryBrillant, markGuilty) IS
        accuses
   FROM M
)
```

Having such defined s-modules M_1 i M_2 we are able to analyze different scenarios. To perform such analyses we create a new s-module $M_0 = M_1$ UNION M_2.

1. We may assume that `henryBrillant` is not a trusted witness:

```
SELECT Murderer
   ADD NOT TrustedWitness(henryBrilliant)
   ADD Murderer EQ EXIST murdered {victim}
   FROM M1 UNION M2
```

We ask for a murderer. Therefore we define `Murderer` as a person who `murdered` a `victim`. The resulting s-module will return exactly one individual (`tedInnocent`) as an instance of `Murderer` concept.

2. We may assume that `markGuilty` is a murderer

```
SELECT TrustedWitness, NotTrustedWitness
   ADD (markGuilty, victim) IS murdered
   ADD NotTrustedWitness EQ NOT TrustedWitness
   FROM M1 UNION M2
```

In this case we can conclude that `henryBrillant` is a trusted witness (but this does not mean that `johnShady` is not a trusted witness). The s-module created within the query defines `NotTrustedWitness` concept (the complement of `TrustedWitness` concept defined in s-modules M_1 and M_2). The s-module which is the result for this query will return a single instance of `TrustedWitness` concept and will return no instance of `NotTrustedWitnesss` concept.

3. We may assume that `johnShady` was not present at the crime scene

```
SELECT TrustedWitness, Murderer
   ADD johnShady IS EXIST presentAt CrimeScene
   ADD Murderer EQ EXIST murdered {victim}
   FROM M1 UNION M2
```

In this case we can conclude that `johnShady`$\psi$is not a trusted witness and `markGuilty`$\psi$is the murderer. To do this we define `Murderer` concept similarly to the second example.

Example 4 (terminological queries)

Let us define the following s-module in which we have defined `Students`, `Teachers` and `WorkingPeople`.

```
CREATE CONGLOMERATION M
   ADD CONCEPT Students
   ADD CONCEPT Teachers
   ADD CONCEPT WorkingPeople
```

```
CONSTRAIN CONGLOMERATION M
   ADD Teachers ISSUB WorkingPeople
   ADD Students ISSUB WorkingPeople
```

We ask whether the concept `WorkingPeople` subsumes the concept `Teachers`:

```
SELECT CONCEPTS NamedConcept, ROLES subsumer
FROM META (CONGLOMERATION M)
WHERE { Teachers, WorkingPeople }
```

In KQL the terminological query is, as stated above, issued against a meta s-module that is created with the META tag. As a result we receive an s-module with a single concept defined named `NamedConcept` and a single role `subsumer`. The individuals of names `Teachers` and `WorkingPeople` are instances of `NamedConcept` concept and are in subsumption relation: the pair (`Teachers`, `WorkingPeople`) is instance of `subsumer` role. Otherwise (e.g. if we would ask for concepts `Students` and `Teachers`), the resulting s-module will not contain any instance of `subsumer` role. In this example the concept `Students` and the concept `Teachers` are named concepts. In the case of unnamed concepts a new s-module should be created in which names are assigned to the complex concepts. For instance, let us ask whether the complex concept being the intersection of `Students` and `Teachers` is subsumed by the concept `WorkingPeople`.

```
SELECT CONCEPTS NamedConcept, ROLES subsumer
  FROM META (
      SELECT CONCEPTS WorkingPeople, StudentsAndTeachers
      ADD (StudentsAndTeachers EQUALS Students AND Teachers)
      FROM CONGLOMERATION M
    )
  WHERE { Students, Teachers }
```

In the query above there is created a new conglomerate where the complex concept being an intersection of concepts `Students` and `Teachers` was named `StudentsAndTeachers`.

6 Summary

The paper presents the concept of a s-modular knowledge base along with a language for accessing the knowledge stored in such a base. Existing languages for accessing the knowledge bases focus either on defining knowledge bases, or on the querying a knowledge base [2], [4], [5]. On the contrary, the KQL language does cover all aspects of accessing a knowledge base, i.e. it offers both ability of creating and populating knowledge bases, as well as ability of querying them. The same language constructs are used in commands of writing to and reading from a knowledge base. In addition, due to the incorporation of s-module algebra which is the theoretical basis for KQL, ability of modularizing knowledge bases is smoothly integrated into the language. Different languages provide different ways for handling terminological queries. For KQL it has been chosen a very flexible and conceptually coherent method of asking terminological queries through a metaontology. Detailed comparison of KQL language with other languages can be found in [13].

A formal model for an s-modular knowledge base from the very beginning has been designed with a special attention for its use to handle terminological queries. This distinguishes it from other metamodels, such as Ontology Definition Metamodel published by OMG [14], created for the sake of generality and widest range of applications. KQL is also being actively and continuously developed, primarily for

various practical extensions, such as operations on concrete domains. The language authors' goal is to provide knowledge engineers with a comfortable, but simultaneously based on solid theoretical foundations, set of tools to create and manipulate knowledge bases.

References

1. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member, Submission May 21 (2004), http://www.w3.org/Submission/SWRL/
2. Prud'hommeaux, E.: SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparqlquery/
3. Fikes, R., Hayes, P., Horrocks, I.: OWL-QL - A Language for Deductive Query Answering on the Semantic Web. Knowledge Systems Laboratory. Stanford University, Stanford (2003), http://ksl.stanford.edu/KSL_Abstracts/KSL-03-14.html
4. Kubias, A., Schenk, S., Staab, S.: (Demonstration at ESWC), SAIQL Query Engine - Querying OWL Theories for Ontology Extraction, http://www.eswc2007.org/pdf/demopdf/SaiqlPosterProposal.pdf
5. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: Description Logic Handbook. Cambridge University Press, Cambridge (2002)
6. Goczyła, K., Waloszek, A., Waloszek, W.: S-Modules – An Approach to Capture Semantics for Modularized DL Knowledge Bases. In: KEOD 2009, Funchal-Madeira, Portugalia, pp. 117–122 (2009)
7. Goczyła, K., Waloszek, A., Waloszek, W.: Algebra of ontology modules for semantic agents. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 492–503. Springer, Heidelberg (2009)
8. Codd, E.F.: Relational Completeness of Data Base Sublanguages. Database Systems 6, 65–98 (1979)
9. Bouquet, P., Ghidini, C., Giunchiglia, F., Blanzieri, E.: Theories and uses of context in knowledge representation and reasoning. Journal of Pragmatics 35(3), 455–484 (2003); Intelligence, Third International Atlantic Web Intelligence Conference, pp. 163–169. Springer, Heidelberg (2005)
10. Ghidini, C., Giunchiglia, F.: Local Model Semantics, or Contextual Reasoning = Locality + Compatibility. Artificial Intelligence 127(2), 221–259 (2001)
11. OWL Web Ontology Language Guide, W3C Recommendation February 10 (2004), http://www.w3.org/TR/owl-guide/
12. Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission May 21 (2004), http://www.w3.org/Submission/SWRL/
13. Goczyła, K., Piotrowski, P., Waloszek, A., Waloszek, W., Zawadzka, T.: Język KQL jako realizacja idei języka SQL dla bazy wiedzy. Studia Informatica 31(2A(89)), 47–61 (2010)
14. Ontology Definition Metamodel (ODM) Version 1.0, Release date (May 2009), http://www.omg.org/spec/ODM/1.0/PDF

Conditional Statements Grounded in Past, Present and Future

Grzegorz Skorupa and Radosław Katarzyniak

Wroclaw University of Technology, Wybrzeże Wyspiańskiego 27
{grzegorz.skorupa,radoslaw.katarzyniak}@pwr.wroc.pl

Abstract. In previous works we defined a model for grounding conditional statements extended with modal operators of possibility and belief. It was assumed that statements were told by an agent to describe objects at current moment. Within this work we extend the model by adding time variable. This extension allows agent to describe future and past events at known time moments.

1 Introduction

One of the important aspects of agent based systems is communication. Agents usually are supposed to communicate between themselves, but it is also often required that they have abilities to effectively communicate with humans. One of the most popular trends in agent communication is tightly connected with speech act theory [6]. It specifies how agents can ask questions, make commitments, negotiate and i.e. inform each other. Speech act theory mainly specifies protocols and types of communicates but pays little attention to internal message meaning. Within this work we concentrate on the usage of conditional statements. Such statements can be embodied within inform predicate of speech act theory.

We assume agent lives in an environment, where she observes objects and their properties. Objects can change properties within time. Agent's task is to describe the viewed environment. Description is constructed from a class of natural language statements. Those statements describe observed objects and cover: simple statements, complex statements having conjunctions like 'and' and 'or' and in particular conditional statements of the form 'If ..., then ...'.

The agent has only partial knowledge about the environment and is unable to view or predict all its properties. Based on her knowledge agent can evaluate how likely different properties are to occur. To effectively communicate its predictions and evaluations agent can enrich its utterance with modal operators of belief and possibility. If agent is not sure of something she can say that she believes in it or finds it possible.

Our primary aim is to imitate human speech abilities when choosing proper statement to describe a situation. It is of high importance that agent chooses sentences that are not only logically true but also consistent with its mental state and knowledge. Agent should pick the statements in a similar way human does. None of the statements should mislead a human listener. For example if

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 112–121, 2010.

one says "If there are clouds, it will rain" he may suggest to the listener that: he is not sure whether there are clouds and he is not sure if it will rain.

This work extends previous works on grounding modal communicates [13] and in particular previous articles on grounding modal conditional statements to describe current situation [18,19]. Within the scope of previous works we have defined conditions on choosing a simple conditional statements to describe currently observed moment. Within this work we redefine these conditions to cover conditional statements referring to fixed moments in future and past. We also add time to agent's model, so that agent has a memory of previous events and is able to imagine future events. This model is required to pick the right sentence. In paragraph 2 we specify what we mean by human behaviour imitation and which of the aspects we consider. Paragraph 3 defines what types of conditional statements are considered. In paragraph 4 we define criteria for telling conditional statements. A lot of attention is paid to modal operators and time. In paragraphs 5 and 6 a formal language is defined. Paragraphs 7 and 8 outline a method of implementing speech abilities in an agent.

2 Properly Using a Statement

When telling a statement, being simply logically consistent, is not enough because a statement may be completely useless to the recipient or even mislead him. In order to be informative and precise when telling a statement agent has to consider many criteria. She has to evaluate if a hearer needs the information, if she wants to pass this information or what she herself needs to know.

The difference between a true statement and a properly used statement has been a debate for a long time. Grice [8] claims that what is said is simply not the same as what is meant. Ajdukiewicz [3] differentiated statements classical logical truth from a language custom. These are only two examples from many that can be found in literature. The general conclusion is that a statement, when spoken, means more than its pure logical meaning in the terms of classical logic. This in turn implies that some statements can't be used in some of the situations, because they may mislead the listener. Let us take an statement from [8]: "Some athletes smoke." that suggests to a listener that not all athletes smoke. If it wasn't so, a speaker would say "All athletes smoke". On the other hand, logically the statement only says, that there are athletes that smoke.

We are focusing on conditional statements, where the distinction mentioned above is very clear. This has resulted in many works showing out this problem from different points of view [3,17,5,4,16]. Many alternative theories and logics of conditionals have emerged (see [2,14,20] for examples). Trivially speaking, the problem comes from the fact that we tend to use conditionals in many situations with different meanings. So far little attention has been paid on choosing the right conditions defining when a conditional statement can be said and embedding them within an artificial entity such as agent.

Our aim is to formulate conditions that rule out improper usages of most typical and simple conditional statements. We also show that these conditions can

be successfully implemented within an agent. We do not consider all criteria of choosing the right and the final statement. Our aim is to filter out all statements that would not be used by a human, if he knew what agent observes and knows.. We do not consider complicated dialogue context. In our opinion resulting filtered set of 'approved' statements can be later used to choose the final agents utterance based on her needs, knowledge of the listener and his needs.

3 Considered Conditional Statements

Conditional statements can be divided into categories in many ways. The most general categorisation is into 'indicative conditionals' and 'subjunctive' or 'counterfactual conditionals'. See [1] for more detailed description. First group refers to conditional statements describing situations that are possible to happen or one is able to verify them. The second group includes statements that are impossible or improbable to happen. Usually it contains 'would' word in the consequent. Example[1]:

- "If it rains, the match will be cancelled." - the speaker thinks that the rain is possible - an indicative conditional.
- "If it were to rain, the match would be cancelled." - the speaker thinks it will not rain - a subjunctive conditional.

We shall be dealing only with typical indicative conditionals. We also rule out so called 'Thomason Conditionals' like "If Sally is deceiving me, I do not believe it." [21].

We assume agent uses an indicative conditional statement to describe objects properties in fixed time moments in past, present or future. The properties are crisp (object can have or lack them).

Additionally agent's utterance may be enriched by an operator of belief or possibility. We assume the operator is always connected to the consequent, not the antecedent. Exemplary statements we are considering are:

- If the apple red then I believe it is ripe.
- If you don't study today, you will fail the exam next week.
- If she took that flight yesterday, she is now somewhere in town.
- If she took that flight yesterday, I will meet her tomorrow.

4 Using the Indicative Conditional

We are considering conditional statements used to describe possible situations within fixed time points in past, present or future. Suppose a statement of the form "If ϕ, then ψ" is told.

In [18] and [19] we have pointed out that such conditional statements, when used to present, can be used only when a speaker does not know whether ϕ holds.

[1] Example comes from [5].

The speaker can't also know that ψ holds. Such point of view is consistent with most, if not all, of works that can be found in the literature [3,7,17]. If the speaker knew any of them there would be no point to use an indicative conditional.

The another quite common remark is that there must be some dependency between ϕ and ψ. There are many questions on how this dependency should be modelled and so far surprisingly little has been done in this subject. We claim a speaker informs, that he has reasoned about both situations (where ϕ holds or not) and found out that ψ is guaranteed to hold only when ϕ holds. In fact the speaker has reasoned about four possible situations, and found out that situation where ϕ holds and ψ does not hold is impossible. Hence the speaker is ready to infer ψ, if he finds out that ϕ holds. But as long as he does not know ϕ he is unable to tell much more about ψ. This assumption is important, because it rules out meaningless statements like: "If 2+2=4, then a square is a rectangle." or "If moon is a piece of cheese, then I will die on a day with an even date"[2].

4.1 Belief and Possibility Operators

Within the scope of this work we analyse operators of belief and possibility which should be understood according to Hintikka's work [10]. We analyse usage of these operators within the consequent of indicative conditional statement.

In previous work [19] we pointed out that modal operator applied to consequent changes the understanding of relation between the antecedent and the consequent. In a conditional statement without modal operator, the consequent is required in case antecedent holds. When a modal operator of possibility or belief is used the consequent should be only more probable in case antecedent holds. We claim that in the considered use of statements the chance of consequent must rise in case antecedent holds. Suppose a statement "If the apple is green, then I believe it is not ripe" is said. One can conclude from such a statement that green apples are usually not ripe. On the other hand one can conclude that red (not green) apples are usually ripe. One wouldn't say "If the apple is green, then it is possible for it to be ripe."[3] without a specific conversational context such as a question "Can green apples be ripe?". Such conversational context is not considered here.

4.2 Time Reference

Previous conclusions presented in this paragraph were applicable to conditional statements about currently viewed time moment. The question is whether these conditions hold in case of indicative conditionals referring to the past and future.

We claim all of these remarks stay valid regardless of time a conditional considers. Suppose somebody says: "If Oswald didn't kill Kennedy, somebody else did"[4]. It means, that (1) he has doubts about Oswald being the killer, (2) he

[2] Second example is from [3].
[3] There are kinds of apples that can be green and ripe.
[4] Example from [1].

does not know who killed the president and (3) it is sure somebody else did it, if it wasn't Oswald. (1) means that we do not know if the antecedent is true, (2) that we do not know if the consequent is true. (3) means that if we new it wasn't Oswald, we would be sure it was somebody else. Similar reasoning can be made in case of statements referring to the future such as "If you don't study, you will fail the exam.".

What poses problems in deciding if conditional statement can be said, is the unclear time reference. When one says "If there is ϕ, then there will be ψ." does it mean that ϕ is now or will be in nearest future? Similarly will ψ be in the nearest future, some time later, how long will it hold? The interpretation varies according to the meaning of ϕ and ψ. In a statement "If you push that button, the light will lit." usually means that you can push the button, and light will lit immediately after you push the button and stay lit for some time. On the contrary a statement "If it rains, there will be mushrooms to pick" may say that it rains now, will rain for some time and some time after there will be mushrooms. Although often omitted, time reference always plays a crucial role in the meaning of a conditional statement. Usually this reference is added back by a listener when he interprets the statement. This problem has been analysed for example by [15]. The situation is easier when a speaker does not strip the exact time from the statement. In such case there is no need to reproduce the time span of a statement validity. We assume agent tells what time she refers to. The considered statements must contain time points of the events such as: yesterday, tomorrow, next minute.

5 Formal Language

We define only modal conditional statements with time reference, because these will be analysed further in this paper.

The alphabet of the language L consist of the following classes of symbols:

- $O = \{o_1, o_2, \ldots, o_M\}$ to represent atomic individuals (objects),
- $Q = \{q_1, q_2, \ldots, q_K\}$ for predicates (objects properties),
- Natural numbers $\mathbb{N}$ to represent consecutive time moments,
- symbol '$\neg$' for negation, symbol '$\rightarrow$' for conditional statements, additional bracket symbols '(' and ')',
- symbols Pos, Bel for modal operators of possibility, belief and knowledge.

The formulas are divided into two classes: atomic formulas (L_A), and complex formulas (L_C).

Atomic formulas L_A:

Let: $k \in \{1, 2, \ldots, K\}$, $m \in \{1, 2, \ldots, M\}$ and $t \in \mathbb{N}$. Any statement of the form $q_k(o_m, t)$ or $\neg q_k(o_m, t)$ is a proper statement of the language L.

Complex formulas L_C:

Let $\phi, \psi \in L_A$. Any statement of one of the following forms: $\phi \rightarrow \psi$, $\phi \rightarrow Bel(\psi)$, $\phi \rightarrow Pos(\psi)$ is a proper statement of the language L.

6 Intuitive Language Semantics

Below we specify intuitive semantics for language L. The semantics involve different grammatical times. The relation between current time and times of described features determines the usage of a particular grammatical time. We denote the actual time moment as t_{now}, where $now \in \mathbb{N}$.

Formula $[\neg]q_k(o_m, t)$ meaning:

- "o_m is [not] q_k now." when $t = t_{now}$.
- "o_m will [not] be q_k at moment t." when $t > t_{now}$.
- "o_m was [not] q_k at moment t." when $t < t_{now}$.

[¬] and [not] means respective negations are optional.

Formula $[\neg]q_k(o_m, t_1) \rightarrow [\neg]q_l(o_n, t_2)$ meaning:

- "If o_m is [not] q_k now, then o_n is [not] q_l now."
  when $t_1 = t_2 = t_{now}$
- "If o_m is [not] q_k now, then o_n will [not] be q_l at moment t_2."
  when $t_1 = t_{now} < t_2$
- "If o_m is [not] q_k at moment t_1, then o_n will [not] be q_l at moment t_2."
  when $t_{now} < t_1 \leqslant t_2$
- "If o_m is [not] q_k at moment t_1, then o_n was [not] q_l at moment t_2."
  when $t_{now} \leqslant t_1 \wedge t_2 < t_{now}$
- "If o_m was [not] q_k at moment t_1, then o_n is [not] q_l now."
  when $t_1 < t_{now} = t_2$
- "If o_m was [not] q_k at moment t_1, then o_n was [not] q_l at moment t_2."
  when $t_1 < t_{now} \wedge t_2 < t_{now}$

Formula $[\neg]q_k(o_m, t_1) \rightarrow Bel([\neg]q_l(o_n, t_2))$ means the same as formula $[\neg]q_k(o_m, t_1) \rightarrow [\neg]q_l(o_n, t_2)$ except that agent believes the consequent. Example:

- "If o_m is [not] q_k now, then I believe o_n is [not] q_l now."
  when $t_1 = t_2 = t_{now}$

Formula $[\neg]q_k(o_m, t_1) \rightarrow Pos([\neg]q_l(o_n, t_2))$ means the same as formula $[\neg]q_k(o_m, t_1) \rightarrow [\neg]q_l(o_n, t_2)$ except that agent finds the consequent possible. Example:

- "If o_m is [not] q_k now, then it is possible o_n will [not] be q_l at moment t_2."
  when $t_1 = t_{now} < t_2$

Examples:

- "The light will be lit at 5 pm". Can be formally written as $q_1(o_1, t_1)$ where q_1 is the property of being 'lit', o_1 refers to the described light and t_1 means '5 pm'.
- "If the apple is red now, then it is ripe". Formally: $q_1(o_1, t_{now}) \rightarrow q_2(o_1, t_{now})$ where q_1 means 'red', q_2 'ripe' and o_1 is the described apple.
- "If the apple is green now, then I believe it is not ripe". Formally written as $q_1(o_1, t_{now}) \rightarrow Bel(\neg q_2(o_1, t_{now}))$.

7 Mental State Model

In order to decide when agent can tell a statement she has to posses internal representation of its knowledge about the world. We assume such representation consist of a set of possible flows of events within time. Each flow of events is called possible world and represents one possibility on how the world might look like according to an agent. This representation has to be generated autonomously by an agent based on her environment observations and knowledge.

$$W = \{(w^{(1)}, p^{(1)}), (w^{(2)}, p^{(2)}), ..., (w^{(S)}, p^{(S)})\} \tag{1}$$

where $w^{(s)}$, $s = 1, 2, ..., S$ is a possible world and $p^{(s)}$ is a chance of this world being an actual, unknown to the agent, world. Mental model represents agent's knowledge on how the world may be at a given time moment. Every possible world represents some possible flow of events in the nearest past and future. One of possible worlds should the actual, real world flow. Agent is not omnipotent and does not know which of the worlds is the real one. She can only evaluate how probable the world is to be the real one.

It is assumed that $\sum_{s=1}^{S} p^{(s)} = 1$ and $p_s = P(w^{(s)})$ $(s = 1, 2, ..., S)$ defines a probability distribution over W. We assume that every world has positive probability ($\forall_{s \in \{1,2,...,S\}}\ p_s > 0$). It is not necessary for $p^{(s)}$ to be probability in its strict definition. It should be some estimation of probability created by the agent. The greater p_s, the more probable w_s is.

Each possible world is a function defining objects properties at past, current and future time moments.

$$w^{(s)} : Q \times \mathbb{N} \rightarrow 2^{O}, \quad s = 1, 2, ..., S \tag{2}$$

Function $w^{(s)}(q_k, t)$ returns a set of objects that contain given property $q_k \in Q$ at time moment $t \in \mathbb{N}$. Function describes one possible world predicted by an agent based on her knowledge. Assume that t_{now} is current time moment. All values of function $w^{(s)}$ for $t > t_{now}$ are a result of agent's prediction. If agent has observed the object's feature the value of function $w^{(s)}$ for $t \leqslant t_{now}$ is agent past (or current) observation. If agent has not seen the feature the value is an evaluation on how the world could have been. The observations gathered by agent so far stay constant among all possible worlds, while evaluations can change between worlds. In working implementation the $w^{(s)}$ domain will be bounded by time because it is impossible to predict everything in an infinite future.

Let $o \in O$ be an object and $t \in \mathbb{N}$ be a time moment. We say that:
If $o \in w^{(s)}(q_k, t)$, then object o is assumed to exhibit property q_k in world s at time moment t.
If $o \notin w^{(s)}(q_k, t)$, then object o is assumed to not exhibit property q_k in world s at time moment t.

For example, if $o \in w^{(s)}(q_k, t)$ for all possible worlds, agent predicts object o will surely have property q_k at moment t. If there are some worlds where $o \in w^{(s)}(q_k, t)$ and some where $o \notin w^{(s)}(q_k, t)$ then agent is not sure whether object will have property q_k at moment t. Let us now assume that, in all worlds,

such that $o \in w^{(s)}(q_k, t)$, it is also true that $o \in w^{(s)}(q_l, t+1)$. In such case agent knows, that object having property q_k at moment t, will have property q_l at next moment $t+1$.

8 Grounding Conditional Statements

In the following subsections we define an epistemic satisfaction relation $\models^E$ that tells when a statement can be said. When this relation holds, we say, the statement is properly grounded in agent's mental state. In other words we say the statement is properly connected to the world on the basis of agent's internal representation of that world [9,13]. The definitions in following subsections have been remodelled from [18] to cover time variable.

Let us assume that: $\phi = q_k(o_m, t_1)$ is an arbitrary chosen property $q_k \in Q$ for object $o_m \in O$ at moment t_1 and $\psi = q_l(o_n, t_2)$ is an arbitrary chosen property $q_l \in Q$ for object $o_n \in O$ at moment t_2. We shall analyse three basic types of sentences: $\phi \rightarrow \psi$, $\phi \rightarrow Bel(\psi)$ and $\phi \rightarrow Pos(\psi)$,

In order to decide if given conditional statement can be told agent has to compare its meaning with her mental model of the environment. There are three crucial sets that are required to check the conditional statement:

$$G^+(\phi) = \{s : o_m \in w^{(s)}(q_k, t_1)\} \tag{3}$$

$$G^-(\phi) = \{s : o_m \notin w^{(s)}(q_k, t_1)\} \tag{4}$$

$$G^+(\psi) = \{s : o_n \in w^{(s)}(q_l, t_2)\} \tag{5}$$

The first set in equation 3 contains all indices of worlds where antecedent holds and the second in equation 4 contains indices of worlds where antecedent does not hold. Third set in equation 5 holds all worlds where consequent holds.

All the probabilities of sets like $P(G^+(\phi))$ and conditional probabilities like $P(G^+(\psi) \,|\, G^+(\phi))$ can be easily calculated from known probability distribution $p_s = P(w^{(s)})$ $(s = 1, 2, ..., S)$.

8.1 Grounding Statements without Modal Operator

Definition 1. *Epistemic relation* $\models^E \phi \rightarrow \psi$ *holds iff all following conditions are met:*

a. $\underline{\alpha} < P(G^+(\phi)) < \overline{\alpha}$

b. $P(G^+(\psi) \,|\, G^+(\phi)) = 1$

c. $P(G^+(\psi) \,|\, G^+(\phi)) > \beta\, P(G^+(\psi) \,|\, G^-(\phi))$

where $0 < \underline{\alpha} < \overline{\alpha} < 1$, $\beta > 1$ *are fixed parameters.*

Condition *a* limits conditional statements to indicative conditionals. It means that agent does not know whether antecedent holds or not.

Condition *b* tells that consequent must hold in case where antecedent holds.

Condition *c* ensures that consequent must be more probable when antecedent holds, compared to a situation where antecedent does not hold. The greater the β parameter the more ϕ has to influence the chance of ψ. For example, when $\beta = 2$, ψ has to happen at least twice as often in case of ϕ.

8.2 Grounding Statements with Belief Operator

Definition 2. *Epistemic relation* $\models^E \phi \rightarrow Bel(\psi)$ *holds iff all following conditions are met:*

a. $\underline{\alpha} < P(G^+(\phi)) < \overline{\alpha}$

b. $\underline{\alpha}_{Bel} \leqslant P(G^+(\psi) \,|\, G^+(\phi)) < \overline{\alpha}_{Bel}$

c. $P(G^+(\psi)|G^+(\phi)) > \beta\, P(G^+(\psi) \,|\, G^-(\phi))$

where $0 < \underline{\alpha} < \overline{\alpha} < 1$, $0 < \underline{\alpha}_{Bel} < \overline{\alpha}_{Bel} < 1$, $\beta > 1$ *are fixed parameters.*

Conditions *a* and *c* are the same as in case of $\phi \rightarrow \psi$ statement.

Condition *b* requires that ψ must be greatly probable in case of ϕ. Parameter $\overline{\alpha}_{Bel}$ should be close to 1 and $\underline{\alpha}$ should be quite high to ensure the correct understanding of the belief operator. The higher the $\underline{\alpha}$, the less willing agent is to say that she believes in something.

8.3 Grounding Statements with Possibility Operator

Definition 3. *Epistemic relation* $\models^E \phi \rightarrow Pos(\psi)$ *holds iff all following conditions are met:*

a. $\underline{\alpha} < P(G^+(\phi)) < \overline{\alpha}$

b. $\underline{\alpha}_{Pos} \leqslant P(G^+(\psi) \,|\, G^+(\phi)) < \overline{\alpha}_{Pos}$

c. $P(G^+(\psi)|G^+(\phi)) > \beta\, P(G^+(\psi) \,|\, G^-(\phi))$

where $0 < \underline{\alpha} < \overline{\alpha} < 1$, $0 < \underline{\alpha}_{Pos} < \overline{\alpha}_{Pos} < 1$, $\beta > 1$ *are fixed parameters.*

Conditions *a* and *c* are the same as in case of $\phi \rightarrow \psi$ statement.

Condition *b* is almost the same as in case of $\phi \rightarrow Bel(\psi)$. The only difference is that parameters for possibility operator should be smaller than those for belief operator. It is recommended to set $\overline{\alpha}_{Pos} = \underline{\alpha}_{Bel}$. In such case there is no situation where agent can say that she at the same time moment only finds something possible and believes in it (can happen when $\overline{\alpha}_{Pos} > \underline{\alpha}_{Bel}$). On the other hand, there is no situation where agent can't say that she believes something, because it is too little possible and at the same time moment is unable to say that something is possible, because she finds it too possible (can happen when $\overline{\alpha}_{Pos} < \underline{\alpha}_{Bel}$).

9 Summary

We have introduced a time variable to conditional statements with modal operators of belief and possibility. Such extension allows agent to describe past and future time moments. We showed that conditional statement with time variable can be successfully grounded in agent's mental model.

The statements always have crisply defined references to points is time. The proposed analysis has shown, that omitting time reference from a statement will impose serious interpretation problems, because listener has to reproduce the meaning of a statement based on its contents. On the other hand grounding such statements within an agent will lead to many implementations of epistemic satisfaction relation based on the character of described variables. We are planning in the nearest future the extension of language to allow such statements for most typical types of events such as achievements, continuous events and state changes.

References

1. Adams, E.W.: Subjunctive and Indicative Conditionals. Foundations of Language 6, 89–94 (1970)
2. Adams, E.W.: The Logic of Conditionals. Reidel, Dordrecht (1975)
3. Ajdukiewicz, K.: Conditional sentence and material implication. Studia Logica 4(1), 135–153 (1956)
4. Bogusławski: More on Ajdukeiwicz's Conception of «Expressing». Studia Semiotyczne XIV-XV, 249–270 (1986) (in Polish)
5. Clark, M.: Ifs and Hooks. Analysis 32(2), 33–39 (1971)
6. Cohen, P., Levesque, H.: Communicative Actions for Artifical Agents. In: Proc. of the 1st International Conference on Multi-agent Systems, San Francisco (1995)
7. Arló Costa, H., Levi, I.: Two notions of epistemic validity. Synthese 109(2), 217–262 (1986)
8. Grice, H.P.: Meaning. Philosophical Review 66, 377–388 (1957)
9. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
10. Hintikka, J.: Knowledge and belief. In: An Introduction to the Logic of the Ttwo Notions. Cornell University Press, Ithica (1962)
11. Huhns, N., Signh, M.: Cognitive Agents. IEEE Internet Computing 2(6), 87–89 (1998)
12. Katarzyniak, R.: The language grounding problem and its relation to the internal structure of cognitive agents. Journal of Universal Computer Science 11(2), 357–374 (2005)
13. Katarzyniak, R.: Gruntowanie modalnego języka komunikacji w systemach agentowych. Exit, Warsaw (2007) (in Polish)
14. Lewis, D.: Probability of Conditionals and Conditional Probabilities. Philosophical Review 85, 297–315 (1976)
15. Moens, M., Steedman, M.: Temporal ontology and temporal reference. Journal of Computational Linguistics 14(2), 15–28 (1988)
16. Pelc, J.: If, then. Studia Semiotyczne XIV-XV, 271–286 (1986) (in Polish)
17. Ramsey, F.: General Propositions and Causality. Ramsey, 145–163 (1990)
18. Skorupa, G., Katarzyniak, R.: Extending modal language of agents' communication with modal implications. Information Systems Architecture and Technology, 127–136 (2009)
19. Skorupa, G., Katarzyniak, R.: Applying Possibility and Belief Operators to Conditional Statements. LNAI (2010); in Print Information Systems Architecture and Technology, 127–136 (2009)
20. Stalnaker, R.: A Theory of Conditionals. Studies in Logical Theory. American Philosophical Quarterly, Monograph 2, 98–112 (1968)
21. Willer, M.: New surprises for the Ramsey Test. Synthese. Springer, Heidelberg (2009) (online)

Automatic Ontology Evolution in Open and Dynamic Computing Environments

Edgar Jembere, Sibusiso S. Xulu, and Matthew O. Adigun

Centre for Mobile e-Service for Development, University of Zululand, RSA
ejembere@yahoo.co.uk, ssxulu@pan.uzulu.ac.za,
madigun@pan.uzulu.ac.za

Abstract. Automated computing in open and dynamic computing environments requires automatic update and revision of the Knowledge Bases (KBs) to keep the KBs up to date with the dynamics in the environment and correct incorrect knowledge held in the KBs respectively. Furthermore, the truthfulness, applicability and validity of this knowledge depend on the context under which the knowledge is to be used. This then calls for the development of solutions to enable KBs to (i) be evolved over time enabling them to keep up to date with the evolving world or changes in the world's conceptualisation, (ii) allow situational reasoning, and (iii) reasoning under uncertain, incomplete and inconsistent knowledge. The emerging fielding of probabilistic ontologies is impregnated with promises to resolve such issues. However, an investigation on how such knowledge representations can be objectively and rationally evolved is needed. This paper presents issues, methods and ideas towards rational probabilistic ontology evolution in open and dynamic computing environments.

Keywords: Ontology evolution, Belief Change, Probabilistic ontology.

1 Introduction

In net centric and collaborative computing environments, ontologies have become ubiquitous. They are the means through which the information overload is evaded and autonomous computing is achieved. Manual evolution of ontologies is both difficult and impractical given the size of present day ontologies and subjectivity of the manual ontology evolution process [1]. This has prompted researchers to look for ways of objectively, rationally and automatically evolving ontologies (e.g. [1], [2], [3]). Addressing ontology evolution at knowledge change level, while ignoring Knowledge representation, does not suffice given the openness and the dynamic nature of the next generation computing environments, such as the semantic web and semantic grids. Uncertainty, inconsistency and complexity are typical characteristics of open and dynamic computing environments which cannot be avoided [1], [4], [5], [6], [7]. They need to be captured at knowledge representation level. Classical ontology modelling approaches do no cater for these characteristics. This calls for development of mechanisms that enable next generation computing environments to account for these characteristics in representing and reasoning upon the knowledge in

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 122–132, 2010.

their KBs. Probabilistic approaches to knowledge representation are a promising candidate towards this goal.

The rest of this paper is organised has follows. Section 2 presents a use case we used as the object of analysis to abstract the peculiarities of open and dynamic computing environments. Using the use case presented in Section 2 and a Literature Survey, Section 3 presents some key design constructs for an ontology representation and management framework for open and dynamic computing environments. Section 4 discusses the current research trends towards handling uncertainty, incompleteness and inconsistencies in ontologies. In Section 5, we present a knowledge representation framework for open and dynamic computing environments. Thereafter we give a review of how belief change is handled in classical logic KBs (Section 6). The two main approaches to belief change are discussed and a theoretical evaluation of how they can be used in ontology change is given. Section 7 uses the results of Section 5 to address iterated belief change (evolution) in probabilistic ontologies. Section 8 concludes the paper and gives our future research directions towards rational ontology evolution in open and dynamic computing environments..

2 Use Case Scenario

Taking inspiration from the Next-Generation Grid vision envisaged in [8], [9], [10] and [11], we considered a next generation grid environment, such that based on a user request for a composite application, the grid infrastructure can automatically discover, select and bind to all the grid services that make up an application that satisfies the user's requirements and preferences and optimises the user's utility. Such a task requires the infrastructure to have the user knowledge, web/grid services knowledge and personalisation function knowledge [12]. Typically, from a semantic web point of view, such information should be captured in ontologies. The user knowledge will include user demographics, interests, context and preferences etc. The service knowledge will include service functionality, features, Quality of Service properties etc. Owing to the somehow chaotic nature of open computing environment this information is likely to come from unreliable sources resulting in uncertainties. Due to changes in the domain or changes in its conceptualisation, the knowledge is bound to change over time and the system needs to capture these changes for it to efficiently serve the user. Further to this, the computing environment is highly dynamic and the system needs to make decisions on which services to bind to, which personalisation function to use, etc as the computing environment changes. Given this, coupled with the uncertainties, inconsistencies, incompleteness and complexities that characterise such computing environments, the system needs a mechanism for handling uncertain and inconsistent knowledge in its decision making. Allowing situational reasoning is one approach which is used to reduce inconsistencies through focusing inferences only to relevant portions of the KB [13, 14].

3 Design Considerations

An analysis of the use case presented in Section 2, and a survey of related research works, found the following challenges towards realising the goal of supporting knowledge-based decision making in open and dynamic computing environments.

Certainty of uncertain and inconsistent knowledge. Much of the information in open and dynamic computing environments is uncertain, incomplete, inconsistent, often incorrect or only partially correct raising credibility issues over inferences drawn from such information [13]. Using determinstic ontologies to represent such knowledge leaves a lot to be desired. Representation of uncertainty in such environments has the promises of discounting the effect of these knowledge imperfections and provides a proof theory over KBs holding such information.

The dynamic nature and complexity of the computing environment. Knowledge representations in open and dynamic environments need mechanisms for incorporation of new knowledge in to the KBs. This is due to the following two reasons: (i) owing to the complexity of the environment and incredibility of the knowledge sources, the system beliefs about the world may simply be mistaken or incomplete, (ii) the system's beliefs about the environment might have been correct at some time, but the knowledge may become inaccurate due to the changes in the world, rendering certain facts true and falsifying some [15]. Further to this, in dynamic environments, the truthfulness of the system's beliefs might be situational, which will require situational knowledge representation and reasoning [6, 14, 16, 18].

Impracticality of manual management of knowledge change. The complexity of the environment and the size of ontologies in open environments make manual ontology evolution impossible owing to the following reasons: (i) ontology change generally requires that the person effecting the changes to the KB be both a knowledge engineer and domain expert and very few people can be both, (ii) due to collaborative nature of the environment manual belief changes by different knowledge engineers is likely to result in different KBs. This is because knowledge engineers have different views on how a certain change should be implemented resulting from differences in background knowledge, personal preferences, subjective opinions etc [2], (iii) in highly dynamic environments, ontology changes are so frequent that by the time the engineer finish effecting a change the updated KB may already be lagging behind current state of knowledge in the domain.

From the above analysis the following design criteria were deduced for knowledge representation and reasoning in open and dynamic computing environments:

i. Knowledge Representation should be able to represent and reason about entities that are related to each other and their uncertainties, which might have resulted from uncertain and inconsistent data,
ii. Support for a formal, rational and objective operators for iterated ontology change (ontology evolution),
iii. Support for situational knowledge representation and reasoning, and
iv. Support for automatic distributed knowledge learning and subsequent ontology revision and update by the learnt knowledge based on a rational and objective ontology change operator.

This paper, however, focuses on the current research trends towards meeting the first three design requirements and gives our view towards realising these requirements. The forth requirement is left as part of our future work.

4 Uncertainty in Ontologies

Probability has emerged as the natural candidate to represent uncertain phenomena. As a result, a lot of research efforts have been directed towards introduction of probabilities in knowledge representations [13]. An investigation of the state of the art revealed that there are three (3) ways of introducing probabilities into ontologies:

Uncertainty modelled in Data only (e.g. Learned Ontology Model (LOM) [5]): In this case uncertainty is modelled in the data model called the LOM which is independent of the concrete ontology. The LOM does not have logical semantics, and hence it does not consider logical inconsistencies. Logically inconsistent axioms are kept in the model with uncertainty represented as annotation capturing the confidence about their correctness. Axioms with higher confidence and are coherent with other axioms in the LOM are returned into the concrete ontology with no uncertainty represented.

Directly extending ontologies by a probabilistic model (e.g. Probabilistic Description Logics [6] and [16], Markov Logic [17]): This approach extends classical ontologies with probabilistic annotations to the knowledge represented in the ontologies without changing the structure of knowledge representations.

Transformation of ontologies in to a structure that handles probabilities (e.g. PR-OWL [18] and OntoBayes [19]): This approach is motivated by the fact that ontologies represent classes, individuals, properties and relationships, whereas probability deals with random variables. So in order to incorporate probabilities in to the ontology, the ontology have to be transformed in to a form that allows probabilistic knowledge representation to be incorporated.

Table 1 shows an evaluation of the above three approaches relative to the design criteria discussed in Section 3. Modelling uncertainty in data only, results in a classical ontology. This guarantees the use of knowledge change techniques that have been adapted from belief change analysis (AGM [20] and KM [21] theories) for ontology evolution. The advantage of using these techniques springs from the fact that these theories have been formalised and have matured over the years. However, iterated belief change is still a challenge in these techniques due to the fact that these techniques were designed for a single change operation and the outcome only captures the resulting KB with no resultant relative importance of the axioms in the KB, which results in loss of rationality for the next change to be effected. As a result objectivity in the change operation is lost after the first change operation.

Research works on ontology evolution that adopt direct extension of classical ontologies with probabilistic models (PMs) focus on resolving inconsistencies in the resulting ontology by relaxing the axioms in the KB with probabilistic weights. Formalisation of belief change process is not discussed in these ontologies. However, their knowledge representation (KR) formalism suggests that probabilistic conditioning can be used as the ground theory for belief change.

Transformation of ontologies into graphical structures results in Knowledge structures such as Object-Oriented Bayesian Network (OOBN) [22], Probabilistic Relational Models (PRMs) [23] and Multi Entity Bayesian Networks (MEBN) [13], that allow representation of uncertainty about the attributes of instances of different

types of objects. We call these knowledge representation structures Probabilistic Relational Ontologies (PROs). Although Scharrenbach and Bernstein [6] point out that transformation of ontologies into PROs causes an ontology evolution scheme to be complex, it can be argued that evolution of such ontologies can only be as complex as that of classical ontologies (assuming that the initial ontology is probabilistic). The argument is based on the fact that evolution of PROs takes advantage of the algorithms that have been tried and tested in the mature field of Graphical Probability and Decision models. In view of the forgoing discussion, we view PROs as the best knowledge representation structures for open and dynamic environments. Hence in Section 5 we are going to discuss a generalised PRO knowledge representation model before we discuss their evolution in the subsequent sections.

Table 1. Evaluation of Approaches to probabilistic knowledge representation

Design Requirement	Uncertainty modelled in data only	Directly extending ontologies by a PM	Transformation of ontologies to another structure
Representation of uncertain & inconsistent knowledge	Only handles uncertainty in the data model and not in the ontology	Handles uncertainty in the ontology, but cannot cater for structural uncertainty	Handles uncertainty in the ontology, including structural uncertainty.
Formalisation of belief change	Belief change theories designed for Classical logic (e.g. AGM [20] theory) are used.	No ontology change theory is discussed. Conditioning can be used.	Bayesian theory on non-structural knowledge. No approach has been proposed for structural knowledge.
Rationality of Belief Change	Rational	N/A	Rational
Objectivity of belief change	Objective only for a single change & not over iterated belief change.	N/A	Objective belief change through Bayesian theory on non-structural knowledge
Easy of evolution	Belief change theory only for single belief change and not iterated belief change.	Might be easy for non-structural knowledge if conditioning is adopted	More complex at structural level but Bayesian theory handles iterated belief change of non-structural knowledge
Situational KR	Not explicit in the knowledge model	Not explicit in the knowledge model	Not explicit in the knowledge model

5 KR for Open and Dynamic Environments

Taking inspiration from MEBN theory [13], a generalised PRO is defined as a set of Frames, known as MFrags, consisting of Directed Acyclic Graphs (DAGs) satisfying some consistency constraints. Each MFrag, $\mathcal{F}$, is defined as follows: $\mathcal{F}=(\mathcal{C},\mathcal{I},\mathcal{R},\mathcal{G},\mathcal{D})$, where $\mathcal{C}$ is a finite set of context value assignments, $\mathcal{I}$ is a finite set of independent input random variables, $\mathcal{R}$ is a finite set of dependent random variables known as

Resident random variables, $\mathcal{G}$ is a DAG having elements from $\mathcal{I}$ as root (input) nodes and elements in $\mathcal{R}$ as child (Resident random variable) nodes, and a set $\mathcal{D}$ of local distributions one for each member of $\mathcal{R}$. The local distribution specifies conditional probability distributions for the dependent random variables represented in Conditional Probability Tables (CPTs). Random variables are First Order Logic (FOL) statements which take entities that exist in the domain as arguments. *A PRO is said to be inconsistent if it does not conform to the following consistency constraints: (i) there should be no cyclic influences in the DAGs and, (ii) there should be no multiple conflicting distributions for random variables in different MFrags* [13]. The set of context value assignments give the contextual constraints under which a hypothesis represented by the DAG, $\mathcal{G}$, holds. This makes MEBN a natural choice for knowledge representation in dynamic computing environments where context data influences the truthfulness and validity of knowledge axioms in the KB. The effect of context to the ontology can be either be: (i) on the statistical regularities in a given MFrag or (ii) on the structure of the DAG in the MFrag. In each of these cases the respective context data is represented differently. In the first case, the context data have to be represented as input nodes in the MFrag, where it can only affect local distribution of the resident random variables. In the second case, because the structure of the MFrag is changed the context has to be represented as context nodes. Figure 1 shows Service Cost MFrag for services domain ontology. In Figure 1, it is assumed that *network context* and *access device type* are context variables that affect the structure of the Service Cost MFrag and hence are represented as context nodes. *Time of the day* is assumed not to have an effect on the structure of the DAG and hence it is represented as an Input Node.

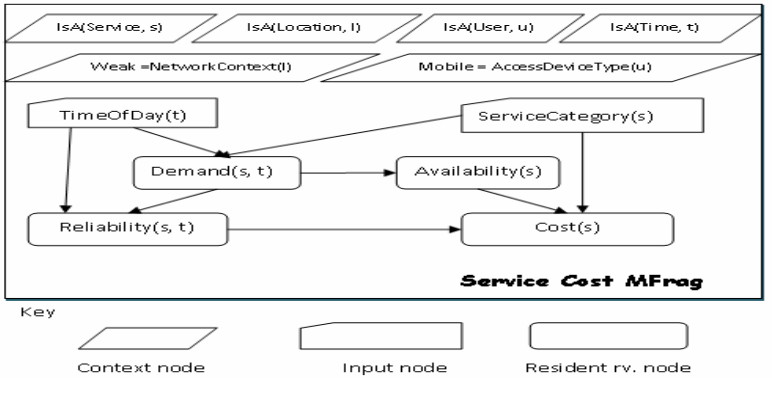


Fig. 1. Probabilistic Relational Ontology MFrag

Taking inspiration from the semantics of description logics we separate a probabilistic ontology into two components the *intensional* and the *extensional* component. The intensional components express the general knowledge about the concepts and their relationships in a given domain. That is, it describes what input and dependent random variables should be in the DAGs and the dependences between the variables. The extensional component describes the state of affairs regarding the

instances of concepts and relationships in a domain. This distinction is very important for characterisation of ontology evolution as will be discussed in the Section 7.

6 Rational Belief Change and Ontology Evolution

Ontology evolution is defined as, "*the process of modifying ontology in response to a change in the domain (first kind of change) or its conceptualisation (second kind of change)*" [25]. This view to ontology evolution is borrowed from belief change theory. In belief change theory, belief change is necessitated by two factors: (i) the world may have changed, or (ii) the system's beliefs about the world may be simply mistaken or incomplete. Changes necessitated by changes in the world are addressed by belief update while belief revision addresses changes necessitated by mistaken or incomplete knowledge about the world.

Belief revision is used to effect belief changes in static domains. In such domains, belief changes are only necessitated by the fact that the system's beliefs about the world are mistaken or incomplete since there will be no changes in the domain. AGM theory [20] is the flagship implementation of belief revision. AGM takes a coherence view to knowledge change, which advocates that credibility of an axiom depends on how coherent it is with other axiom in the KB. The premise behind the coherence view is, if a belief revision operation calls for some beliefs to be retracted (in order to keep the KB consistent after the belief change operation), the relative entrenchment of a belief depends on how coherent the belief is with other beliefs in the belief set.

Belief update is used for belief change in dynamic worlds. The work by Katsuno and Mendelzon [21], known as the KM theory is the most popular theory for belief update. The KM theory uses the event model to capture knowledge changes in a dynamic world. The KM theory takes observation of a new axiom as evidence that there have been a change in the world. The assumption here is the existence of new axioms is the least of what could have possibly changed in the world. In order to capture what have transpired the KM theory models the most plausible transition of a given world, w, into a world, v, that satisfies the observation of A. A set of pre-orders that are reflexive and transitive are defined over a set of worlds. A pre-order $u \leq_w v$ implies that u is at least as plausible a relative change relative to w as v.

In KM, unlike in AGM, once an inconsistency is introduced in the KB there is no way to eliminate it using the update operator because update assumes that the world has changed. This implies that the update operator does not allow an observation to force revision of the system's beliefs about the state of the world prior to the observation. In short the KM theory can be considered to be concerned with states of a changing world at given time stamps and the transition between these states necessitated by an event model, where as the AGM is concerned with belief states and their possible revision necessitated by the epistemic entrenchment of axioms in the belief set. A lot of effort has gone into generalising the KM theory to cater of revision (e.g. [24] and [27]). While such approaches can cater for both belief revision and update, the resultant KB is simply a belief set with no updated epistemic states. The resultant belief state provides no guidance for changes in belief due to subsequent changes. This makes such approaches less promising candidates for automatic ontology evolution. Another key challenge to adoption of AGM and KM theories is

that in both theories interpretation and assignment of rankings cannot be done in natural way that will allow automation of iterated belief change.

Generally, belief change is guided by the following six (6) principles [1], [2] (i) principle of primacy of new information, (ii) principle of irrelevance of syntax, (iii) principle of consistency maintenance, (iv) principle of adequacy of representation, (v) principle of fairness, (vi) principle of minimal change. Some of the principles need to be relaxed to be applicable to ontology evolution in open and dynamic environment. An analysis of our application environment revealed that the principle of primacy of new information and the principle of consistency maintenance will need to be relaxed. New information may need to be rejected or partially rejected due to the distributed nature, uncertainty and incompleteness of data, and unreliability of information sources in open and dynamic computing environments. In such a case, an ontology change operator needs to consider techniques from non-prioritised belief change [25]. Apart from considering such techniques, probabilistic ontologies handle non-prioritised belief change in a natural way through assignment of a degree of belief to all axioms, including new axioms before factoring them into the KB. The challenge with consistency maintenance principle is that inconsistency in probabilistic ontologies cannot be defined as it is defined in determinstic ontologies, because inconsistencies such as contradictory axioms do not exist in Probabilistic ontologies.

7 Ontology Evolution in Dynamic Environments

The current work on ontology evolution (e.g. [1], [2], [26], [28]) adapt the AGM theory for ontology evolution. To the best of our knowledge, no work has tried to use KM theory in ontology evolution. Apart from the non-availability of formal, rational and objective mechanisms for evolution of probabilistic ontologies, no work exist on automatic ontology update even for classical ontologies. This is mainly because classical ontologies separate ontology and epistemology. Ontologies capture what kinds of entities exist, where as epistemology is about how knowledge of what exists is acquired, and the justification of such knowledge. Due to the fact that in dynamic environments, knowledge changes as data accrues, epistemology cannot be separated from ontology. Probabilistic ontologies based on bayesian theory handles this combination of ontology and epistemology in a more natural way. Further to this, Bayesian theory provides a rational and objective solution to belief change for non structural knowledge. However, formalisation of structural evolution of probabilistic ontologies has not been addressed and cannot be supported in a natural way through Bayesian semantics. We believe that techniques from belief change theories can be adapted for this purpose.

Our approach to ontology evolution takes advantage of our view of separating extensional and intensional knowledge discussed in Section 5. Evolution at extensional level will involve iterated belief change on individuals and instances, which will only affect the values in the CPTs. Evolution at intensional level will involve iterated change on the structural knowledge. This distinction is important because knowledge evolution at these two levels pragmatically have different requirements and nature. Knowledge at structural level is more persistent only changing after some time where as knowledge at extensional level changes rapidly responding to the changes in the environment.

Belief change at extensional level is handled by bayesian theory. Bayesian theory has capabilities of handling iterated belief revision through conditioning on evidence iteratively as newer evidence accrues (assuming the world is static). Mature parametric learning algorithms exist for this purpose. The key challenge that arises here is, how can belief update be catered for at extensional level in a rational way as the world changes. In belief change theory this is addressed by introducing a temporal component into the knowledge representation. Against this background we are envisaging taking advantage of the techniques emanating from the field of Dynamic Bayesian Networks (DBNs).

In order to ensure rational belief change and consistency maintenance at extensional level, ontologies resulting from a change must conform to the consistency constraints defined in Section 5. Though quite a number of algorithms for structure learning in Bayesian networks exist, formalisation of how these algorithms can be used for iterated structural change in PROs ensuring conformance to belief change principles has not been explored. Structural ontology changes include (i) introduction/retraction of attribute values for random variables, (ii) expansion of KB with new variables, (iii) revision/update of structural dependencies between variables [13]. Ontology Change due to existence of new possible values for a random variable in the DAG can be treated as an ontology expansion problem not requiring anything to be removed from the KB to accept the change. The attribute value should just be added in all Conditional Probability Tables (CPTs) in all MFrags in the KB the random variable appears. *Expansion of a KB by a new random variable* and *revision/ update of structural dependencies between variables* may result in an inconsistent ontology. To ensure consistency this will require that some already existing structural dependency be retracted from the ontology. This then calls for the development of objective and rational belief change principles for managing iterated change at intensional level balancing the propensities of consistency maintenance and minimal change.

8 Conclusion and Future Work

This paper presented techniques, ideas, and issues towards design of knowledge representation and evolution in open and dynamic computing environments. Classical ontologies fall short in representing and reasoning about knowledge in open and dynamic environments due their lack of a principled way for handling uncertain, inconsistent and incomplete knowledge, which are unavoidable in open and dynamic computing environment. Owing to the fact that the computing environment is dynamic, the knowledge about the environment is bound to change over time. Manual evolution of the knowledge representation in such environments is impractical due to the complexity and size of the ontologies, hence, the need for intuitive, objective and automatic ontology evolution algorithms for open and dynamic environments.

We advocated for the adoption of probabilistic ontologies because of their ability to represent uncertain phenomena and resolving inconsistencies in knowledge representation through relaxation of conflicting axioms. Belief change theory and Bayesian techniques are adapted into the formalisation of probabilistic ontology evolution. This way, most of the techniques, ideas, intuitions and algorithms from belief change and Bayesian theory can be leveraged on in developing a rational

probabilistic ontology evolution solution. This paper did not provide a concrete solution to the problem at hand; our aim was to lay the theoretical foundation upon which deeper results can be based on, thus paving the way for the development of effective solutions to knowledge representation and evolution in open and dynamic computing environments.

This work is being done in the context of Grid intelligence were the results are to be used in building intelligence in grid computing environments to support coordinated resource sharing and problem solving and on the grid to support knowledge-based applications that run on the grid. Context-awareness and personalisation are some of the knowledge intensive functions that need to be supported by our solution approach in grid environments. Owing to the distributed nature of these environments, support for automatic distributed knowledge discovery and subsequent ontology revision and update is needed. Future work should address how our proposed solution approach can be incorporated into grid environments.

Acknowledgments. This work was done under the Centre for Mobile e-Services for Development at the University of Zululand. The centre is funded by THRIP, Telkom, NRF and Huawei, and Alcatel.

References

1. Flouris, G.: On Belief Change and Ontology Evolution. PhD thesis, University of Crete, Greece (2006) (unpublished)
2. Flouris, G., Plexousakis, D., Antoniou, G.: Evolving Ontology Evolution. In: Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., Štuller, J. (eds.) SOFSEM 2006. LNCS, vol. 3831, pp. 14–29. Springer, Heidelberg (2006)
3. Bundy, A., Chan, M.: Towards ontology evolution in physics. In: Hodges, W., de Queiroz, R. (eds.) Logic, Language, Information and Computation. LNCS (LNAI), vol. 5110, pp. 98–110. Springer, Heidelberg (2008)
4. Yang, Y., Calmet, J.: OntoBayes: An Ontology-Driven Uncertainty Model. In: CIMCA/IAWTIC 2005, pp. 45–463 (2005)
5. Haase, P., Völker, J.: Ontology Learning and Reasoning - Dealing with Uncertainty and Inconsistency. In: ISWC-URSW 2005, pp. 45–55 (2005)
6. Scharrenbach, T., Bernstein, A.: On the Evolution of Ontologies using Probabilistic Description Logics. In: Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (2009)
7. Ding, Z., Peng, Y.: A probabilistic extension to ontology language owl. In: Proceedings of the International Conference on System Sciences (HICSS 2004), vol. 4, pp. 1–15. IEEE Computer Society, Washington (2004)
8. Cannataro, M., Talia, D.: Semantic and Knowledge Grids: Building the Next -Generation Grid. IEEE Intelligent Systems - Special Issue on e-Science 19(1), 56–63 (2004)
9. De Roure, D., Jennings, N.R., Shadbolt, N.: The Semantic Grid: A Future e-Science Infrastructure. In: Berman, F., Hey, A.J.G., Fox, G. (eds.) Grid Computing: Making the Global Infrastructure a Reality, pp. 437–470. John Wiley & Sons, Chichester (2003)
10. Brezany, P., Goscinski, A., Janciak, I., Tjoa, A.M.: The development of a Wisdom Autonomic Grid. In: Proceedings of the Workshop on Knowledge Grid and Grid Intelligence 2004, China (2004)

11. Zhuge, H.: Semantics, Resource and Grid. Future Generation Computer Systems 20(1), 1–5 (2004)
12. Liu, P., Nie, G., Chen, D., Fu, Z.: The knowledge grid based intelligent electronic commerce recommender systems. In: Proceedings of the IEEE International Conference on Service-Oriented Computing and Applications, Newport Beach, pp. 223–232 (2007)
13. Laskey, K.B.: MEBN.: A Logic for Open-World Probabilistic Reasoning. The Volnegau School of Information Technology and Engineering. George Mason University, Fairfax, VA, USA (2005)
14. Ngo, L., Haddawy, P.: Answering queries from context-sensitive probabilistic knowledge bases. Theoretical Computer Science 171(1-2), 147–177 (1997)
15. Costa, P.C.G.D., Laskey, K.B., Laskey, K.J., Pool, M.: Uncertainty Reasoning for the Semantic Web. In: Proceedings of the International Semantic Web Conference, ISWC 2005, Workshop 3, ISWC-URSW, Galway, Ireland (2005)
16. Lukasiewicz, T.: Probabilistic Default Reasoning with Conditional Constraints. Ann. Math. Artif. Intell. 34(1-3), 35–88 (2002)
17. Domingos, P., Kok, S., Lowd, D., Poon, H., Richardson, M., Singla, P.: Markov Logic. In: De Raedt, L., Frasconi, P., Kersting, K., Muggleton, S.H. (eds.) Probabilistic Inductive Logic Programming. LNCS (LNAI), vol. 4911, pp. 92–117. Springer, Heidelberg (2008)
18. Costa, P.C.G.D.: Bayesian Semantics for the Semantic Web. Doctoral Dissertation. Department of Systems Engineering and Operations Research, George Mason University, Fairfax, VA, USA (2005)
19. Yang, Y., Calmet, J.: OntoBayes: An Ontology-Driven Uncertainty Model. In: CIMCA/IAWTIC 2005, pp. 457–463 (2005)
20. Alchourron, C., Gärdenfors, P., Makinson, D.: On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. Journal of Symbolic Logic 50, 510–530 (1985)
21. Katsuno, H., Mendelzon, A.O.: On the Difference between Updating a Knowledge Base and Revising It. In: KR 1991, pp. 387–394 (1991)
22. Koller, D., Pfeffer, A.: Object-oriented Bayesian networks. In: Geiger, D., Shenoy, P.P. (eds.) Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, pp. 302–313. Morgan Kaufmann, San Francisco (1997)
23. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning Probabilistic Relational Models. In: IJCAI 1999, pp. 1300–1309 (1999)
24. Boutilier, C.: A Unified Model of Qualitative Belief Change: A Dynamical Systems Perspective. Artif. Intell. 98(1-2), 281–316 (1998)
25. Mazzieri, M., Dragoni, A.F.: Ontology Revision as Non-Prioritized Belief Revision. In: ESOE 2007, pp. 58–69 (2007)
26. Flouris, G., Plexousakis, D.: Bridging Ontology Evolution and Belief Change. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 486–489. Springer, Heidelberg (2006)
27. Friedman, N., Halpern, J.Y.: Modeling belief in dynamic systems part II: revision and update. Journal of Artificial Intelligence Research 10(1), 117–167 (1999)
28. Guelfi, N., Pruski, C., Reynaud, C.: Understanding Supporting Ontology Evolution by Observing the WWW Conference. In: ESOE 2007, pp. 19–32 (2007)

Diagnostic Tests Based on Knowledge States

Sylvia Encheva[1] and Sharil Tumin[2]

[1] Stord/Haugesund University College, Bjørnsonsg. 45, 5528 Haugesund, Norway
sbe@hsh.no
[2] University of Bergen, IT-Dept., P.O. Box 7800, 5020 Bergen, Norway
edpst@it.uib.no

Abstract. In this paper we investigate stabilities of concepts resulting from extending partial ordering of a set of elements to partial ordering of a larger set of elements. This approach can be applied for facilitating the process of decision making related to automated tests based on knowledge states.

Keywords: Stabilities of concepts, uncertainty, orderings, automated tests.

1 Introduction

Operating with sufficient amount of information about a student's level of knowledge is very important with respect to creating an efficient and personalized learning environment. Such information however is often obtained by examining Web-based tests where most of the available technical solutions are not supporting an optimization of the quantity and quality of formative assessments. This can be resolved by employing methods from the theories of Formal Concept Analysis, [12] and knowledge spaces, [9].

A knowledge state of a student is a set of problems that the student can solve and a domain of knowledge is a collection of items (e.g., learning objects, problems, questions, exercises, examples, etc.) in a given field of knowledge (e.g., mathematics, physics, chemistry, biology, etc.) [10]. A knowledge structure of a subject is the set of all knowledge states in a subject. Knowledge structures illustrate the implications between items in a set of problems.

A skill map provides the basis for mapping the performance of a learner (represented by her knowledge state) into the competence level of the skills that the learner needs for mastering the content of the items contained in her knowledge state, [1]. A skill map is a cognitive model specifying the skills required to handle a particular set of items. The model requires relations between items and skills to be validated, [10]. Skill maps in a distributed skill map are generated by several individuals, where each one designs an item and points to skills required to work with that item. Thus a single item is often related to different sets of skills and the final set of sufficient skills has to be validated. Once a skill map is found reliable it has to be applied in real situations.

Since concepts are necessary for expressing human knowledge, any knowledge management process benefits from a comprehensive formalization of concepts,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 133–141, 2010.

[24]. Formal Concept Analisis offers such a formalization by mathematizing the concept of 'concept' as a unit of thought constituted of two parts: its extension and its intension, [12].

In this paper we investigate stabilities of concepts resulting from extending partial ordering of three items to partial ordering of four items belonging to knowledge structures. This can be applied for facilitating the process of decision making related to tests based on knowledge states.

The rest of the paper is organised as follows. Related work and supporting theory may be found in Section 2. The main results are presented in Section 3. Conclusions can be found in Section 4.

2 Background

2.1 Posets

Let P be a non-empty ordered set. If $sup\{x, y\}$ and $inf\{x, y\}$ exist for all $x, y \in P$, then P is called a *lattice* [7]. In a lattice illustrating partial ordering of knowledge values, the logical conjunction is identified with the meet operation and the logical disjunction with the join operation.

A *context* is a triple (G, M, I) where G and M are sets and $I \subset G \times M$. The elements of G and M are called *objects* and *attributes* respectively [7], [26]. For $A \subseteq G$ and $B \subseteq M$, define $A' = \{m \in M \mid (\forall g \in A) \ gIm\}$, $B' = \{g \in G \mid (\forall m \in B) \ gIm\}$ where A' is the set of attributes common to all the objects in A and B' is the set of objects possessing the attributes in B.

A *concept* of the context (G, M, I) is defined to be a pair (A, B) where $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. The *extent* of the concept (A, B) is A while its *intent* is B. A subset A of G is the extent of some concept if and only if $A'' = A$ in which case the unique concept of the which A is an extent is (A, A'). The corresponding statement applies to those subsets $B \in M$ which is the intent of some concepts.

The set of all concepts of the context (G, M, I) is denoted by $\mathfrak{B}(G, M, I)$. $\langle \mathfrak{B}(G, M, I); \leq \rangle$ is a complete lattice and it is known as the *concept lattice* of the context (G, M, I).

2.2 Ordered Sets

Determining a consensus from a group of orderings and making statistically significant statements about orderings have been discussed in [4].

A relation I is an *indifference* relation when given AIB neither $A > B$ nor $A < B$ has place in the componentwise ordering. A partial ordering whose indifference relation is transitive is called a *weak ordering*.

Let w_1, w_2, w_3 be weak orderings. Then w_2 is between w_1 and w_3 if each decision made by w_2 is made by either w_1 or w_3 and any decision made by both w_1 and w_3 is made by w_2, i.e. $w_1 \cap w_3 \subseteq w_2 \subseteq w_1 \cup w_3$.

The distance $d(w_1, w_3)$ is defined as $d(w_1, w_2) + d(w_2, w_3) = d(w_1, w_3)$. The distance is a metric in the usual sense, it is invariant under permutation of alternatives, and the minimum positive distance is 1, Fig 1.

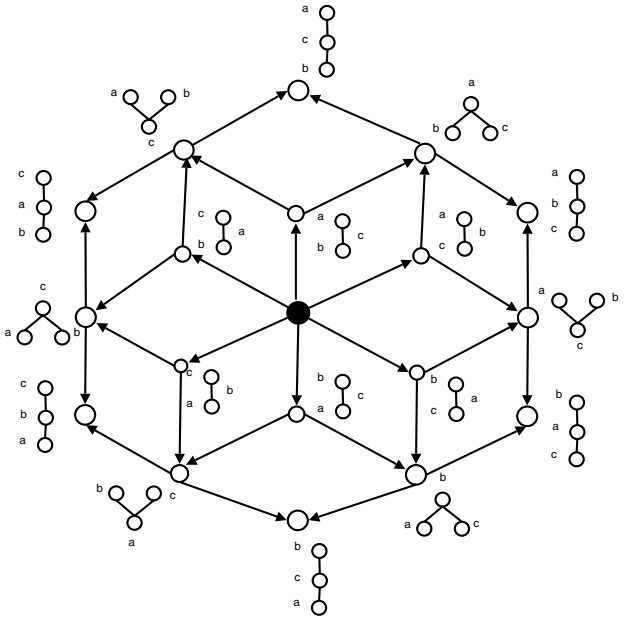


Fig. 1. The graph of all partial orderings of a 3-element set

2.3 Knowledge Spaces

Within the theory of knowledge spaces, a knowledge domain is described by a set of problems Q, [10]. Then the knowledge state of a student is a set $Q' \subseteq Q$ that the student can solve. A knowledge structure of a subject is a set of all knowledge states. A knowledge space is a closure system, where closure system on a finite set M is a set F of subsets of M such that $M \in F$ *and* C, $C^1 \in F \Rightarrow C \bigcap C^1 \in F$.

The complements of the intents of a formal context is a knowledge space and a response pattern corresponds to a set of attributes, [20]. Thus a knowledge space can be obtained from a formal context.

Existence of a knowledge space indicates that an item can be related to several sets of pre- requisites, [9]. Therefore, a solution for an item can be obtained via different strategies.

A *skill map* is a triple (Q, S, f) where Q is a non-empty set of items, S is a non-empty set of skills and f is a mapping from Q to $2^S \backslash \emptyset$, [10], [14]. The set of skills assigned to q is the set $f(q) \subset S, \forall q \in Q$. Thus a correct solution of item q implies possesion of all the skills belonging to $f(q)$.

A *distributed skill* map, [22] is a pair $(\mathcal{D}, \vee)$ where $\mathcal{D}$ is a collection of skill maps, and $\vee$ (read: "join") is a binary operator such that, given any three skill maps $A := (Q_A, S_A, \sigma_A)$, $B := (Q_B, S_B, \sigma_B)$ and $C := (Q_C, S_C, \sigma_C)$ and $\mathcal{D}$, $C = A \vee B$ if and only if

1. $Q_C = Q_A \cup Q_B$,
2. $S_C = S_A \cup S_B$,
3. $\sigma_C(q) = \sigma_A(q) \cup \sigma_B(q) \quad \forall \quad q \in Q_C$.

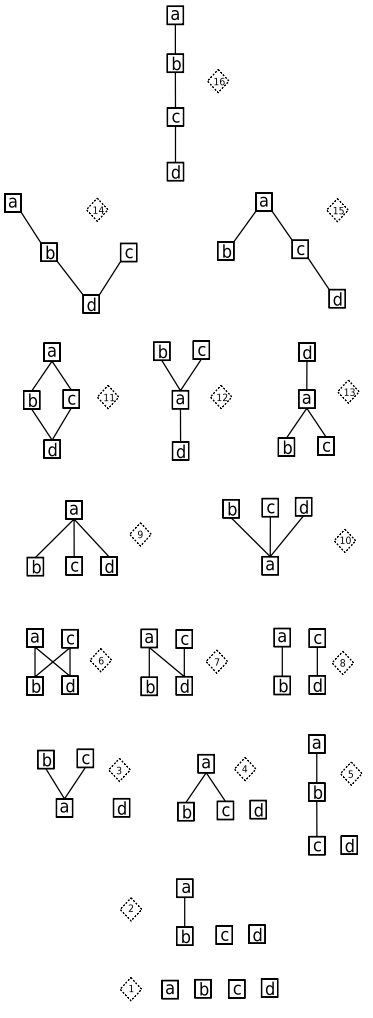


Fig. 2. Orderings of four elements

3 Posets of Elements

In this section we consider changes within various concepts when an object is added or removed and/or when an attribute is added or removed.

1) Suppose the system offers three elements. The following orderings are to be considered: the three elements are ordered consecutively ($G3l$), one of the elements is placed higher than the other two elements and no preference is shown with respect to those two elements ($G31$), one of the elements is placed lower than the other two elements and no preference is shown with respect to these two elements ($G32$), only two of the elements are considered ($G2r$), none of the elements is considered (Gn).

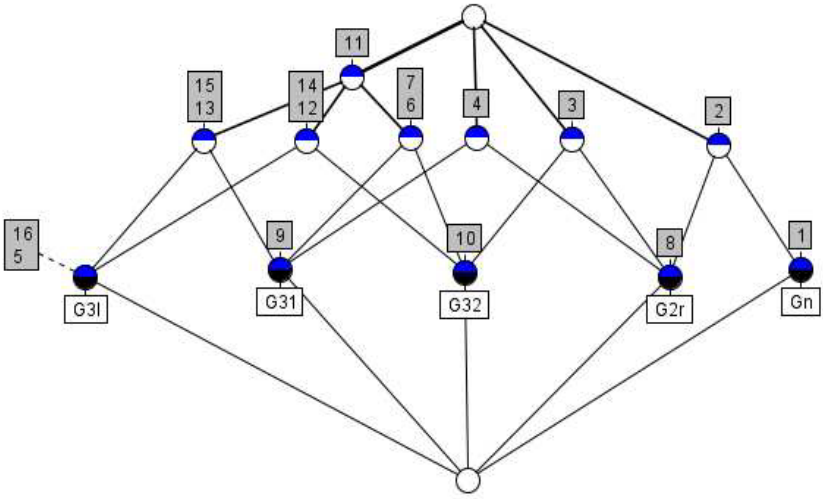


Fig. 3. Concept lattice of the context in Table 1

2) The following orderings are to be considered in the case of four elements: the four elements are not ranked, only two elements are ranked, two couples of elements are ranked, three couples of elements are ranked, four couples of elements are ranked, one element is ranked higher than the other three, three elements are ranked higher than the forth element, three elements are ranked linearly and the forth element is not ranked, three elements are ranked linearly and the forth element is ranked higher than the last element, three elements are ranked linearly and the forth element is ranked below than the top element, one element is compared to the other three elements and another element ranked higher than the rest, one element is compared to the other three elements and two elements are ranked higher than the rest, one element is compared to the other three elements and two elements are ranked higher than the rest, every element is compared to two of the other elements, and all elements are ranked linearly. Representatives of these posets are summarized in Fig. 2.

The objects in Table 1 are the orderings of three elements like $G3l$, $G32$, etc. The attributes in Table 1 illustrate whether a particular ordering of three elements can be extended naturally to an ordering of four elements denoted by $1, 2, ..., 16$.

3.1 Iceberg Concept Lattices

Conceptual hierarchies that are inherent in data can be represented by concept lattices. Iceberg lattices are a conceptual clustering method and are based on the theory of Formal Concept Analysis. They illustrate frequent item sets and visualise association rules, [24].

Table 1. Relationships between an ordered set of 3-elements and an ordered set of 4-elements

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gn	×	×														
G2r			×	×				×								
G31				×		×	×		×		×		×		×	
G32			×			×	×			×	×	×		×		
G3l											×	×	×	×	×	×

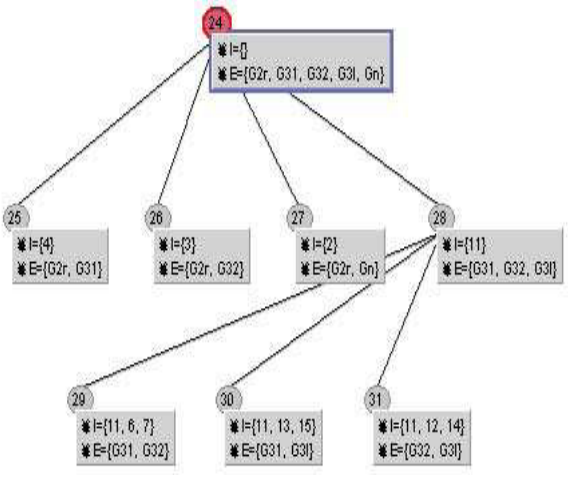


Fig. 4. Iceberg concept lattice

Iceberg lattices are represented below as in [18].

Definition 1. *Let $B \subseteq M$. The support count of the attribute set B in $\mathbb{K}$ is*

$$\sigma(B) = \frac{|B'|}{|G|}$$

Let minsupp be a threshold $\in [0, 1]$, then B is said to be a frequent itemset if $\sigma(B) \geq$ minsupp. A concept is called frequent concept if its intent is frequent.

Definition 2. *The set of all frequent concepts of a context $\mathbb{K}$ is called iceberg lattice of the context $\mathbb{K}$.*

Iceberg Lattices can be used to discover and visualize association rules, [17]. Within a formal context $K = (G, M, I)$, the task of mining association rules is to determine all pairs $X \rightarrow Y$ of M such that $\sigma(X \rightarrow Y) = \sigma(X \cup Y) \geq minsupp$, and the confidence $conf(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$ is above a given threshold minconf $\in [0, 1]$.

For example, the iceberg $\mathcal{L}_1^{0.4}$ obtained from the complete lattice of Fig. 3 with $\alpha \geq 0.4$ is shown in Fig. 4. The number of important concepts in Fig. 4 is reduced to 7.

3.2 Stability

Stability of a formal concept is discussed in [15] and [16].

Definition 3. *Let* (A, B) *a formal concept of* $\mathfrak{B}(G, M, I)$. *Stability of* (A, B) *is*

$$\gamma(A, B) = \frac{|\{C \subseteq A | C' = A' = B\}|}{2^{|A|}}$$

The stability index of a concept indicates how much the concept intent depends on particular objects of the extent. Given a concept (A, B), the stability index measures the number of elements of G that are in the same equivalence class of A, where an equivalence class is defined as follows.

Definition 4. *Let* $X \subseteq G$, *we denote by* X *the equivalence class of* X *where* $\langle X \rangle = \{Y \subseteq G | Y' = X'\}$.

If X is closed then $Y \subset X, \forall \ \ Y \in X$, and $\gamma(A, B) = \frac{|\langle A \rangle|}{2^{|A|}}$.

Concepts' stabilities from the complete lattice in Fig. 3 can be seen in Fig. 5.

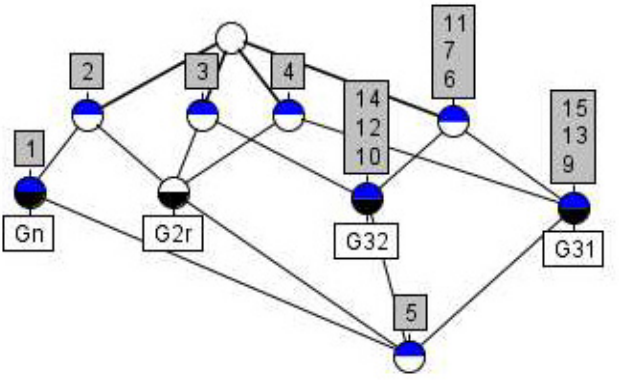


Fig. 5. A lattice without attribute 8

3.3 Cognition

A cognitive model represents the knowledge that an ideal student would possess about a particular subject. Developing a cognitive tutor involves creating a cognitive model of students' problem solving by writing production rules that characterize the variety of strategies and misconceptions students may acquire, [19].

Taxonomy for automated hinting is developed in [25]. The role of hints in a Web based learning systems is considered in [5]. A model for detecting students' misuse of help in intelligent tutoring systems is presented in [3]. A proliferation of

hint abuse (e.g., using hints to find answers rather than trying to understand) was found in [2]. However, evidence that when used appropriately, on-demand help can have a positive impact on learning was found in [23]. Various instructional explanations are discussed in [21].

The approach presented in [6] can be applied for better understanding of the factors that are common to final examination performance. Such factors include student effort level, level of student motivation, study approach and personality, [8], and [11].

4 Conclusion

Application of iceberg lattices in the process of knowledge assessment can considerably improve the process of selecting the most important information without inclusion of unnecessary details. It is a step forward in the process of extracting the minimal amount of items indicating existence of a predefined set of skills. This can considerably minimize the amount of manpower required in the process of creating and evaluating items' sets to be enclosed in automated tests. Another advantage is related to students' learning. Well chosen test items can contribute for avoiding the problem of over testing and thus discouraging students from active participation in the learning process.

References

1. Albert, D., Stefanutti, L.: Ordering and Combining Distributed Learning Objects through Skill Maps and Asset Structures. In: Proceedings of the ICCE 2003, pp. 1–9 (2003)
2. Aleven, V., Koedinger, K.R.: Limitations of Student Control: Do Student Know when they need help? In: Gauthier, G., Frasson, C., VanLehn, K. (eds.) ITS 2000. LNCS, vol. 1839, pp. 292–303. Springer, Heidelberg (2000)
3. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 531–540. Springer, Heidelberg (2004)
4. Bogart, K.P.: Some social sciences applications of ordered sets. In: Rival, I. (ed.) Ordered Sets, pp. 759–787. Reidel, Dordrecht (1982)
5. Brunstein, A., Krems, J.: Helps and Hints for Learning with Web Based Learning Systems: The Role of Instructions. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 794–796. Springer, Heidelberg (2004)
6. Chamorro-Premuzic, T., Furnham, A.: Personality predicts academic performance: Evidence from two longitudinal university samples. Journal of Research in Personality 37, 319–338 (2003)
7. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order. Cambridge University Press, Cambridge (2005)
8. Davidson, R.A.: Relationship of study approach and exam performance. Journal of Accounting Education 20, 29–44 (2002)
9. Doignon, J.-P., Falmagne, J.-C.: Spaces for the assessment of knowledge. International Journal of Man-Machine Studies 23, 175–196 (1985)
10. Doignon, J.-P., Falmagne, J.-C.: Knowledge Spaces. Springer, Heidelberg (1999)

11. Furnham, A., Chamorro-Premuzic, T.: Personality and intelligence as predictors of statistics examination results. Personality and Individual Differences 37, 943–955 (2004)
12. Ganter, B., Wille, R.: Formal Concept Analysis. Springer, Heidelberg (1999)
13. Godin, R., Mili, H.: Building and Maintaining Analysis-Level Class Hierarchies Using Galois Lattices. In: Proc. OOPSLA 1993, Washington, DC, USA (1993); Special issue of Sigplan Notice 28(10), 394–410 (1993)
14. Hockemeyer, C., Conlan, O., Wade, V., Albert, D.: Applying competence prerequisite structures for e-Learning and skill management. Journal of Universal Computer Science 9, 1428–1436 (2003)
15. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence 49, 101–115 (2007)
16. Kuznetsov, S., Obiedkov, S., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In: Priss, U., Polovina, S., Hill, R. (eds.) ICCS 2007. LNCS (LNAI), vol. 4604, pp. 241–254. Springer, Heidelberg (2007)
17. Nehme, K., Valtchev, P., Hacene, M.R., Godin, R.: On Computing the Minimal Generator Family for Concept Lattices and Icebergs. In: Ganter, B., Godin, R. (eds.) ICFCA 2005. LNCS (LNAI), vol. 3403, pp. 192–207. Springer, Heidelberg (2005)
18. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 398–416. Springer, Heidelberg (1999)
19. Pecheanu, E., Segal, C., Stefanescu, D.: Content modeling in Intelligent Instructional Environment. In: Luo, Y. (ed.) CDVE 2004. LNCS (LNAI), vol. 3190, pp. 1229–1234. Springer, Heidelberg (2003)
20. Rusch, A., Wille, R.: Knowledge Spaces and Formal Concept Analysis. In: Bock, H.H., Polasek, W. (eds.) Data Analysis and Information Systems. Statistical and Conceptual Approaches, pp. 427–436. Springer, Berlin (1996)
21. Schworm, S., Renkl, A.: Learning by solved example problems: Instructional explanations reduce self-explanation activity. In: Gray, W.D., Schunn, C.D. (eds.) Proceeding of the 24th Annual Conference of the Cognitive Science Society, pp. 816–821. Erlbaum, Mahwah (2002)
22. Stefanutti, L., Albert, D., Hockemeyer, C.: Derivation of Knowledge Structures for Distributed Learning Objects. In: Ritrovato, P., Allison, C., Cerri, S.A., Dimitrakos, T., Gaeta, M., Salerno, S. (eds.) Towards the Learning Grid: Advances in Human Learning Services. Frontiers in Artificial Intelligence and Applications, vol. 127, pp. 105–112. IOS Press, Amsterdam (2005)
23. Stewart, K.L., Felicetti, L.A.: Learning styles of marketing majors. Educational Research Quarterly 15, 15–23 (1992)
24. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing Iceberg Concept Lattices with Titanic. Data and Knowledge Engineering 42(2), 189–222 (2002)
25. Tsovaltzi, D., Fiedler, A., Horacek, H.: A Multi-dimensional Taxonomy for Automating Hinting. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 772–781. Springer, Heidelberg (2004)
26. Wille, R.: Concept lattices and conceptual knowledge systems. Computers Math. Applications 23(6-9), 493–515 (1992)

An Ontology-Supported Ubiquitous Interface Agent for Cloud Computing - Example on Zigbee Technique

Sheng-Yuan Yang[1], Dong-Liang Lee[2], and Chun-Liang Hsu[3]

[1] Department of Computer and Communication Engineering, St. John's University, Taiwan
[2] Department of Information Management, St. John's University, Taiwan
[3] Department of Electrical Engineering, St. John's University, Taiwan
{ysy,lianglee,liang}@mail.sju.edu.tw

Abstract. An ontology-supported ubiquitous interface agent and its interaction diagram with the backend information agent system, i.e., OntoIAS, in cloud computing environments were proposed. The agent employs the CURRL to transform user commands into internal canonical format to conveniently process those commands by OntoIAS, which can avoid numerous, jumbled, and incorrect information torrents that results in misunderstanding of the information intention of users. In this paper, we preliminarily proposed the agent with the Zigbee techniques and related interaction diagrams with OntoIAS in cloud computing environments. The system prototype and experimental outcomes can also reveal the feasibility of the system architecture.

Keywords: Ubiquitous Interface Agent, Ontology, Zigbee, Cloud Computing.

1 Introduction

With increasing popularity of computers, network techniques, and WWW, information shows the multiple appearances and huge amount explosion. Therefore, the way of helping the user to quickly, precisely, and effectively get profoundly, relevantly, and up-to-dated useful information has quickly become the critical topic that the industrial, official government, and academic groups strived for in last ten years. A variety of information retrieval tools has been thus created by information providers, including information portals, search engines, etc., which could help users to filter, search for, organize, and represent related query information. Information agents are software products for assisting and guiding users to reach the goal of information retrieval. Not only can the agent possess the four main functions: information searching, information extracting, information classifying, and information representing/ranking, but also it can really and effectively up-rise the performance of information query to the user and collocate the factors in user interfaces, network speed, amount of the backend databases, and usage scenarios. Up to now, however, most of Web information agent systems are closely related to the traditional information equipments that can not directly apply to the modern mobile equipments resulting from the core part of information agent in ubiquitous environments. This study exactly focused on how to construct a ubiquitous interface agent with mobile equipments in ubiquitous environments.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 142–153, 2010.

Cloud computing is a technique of Internet- ("cloud-") based development and use of computer technology. In other words, it will set up the necessary operating resources and related data into the Internet and then users can directly use them whenever they can access the Internet. Ubiquitous computing is a post-desktop model of human-computer interaction in which information processing has been thoroughly integrated into everyday objects and activities. Computers will exist in our lives in hidden, popularized, and ubiquitous ways. Even though many people already know we have entered a ubiquitous environment, this kind of natural interaction mode has not appeared yet. Nevertheless, there is lots of traditional equipment that is suitable for these natural interaction modes, for example, mobile phones, RFID (Radio Frequency Identification), Bluetooth, Zigbee, etc. Such devices operate through wireless sensor techniques to recognize related users and let users naturally interact with relevant network services, thus, reaching the goal of ubiquitous computing. Furthermore, how to construct an interaction diagram of cloud computing for extensively and seamlessly entering related web information agent systems through modern mobile equipments in ubiquitous environments is our major investigation.

To sum up, this study focused on designing a ubiquitous interface agent based on the ontology technology and interaction diagram with the backend information agent system, i.e., OntoIAS (Ontology-supported Information Agent Shell), in cloud computing environments. The agent employs the CURRL (Canonical User Request Representation Language) to transform user commands into internal canonical format to conveniently process those commands by OntoIAS, which can avoid numerous, jumbled, and incorrect information torrents that result in misunderstanding of the information intention of users. The system creates an interaction diagram that both solves the congenital defect problems of mobile equipments and adequately elaborates the powerful functions of backend information systems. In this paper, we preliminarily proposed a ubiquitous interface agent with the Zigbee techniques and related interaction diagrams with OntoIAS in cloud computing environments. The system prototype and experimental outcomes can also reveal the feasibility of the system architecture.

2 Background Knowledge and Techniques

2.1 Ontology

Ontology was one theory in philosophy and primarily to explore knowledge features of life and real objects, which can provide complete semantic models with sharing and reusing substances. To use the concept of ontology can accomplish the knowledge core in a specified domain and automatically learn related information, communication, accessing and even induce new knowledge; hence, ontology is a powerful tool to construct and maintain an information system [15]. Fig. 1 illustrates the ontology structure of Scholars, which defines related basic knowledge of scholars and its conceptual hierarchy relationship and relevant features.

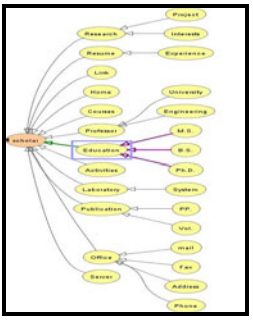


Fig. 1. The ontology structure of Scholars

2.2 Ubiquitous Computing

Mark Weiser created the phrase "ubiquitous computing" around 1988, during his tenure as Chief Technologist of the Xerox Palo Alto Research Centre. This concept pointed out that the third-wave revolution of computers had already come. Computers will exist in our lives in hidden, popularized, and ubiquitous ways. The human-machine interaction modes, which operate regardless of Command-Line, Menu-Driven, or even GUI-based ones, are improper and unsuitable for ubiquitous environments. This is a clear suggestion of the necessity of more natural human-machine interaction modes to support ubiquitous computing. There are many examples of applications in previous studies [2,7,8,9,11]. In summary, the scopes of discussion related to ubiquitous computing include Pervasive Computing, Sentient Computing, Simulated Reality, Wearable Computers, Context-Aware Pervasive Systems, Ambient Intelligence, Virtual Reality, Human-Centred Computing, etc. Relevant research contents include ubiquitous software and hardware infrastructures, protocols, components, access security, etc. There is a lack of ubiquitous research in software system applications in Taiwan, especially. The only study of its kind involving ubiquitous research on agent systems was the 2008 project that explored a ubiquitous service system with an embedded intelligent interface by Director Dr. C.C. Hsu, dept. of computer science and information engineering, Fu Jen Catholic University, Taiwan. This was an influential and significant study related to web information systems.

2.3 Cloud Computing

In concept, cloud computing is an information technology that makes users utilize the information services when they can only access the Internet, and even cannot completely understand the complex information service structure and possess any professional knowledge. Cloud computing earlier borrowed from the techniques of Grid Computing and Utility Computing in the early 1990s. In the 21st century, the related network services vigorously develop based on the improvement of network techniques. In 2007, Google proposed the concept of cloud computing that also start the huge business opportunity of cloud computing, including IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service). It still more achieves the concept of new 3C, i.e., Cloud Computing, Connecting, and Client Devices [4].

In this paper, we exactly investigated a ubiquitous interface agent with the Zigbee technique and related interaction diagrams with OntoIAS in cloud computing environments. That means, in cloud computing environment, the ubiquitous interface agent is responsible for the role of client device; the Zigbee technique is responsible for communication mechanism; finally, the backend information agent system, OntoIAS is responsible for the role of the provider of cloud computing. Furthermore, this study can reach the investigation goal of constructing an interaction diagram of cloud computing for extensively and seamlessly entering related web information agent systems through modern mobile equipments.

2.4 Zigbee Technique with C/C++

Zigbee explored related techniques of disorderly and unsystematic information transmission that is similar to the behavior of honey bees after return to the beehive. The Zigbee specifications of hardware and software were announced and completed by the Zigbee Alliance and the IEEE 802.15.4 standard, its protocol stack as shown in Fig. 2. It was planned to become a low-speed (250kbps), short-distance, low-power, and simple-architecture wireless mesh networking technology. Currently, ZigBee operates in the industrial, scientific and medical (ISM) radio bands; 868 MHz in Europe, 915 MHz in the USA and Australia, and 2.4 GHz in most worldwide. In the 2.4 GHz band there are 16 ZigBee channels, while the channel quantities are 10 and 1 in the 915MHz and 868MHz band, respectively. Zigbee supports the client-server and point-to-point modes and has the higher extensibility, which can simultaneously have over 65,000 device connections in a network. The main applications focus on the information transmission of home appliances automation, environment security and control, individually medical treatment, etc. [6].

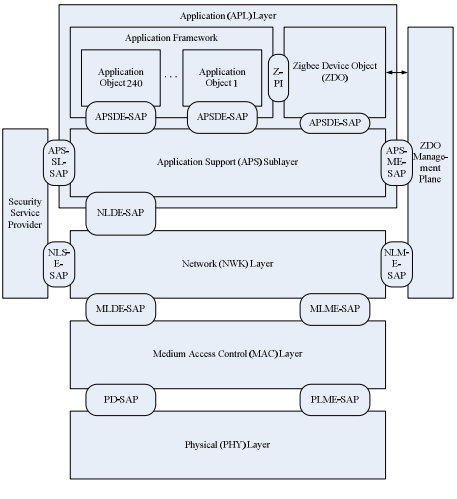


Fig. 2. Zigbee protocol stack

C/C++ is a general-purpose computer language. It is usually and popularly used to develop application software and system exploitation, which possesses characteristics of high efficiency, high flexibility, abundant functions, strong expression, high transplantation and compatibility, etc. Currently, the compiler of C/C++ universally exists in variously different operating systems, such as Microsoft Windows, Linux, UNIX, etc. [12]. In this paper, we employ the JN5121 Zigbee module of Jennic Ltd., which uses the C/C++ to develop its applications of ROM in the module and Jennic names the program Codeblocks [1]. Its main function is the code editing and compiling, and then employs the Flash Programmer to burn the compiled codes into the JN5121 module. Owing to the Codeblocks and Flash Programmer are developed by Jennic Ltd. according to the JN5121 module. The system developers must install their official drivers for application development.

2.5 Developing Techniques

This system adapted MS SQL Server as backend knowledge-database sharing platform based on ontology. MS SQL Server is one broadly used relational database management system [14]. SQL (stands for Structured Query Language) is one query language to get the data in the relational database. The agent system itself was developed with Java SE and ME, and the Zigbee technology with C/C++ mentioned above. The ontology construction tool, Protégé, was an ontology freeware developed by SMI (Stanford Medical Informatics). Protégé not only was one of the most important platforms to construct ontology but also the most frequently adapted one [3]. Its most special feature is that it used multi components to edit and make ontology and led knowledge workers to construct knowledge management system based on ontology; furthermore, users could transfer to different formats of ontology such as RDF(S), OWL, XML or directly inherit into database just like MySQL and MS SQL Server, which have better supported function than other tools [16].

3 Architecture of the Ubiquitous Interface Agent and Interaction Diagram with OntoIAS

To reach ubiquitous research goals of this study, users can employ the Ubiquitous Interface Agent to use the backend information agent system: OntoIAS via related mobile equipments, such as mobile phones, PDA-related equipments, lap-top computers, related equipments with Zigbee interfaces, or another related information systems that fit in with communication protocols. Therefore, the agent has to provide the communication bridge between the mobile and wireless equipments and related web information systems, as shown in Fig. 3. This must be done to satisfy the basic requirements of seamless information services in ubiquitous computing, whose related interaction diagrams contain the following actions: users key in specific information requirements to trigger OntoIAS to return query information, users directly query OntoIAS to provide commonly used hot information, etc.

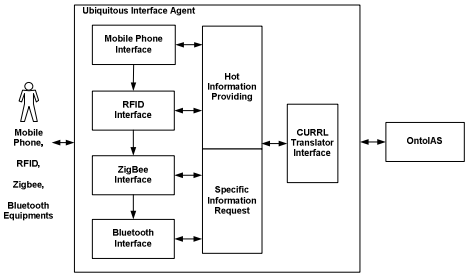


Fig. 3. System architecture of the Ubiquitous Interface Agent

In general, mobile equipments possess small screens, small keyboards, and poor system efficiency. Therefore, the system employs simple codes or hot keys to describe common or important information requirements for dealing with complex user commands. Through the techniques of packet decoding and recognizing, the Ubiquitous Interface Agent employs CURRL to transform user commands into internal canonical format to conveniently process those commands by OntoIAS, which can avoid numerous, jumbled, and incorrect information torrents that result in misunderstanding of the information intention of users.

When users key in specific information queries, the agent divides the user queries into three types of commands for fast processing. The three types include Query, Simple Command, and Conditional Command. We modified FURRL (Formalized User Request Representation Language) [5] to design a CURRL to represent the above user commands. The CURRL is a frame-based command representation that makes it easy to map users' command intentions, objects, and goals into corresponding frame slots. Table 1 illustrates some examples of user commands.

Table 1. Examples of user commands with CURRL

User command	User command	CURRL
Query	What is Ubiquitous Computing?	Query [Theme = + Computing, aTheme = Ubiquitous, tSpace = At (WWW)]
Simple Command	Anything else?	Command [Theme = +Anymore, tTime = Now, object = Related, oSpace = At (Last-one)
Conditional Command	Retrieving Fuzzy webpages with the exception of AI	ConditionalCommand [Condition [Theme = -Fuzzy, tTime = Now, tSpace = At (WWW)], Command [Theme = +AI, tTime = Now, tSPace = At (WWW)]]

In short, we simplify the design of the agent into a data decoding controller of related communication equipment; and then employ CURRL to transform them into an internal canonical format; finally, we trigger OntoIAS to provide information solutions. The interaction diagram can not only solve the congenital defects of the

mobile equipment mentioned above, but can also adequately elaborate the powerful functions of the backend system, OntoIAS. The application environment of the system uses related mobile equipment and constructs them with the architecture of the World Wide Web. The operating mode adopts a three-tier architecture, including the user end, the ubiquitous interface Agent end, and the OntoIAS end. The interaction system prototype that includes the three-tier multi-agent architecture can solve the congenital defect problems of mobile equipment encounter in satisfying users' basic information requirements in ubiquitous environments. Detailed explanations of the interaction diagrams are provided as follows:

(1) The user keys in specific information queries to trigger OntoIAS to return information solutions: after the user enters his/her account, the Ubiquitous Interface Agent provides the interface for entering the information query, and then employs CURRL to transform the query into the internal query format, finally, triggering OntoIAS to process the query and return its solutions. After the user finishes browsing, the Ubiquitous Interface Agent returns relevant feedback to OntoRecommander to act as the calculation base of hot information and records them in the user's profile;
(2) The user directly queries OntoIAS to provide commonly used hot information: after the user enters his/her account, Ubiquitous Interface Agent directly triggers the OntoRecommander of OntoIAS, according to the user's account, and returns commonly used hot information to the user. That is, the user directly uses the information but has to do nothing.

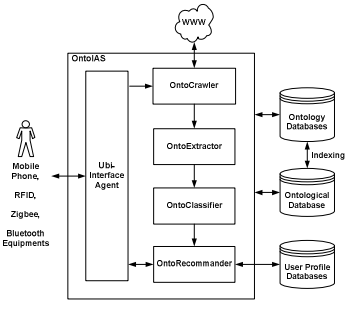


Fig. 4. Conceptualized architecture of the backend information agent system: OntoIAS

Fig. 4 illustrates the OntoIAS architecture diagram [17]. It contains the four main modules of information agents, including information searching, information extracting, information classifying, and information presenting/ranking, corresponding to OntoCrawler, OntoExtractor, OntoClassifier, and OntoRecommander, respectively. An Ontological Database (OD) is a stored structure designed according to the ontology structure, serving as an ontology-directed canonical format for storing webpage information processed by OntoIAS. Users can employ the ubiquitous interface agent to use the OntoIAS via related mobile equipments, or other related information systems fitted with communication protocols. Therefore, the proposed method can reach the

goals of ubiquitous research. User profile databases are responsible for recording relevant user models. The system can trigger OntoRecommander to provide relatively personal information services.

4 System Display and Evaluations

4.1 System Prototype

Our system is developed using C++Builder on Windows XP® Service Pack 3 of Professional Version with Intel® Core 2 Duo CPU at 2.53GHz and 2GB memory, and the Zigbee device with JN5121 (Jennic Ltd.) chip. To use Zigbee devices with other chips such as TI-CC2431 (Texas Instruments), etc., the system must install their official drivers for normal action.

The system employed Java ME and SE to develop the Ubiquitous Interface to simulate the communication between cellular phones and the computer end. Because of the communication of Zigbee JN5121 module has to use the C/C++ language. For this reason, we need to develop related middle-wares; that is to go through the serial port transferred circuit via the RS232 port for connecting to Zigbee JN5121 module, as shown in Fig. 5.

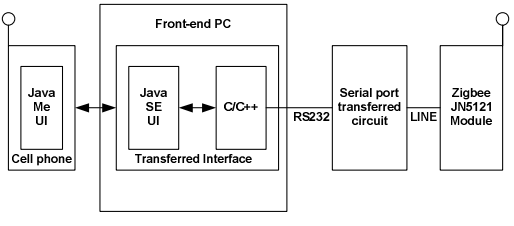


Fig. 5. Front-end of the ubiquitous interface agent

By way of the front-end Zigbee JN5121 module to deliver data, the system prototype employs the back-end Zigbee JN5121 module to receive data, and then goes through the reversed operation mentioned above to connect to the back-end server system for carrying out related processes, as shown in Fig. 6. After back-end system processing, the system prototype orderly returns and presents the results into the cell-phone simulator.

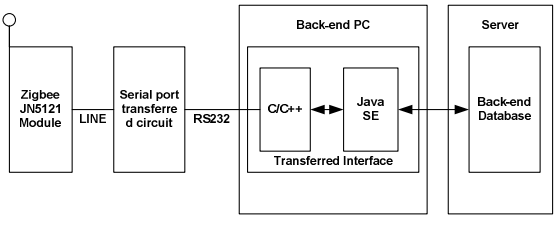


Fig. 6. Back-end system of the ubiquitous interface agent

Fig. 7. Screen of client device end

The execution steps of the ubiquitous interface agent system prototype with the Zigbee technique and related interaction diagrams in cloud computing environments are explained with the ASUS cell phone P552w and detailed follows:

(1) Client device end: enters the query keywords in the cell phone and assigns related search engines, the testing screen as shown in Fig. 7;
(2) Connecting technology: starts the Zigbee transmitted program as shown in Table 2 and sends related query information from step 1 to the cloud computing provider OntoIAS;
(3) Cloud computing end: OntoIAS starts the received program as shown in Table 3, then executes related query processes, and finally communicate the query results to the cell phone of the client through a series of Request-Response manner in the Client-Server mode, as shown in Fig. 8.

Table 2. Processing procedure and related function description of the transmitted end

Action Description	Related Functions
Local Device	JZS_vStartStack
Devices Discovered	JZS_vStartNetwork
Services Discovered	JZS_vJoinNetwork
Client Session	JZA_vStackEvent

Table 3. Processing procedure and related function descriptions of the received end

Action Description	Related Functions
Local Device	JZS_vStartStack
Devices Discovered	JZS_DiscoverNetworks
Server Connection	JZS_vPollParaent
Server Request Handler	JZA_u8AfMsgObject

Fig. 8. Received screen

4.2 Performance Evaluation

The information recommendation means that the optimal recommendation is chosen from a group of related information sets, just like the concept of sampling. In the sampling survey domain, the reliability was generally employed to measure the degree of precision of the sampling system itself, while the validity emphasized whether it could correctly reflect the properties of the appearance of things or not. In this study, we employed the aid of a mathematic model, provided by J.P. Peter [10] in 1979, and cited by numerous other studies to represent the definitions of reliability and validity [17].

Table 5. Results of the 3 recommendations

No	Information Recommendation	Fuzzy Theory		Artificial Neural Network		Artificial Intelligence	
		r_{t1}	V_{a1}	r_{t1}	V_{a1}	r_{t1}	V_{a1}
1	Course Information	0.93	0.92	0.88	0.86	0.96	0.75
	Academic Activities	0.94	0.93	0.81	0.76	0.98	0.93
	Website Recommendation	0.91	0.83	0.93	0.78	0.97	0.88
Average		**0.93**	**0.89**	**0.87**	**0.8**	**0.97**	**0.85**
2	Course Information	0.94	0.9	0.86	0.65	0.83	0.85
	Academic Activities	0.92	0.78	0.78	0.58	0.94	0.88
	Website Recommendation	0.92	0.88	0.91	0.89	0.98	0.96
Average		**0.93**	**0.85**	**0.85**	**0.71**	**0.92**	**0.9**
3	Course Information	0.78	0.63	0.88	0.68	0.99	0.89
	Academic Activities	0.94	0.88	0.79	0.72	0.96	0.92
	Website Recommendation	0.96	0.93	0.88	0.78	0.97	0.9
Average		**0.89**	**0.81**	**0.85**	**0.73**	**0.97**	**0.9**

Table 5 illustrates the reliabilities and validities of information recommendation in different professional domains, while the total average results are shown in Table 6. The average values of reliability and validity were 0.91 and 0.83, respectively. In this experiment, we randomly chose 100 data from the personal webpages of the members of the Taiwanese Association for Artificial Intelligence to carry out different 3 separate recommending experiments. The significant information recommendation of these experiments were asserted by the domain experts, including observed values, true values, error values, and related variances. From previous technical studies [13], we know that the regular-level values of reliability and validity are 0.7 and 0.5, respectively, which verifies and validates that our experiment results have high-level outcomes of information recommendation of the proposed system.

Table 6. Total average results

Performance	Fuzzy Theory	Artificial Neural Network	Artificial Intelligence	Total Average
Average Reliability	0.92	0.86	0.95	**0.91**
Average Validity	0.85	0.75	0.88	**0.83**

5 Conclusions

In this paper, an ontology-supported ubiquitous interface agent and interaction diagram with the backend information agent system in cloud computing environments were proposed. The agent adopts the CURRL to fast and precisely deal with user query commands for conveniently processing those commands by OntoIAS. The

system also creates an interaction diagram that both solves the congenital defect problems of mobile equipments and adequately elaborates the powerful functions of backend information systems. In this paper, we preliminarily proposed a ubiquitous interface agent with the Zigbee technique and related interaction diagrams with OntoIAS in cloud computing environments. The system prototype and experiment outcomes can not only reveal the feasibility of the system architecture, but also have high-level outcomes of information recommendation. Continuously improving the performance efficiency, expanding database of ontology and its related linking interface, and developing the middle programs, for example, RFID, Bluetooth, etc., with backend systems for truly completing the wireless communication functions of the ubiquitous interface agent would be the everlasting research in the future.

Acknowledgement

The authors would like to thank Ssu-Hsien Lu, Ting-An Chen, Chi-Feng Wu, and Zhe-Min Ni for their assistance in system implementation and experiments. This partial work was supported by the National Science Council, R.O.C., under Grants NSC-99-2221-E-129-012 and NSC-99-2623-E-129-002-ET, and the Ministry of Education, Taiwan, R.O.C., under Grant Skill of Taiwan (1) Word No. 0990045921s.

References

1. Chang, H.L., Wang, S.K.: Jennic Software Developing Guide for Jennic JN51XX. Beijing Boccn Tech Co., Ltd., Beijing (2008)
2. Chang, Y.C.: Clandestine Service Discovery Protocols for Pervasive Computing. Master Thesis, Dept. of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan (2008)
3. Grosso, W.E., Eriksson, H., Fergerson, R.W., Gennari, J.H., Tu, S.W., Musen, M.A.: Knowledge Modeling at the Millennium: the Design and Evolution of Protege-2000. SMI Technical Report, SMI-1999-0801, Stanford University, NY, USA (1999)
4. Gu, S.R.: Cloud Computing Robs Their Turfs, A New Warring Age of IT Industry. Common Wealth Magazine 423, 178–181 (2009)
5. Kim, J., Jang, M., Sohn, J.C.: An Ontological Approach for Natural Language Command Interpretation and Its Application in Robotics. In: Proc. of International Joint Conference on SICE-ICASE, Busan, Korea, pp. 3874–3878 (2006)
6. Lee, J.S.: Compliant Testing and Certification for ZigBee Protocol Stacks and Application Profiles. ICL Technical Journal, ITRI, Hsinchu, Taiwan 123, 98–107 (2008)
7. Li, W.Y.: An example of Java Programming in a Semantic Grid-Based Ubiquitous Learning Environment. Master Thesis, Dept. of Management Information Systems, Chun Yuan Christian University, Taoyuan, Taiwan (2007)
8. Liang, B.S.: Ubiquitous Messaging and Presence Service. Master Thesis, Dept. of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan (2007)
9. Partridge, K., Golle, P.: On Using Existing Time-Use Study Data for Ubiquitous Computing Applications. In: Proc. of Tenth International Conference on Ubiquitous Computing, Seoul, South Korea (2008)

10. Peter, J.P.: Reliability: A review of psychometric basics and recent marketing practices. Journal of Marketing Research 16, 6–17 (1979)
11. Smetters, D., Balfanz, D., Durfee, G., Smith, T., Lee, K.H.: Instant Matchmaking: Simple, Secure Virtual Extensions to Ubiquitous Computing Environments. In: Proc. of Eighth International Conference on Ubiquitous Computing, Orange County, California, USA (2006)
12. Wikipedia, C Language, http://zh.wikipedia.org/zh-tw/C%E8%AA%9E%E8%A8%80 (visited on March 12, 2010)
13. Wu, T.X.: The Reliability and Validity of Attitude and Behavior Research: Theory, Application, and Self-examination, pp. 29–53. Public Opinion Monthly, Taiwan (1985)
14. Yang, S.Y.: Developing of an Ontological Interface Agent with Template-based Linguistic Processing Technique for FAQ Services. Expert Systems with Applications 36(2), 4049–4060 (2009)
15. Yang, S.Y., Hsu, C.L.: An Ontological Proxy Agent with Prediction, CBR, and RBR Techniques for Fast Query Processing. Expert Systems with Applications 36(5), 9358–9370 (2009)
16. Yang, S.Y.: OntoPortal: An Ontology-Supported Portal Architecture with Linguistically Enhanced and Focused Crawler Technologies. Expert Systems with Applications 36(6), 10148–10157 (2009)
17. Yang, S.Y.: Research and development of an ontology-supported information agent shell for ubiquitous services. In: Proc. of the Fourteen Conference on Artificial Intelligence and Applications, Taichung, Taiwan, p. 15 (2009)

Semantic Optimization of Query Transformation in Semantic Peer-to-Peer Networks

Jason J. Jung

Knowledge Engineering Laboratory
Department of Computer Engineering
Yeungnam University
Gyeongsan, Korea 712-749
j2jung@{ynu.ac.kr,intelligent.pe.kr}

Abstract. Ontologies have been an important role of supporting efficient interoperability among information systems in distributed environment. In this paper, we propose a query transformation method to efficiently collect as many relevant resources from the distributed information systems as possible. More importantly, the query from the source peer can be decomposed and propagated as maintaining the original contexts. Through the experiments, we have shown that query-activated concept (QAC)-based schemes have fulfilled an efficient query decomposition process.

Keywords: Collective intelligence; Tag matching; Multilingual tagging; Semantic grounding; Social tagging; Folksonomy.

1 Introduction

Mapping between heterogenous ontologies has been an important challenge on supporting efficient interoperability between distributed information systems. Once mappings between the two ontologies have been found, a pair of the corresponding systems in the distributed environments (e.g., a semantic social network [1, 2] or a semantic peer-to-peer network [3]) can exchange relevant information between them. However, ontology mapping by human experts is usually an expensive (i.e., time-consuming) process. (Of cause, ontology matchers (i.e., software tools) can do it automatically but their precision is relatively low [4].) Additional problem is low scalability. As the number of ontologies in a distributed network increases, the cost of ontology mapping for interoperability might exponentially increase.

In order to deal with the scalability problem with a large number of ontologies, in the previous studies [5–8], we have been focusing on sharing and composing the ontology mappings which already exist. In this work, such distributed environment is referred to a mapping network N_{MAP}. For example, as shown in Fig. 1, a mapping network is composed of 7 peer systems (i.e., S_1 to S_7). Only solid arrows indicate the manual (or semi-automatic) mappings between corresponding ontologies Hence, mapping between source peer S_1 and destination peer S_4 can be composed by combining two mappings i) between S_1 and S_3 and ii) between S_3 and S_4.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 154–162, 2010.

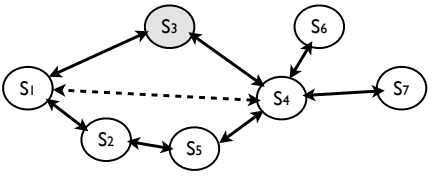


Fig. 1. Example of ontology mapping composition in a mapping network

There can be a number of ways of interoperating between the systems. In this work, we are simply interested in a query-answer model [9]. A query from a source peer can be transformed for making a destination peer more understandable by referring to the mapping between two corresponding ontologies [10–12]. Through the empirical study on query transformation in a distributed information retrieval system [8, 13], we have already found the following implications;

Precision of mapping. The existing mappings should be precise so as to guarantee the precision of the composed mappings.

Peer distance on a mapping network. The peer distance between two peers indicates the number of composition of the existing mappings. As this distance between them is longer, information loss of their interoperability might exponentially increases (i.e., performance, e.g., precision, of their interoperability decreases).

However, we find that the mapping composition in the distributed network has to be reconfigured for better performance, depending on not only the previous two factors (i.e., precision of mapping and peer distance) but also context of a query given from a source peer. Thereby, in this study, we focus on making the context information of the query sustainable until the destination even though the query is transformed and distributed. We refer to this process as *semantic optimization* on query transformation. Three issues have arisen for optimizing the interoperability between peers, as follows.

Multiple paths between two peers. There can be more than one path between a source and a destination. In Fig. 1, an indirect mapping between S_1 and S_4 can be done with two possible compositions; i) via S_3 and ii) via S_2 and S_5. Thus, the best path should be selected for semantic optimization.

Context of a query from a source peer. A context can be decomposed into several sub-contexts [14]. Thus, it means that a query from a source can be divided into several sub-queries, and each of these sub-queries should be transformed according to the best mapping composition paths, respectively. For instance, two queries q' and q'' can be generated from a query q from S_1, and they are transformed by difference compositions.

Semantic centrality. Given a mapping network, a centrality of each peer can be measured by considering how the mappings between peer ontologies exist. For example, if a peer (i.e., S_4) has more direct mappings compared to the others, then the peer might be playing an important role of bridging other peers.

The outline of this paper is as follows. In Sect. 2, we explain formal notations and the problem that we want to deal with. Sect. 3 and Sect. 4 address how to divide a query,

and how to choose the best composition path from the multiple paths, by referring to the ontology mapping patterns. In Sect. 5, we show an example of semantic optimization, and experimental results. Sect. 6 discusses some issues of the proposed approach, and compares this study with the existing approaches. Finally, in Sect. 7, we draw the conclusion and mention future work.

2 Formal Notations and Problem Description

In this work, the ontologies in a distributed network and the mapping between the ontologies are formalized, as follows;

Definition 1 (Peer ontology [6]). *A peer ontology is defined as* $\mathcal{O} := \langle \mathcal{C}, \mathcal{R}, \mathcal{E}_{\mathcal{R}}, \mathcal{I}_{\mathcal{C}} \rangle$ *where* $\mathcal{C}$ *and* $\mathcal{R}$ *are a set of classes (or concepts), a set of relations (e.g., equivalence, subsumption, disjunction, etc), respectively.* $\mathcal{E}_{\mathcal{R}} \subseteq \mathcal{C} \times \mathcal{C}$ *is a set of relationships between classes, represented as a set of triples* $\{\langle c_i, r, c_j \rangle | c_i, c_j \in \mathcal{C}, r \in \mathcal{R}\}$. $\mathcal{I}_{\mathcal{C}}$ *is a power set of instance sets of a class* $c_i \in \mathcal{C}$.

Definition 2 (Mapping [6, 8, 13]). *Given two ontologies* $\mathcal{O}_i$ *and* $\mathcal{O}_j$*, a mapping* $\mathcal{M}_{ij}$ *between them is represented as*

$$\mathcal{M}_{ij} = \{\langle e, e', r, CF \rangle | e \in \mathcal{O}_i, e' \in \mathcal{O}_j, r \in \mathcal{R}, CF \in [0, 1]\} \tag{1}$$

where e *and* e' *are a pair of ontology elements in the two ontologies, respectively. In addition,* r *is one of the semantic relationships (i.e., equivalence, subsumption, disjoint, and so on), and* CF *indicates a confidence value of each correspondence in the mapping.*

In this work, the context of the source peer is implicitly contained in the query. Thus, we assume that each peer can generate the query by deriving a set of concepts from its own peer ontology.

Definition 3 (Concept-based query [6]). *For simplicity, a query is represented as*

$$q ::= c | \neg q | q \vee q' | q \wedge q' \tag{2}$$

where c *is a concept* $\mathcal{C}$ *in a peer ontology* $\mathcal{O}$.

Definition 4 (Query-activated concept (QAC) [13]). *Given a query* q_i *from a peer* S_i *and a mapping* $\mathcal{M}_{ij}$*, query-activated concept* $C_Q^j(q_i)$ *can be extracted as*

$$C_Q^j(q_i) = \{c | c \in q_i, c \in \mathcal{C}_i \cap \mathcal{M}_{ij}\} \tag{3}$$

where S_j *is a destination peer.*

For example, in Fig. 2, the dotted circle indicates a query, while two pairs of filled ones linked with arrows are the mappings. We can see that the context of the query can be relevant to several mappings (i.e., $\mathcal{M}_{13}$ and $\mathcal{M}_{12}$), simultaneously.

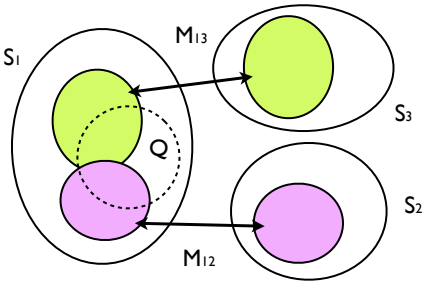


Fig. 2. Example of query-activated concept

Definition 5 (Semantic coverage ratio). *Given a query, a semantic coverage ratio τ_Q is computed by the ratio of the common mapping elements to the query-activated concepts. The τ_Q can be formulated as*

$$\tau_Q(q_i, S_j) = \frac{|C_Q^j(q)|}{|q_i|} \tag{4}$$

where S_j is the destination peer. It means how many mapping elements can cover the query-activated concepts.

Thus, the main goal of this work is to maximize the summation of all semantic coverage ratios until a query q_i from a source peer S_i is reached to all available peers in the mapping networks N_{MAP}. It can be formulated as

$$\max \sum_{S_j \in N_{MAP}} \tau_Q(q_i, S_j), \tag{5}$$

and then, we can determine how the direct mapping should be composed for the given query. Differently from the previous work [8], this work proposes a query decomposition scheme (more descriptions are in Sect. 3). Hence, Eq. 5 is rewritten to

$$\max \sum_{S_j \in N_{MAP}} \sum_{q_i' \in \sqcup(q)} \tau_Q(q_i', S_j) \tag{6}$$

where $q \equiv \bigcup_{q' \in \sqcup(q)} q'$. The semantic decomposition operator $\sqcup(q)$ can return the set of sub-queries, and $\max |\sqcup(q)|$ can not be more than all possible paths between source and destination peers.

3 Semantic Decomposition

According to the context of the given query, the query should be decomposed, and each of the modular queries is transformed by referring to the corresponding mapping compositions in mapping network. Thus, we want to list up all possible heuristics for

the query decomposition, and empirically justify whether each of the heuristics works or not, as follows.

$$\begin{aligned} q' = \sqcup(q) &= q, \text{ no decomposition } (D_{Generic}) \\ &= C_Q^j(q), \text{ with QAC } (D_{QAC}) \\ &= C_Q^j(q) - \bigcup_{S_k \in N_{MAP}, k \neq j} C_Q^k(q), \text{ with unique QAC } (D_{UQAC}) \\ &= WN(q), \text{ with a WordNet } (D_{WN}) \\ &= WN(C_Q^j(q)), \text{ with a WordNet } (D_{WN+QAC}) \end{aligned}$$

While $D_{Generic}$ is the simplest one without any decomposition, D_{QAC} and D_{UQAC} can extract a part of concepts related to the mappings. Especially, D_{WN} takes into account semantic relationship between concepts in a query by using background knowledge (e.g., WordNet[1]). Finally, D_{WN+QAC} is a hybrid approach combining the two previous ones.

4 Composition Path Based on Semantic Centrality

Once the decomposed query is obtained, we have to find the optimized path of mapping composition. In this paper, we want to apply social centrality measurements to do this. As a matter of fact, there have been several centrality indices to measure the power of structural position on social network [15]. However, they are not appropriate to reflect the centrality among the semantic relationships (i.e., ontology mappings) between distributed information peers (i.e., in a mapping network).

Thereby, we define a semantic centrality as the power of semantic bridging on the mapping network. Suppose that two peers s and t are not able to communicate with each other, due to the semantic heterogeneity between their ontologies $\mathcal{O}_s$ and $\mathcal{O}_t$. We need to search for the peer ontology $\mathcal{O}_i$ of which semantic centrality is high enough to reconcile these ontologies. It means $\mathcal{O}_i$ is containing some classes matched with the consensual ontology $\mathcal{CO}$. We intuitively assume that a peer is assigned higher semantic centrality, as his ontology includes more consensual classes in common. Thus, we formulate a semantic centrality of i-th peer $C^{\diamond}(i)$ as

$$C^{\diamond}(i) = \frac{|\mathcal{O}_i \cap \mathcal{CO}|}{|\mathcal{O}_i|} \sum_{s \neq t \neq i \in N} \frac{\sigma^{\diamond}_{\mathcal{O}_s, \mathcal{O}_t}(\mathcal{O}_i)}{|SP^{\diamond}(s,t)|} \tag{7}$$

which means the semantic closeness (or coverage) of the peer ontology $\mathcal{O}_i$ to the discovered consensual ontology $\mathcal{CO}$. The denominator $|\mathcal{O}_i|$ is for the normalization by the total number of classes organizing the peer ontology. $SP^{\diamond}$ is a pair of peers whose peer ontologies are not semantically interoperable directly. So, C_B can be replaced by C_C or others. More importantly, function $\sigma^{\diamond}$ is to determine the efficiency of reconciliation, and it is given by

$$\sigma^{\diamond}_{\mathcal{O}_s, \mathcal{O}_t}(\mathcal{O}_i) = \frac{|\mathcal{O}_s \cap \mathcal{O}_i| \cdot |\mathcal{O}_t \cap \mathcal{O}_i|}{|\mathcal{O}_s \cap \mathcal{CO}| \cdot |\mathcal{O}_t \cap \mathcal{CO}|} \tag{8}$$

[1] http://wordnet.princeton.edu/wordnet/download/

which expresses that the number of matched classes between two ontologies is in linear proportion, in contrast of that of matched classes with consensus ontologies. Additionally, in Equ. 7 and 8, the counting computation of union sets is done by

$$|A \cap B| = count(\langle c, c' \rangle)_{\langle c,c' \rangle \in Pairing(A,B), Sim_C(c,c')=1}. \quad (9)$$

Consequently, the semantic centrality can be used for determining composition path in the mapping network.

5 Experimental Results

In order to evaluate the proposed query transformation in distributed information systems, we have built seven ontology-based information systems (i.e., S_A to S_G) with linkages, as shown in Fig. 3. All of the mapping results have been collected by human experts.

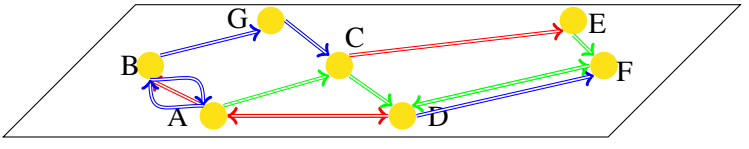


Fig. 3. A mapping network as a testing-bed

We have focused on two evaluation issues (i.e., mapping composition and transformation path selection), and have collected experimental results.

5.1 Evaluation on Mapping Composition

By using OLA API [16], we have automatically collected the direct mapping results (i.e., $\mathcal{M}$). The mapping results have been composed in all possible cases (i.e., $\widetilde{\mathcal{M}}$). More importantly, we have measured the semantic centrality of all peers, as shown in Table 1.

Table 1. Semantic centrality of peers in Fig. 3

Peers	S_A	S_C	S_D	S_E	S_F	S_G
$C^{\diamond}$	0.875	0.73	0.802	0.152	0.615	0.175

The performance of mapping composition has been tested by two well-known IR criteria, precision and recall.

$$Precision = \frac{|\mathcal{M} \cap \widetilde{\mathcal{M}}|}{|\widetilde{\mathcal{M}}|} \text{ and } Recall = \frac{|\mathcal{M} \cap \widetilde{\mathcal{M}}|}{|\mathcal{M}|} \quad (10)$$

In average, compared to results in the previous work [8] (i.e., 73% recall and 79% precision), we have obtained significantly better results (i.e., 82% recall and 88.5% precision) thanks to measuring the semantic centrality. We note that as the mapping

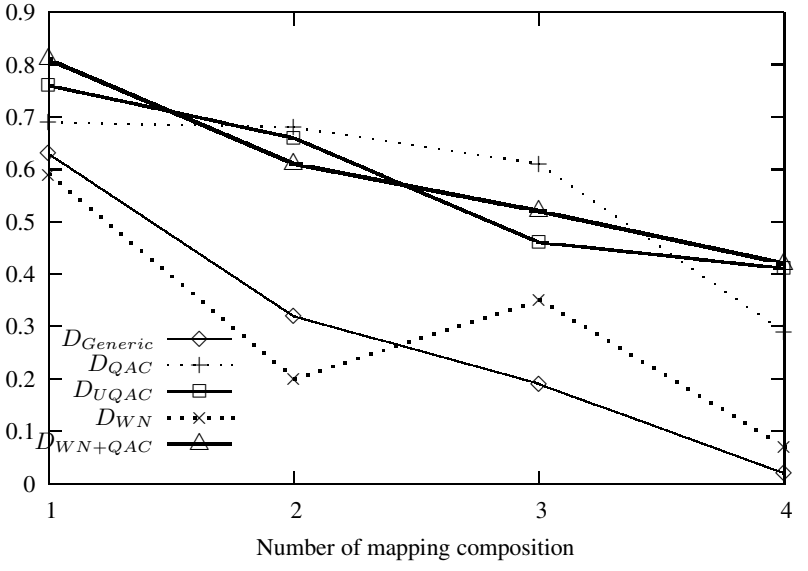


Fig. 4. The performance of recall of transforming the decomposed queries

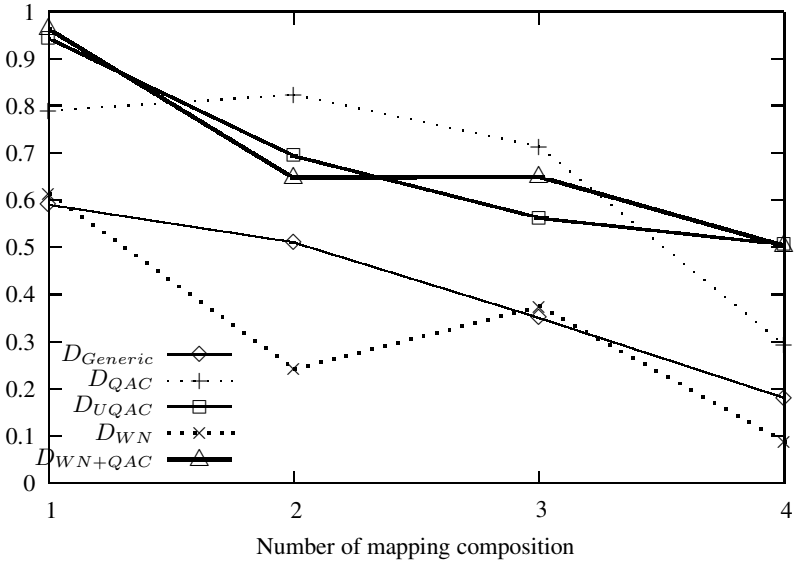


Fig. 5. The performance of precision of transforming the decomposed queries

results are composed (i.e., the number of mapping composition is increased) in all cases, the recall and precision is getting decreased by nature. This is the information loss cased by mismatching problem of ontology mapping algorithms.

5.2 Evaluation on Query Decomposition

In second issue, we have tested the performance of transformation path selection resulting from five query decomposition heuristics (i.e., $D_{Generic}$ to D_{WN+QAC}) by asking real users to participate to peer S_A in Fig. 3. Thirty users were asked to generate 10 queries with SparQL to search for a certain information, depending on their contexts. These queries have been able to be sent to only three system S_B, S_C, and S_D, for considering query decomposition along with all possible linkages. The performance of recall and precision of transforming the sub-queries decomposed the five heuristics are shown in Fig. 4 and Fig. 5, respectively. We can find out that, in general, query-activated concept-based decomposition schemes (i.e., D_{QAC}, D_{UQAC}, and D_{WN+QAC}) outperforms the others

6 Discussion and Related Work

The proposed method has been tackling semantic query decomposition for efficient query transformation on heterogeneous distributed network. Through the experiments, we have checked that as repeating the mapping composition, there is anyhow a certain amount of information loss. More importantly, the transformed queries after decomposition have shown relatively higher precision performance, compared with recall. It means that the query decomposition has positive effect on transformation meaning retaining the original contexts from the source.

There have been various related work. Traditionally, many studies [17] have been done on query expansion and rewriting for distributed databases. In [18], complex query for RDBMS has been investigated how to conduct query modulation.

Particularly, in semantic web community, semantics and knowledge can support the query processes [10–12, 19]. Also, OWL-QL [9] has proposed a number of issues on ontology-based queries.

7 Concluding Remark and Future Work

We want to draw a conclusion in this section. Mainly, we emphasize that there should be an efficient mechanism to collect relevant resources from the distributed sources. Mappings between ontologies can be applied to transform the queries. Moreover, in economical reason, the mappings were reused and shared for generating indirect mapping.

As future work, we have three main plans to investigate the followings issues i) *semantic subgroup discovery*, to organize the sophisticated user groups with enhancing the designed discovery methods, and ii) *semantic synchronization,* to maximize the efficiency interoperability by information diffusion. Furthermore, we have to consider to enhance the semantic centrality measurement $C^{\diamond}$ by combining with i) authoritative and hub centrality measurement, and ii) the modified shortest paths $spd(n,t) = \frac{1}{C^{\diamond}(n)+C^{\diamond}(t)}$.

Acknowledgement

This work was supported by the Korean Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST) (2009-0066751).

References

1. Jung, J.J., Euzenat, J.: Towards semantic social networks. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 267–280. Springer, Heidelberg (2007)
2. Mika, P.: Social Networks and the Semantic Web. Semantic Web And Beyond Computing for Human Experience, vol. 5. Springer, Heidelberg (2007)
3. Haase, P., Siebes, R., van Harmelen, F.: Expertise-based peer selection in peer-to-peer networks. Knowledge and Information Systems 15(1), 75–107 (2008)
4. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
5. Lumineau, N., Doucet, A., Gançarski, S.: Thematic schema building for mediation-based peer-to-peer architecture. Electronic Notes in Theoretical Computer Science 150(2), 21–36 (2006)
6. Jung, J.J.: Query transformation based on semantic centrality in semantic social network. Journal of Universal Computer Science 14(7), 1031–1047 (2008)
7. Brzykcy, G., Bartoszek, J., Pankowski, T.: Schema mappings and agents' actions in p2p data integration system. Journal of Universal Computer Science 14(7), 1048–1060 (2008)
8. Jung, J.J.: An empirical study on optimizing query transformation on semantic peer-to-peer neworks. Journal of Intelligent & Fuzzy Systems 21(3), 187–195 (2010)
9. Fikes, R., Hayes, P., Horrocks, I.: Owl-ql—a language for deductive query answering on the semantic web. Journal of Web Semantics 2(1), 19–29 (2004)
10. Owei, V., Navathe, S.B.: Enriching the conceptual basis for query formulation through relationship semantics in databases. Information Systems 26(6), 445–475 (2001)
11. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Information Processing & Management 43(4), 866–886 (2007)
12. Zenz, G., Zhou, X., Minack, E., Siberski, W., Nejdl, W.: From keywords to semantic queries—incremental query construction on the semantic web. Journal of Web Semantics 7(3), 166–176 (2009)
13. Jung, J.J.: Reusing ontology mappings for query segmentation and routing in semantic peer-to-peer environment. Information Sciences 180(17), 3248–3257 (2010)
14. Jung, J.J.: Ontology-based context synchronization for ad-hoc social collaborations. Knowledge-Based Systems 21(7), 573–580 (2008)
15. Freeman, L.: Centrality in social networks: Conceptual clarification. Social Networks 1, 215–239 (1979)
16. Euzenat, J.: An API for ontology alignment. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 698–712. Springer, Heidelberg (2004)
17. Ramesh, A., Ali, D.L.: Query transformation in heterogeneous distributed database systems. Computers & Industrial Engineering 31(1-2), 323–326 (1996)
18. Chatziantoniou, D., Ross, K.A.: Partitioned optimization of complex queries. Information Systems 32(2), 248–282 (2007)
19. Albayrak, S., Milosevic, D.: Multi-domain collaborative exploration mechanisms for query expansion in an agent-based filtering framework. Electronic Commerce Research and Applications 6(4), 399–412 (2007)

Comparative Studies of Parallel and Vertical Stereo Vision-Based 3D Pneumatic Arms

Ray-Hwa Wong[1,2,*], Y. Wang[1], and Chao-Yi Liu[1]

[1] Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taiwan, Republic of China
[2] Department of Mechanical Engineering, Hwa Hsia Institute of Technology, Taiwan, Republic of China

Abstract. Two parallel or vertical installed CCDs are used to capture the stereo photo vision of 3D pneumatic arm. CCD is used to obtain the planer image. Through parallel and vertical stereo triangulation and coordinate transformation, the stereo photo vision signals can be adapted for 3D pneumatic arm's feedback control signals. Since the imaging process restricts the sampling rate, the self-organizing sliding-mode fuzzy controller is implemented to simplify fuzzy rules to reduce the computer load and its learning mechanism can optimize fuzzy rules on-line to improve the control performance. The objective of this paper is to compare the measuring accuracy of parallel stereo vision and vertical stereo vision, and also study their control performances of variety trajectory tracking experiments.

Keywords: 3D pneumatic arm, parallel stereo vision, vertical stereo vision, self organizing sliding mode fuzzy controller.

1 Introduction

Pneumatic muscle and rotational actuators are developed to take the place of conventional pneumatic linear actuators in applications of rotational, non-aligned and complicated mechanisms. In industrial applications, pneumatic robots are used widely because of their simplification and efficiency [1]. In this paper, two pneumatic muscle actuators associated with pressure type servo-valves and one rotational actuator associated with flow type servo-valve are used to setup a 3D pneumatic arm control system to simulate the motion of an excavator.

Vision-based pneumatic arms have been applied in the industries lately [2]. Instead of the contacted angular displacement sensors, this paper implements two parallel or vertical installed CCDs to capture the stereo photo vision of 3D pneumatic arm. It can determine the arm's location in the three-dimensional Cartesian coordinate by the stereo triangulation, and then transform into the three-dimensional angular displacement signals for arm's control. Encoders and CCDs are included in this 3D pneumatic arm control system and their displacement measurements can be simultaneously recorded for comparisons. It thus can study the measuring accuracy of parallel stereo vision and

* Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 163–173, 2010.

vertical stereo vision based on the classical contacted sensors. Then, it can compare the control performance of parallel and vertical stereo vision-based 3D pneumatic arms.

This 3D pneumatic arm is a non-linear and non-coupled control system. Without the detailed model, fuzzy control algorithms have been found to be effective in dealing with non-linear, complicated and ill-defined systems. The sliding-mode controller [3] and the self-learning fuzzy controller [4] have been widely applied for pneumatic control systems. To integrate the sliding-mode and self-learning fuzzy controllers, this paper proposes the self-organizing sliding-mode fuzzy controller for the trajectory tracking control of 3D pneumatic arm. The sliding surface function is used to reduce the two-dimensional into one-dimensional system variables. The one-dimensional self-learning mechanism provides the optimized fuzzy rules online.

This paper is mainly to compare the static and dynamic measuring errors of parallel and vertical stereo visions, and to compare their trajectory tracking performances of variety trajectory tracking experiments. Thus, this paper could justify the implementations of parallel and vertical stereo vision-based 3D pneumatic arms to replace the encoder-based 3D pneumatic arm.

2 System Descriptions

Fig. 1 is a 3D pneumatic arm control system which is used to simulate the excavator's motion. M_1 and M_2 are muscle actuators and their specifications are $20^{\phi} \times 189\,mm$ and $20^{\phi} \times 181\,mm$ respectively. They are driven by pressure type servo valves FESTO MPPES-3-1/8-010. The contraction range of muscle actuator is -3 ~ 20%. R is a rotational actuator and its specification is $40^{\phi} \times 270^{\circ}$. It is driven by flow type servo valve FESTO MPYE-5-1/8-CF-010-B. The work space of this 3D pneumatic arm is limited due to the limitations of contraction range of muscle actuators. The first arm's length l_1 is 322*mm* and its mass including the encoder is 1.3*kg*. The second arm's length l_2 is 460*mm* and the mass is 1.8*kg*. The eccentric length l_3 is 35*mm*. The height h is 42*mm*. The loading is 4*kg*.

Both CCDs and encoders are installed to measure the three-dimensional arm's displacement signals. CCDs are parallel or vertical aligned. $\theta_1 \sim \theta_3$ are angular displacements measured by encoders $E_1 \sim E_3$ which resolutions are 2048*pulses/cycle*. The CCD camera is SONY LV-75 which is to capture the planer image and its resolution is 640×480*pixels*. The calibration contains a pattern of 16×12 circular grids. The size of rectangular pattern is $240mm \times 180mm$. The distance between camera and model plane is 225*mm*. Via the calibration mode provided by Matrox image library, the camera calibration process [5] is applied to eliminate the errors caused by optical and planar distortions. The scaling rate of camera is 59.94*Hz*. The video capture board is a photo vision decoder to transfer the vision signal and the Matrox image library is used to determine the two-dimensional coordinates in the image plane. Then, parallel or vertical stereo vision signals can determine the 3D pneumatic arm's location by the stereo triangulation technique. Since this image processing is restricted by the capacities of Matrox image library and CCD camera, the maximum sampling rate of this vision-based control system is 30*Hz*. The

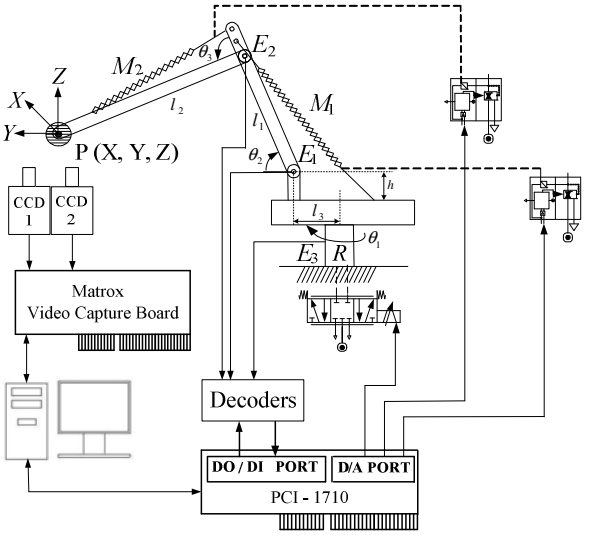


Fig. 1. Schematic diagram of a 3D pneumatic arm control system

resolution of D/A port is 12*bits*. The personal computer (PC) is a 80586 microcomputer system. The control program is developed by Visual C++.Net.

3 Parallel Stereo Triangulation

Two parallel aligned CCDs with the same focal length (f) are installed to provide disparity images and to determine the loading point $P(X_p, Y_p, Z_p)$. This point's location can be reconstructed from the perspective projections on the image planes of two CCDs. Its parallel stereo triangulation is shown as Fig.2. The baseline of two CCDs is H which is perpendicular to the optical axes. y_1 and y_2 are the perspective projections on CCD1 and CCD2 image planes. The relationship of parallel stereo triangulation is

$$\frac{y_1}{Y_p} = \frac{f}{X_p} = \frac{y_2}{H - Y_p} \tag{1}$$

So, the X_p and Y_p locations of object are

$$Y_p = \frac{y_1 \cdot H}{y_1 + y_2} \tag{2}$$

$$X_p = \frac{H \cdot f}{y_1 + y_2} \tag{3}$$

$$Z_p = \frac{z_1 \cdot H}{y_1 + y_2} \tag{4}$$

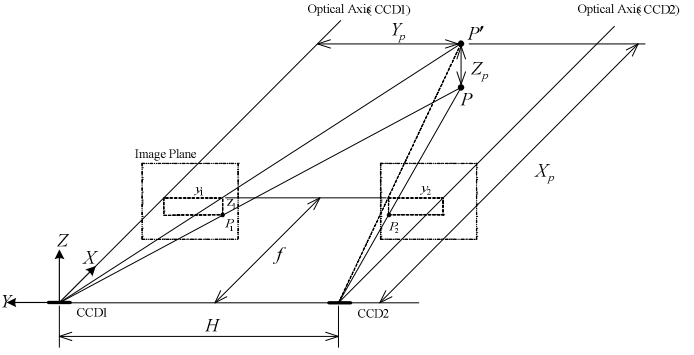


Fig. 2. Stereo triangulation of parallel stereo-vision

4 Vertical Stereo Triangulation

Two vertical installed CCDs are used to capture disparity images and to determine the loading point $P(X_p, Y_p, Z_p)$. The vertical stereo triangulation is shown as Fig.3.

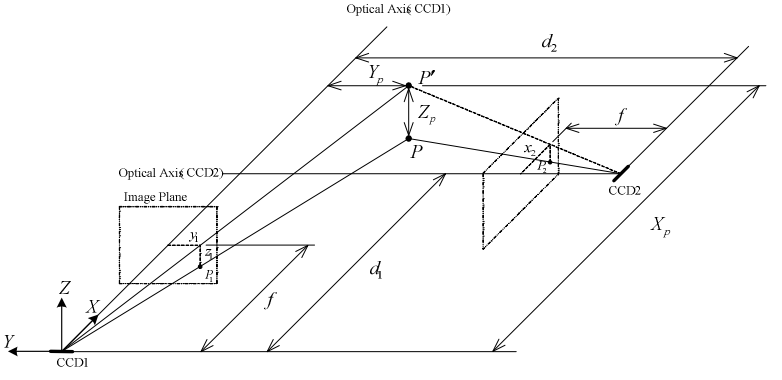


Fig. 3. Stereo triangulation of vertical stereo-vision

d_1 and d_2 are the respective distances of optical axes' intersection to CCD1 and CCD2. y_1 and z_1 are the projections on the CCD1 image plane. x_2 is the projection on the CCD2 image plane. The relationship of vertical stereo triangulation is

$$\frac{X_p - d_1}{d_2 - Y_p} = \frac{x_2}{f} \tag{5}$$

$$\frac{Y_p}{X_p} = \frac{y_1}{f} \tag{6}$$

So, X_p, Y_p and Z_p locations of object are

$$X_p = \frac{f(d_1 f + d_2 x_2)}{x_2 y_1 + f^2} \tag{7}$$

$$Y_p = \frac{y_1(d_1 f + d_2 x_2)}{x_2 y_1 + f^2} \tag{8}$$

$$Z_p = \frac{z_1 \sqrt{X_p{}^2 + Y_p{}^2}}{\sqrt{y_1{}^2 + f^2}} \tag{9}$$

5 Control Scheme

The 3D pneumatic arm control system shown in Fig. 1 is used to simulate the excavator's motion. Fig. 4 is the functional block diagrams of control structures using parallel or vertical stereo-vision. The kinetic and inverse kinetic transformations are the relationship between angular displacements ($\theta_1, \theta_2, \theta_3$) and absolute location $P(X_p, Y_p, Z_p)$.

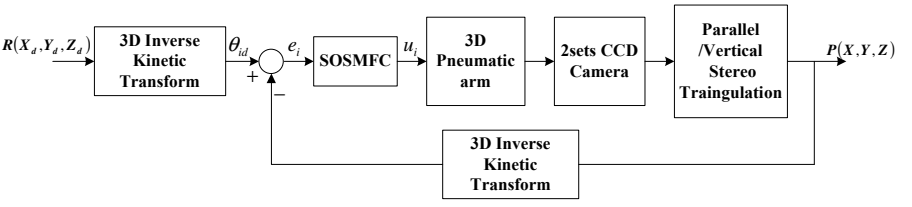


Fig. 4. Functional block diagram of vision-based control structure

This 3D pneumatic arm control system is a non-coupled three-input three-output system. θ_1, θ_2 and θ_3 could be individually driven by M_1, M_2 muscle actuators and R rotational actuator respectively. And thus, the controller can be designed for each sub-system individually.

In comparison with other control algorithms, the fuzzy control approach has been found to be an effective algorithm to deal with this highly nonlinear 3D pneumatic arm control system. In general, the fuzzy rule of fuzzy logic control is two-dimensional which depends on system variables (e, $\dot{e}$). To optimize three two-dimensional fuzzy rule tables of three dimensional actuating subsystems requires a lot of effort. Therefore, the self-organizing sliding-mode fuzzy controller is proposed to simplify fuzzy rules and reduce the computer load in the fuzzy rule learning mechanism.

Fig. 5 is the configuration of the self-organizing sliding-mode fuzzy controller and the error signal e represents θ_1, θ_2 or θ_3 respectively. The sliding surface is designed to simplify system variables and reduce two-dimensional fuzzy rules into one-dimensional. It is described as

$$s = \alpha \cdot e + \dot{e} \tag{10}$$

where α is a positive constant. The gains G_s and G_u are used to normalize between system variables and the universal of fuzzy sets. The fuzzy sets are finely divided into 13 linguistic fuzzy subsets. Fuzzification is adopted the triangular-type membership function to obtain linguistic variables. The fuzzy inference is based on the Max-Min product composition and is used to operate fuzzy control rules. The height method [6] is used to defuzzify the fuzzy sets to attain the control signal. In the fuzzy rule learning mechanism, the input reinforcement is defined by corrections to compensate for error and error change, which weighting factors are $\frac{\alpha T}{1+\alpha T}$ and $\frac{1}{1+\alpha T}$ respectively. Thus, the linguistic approach rule base of self-organizing learning mechanism can be modified as

$$RULES(nT+T) = RULES(nT) + \frac{\gamma T}{M(1+\alpha T)} \cdot s(nT) \tag{11}$$

Here, γ is the learning rate to compensate the output error. M is a ratio to simulate the relationship between input signal of servo-valve and angular displacement output. T is the sampling time.

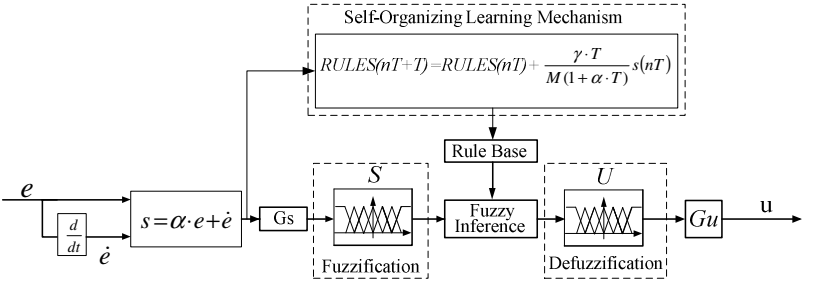


Fig. 5. Configuration of self-organizing sliding mode fuzzy controller

6 Measuring Errors of Stereo Visions

The difference between encoders' signals and stereo vision's signals is defined as the measuring error. The average static measuring error within the working space of experiments in Section 7 of parallel stereo vision is 0.71*mm*. The average static measuring error of vertical stereo vision is 0.81*mm.*

Fig. 6 is their dynamic measuring errors versus the object's velocity. It indicates that the dynamic measuring error of parallel stereo vision is relatively sensitive to the object's velocity. While the object's velocity is under 10*mm/s*, the dynamic measuring error of parallel stereo vision is 0.57~0.64*mm*, which is near to the static measuring error. Since the imaging process induces the time-delay effects, the dynamic measuring error is sensitive to the high object's velocity. The measuring error is up to 3.5~7.8*mm* at 50~60*mm/s*. It means that the parallel stereo vision-based system is reliable at low speed and quasi-static applications only. The dynamic measuring error of vertical stereo vision is less than the parallel stereo vision, and is less sensitive to the object's velocity. The dynamic measuring error of vertical stereo vision is below 1.0*mm* while the object's velocity is less than 40*mm/s*. The vertical stereo vision can be implemented for wider object's velocity range.

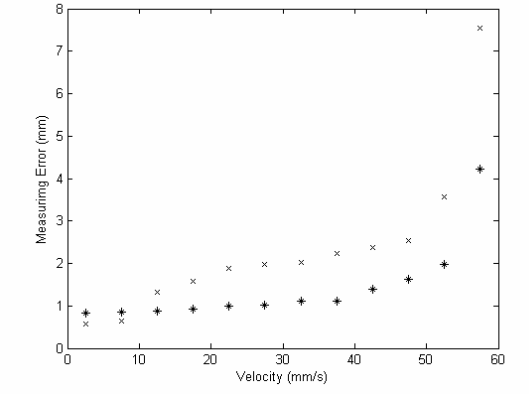


Fig. 6. Measuring error of various object's velocity （* : vertical, × : parallel）

7 Experimental Results

The 3D pneumatic arm control system in this paper is applied for variety trajectory tracking applications including step, ramp and parabolic commands. The sampling rate of vision-based system is valid within 30*Hz* due to the limitations of photo vision capture process and computational loading. Although the encoder-based system could apply in the high sampling rate, the sampling time of following experiments of both vision-based and encoder-based systems is fixed at 0.035 sec. in order to ignore the sampling effects. The focal length *f* of CCD is 6*mm*. The baseline *H* is 41.5*mm*. The distances of optical axes' intersection to CCDs are d_1=225*mm* and d_2 =225*mm*. Parameters of the self-organizing sliding-mode fuzzy controller are properly chosen as G_s=0.3, α=4, γ=0.1, M=1, G_u=4 and G_u=1 for muscle actuators and rotational actuator respectively. Initial values of one-dimensional fuzzy rules are (-1.0, -0.8, -0.6, -0.4, -0.2, -0.1, 0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0).

To evaluate the accuracy, absolute and average position errors are defined as:

$$e(k) = \sqrt{(X_d(k) - X_p(k))^2 + (Y_d(k) - Y_p(k))^2 + (Z_d(k) - Z_p(k))^2} \tag{12}$$

$$\bar{e} = \frac{\sum_{k=1}^{N} e(k)}{N} \tag{13}$$

The following experiments are included to compare the control performances of parallel vision-based, vertical vision-based and encoder-based systems. Case 1 is a point-to-point control experiment and its commands are $X_d = Y_d = Z_d = 80u(t)$. Fig. 7 is the time response. Due to the characteristics of imaging process, the time delay of vision-based system is slightly greater than encoder-based system. The overall behaviors in time responses are similar. To compare with the parallel vision-based system, the vertical vision-based system matches well with the encoder-based system. The error of the vertical vision-based system is smaller than the parallel vision-based system. In the steady state performance, the steady state error e_{ss} of parallel vision-based system is 0.89*mm*, the vertical vision-based system is 0.49*mm* and the encoder-based system is 0.42*mm*, which are summarized in Table 1. Thus, both e_{ss} of parallel and vertical vision-based systems are less than the static measuring errors, and are reasonable to take the place of encoders in the 3D pneumatic arm control system.

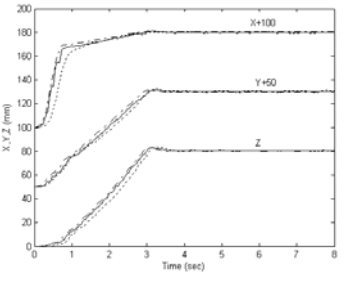


Fig. 7. Time response of Case 1 (—— : vertical, – – : parallel, – · – : encoder)

Table 1. Average errors and steady state errors of Cases1~3 (unit: *mm*)

	$\bar{e}$			e_{ss}		
	Parallel	Vertical	Encoder	Parallel	Vertical	Encoder
Case 1	28.91	26.60	26.22	0.89	0.49	0.42
Case 2	3.40	3.71	3.36	0.90	0.73	0.77
Case 3	3.36	3.61	3.03	0.79	0.52	0.39

Case 2 is a trajectory tracking control of a ramp input. Fig. 8 is the trajectory tracking performance of parallel vision-based, vertical vision-based and encoder-based systems. Their performances are similar. Their average errors $\bar{e}$ and steady

state errors e_{ss} are summarized in Table 2. Since this case is to trace a ramp command, measuring errors of vision-based system are mainly induced in the low-velocity range. $\bar{e}$ of parallel vision-based and vertical vision-based systems are 3.40*mm* and 3.71*mm* respectively. $\bar{e}$ of the parallel vision-based system match well with the encoder-based system, and is slightly better than the vertical vision-based system. Again, except the slight oscillations of vision-based system in the steady state, their overall performances have no significant difference. Both parallel and vertical visions applied for measuring the 3D pneumatic arm's displacement are acceptable.

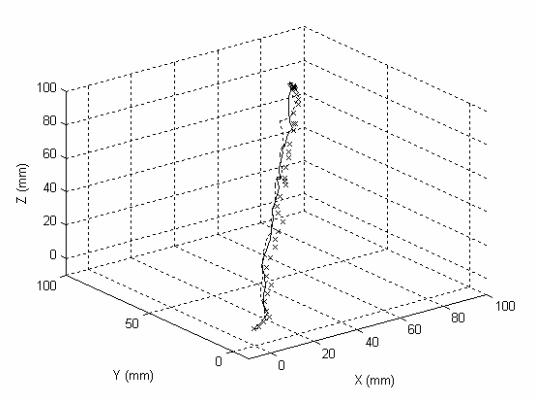


Fig. 8. Trajectory tracking performance of Case 2 (—— : vertical,– – : parallel, × : encoder)

To compare the difference of Case 1 and Case 2, the difference $\bar{e}$ of parallel vision-based and encoder-based systems is 0.04*mm* in Case 2, which is much less than 2.69*mm* in Case 1. The difference e_{ss} of parallel vision-based and encoder-based systems is 0.13*mm* in Case 2, and again which is less than 0.43*mm* in Case 1. Both are significantly less than the static measuring error. To compare with vertical vision-based and encoder-based systems, there are no significant difference $\bar{e}$ and e_{ss} between Case 1 and Case 2. The dynamic measuring effects are less influenced in the vertical vision-based system.

Case 3 is a quadratic trajectory tracking control. Fig. 9 is the trajectory tracking performance. Due to the characteristics of parabolic function, the system delay is relatively decreased in Case 3. Again, the tracking performances of both parallel and vertical vision-based systems match well with the encoder-based systems. $\bar{e}$ and e_{ss} of both parallel and vertical vision-based systems are similar to the encoder-based system indicated in Table 1. Thus, it is reasonable to replace encoders by parallel and vertical stereo-visions in the quadratic trajectory tracking control of 3D pneumatic arm control system.

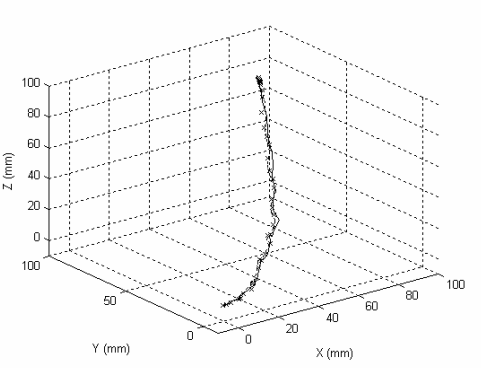


Fig. 9. Trajectory tracking performance of Case 3 (——: vertical, – – : parallel, ×: encoder)

8 Conclusions

Comparisons of the previous experimental studies lead to following conclusions.

1. Both parallel and vertical stereo-vision techniques based on stereo triangulation provide accurate 3D displacement measurement. But, it requires a lot of computational time in the measuring process, and thus, the vision-based system is valid for low sampling frequency applications only.

2. The steady state errors of variety applications are near to the static measuring error. Thus, the parallel and vertical stereo-vision based 3D pneumatic arms have accurate steady state and quasi-static performance.

3. The self-organizing sliding-mode fuzzy controller can simplify the fuzzy rules' optimizing process and can fit for vision-based and encoder-based 3D pneumatic arms.

4. To compare the parallel vision-based system with the vertical vision-based system, their static measuring errors have no significant difference. In the dynamic measuring error analysis, the vertical vision-based are less sensitive to the object's velocity and relatively smaller than the parallel vision-based system in the high velocity range.

5. Experimental results indicate that the trajectory tracking performance of parallel vision-based, vertical vision-based and encoder-based 3D pneumatic arms match well with each other. Therefore, parallel and vertical stereo-visions are options to replace encoders in the trajectory tracking applications of 3D pneumatic arms.

References

1. Tondu, B., Ippolito, S., Guiochet, J., Daidie, A.: A Seven-degree-of-freedom Robot-arm Driven by Pneumatic Artificial Muscles for Humanoid Robots. The International Journal of Robotics Research 24(4), 257–274 (2005)
2. Hutchinson, S., Hager, G.D., Corke, P.I.: A Tutorial on Visual Servo Control. IEEE, Robotics and Automation 12(5), 651–670 (1996)

3. Errahimi, F., M'Sirdi, N.K., Abarkane, H.: Pneumatic Robot Leg Control via Second Order Sliding Mode Technique. Advances in Modelling and Analysis 60(1-2), 63–72 (2005)
4. Wang, Y.T., Wong, R.H., Yu, C.H., Huang, W.C.: Fuzzy Control: 2D Pneumatic Muscle Actuator's Arm. The Journal of the Institute of Measurement and Control 42, 24–27 (2009)
5. Fiale, M., Shu, C.: Self-Identification Patterns for Plane-Based Camera Calibration. Machine Vision and Applications 19(4), 209–216 (2008)
6. Lee, C.C.: Fuzzy Logic in Control Systems: Fuzzy Logic Controller- Part I and II. IEEE Transactions on Systems, Man and Cybernetics 20(2), 404–435 (1990)

An Effective Image Enhancement Method for Electronic Portal Images

Mao-Hsiung Hung[1,2], Shu-Chuan Chu[3], John F. Roddick[3], Jeng-Shyang Pan[1,2], and Chin-Shiuh Shieh[1]

[1] Dept. of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 807, Taiwan
[2] Innovative Information Industry Research Center (IIIRC), Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
[3] School of Computer Science, Engineering and Mathematics, Flinders University of South Australia, Adelaide, 5001, Australia
mhhung0502@gmail.com, chu0027@csem.flinders.edu.au, jspan@cc.kuas.edu.tw, csshieh@cc.kuas.edu.tw

Abstract. Due to the inherent low-contrast in Electronic Portal Images (EPI), the perception quality of EPI has certain gap to the expectation of most physicians. It is essential to have effective post-processing methods to enhance the visual quality of EPI. However, only limited efforts had been paid to this issue in the past decade. To this problem, an integrated approach featuring automatic thresholding is developed and presented in this article. Firstly, Gray-Level Grouping (GLG) is applied to improve the global contrast of the whole image. Secondly, Adaptive Image Contrast Enhancement (AICE) is used to refine the local contrast within a neighborhood. Finally, a simple spatial filter is employed to reduce noises. The experimental results indicate that the proposed method greatly improves the visual perceptibility as compared with previous approaches.

Keywords: electronic portal image, contrast enhancement, gray-level grouping, adaptive image contrast enhancement.

1 Introduction

In radiation therapy, megavoltage X-rays of the linear accelerator (LINAC) are often used to develop for inspection. As shown in Fig. 1, Electronic Portal Imaging Device (EPID) uses digital imaging, such as a CCD video camera, liquid ion chamber and amorphous silicon (a-Si) flat panel detector are used to capture a digital image. The electronic portal images almost have replaced the traditional portal films.

However, portal images which are acquired from an EPID, still suffer from poor visual contrast. Because of the using of the megavoltage X-rays, the most of photon interactions are caused by Compton Effect [1]. The megavoltage X-rays cause very small difference of mass attenuation between bone and soft tissue in portal images. As a result, it is difficult to verify the treatment region of the patient with raw portal images. Therefore, digital image processing is often applied to improve the visualization of anatomical structure in EPIs [2].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 174–183, 2010.

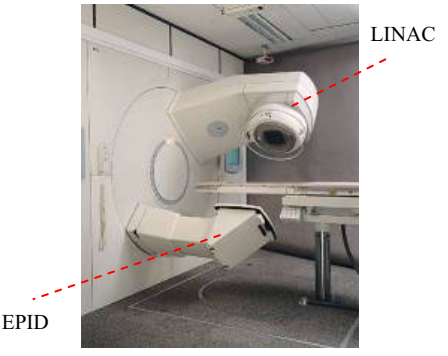


Fig. 1. LINAC and EPID

In the past decade, very few papers related to image enhancement for EPI have been published. Crooks and Fallone [3] propose a selective histogram equalization which enhances the contrast in both the field and surrounding regions in the double-exposure portal images. Shalev et al. [4] propose a method of three sequential procedures of contrast enhancement, noise reduction and edge sharpening. Following the idea of the three sequential procedures, Koutsofios et al. [5] conduct and verify 12 different combinations of the three procedures for the clinical usefulness.

Adaptive Histogram Equalization (AHE) [6] is one of the most widespread algorithms to improve various low-contrast medical images, such as portal images, ultra-sonogram and mammogram. However, AHE often generates over-enhancement and introduces serious artifacts such as blurring and noise amplification, to influence image visualization. Contrast Limitation Adaptive Histogram Equalization (CLAHE) [6] is suggested to control the contrast strength of the AHE using a clipping level. The clipping level is still sensitive to control in CLAHE.

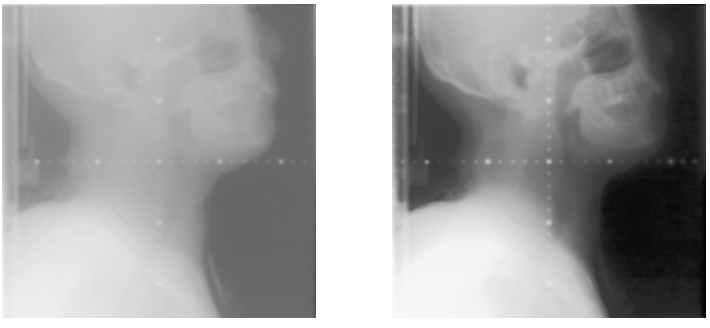
In this paper, we integrate two algorithms to enhance contrast for EPI. Gray-Level Grouping (GLG) [7] is first applied to improve the global contrast of the image. Second, Adaptive Image Contrast Enhancement (AICE) [8] is used to gain the local contrast with the pixel neighborhood. Meanwhile, we propose a novel scheme choosing a suitable parameter with the image gradient in AICE. Finally, a simple spatial filter reduces noise. The experimental results indicate that the proposed method not only effective enhance the low contrast of EPI, but also greatly reduce the artifacts of blurring and noise amplification comparing with the traditional methods.

2 Proposed Method

Our proposed method contains three sequential procedures. The first procedure enhances the global contrast. After that, the second procedure enhances the local contrast. The last procedure reduces the noises which are amplified by the contrast enhancement. The details of the procedures will be described in the following subsections.

2.1 Global Contrast Enhancement

Global contrast enhancement often means that a contrast enhancement algorithm refers all pixels to transform every point. Histogram Equalization (HE) is the most popular algorithm of global contrast enhancement. The histogram of the whole pixels in an image is given at first. Then, according to the cumulative function of the histogram, the point-to-point transformation is applied every point in HE. HE can make the output image getting higher contrast based on the transform.



(a) (b)

Fig. 2. Fast GLG enhancement: (a) Original image, (b) Enhancement result

HE often obtains unsatisfied results in some cases. For an example, if an image has a large area of pixels close to zeros intensity, these pixels are transformed into the quite light intensities by HE and the output image appears washed-out and low visiuality [9]. The washed-out effect often happens in some radiological images [7]. In addition, HE reduces the information of the processed image by redistributing the gray levels [10].

Chen et al. [7] propose a automatic method named by Gray-Level Grouping (GLG) to distribute the gray levels efficiently. In GLG, the first step groups the two smallest non-zeros bins of the histogram. Secondly, the redistribution makes every group uniformly occupying a grayscale segment of the same size. The last step ungroups the previously grouped gray-levels within grayscale segments. The above steps execute iteratively and optimize a contrast criterion.

The experimental results indicate that GLG totally improves the washed-out effect by HE and obtain satisfied results for other images which HE can performs well. In our work, we apply a fast algorithm of GLG suggested by Chen et al. In fast GLG, the three steps of grouping, redistribution and ungrouping reach until a proper number groups (i.e. 20 groups) without maximizing the contrast criterion. Fig. 2 (a)-(b) show the portal image captured from a patient's head and neck and its enhancement result of by the fast GLG. We found that there are many unused bins in two sides of the original histogram. By the fast GLG enhancement, the non-zero bins are completely distributed in all gray levels for the output image.

2.2 Local Contrast Enhancement

Local contrast enhancement means that a contrast enhancement algorithm processes every point by refering the neighborhood pixels. Adaptive Histogram Equalization (AHE) is the most representative algorithm of local contrast enhancement. Stark [8] has presented a generalization form of AHE and designed a cumulation function to control the enhancement strength based on the generalization form. We denote the method as Adaptive Image Contrast Enhancement (AICE). The transformation function of AHE, $z(x, y, u)$, is defined as a convolution as

$$z(x, y, u) = \sum_{v} \hat{h}(x, y, v) f_c(u, v) \tag{1}$$

where $\hat{h}(x, y, v)$ is the local histogram of a $w \times w$ window centering in (x, y), $f_c(u, v)$ is the cumulative function, u is the gray level in (x, y) and v is a gray-level variable. Then, AICE has introduced a signed power-law function, $q(u-v, \alpha)$, as shown in Eq.(2). The signed power-law function varies between a step function (when α=0, i.e. $q(u-v, 0)$) and an identity function (when α=1, $q(u-v, 1)$). Let $f_c(u, v) = q(u-v, 0)$, and then the transformation by Eq.(1) yields the effect as standard AHE. In the other hand, let $f_c(u, v) = q(u-v, 1)$, and then yields the effect of the local mean subtraction. In addition, let $f_c(u, v) = u - q(u-v, 1)$, and then yield the effect of the local mean. To combine the above effects, $f_c(u, v)$ is designed as Eq.(3). In the first term, $q(u-v, \alpha)$ provides the variation between the two effects of standard AHE (when α=0) and local-mean subtraction (when α=1). In the second term, $\beta[u - q(u-v, 1)]$ obtains β times of the image of local mean. α and β are defined in the range of [0,1]. By the controlling of α and β, the different enhancement strengths can be implemented. Stark found that the scheme of α=β often obtains effective results in AICE. By α=β, the parameters of AICE can be simplified into the only one parameter, and we select α to present α and β.

$$q(u-v, \alpha) = q(u-v, \alpha) = \frac{1}{2} \operatorname{sign}(u-v) \left| 2(u-v) \right|^{\alpha} \tag{2}$$

$$f_c(u, v) = q(u-v, \alpha) + \beta[u - q(u-v, 1)] \tag{3}$$

We conducted AICE on a portal image with different α values of 0.9, 0.8,…, 0.0, as shown in Fig. 3. The window size (w) for the determination of the local histogram is assigned 81. We found that smaller α is set, and more details outputs. The output details contain the bone structures and noise. When α<0.5, more noises are gradually amplified to influence visualization. The blurring artifact even occurs when α<0.2. Although AICE can obtain more contrast by decreasing α, the side effects grow worse.

In AICE, the tradeoff between the contrast and noise amplification becomes the most concern problem for the setting of α. The manual setting of α by watching and dragging, e.g. using a slider bar on an interactive interface is common in some applications of AICE. In our work, we propose an automatic scheme to choose α under proper noise amplification.

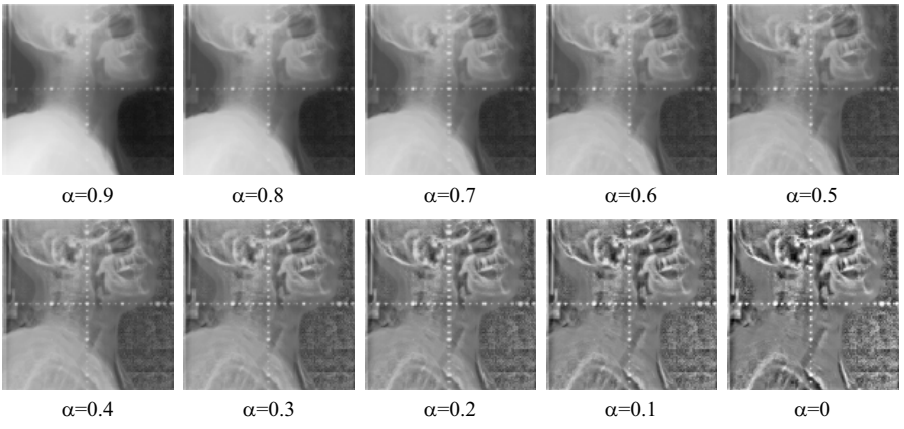


Fig. 3. Enhancement results of AICE under α values from 0.9 to 0

We use a gradient measure to evaluate the noise amplification. Sobel operation is first applied to compute the gradient magnitude of a pixel at (x, y), $\nabla I(x, y)$, as

$$\nabla I(x,y)=\sqrt{\left(G_x * I(x,y)\right)^2+\left(G_y * I(x,y)\right)^2},$$
$$\text{where } G_x=\frac{1}{4}\cdot\begin{bmatrix}1 & 0 & -1\\ 2 & 0 & -2\\ 1 & 0 & -1\end{bmatrix} \text{ and } G_y=\frac{1}{4}\cdot\begin{bmatrix}1 & 2 & 1\\ 0 & 0 & 0\\ -1 & -2 & -1\end{bmatrix}. \tag{4}$$

In Eq.(5), G_x and G_y are respectively the x-direction and y-direction masks, $*$ is the convolution operation and $I(x, y)$ is the enhanced image by AICE. Then, the summation of all gradient magnitudes of the image yields a Total Gradient (*TG*), as follow.

$$TG=\sum_{x,y}\nabla I(x,y) \tag{5}$$

In AICE, we can input a different α to obtain a different enhanced image, so that *TG* can be regarded as a function of α denoted by *TG*(α) which means a total gradient of an enhanced image with a α parameter. Fig. 4(a) shows a curve of *TG*(α) against α.

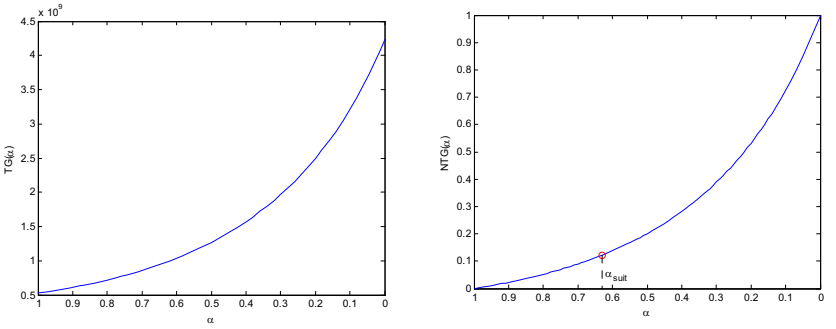
We found that the curve of *TG*(α) is a monotonically increasing and concave-down function along a decreasing α. When α decreases, the *TG*(α) increases and the rate of *TG*(α) also increases. It means that the rate of the noise amplification increases along a decreasing α. We take the characteristic curve of *TG*(α) to control the noise amplification in AICE. To normalize *TG*(α), we define a Normalized *TG*(α), *NTG*(*α*), to replace *TG*(α), as follow.

$$NTG(\alpha)=\frac{TG(\alpha)-TG(1)}{TG(0)-TG(1)}, \tag{6}$$

where *TG*(1) and *TG*(0) mean two *TG*s respectively under α=1 and α=0. Finally, we heuristically choose a suitable α where the derivative of *NTG*(α) reaches a threshold

(*Th*), as shown in Eq.(7). *Th* is experimentally assigned 0.5. Fig. 4(b) shows a curve of *NTG*(α) and the position of α_{suit}. By the determination of the suitable α, we can obtain the balance point in between the contrast and noise amplification in AICE.

$$\alpha_{suit} = \arg_{\alpha}\left(NTG'(\alpha) = Th\right) \tag{7}$$



(a) (b)

Fig. 4. Determination of a suitable α value: (a) *TG*(α) and (b) *NTG*(α) of (a)

After AICE, we found that the histogram of the output image narrows toward the center in the gray level. There are some bins having zero counts (unused bins) in the most right and most left sides in the gray level. To gain the more contrast, we find two limits of two sides for contrast stretch. We first specify the two percentages of the cumulative histogram to saturate at low and high intensities. The two percentages which are set in 0.5% and 99.5%, their corresponding gray levels are r_{min} and r_{max} respectively. Finally, the contrast stretch function is implemented as

$$I'(x,y) = \frac{I(x,y) - r_{min}}{r_{max} - r_{min}} \times (L-1), \tag{8}$$

where L–1 is the maximum of the output image, e.g. 2^{16}–1 for 16-bit images.

2.3 Noise Reduction

After the contrast enhancement, the noise reduction is often applied to obtain more smooth images [11]. Because our contrast enhancement performs in control of the noise amplification, we do not need a powerful and/or complicated algorithm in the procedure. Thus, we use a simple algorithm known by Adaptive Arithmetic Mean Filter [1]. The assumption of a Gaussian noise of zero mean is given in the input image. We lets the variance of the whole image, σ^2_{Noise}, be an estimate of the variance of noise. Let σ^2_{Local} be the variance of the pixel neighborhood. The noise-reduction filter can be implemented as

$$I'(x,y) = I(x,y) - \frac{\sigma^2_{\text{Noise}}}{\sigma^2_{\text{Local}}}\left[I(x,y) - \mu_{\text{Local}}\right], \tag{9}$$

where μ_{Local} is the mean of pixels in the neighborhood. Let $I(x,y) - \mu_{\text{Local}}$ be an estimate of noise. If a noise is present, $\sigma^2_{\text{Local}} \cong \sigma^2_{\text{Noise}}$ and then the noise will be eliminated by Eq.(9). If an edge is present, $\sigma^2_{\text{Local}} >> \sigma^2_{\text{Noise}}$ and then the edge will be remain.

3 Experimental Results

In our experiment, we select eight portal images captured from four bodyparts and/or four positions, denoted by No.(1)-(8). Their descriptions are as shown in Table 1. The linear accelerator is of Precise SLi made by LElekta and the a-Si silicon flat panel of EPID is of XRD 1640 made by Perkin Elmer. The photon energy was set up in 6MeV for all images. The capture images are stored into JPEG Lossless format. The specification is 1024×1024, 16-bit. The suitalbe α values for the eight images are automatically obtained by our suggestion mentioned in Subsection 2.2, as listed in 4th row of Table 1. Most images of the same body part and the same position are very similar. Therefore, in future, these suitable α values will be the useful references for the enhancement of other EPIs without repeating the determination of the suitable α value.

For comparison, we implement Koutsofios' method [5] and conduct it on the eight images. Koutsofios' method contains three procedures, standard AHE, median filter and Laplacian sharpening in sequential order. The standard AHE first enhances images in the contrast. Then, the median filter reduces the noise. Finally, the sharpening based on Laplacian operation is used to enhance the edges.

Fig. 5 shows the experimental results of our proposed method and Koutsofios' method. The results indicate the great improvement of less noise amplification and less blurring compared with Koutsofios' method. These artifacts of noise amplification and blurring often happens in the over enhancement in contrast such as standard AHE. Because we choose the suitable α value, AICE possibly prevents the happening of the over-enhancement in contrast. Moreover, many bone structures and important organ features which are invisible in the original images, are clearly displayed in our enhanced image.

Table 1. The description of eight portal images and their suitable α values for AICE

Image No.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Part	Abdomen	Abdomen	Chest	Chest	Head & neck	Head & neck	Pelvis	Pelvis
Position	AP	lateral	AP	lateral	AP	lateral	AP	lateral
α_{suit}	0.76	0.67	0.67	0.61	0.7	0.63	0.68	0.63

PS: AP means Anterior Posterior.

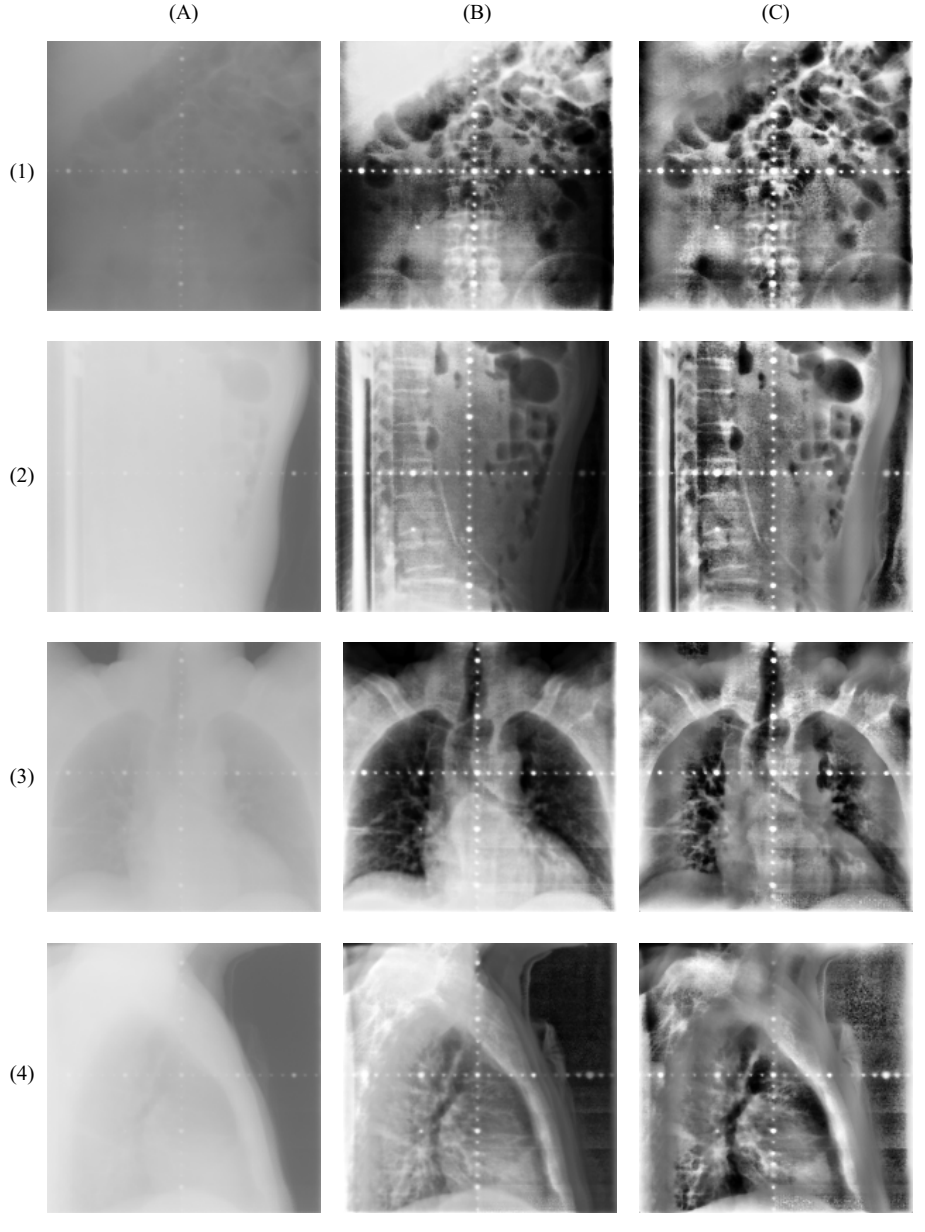


Fig. 5. Experimental results of image No.(1)-(8): original images in column (A), proposed method in column (B), and Koutsofios' method [4] in column (C)

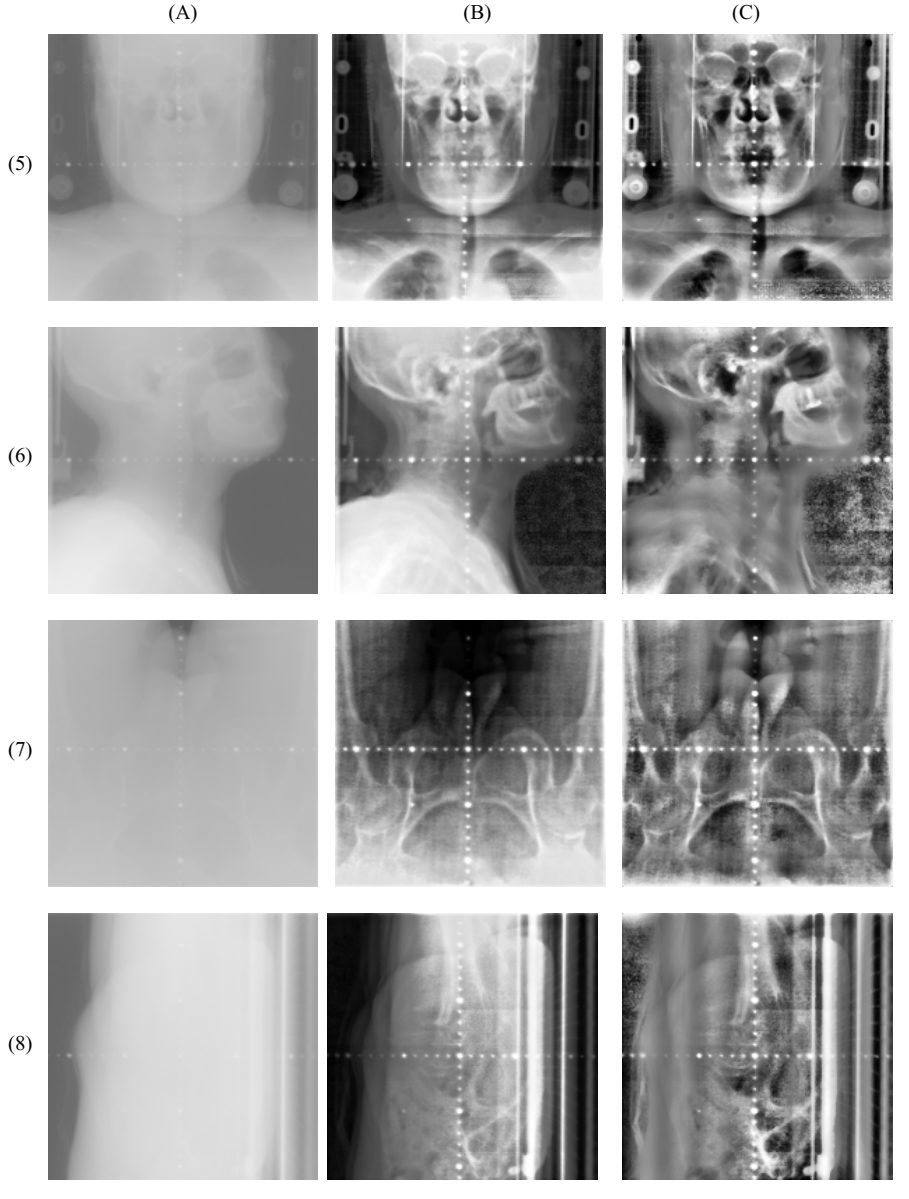


Fig. 5. (*continued*)

4 Conclusion

In this paper, we have presented an effective enhancement method for electronic portal images. We successfully integrate the two contrast enhancement algorithms of

GLG and AICE to produce a strong contrast but fewer artifacts in the enhanced images. Then, the adaptive arithmetic mean filter is applied to remove the enhanced noise. The experimental results indicate the great improvement on the visual perceptibility compared with the traditional method.

Acknowledgment

The authors would like to thank for the courtesy of Radiation Oncology in E-DA Hospital. This work was supported in part by National Science Counsel Granted NSC 98-2221-E-151-036-MY3 and NSC 98-2811-E-214-151-001.

References

[1] Dhawan, A.P.: Medical Image Analysis. John Wiley & Sons Inc., New Jersey (2003)
[2] Cumberlin, R.L., Rodgers, J.E., Fahey, F.H.: Digital image processing of radiation therapy portal films. Computerized Medical Imaging and Graphics 13(3), 227–233 (1989)
[3] Crooks, I., Fallone, B.G.: Contrast enhancement of portal images by selective histogram equalization. Med. Phsy. 20(1), 199–204 (1993)
[4] Shalev, S., Chen, D., Luchka, K.: A comparison of digital enhancement techniques for electronic portal images. Radiotherapy and Oncology 40, suppl. 1, S189 (1996)
[5] Koutsofios, K., Nikoletopoulos, S., Episkopakis, A., Kandarakis, I.: Sequential contrast enhancement of portal images: Study of the influence of image quality and clinical usefulness. In: Proc. of 2006 IEEE Nuclear Science Symposium Conference, vol. 5, pp. 2629–2631 (October-November 2006)
[6] Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B.T.H., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing 39(3), 355–368 (1987)
[7] Chen, Z.-Y., Abidi, B.R., Page, D.L., Abidi, M.A.: Gray-Level Grouping (GLG): An automatic method for optimized image contrast enhancement—Part I: The basic method. IEEE Trans. on Image Processing 15(8), 2290–2302 (2006)
[8] Stark, J.A.: Adaptive Image Contrast Enhancement Using Generalization of Histogram Equalization. IEEE Trans. on Image Processing 9(5), 889–896 (2000)
[9] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice-Hall Inc., New Jersey (2001)
[10] Fu, J.-C., Lien, H.-C., Wong, S.T.C.: Wavelet-based histogram equalization enhancement of gastric sonogram images. Computerized Medical Imaging and Graphics 24(2), 59–68 (2000)
[11] Scharcanski, J., Jung, C.R.: Denoising and enhancing digital mammographic images for visual screening. Computerized Medical Imaging and Graphics 30, 243–254 (2006)

License Plate Tilt Correction Based on the Straight Line Fitting Method and Projection

Kaushik Deb, Andrey Vavilin, Jung-Won Kim, and Kang-Hyun Jo

Intelligent Systems Lab., Dept. of EE and Information Systems, University of Ulsan
Daehak road 100, Nam-gu, 680-749 Ulsan, South Korea
{debkaushik99,andy,jwkim,jkh2010}@islab.ulsan.ac.kr

Abstract. Tilt correction is an integrant part of the automatic vehicle license plate recognition (VLPR) system. In this paper, according to the least square fitting with perpendicular offsets (LSFPO) the VLP region is fitted to a straight line. After the line slope is obtained, rotation angle of the VLP is estimated. Then the whole image is rotated for tilt correction in horizontal direction by this angle. Tilt correction in vertical direction by minimizing the variance of coordinates of the projection points is proposed. Despite the success of VLP detection approaches in the past decades, a few of them can effectively locate license plate (LP), even when vehicle bodies and LPs have similar color. A common drawback of color-based VLP detection is the failure to detect the boundaries or border of LPs. In this paper, we propose a modified recursive labeling algorithm for solving this problem and detecting candidate regions. While conducting the experiment, vehicle images taken under various conditions from traffic stations to evaluate the robustness, the flexibility and effectiveness.

Keywords: Least square fitting with perpendicular offsets (LSFPO), minimum variance, recursive labeling algorithm, shear transform, tilt correction, and vehicle license plate (VLP).

1 Introduction

As license plates can appear at many different angles to the camera's optical axis, each rectangular candidate region is rotated (i.e. correcting tilt) until they are all aligned in the same way before the candidate decomposition. License plate tilt correction and detection are crucial and indispensable components of the character segmentation and automatic recognition of the VLP. One of the major problems in LP tilt correction and detection is determining LP systems. This system must guarantee robust detection under various weather and lighting conditions, independent of orientation and scale of the plate.

As far as tilt correction and detection of the license plate region are concerned, researchers have found various methods of correcting tilt and locating license plate. For example, Karhunen-Loeve (K-L) transformation method has been introduced for correcting a VLP tilt in [1]. However, no explanation of extracting

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 184–193, 2010.

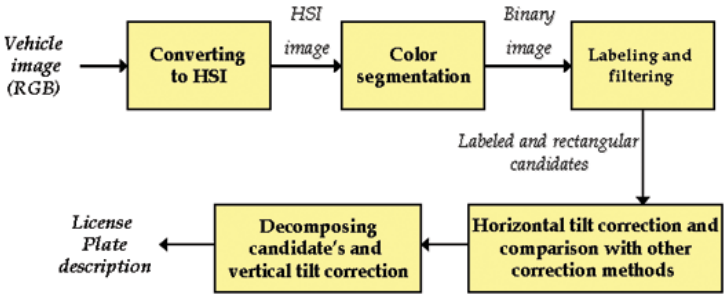


Fig. 1. Proposed vehicle license plate tilt correction and detection framework

LP region has been given in the paper. A region-based LP detection method has been presented in [2], which first applies a mean shift procedure in a spatial-range domain to segment a color vehicle image in order to get LP regions. Fuzzy logic has been applied in detecting license plates in [3].

The emphasis of this paper is on the implementation of a line fitting method based on LSFPO for correcting a VLP tilt in horizontal direction. Tilt correction in vertical direction by minimizing variance of coordinates of the projection points is propose and implement. Horizontal tilt correction performance of LSFPO is evaluated in comparison with other representative method such as, the least square fitting with vertical offsets (LSFVO) of [1]. This paper explores the line fitting method based on LSFPO that outperforms other correction method because of faster processing time, more precise tilt angle, and easily implemented.

2 Proposed Framework

In the author's previous work [4], VLP extraction based on color and geometrical features was presented. We propose in this section, an enhanced version of the framework for VLP tilt correction and detection as shown in Figure 1. To improve the traditional LP detection method, as license plates can appear at many different angles to the camera's optical axis, each rectangular candidate region is rotated until they are all aligned in the same way before the candidate decomposition. For correcting a VLP tilt in horizontal direction, a line fitting method based on LSFPO is introduced. Horizontal tilt correction performance of LSFPO is evaluated in comparison with other correction method such as line fitting based on LSFVO [1]. A common drawback of color-based VLPD is the failure to detect the boundaries or border of LPs. This occurs when vehicle and LP have similar colors. It is important to mention here that some previous researches [2] and [3] doesn't solve this problem and they assert leave these issues to be considered in future study. To overcome this common drawback, we propose and implement a new method named as modified recursive labeling algorithm. Finally, includes performances for candidate's decomposition by using position

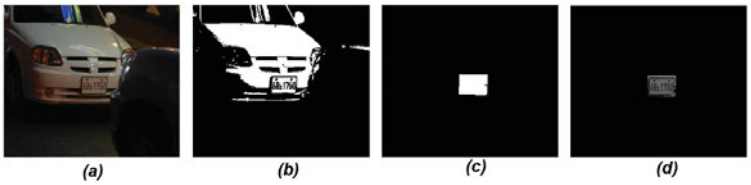


Fig. 2. Successful license plate identification sequence in an unmoving vehicle: (a) an LP images, when vehicle and LP have similar colors, moreover an LP image in the nighttime and also a obstacle located in front of vehicle, (b) color segmentation results, (c) detected candidate after implementation of morphological closing operation and filtering, and (d) extracted candidate after tilt correction in horizontal direction

in the histogram to verify and detect VLP region and vertical tilt correction. A vertical tilt correction method by minimizing variance of coordinates of the projection points is also proposed in the framework.

3 VLP Tilt Correction and Detection Module

In this section, the four primary stages of the proposed VLP tilt correction and detection framework, i.e., color segmentation, labeling and filtering, correcting tilt in horizontal direction, and decomposing candidates and correcting tilt in vertical direction have been discussed in details. Color arrangement of the Korean LPs are well classified. A More detailed explanation for color arrangement and outline of the Korean VLPs could be found in [4].

3.1 Color Segmentation

Color is a distinctive feature because the law decides its usage of the vehicle according to color. Representation of plate color in an invariant manner is of main objectives for our color-based LP detection method. In the proposed framework, input vehicle images are converted into a Hue-Saturation-Intensity (HSI) color images. Then the candidate regions are found by using HSI color space on the basis of using hue, saturation and/or intensity. Many applications use the HSI color model. Machine vision uses HSI color model in identifying the color of different objects. Plate color information is used to detect candidate regions in our experiments, and shape properties of LP allow reducing number of LP-like candidates. A More detailed explanation could be found in [4] for detecting green, yellow, and white license plate pixels. Color segmentation parameters are very sensitive in order to detect as much candidates as possible. All false candidates will be filtered out on the next stages. Examples of proposed color segmentation method is depicted in Figure 2.

3.2 Labeling and Filtering

In the proposed method, a recursive algorithm is implemented for connected component labeling operation. A common drawback of color-based VLPD is the failure to detect the boundaries or border of LPs. This occurs when vehicles and LPs have similar colors. To overcome this common drawback, we proposed and implemented a new method named modified recursive labeling algorithm. If we investigate carefully, when vehicle bodies and LPs have a similar color, we can find there is little color differences between LPs and vehicle color. Based on this idea, we overcome this problem, by trying to find those color difference parameters. To label connected pixels 4-neighbors recursive algorithm was used. Furthermore, connected pixels were grouped if distance in color space was less then predefined threshold D_{min}. Two connected pixels are grouped if $Dist(I_{i,j}, I_{m,n}) < D_{min}$. Color distance between two connected pixels $I_{i,j}$, $I_{m,n}$ is calculated using equation 1.

$$Dist\left(I_{i,j}, I_{m,n}\right) = \sum_{k=\{R,G,B\}} \left|I_{i,j}^{k} - I_{m,n}^{k}\right| \tag{1}$$

where $I_{i,j}^{k=\{R,G,B\}}$ and $I_{m,n}^{k=\{R,G,B\}}$ are pixel belongs to red, green, and blue components of input image I with neighbor coordinate (i, j) and (m, n), respectively. In this step we extract candidate regions that may include LP regions from the binary mask obtained in the previous step. During this step, main geometrical properties of LP candidate such as area, bounding box, and aspect ratio are computed. A more detailed explanation could be found in [4]. These parameters are used in the filtering operation to eliminate LP-like objects from candidate list. Figure 2 portrays the steps for LP segmentation.

3.3 Correcting Tilt

Correcting horizontal tilt by straight line fitting method based on LSFPO: In figure 3(a, b) depicts rotation angle α between the principal axis X' of and the the horizontal axis X of the horizontal tilt VLP region. The least square method minimizes the summed square of residuals or offsets. The residual for the i-th data point d_i is defined as the difference between observed response value y_i and the fitted response value $\hat{y}_i$ and is identified as the error associated with the data. The residuals of the best-fit line for a set of n points using unsquared perpendicular distances d_i of points $(x_i,\ y_i)$ are given by $G = \sum_{i=1}^{n} d_i$. The perpendicular distance from a line $y = ax + b$ to point i is given by $d_i = \frac{|y_i - (ax_i + b)|}{\sqrt{1+a^2}}$. The objective function to be minimized is $G = \sum\limits_{i=1}^{n} \frac{|y_i - (ax_i + b)|}{\sqrt{1+a^2}}$. The absolute value function does not have continuous derivatives, minimizing G is not amenable to analytic solution. However, if the square of the perpendicular distances $G^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \frac{[y_i - (ax_i\ +\ b)]^2}{1+a^2}$ is minimized instead, the problem can be in closed form. G^2 is a minimum when

$$\frac{\partial G^2}{\partial a} = \frac{2}{1+a^2} \sum_{i=1}^{n} (-x_i) \left[y_i - (ax_i\ +\ b)\right] + \sum_{i=1}^{n} \frac{-2a \left[y_i - (ax_i\ +\ b)\right]^2}{(1+a^2)^2} = 0. \tag{2}$$

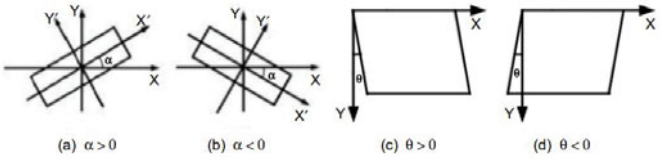


Fig. 3. VLP tilt image in horizontal direction: (a) tilt angle $\alpha > 0$ and (b) tilt angle $\alpha < 0$. VLP tilt image in vertical direction: (c) tilt angle $\theta > 0$ and (d) tilt angle $\theta < 0$.

Solving the above equation, a is obtained:

$$A = \frac{1}{2} \frac{\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)}{n\bar{x}\bar{y} - \sum_{i=1}^{n} x_i y_i}; a = -A \pm \sqrt{A^2 + 1} \tag{3}$$

where $\overline{x}$ and $\overline{y}$ indicates the mean value of x_i and y_i, respectively.

Correcting horizontal tilt by straight line fitting method based on LSFVO: Given a set of data points. It is desired to find the best fitting line from a given set of data points. In principle, deviation between data and fitting line should be minimized. The deviation $d_i = y_i - \hat{y}_i$ is commonly called residue. The vertical distance from a line $y = ax + b$ to point i is given by $d_i = [y_i - (ax_i + b)]$. LSFVO enable the sum of the d_i^2 to achieve the minimum, namely objective function $G^2 = \sum_{i=1}^{n} d_i^2$ is the minimum. The objective function is $G^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$. From the objective function, partial derivative with respect to a is $\frac{\partial G^2}{\partial a} = -2\sum_{i=1}^{n} [y_i - (ax_i + b)](x_i) = 0$. Solving the above equations, a is obtained as

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{4}$$

x_i and y_i are inserted into (3) and (4), to get the fitting slope a. Let $tan\ \alpha = a$, and get the tilt angle α. Rotate the entire image with α from centroid rectangular candidates region of LP image and perform the tilt correction. Figures 4(c1-d1) and 4(c2-d2) portrays a sequence of successful horizontal tilt correction by LSFPO and LSFVO, respectively.

3.4 Decomposing Candidates and Correcting Vertical Tilt

Candidate decomposition: Information is extracted from the image by intensity histograms that play a basic role in image processing, in areas such as enhancement, segmentation and description. In this section, verification and detection of the VLP region as well as character segmentation are considered and discussed. Once the candidate area is binarized the next step is to extract the

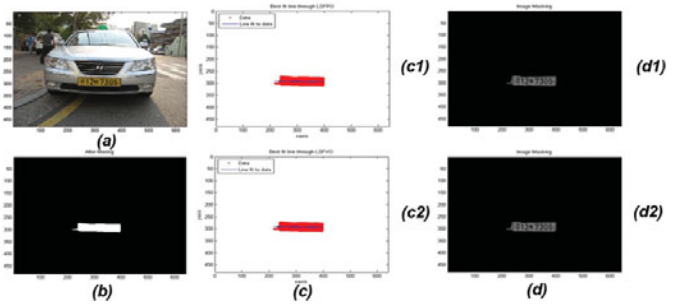


Fig. 4. Illustration of license plate segmentation : (a) an LP image, (b) detected candidate after filtering, (c) finding best fitting line through LSFPO (c1) and LSFVO (c2), and (d) extracted candidate after horizontal tilt correction, respectively

information. At first, regions without interest such as border or some small noisy regions are eliminated; the checking is made by height comparison with other plate characters height. Figure 6 shows the results for verifying predetermined alphanumeric characters.

Vertical tilt correction: After decomposing candidates, the purpose of correcting tilt in vertical direction is to correct VLP shear left and right in vertical direction as shown in Figure 3(c, d) and recognize VLP character accurately.

Let, an LP image is located at point A of coordinates (x, y), will be moved to point A' of coordinates (x', y). The angle between points A and A' is θ degree as shown in Figure 5. This counter clock wise rotation around the vertical direction known as shear transform. On the other hand, point A's clockwise rotation around the vertical direction creates an angle of $-\theta$ degree. Figure 5 depicts shear transformation in x-direction can be written by $x' = x - y \cdot tan\theta$.

When a single character is not vertically tilted, the vertical projection of the character points are distributed intensively in a small range. Thus, the variance is minimal. On the contrary, if a single character is vertically tilted, the character points are widely distributed in a wide range. After investigating this feature, we can say that character point distribution of the vertical projection take the minimum variance when a single character is not in vertical angle of inclination. Let the number of character points N and their coordinate are (x_i, y_i), $i = 1, 2, \ldots N$. Through the shear transformation, each character point (x_i, y_i) is moved to (x'_i, y_i) then the variance of projection point is as follows:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left(x'_i - \frac{1}{N} \sum_{k=1}^{N} x'_k \right)^2 \tag{5}$$

The value of x' is put into equation (5), then the variance of projection point is defined as follows:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left[\left(x_i - \frac{1}{N} \sum_{k=1}^{N} x_k \right) - \left(y_i - \frac{1}{N} \sum_{k=1}^{N} y_k \right) \right.$$

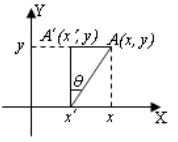


Fig. 5. Schematic diagram for shear transformation

$$\tan\theta]^2 = \frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x}) - (y_i - \bar{y})\tan\theta\right]^2 \tag{6}$$

where $\overline{x}$ and $\overline{y}$ indicates the mean value of x_k and y_k, respectively.

$$if, \begin{cases} u_i = (x_i - \bar{x}) \\ v_i = (y_i - \bar{y}) \end{cases} \tag{7}$$

From equation (7) the value of u_i *and* v_i are put into equation (6), then the variance of the projection point is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left[u_i - v_i.\tan\theta\right]^2. \tag{8}$$

In order to minimize the variance of the projection point (σ^2), the partial derivative of θ is calculate as

$$\frac{\mathrm{d}\sigma^2}{\mathrm{d}\theta} = -\frac{2\sec^2\theta}{N}\sum_{i=1}^{N}\left[u_i - v_i.\tan\theta\right]v_i = 0. \tag{9}$$

As $sec^2\theta \neq 0$, therefore $\sum_{i=1}^{N}\left[u_i - v_i.\tan\theta\right]v_i = 0$. The vertical tilt angle (θ) is computed as follows:

$$\theta = tan^{-1}\frac{\sum_{i=1}^{N} u_i v_i}{\sum_{i=1}^{N} v_i^2}. \tag{10}$$

If we analyze the VLP region carefully, we can find that the vertical tilt direction or angle of characters and the entire LP is basically the same. So we can find various sub-regional average of the vertical tilt angle of the region as a whole, i.e. the vertical tilt angle of the plate. Based on this idea, after character segmentation, we can calculate the vertical tilt angle of individual character through equation (10). Supposed that there are N elements after segmentation of LP region, and then the mean and the standard deviation for individual segmented regions or elements are defined by,

$$\mu_\theta = \frac{\sum_{i=1}^{N}\theta_i}{N};\ \sigma_\theta = \sqrt{\frac{\sum_{i=1}^{N}\left(\theta_i - \mu_\theta\right)^2}{N}} \tag{11}$$

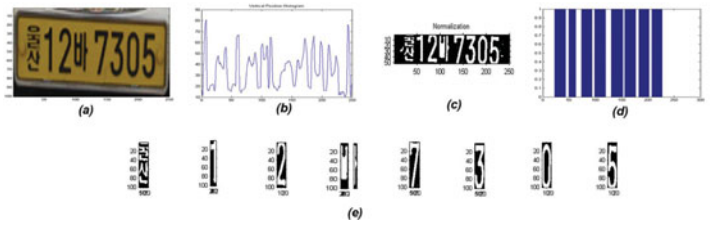


Fig. 6. Steps for verifying predetermined alphanumeric characters: (a) extracting the candidate region, (b) vertical position histogram with LP border, (c) view of normalization candidate region after removing border and noisy area, (d) vertical position histogram (eight peaks for predetermined eight alphanumeric characters in LP region), and (e) character extraction after correcting vertical tilt

Table 1. Comparison of detection rates

Reference number	Detection rate	Detection method
5	80.4%	MM
Proposed framework	100%	LSFPO method and MPP

Whenever VLP regions contain some noises, for example hyphen, border area, and plate fixing dots etc. They cannot exactly reflect vertical tilt angle. For this reason, filtering the character region is necessary and the filtering rule is described as follows:

$$\begin{cases} \theta_i \ is\ retained\ if\ \ |\theta_i - \mu_\theta| < \sigma_\theta \\ \\ \theta_i \ is\ removed \qquad otherwise \end{cases} \tag{12}$$

Averaging all remaining θ_i, we can find the vertical tilted angle (θ) of a whole plate. The image can be corrected in vertical direction through this angle. The corrected result is depicted in Figure 6(e).

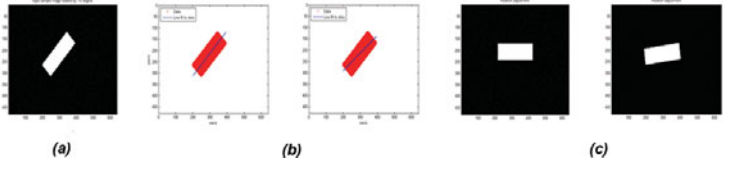


Fig. 7. Illustration of finding best fitting line and principal axis from sample image: (a) input image, (b) finding best fitting line through LSFPO (b1)and LSFVO (b2), and (c) extracted candidate after rotation adjustment

4 Experimental Results and Conclusions

All experiments were done on Pentium-IV 2.4 GHz with 1 GB RAM under MATLAB environment. Images of size 640*480 and 320*240 pixels were used. The image database consists of 200 digital images from different two groups. In order to evaluate the proposed framework, two groups of experiment were conducted. First group was used to compare the proposed framework (LSFPO method and MPP) with mathematical morphology (MM) method of [5]. A comparison between the proposed method and a well-reported method in the literature is given in Table I. From Table I it can be seen that the proposed method outperforms the method repot in [5] from the detection rate points of view.

The second group contains 175 images. All images in that group represents South Korean license plates from the natural scenes obtained in the nonuniform outdoor illumination conditions, multi-style and color of license plates, and various angles of vision. They were taken in distance of 3 to 15 m. Under these conditions, the success of LP detection has reached to more than 96%.

In tilt correction experiments, we compare the tilt performance results of LSFPO with those of LSFVO. For comparing tilt performance, initially we take 9 sample images, where rectangular candidate region is rotated in various angle. Figure 7 illustrates the steps for finding best fitting line and principal axis from among one sample image, when rectangular candidate region rotated by 45°. In this implementation, after obtaining line slope and principal axis, rotation angle error is estimated as follows through LSFPO (0.0200°) and LSFVO (6.9187°), respectively. Experimental result indicate that, less tilt error and the top and bottom line are basically horizontal in Figure 7($c1$). However, 7($c2$) (i.e. LSFVO), have some tilt, and the rotation angle is either small or large. The average processing time for tilt estimation are LSFPO (0.0146 s) and LSFVO (0.0174 s), respectively. The average computational time for rotation adjustment are LSFPO (0.2207 s) and LSFVO (0.2209 s), respectively.

To overcome common drawbacks, in this paper, we implemented a modified recursive labeling algorithm. In our experiments, in 27 images vehicle bodies and LPs have similar color. Among them in 26 images, LPs were detected successfully. Figure 2 shows successful plate identification, where vehicle bodies and their LPs possess similar colors. The average computational time for the color segmentation and filtering operations of the proposed method are 0.16 and 0.07 s, respectively.

In conclusion, a method is adopted in this paper for correcting tilt which is a very crucial part of the VLP automatic recognition. In the vehicle horizontal tilt correction process, two correction methods are implemented for comparing the tilt performance results. Analysis and simulation results suggest that LSFPO outperforms than LSFVO because of faster computational time, easily implemented and more precise tilt correction.

In this paper, we proposed and executed of a method by minimizing variance of coordinates of the projection points for correcting tilt in vertical direction. In addition, the emphasis of this paper is on the implementation of a new method to detect candidate regions when vehicle bodies have similar color. Finally, VLP

regions containing predetermined alphanumeric character that are verified and detected by using position in the histogram. Color arrangement and predetermined alphanumeric character of the Korean license plate are important features for verification and detection of license plate regions. While conducting the experiments, different view point, illumination conditions, and varied distances between vehicle and camera often occurred. In such cases, confirmed the result is very effective when the proposed method is used. However, the proposed method is sensitive with motion blur in the input image. We leave these issues for consideration in future studies.

Acknowledgments. This work was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-C7000-1001-0007).

References

1. Pan, M.-S., Xiong, Q., Yan, J.-B.: A new method for correcting vehicle license plate tilt. Int. J. of Automation and Computing 6(2), 210–216 (2009)
2. Jia, W., Zhang, H., He, X.: Region-based License Plate Detection. J. Network and Comput. Applications 30(4), 1324–1333 (2007)
3. Chang, S.-L., Chen, L.-S., Chung, Y.-C., Chen, S.-W.: Automatic license plate recognition. IEEE Trans. Intell. Transp. Syst. 5(1), 42–53 (2004)
4. Deb, K., Lim, H., Jo, K.-H.: Vehicle license plate extraction based on color and geometrical features. In: IEEE International Symposium on Industrial Electronics, pp. 1650–1655. IEEE press, Los Alamitos (2009)
5. Martin, F., Garcia, M., Alba, J.L.: New Methods for Automatic Reading of VLP's (Vehicle License Plates). In: Proceedings of the IASTED Int. Conf. on SPPRA (2002)

Differential Approximation of the 2-D Laplace Operator for Edge Detection in Digital Images

Jakub Pęksiński and Grzegorz Mikołajczak

West Pomeranian University of Technology,
Faculty of Electrical Engineering
26 Kwietnia 10, 71-126 Szczecin, Poland
jakub.peksinski@zut.edu.pl, grzegorz.mikolajczak@zut.edu.pl

Abstract. The paper presents the application of finite difference edge detection in digital images. For edge detection we use several methods. One of them is a method based on Laplace operator. This paper presents a differential approximation of the two-dimensional Laplace operator. The paper proposes a differential approximation, Laplace operator, based on 9-th lattice mask. Coefficients were determined using the Z transform. Optimization is based on the criterion of maximal compatibility differential approximation of Laplace filter with ideal. Mask parameters were chosen based on the analysis of the error function. Activity obtained filter has been tested on a digital image that contains many elements of geometry.

Keywords: edge detector, differential approximation.

1 Introduction

An important problem of the image processing is cambering and detection of edges applied in many fields including cartography and automatic classification of objects in the image. The process of edge detection reduces the image only to the edges included in it. All operations for edge detection suppress the low-frequency image elements. For such purpose, the Laplace operator featuring multi-directionality (detects edges in all directions) is applied in the digital image processing. The image obtained by implementing such a method has sharper edges as compared with other methods [1]. The Laplace operator of the two-dimensional function *f(x,y)* can be expressed by the following formula:

$$L[f(x,y)] = \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2} \tag{1}$$

As for the discrete function applied to digital images, the second partial derivatives can be approximated by differences of the second order; therefore the following expression is obtained:

$$\begin{aligned} L[(x,y)] &\approx -4f(x,y) + f(x+1,y) + \dots \\ &\dots + f(x-1,y) + f(x,y+1) + f(x,y-1) \end{aligned} \tag{2}$$

The above mentioned expression can be presented as the mask of coefficients (14).

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 194–199, 2010.

2 Differential Approximation of the Laplace Operator

The solution of the Laplace equation (1), i.e. $L[f(x,y)]=0$, by means of the finite difference method, and based on 9 nodes of the square mesh, is featured in the paper. The coordinates of nodes of the mesh and their corresponding function values are presented in fig. 1.

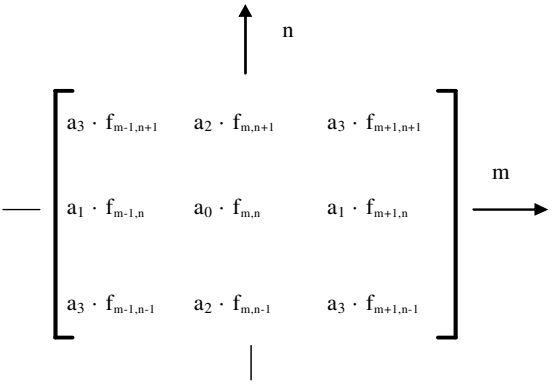


Fig. 1. Coordinates of elements in the 3x3 Laplace filter mask

The differential operator in the central mask point (m, n) is formulated as follows:

$$\begin{aligned} L[f_{m,n}] &= a_0 f_{m,n} + a_1(f_{m+1,n} + f_{m-1,n}) + \dots \\ &\dots + a_2(f_{m,n+1} + f_{m,n-1}) + \dots \\ &\dots + a_3(f_{m+1,n+1} + f_{m-1,n+1} + f_{m-1,n-1} + f_{m+1,n-1}) \end{aligned} \tag{3}$$

The problem is solved through calculation of the following coefficients: a_0, a_1, a_2, a_3. The solution can be obtained by means of a Taylor series expansion of a function of two variables, undetermined coefficient method and transmittance of the 2-D digital filter. The third of the above-mentioned methods has been applied in the paper and each coefficient has been determined on the basis of the condition of the maximum flat characteristics.

The 2-D transform Z of the sequence $\{f_{m,n}\}$ can be expressed as follows [2]:

$$F(z_1, z_2) = \sum_{m=0}^{M} \sum_{n=0}^{N} f_{m,n} z_1^{-m} z_2^{-n} \tag{4}$$

The following formula is obtained after performing transformation Z in relation to the differential operator (3):

$$L(z_1, z_2) = \sum_{m=0}^{M} \sum_{n=0}^{N} L[f_{m,n}] z_1^{-m} z_2^{-n} \tag{5}$$

On the basis of the formula (3), (4) and (5) it can be expressed that:

$$\begin{aligned} L(z_1, z_2) &= F(z_1, z_2) \cdot [a_0 + a_1(z_1 + z_1^{-1}) + \dots \\ &\dots + a_2(z_2 + z_2^{-1}) + \dots \\ &\dots + a_3(z_1 z_2 + z_1^{-1} z_2 + z_1^{-1} z_2^{-1} + z_1 z_2^{-1})] \end{aligned} \tag{6}$$

After the substitution is performed:

$$z_1 = e^{j\omega_1} \qquad z_2 = e^{j\omega_2} \tag{7}$$

the following transmittance of the digital filter based on the differential operator (3) is obtained:

$$\begin{aligned} H(\omega_1, \omega_2) &= \frac{L(e^{j\omega_1}, e^{j\omega_2})}{F(e^{j\omega_1}, e^{j\omega_2})} = \\ &= a_0 + 2a_1 \cos\omega_1 + 2a_2 \cos\omega_2 + 4a_3 \cos\omega_1 \cos\omega_2 \end{aligned} \tag{8}$$

The transmittance of the perfect Laplace operator for the accepted node mesh (fig.1) can be expressed by the following formula:

$$HL(\omega_1, \omega_2) = -(\omega_1^2 + \omega_2^2) \tag{9}$$

In order to obtain the coefficients of the differential approximation (3), the cosine function is expanded into the power series in the transmittance (8). Then, the transmittance can be presented as follows:

$$\begin{aligned} &H(\omega_1, \omega_2) \approx a_0 + 2a_1 + 2a_2 + 4a_3 - \omega_1^2(a_1 + 2a_3) - \omega_2^2(a_2 + 2a_3) + \dots \\ &\dots + \frac{\omega_1^4}{12}(a_1 + 2a_3) + \frac{\omega_2^4}{12}(a_2 + 2a_3) + \omega_1^2\omega_2^2 a_3 + \dots \end{aligned} \tag{10}$$

By means of the criterion of transmittance consistence (10) and the perfect transmittance (9), the following relations between every coefficient a_0, a_1, a_2, a_3 can be determined:

$$\begin{aligned} &a_0 + 2a_1 + 2a_2 + 4a_3 = 0; \\ &a_2 + 2a_3 = 1; \\ &a_1 + 2a_3 = 1; \end{aligned} \tag{11}$$

Assuming that $a_3=p$ (the parameter), the solution of the system of equations can be presented as follows:

$$\begin{aligned} &a_0 = -4 + 4p; \quad a_1 = 1 - 2p; \\ &a_2 = 1 - 2p; \quad a_3 = p; \end{aligned} \tag{12}$$

In the digital processing, the calculation of the value of a point in a target image is performed through adding up of the elements of an image with the appropriate weights around the processed point. Thereby, the filter can be defined as an array (mask) of the weight coefficients:

$$L = \begin{bmatrix} p & 1-2p & p \\ 1-2p & -4+4p & 1-2p \\ p & 1-2p & p \end{bmatrix} \tag{13}$$

The Laplace filter (3) known from the literature and based on five points can be obtained for $p=0$, whereas it is featured in equation (2) that can be presented as the mask of the following coefficients:

$$L0 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{14}$$

The function of the error determined as a difference between the perfect transmittance of the Laplace filter (9) and the approximate transmittance (10) is used for the selection of the parameter p. The error including the lowest powers of the circular frequency ω_1, ω_2 can be expressed by the following formula:

$$L0 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{15}$$

Considering that the following inequality is valid for any two real numbers $a,b \in R$:

$$(a^2+b^2)^2 \geq a^4+b^4 \geq (a^2-b^2)^2 \tag{16}$$

the following three values of the parameter p (12) for which the error (15) is expressed as follows, have been analysed:

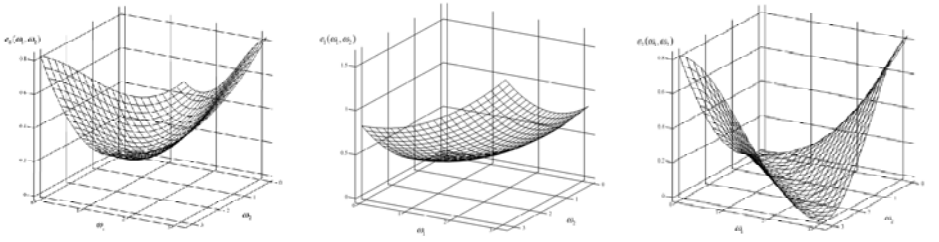
$$\begin{aligned} p_0 &= 0 \quad B0(\omega_1,\omega_2) = \tfrac{1}{12}(\omega_1^4+\omega_2^4) \\ p_1 &= \tfrac{1}{6} \quad B1(\omega_1,\omega_2) = \tfrac{1}{12}(\omega_1^2+\omega_2^2)^2 \\ p_2 &= -\tfrac{1}{6} \quad B2(\omega_1,\omega_2) = \tfrac{1}{12}(\omega_1^2-\omega_2^2)^2 \end{aligned} \tag{17}$$

2 Analysis of the Results

On the basis of the relations (8) and (9), the relative error for each transmittance has been determined depending on the frequency for the fixed parameter p:

$$e(\omega_1,\omega_2)=\left|\frac{H(\omega_1,\omega_2)-HL(\omega_1,\omega_2)}{HL(\omega_1,\omega_2)}\right| \tag{18}$$

The error characteristics (18) for are presented in fig. 2. The analysis of characteristics in fig. 2 proves that the error is the lowest within the range of circular frequency for p_2 ($\omega_1=\omega_2$ - practically equal to zero fig.2c).



a) b) c)

Fig. 2. Error characteristics (18) for the fixed value of p: a) e_0 for $p=0$, b) e_1 for $p=1/6$, c) e_2 for $p=-1/6$

The test images (fig. 3) presenting geometrical figures of various edge thickness filled with grey shades has been generated for the visual evaluation of the obtained results.

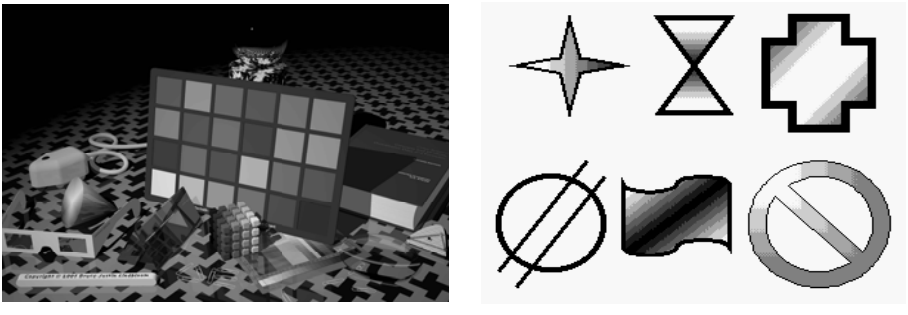
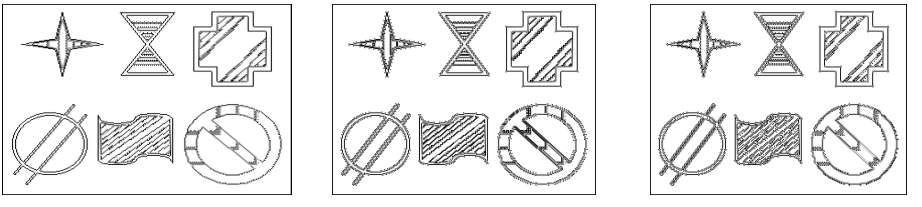


Fig. 3. Test images featuring 256 grey shades and geometrical figures

The images has been subject to filtration by means of the determined Laplace filter (14) for three parameters p (17) which effects are featured in the figure 4 and 5.

a) b) c)

Fig. 4. Test image after filtration by means of the Laplace filter (13) for the fixed value of p: a) for $p=0$, b) for $p=1/6$, c) for $p=-1/6$



a) b) c)

Fig. 5. Test image after filtration by means of the Laplace filter (13) for the fixed value of p: a) for $p=0$, b) for $p=1/6$, c) for $p=-1/6$

Making the analysis of the images, i.e. fig. 4 and 5, it can be concluded that for $p=-1/6$ the edges of figures after filtration are sharper than in other cases. Also, this filter reacts better to any change in grey shade inside the figures.

References

1. Pratt, W.K.: Digital Image Processing, 4th edn. Wiley, New York (2007)
2. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing, 2nd edn. Prentice Hall Signal Processing Series (1999)
3. Dahlquist, G., Bjorck, A.: Numerical Methods. Prentice-Hall, Englewood Cli&s (1974)
4. Ryzyk, I.M., Gradsztejn, I.S.: Tables of integrals. sums and series. PWN, Warsaw (1964)
5. Heath, M., Sarkar, S., Sanocki, T., Bowyer, K.: Comparison of Edge Detectors A Methodology and Initial Study. Computer Vision and Image Understanding 69(1), 38–54 (1998)

ARToolkit-Based Augmented Reality System with Integrated 1-D Barcode: Combining Colorful Markers with Remote Servers of 3D Data for Product Promotion Purposes

Jong-Chih Chien[1], Hoang-Yang Lu[2], Yi-Sheng Wu[3], and Li-Chang Liu[4]

[1] Dept. of Computer Science and Information Engineering, Kainan University
Taoyuan County, Taiwan
jcchien@mail.knu.edu.tw
[2] Dept. of Electrical Engineering, National Taiwan Ocean University
Keelung, Taiwan
HYLU@mail.ntou.edu.tw
[3] Dept. of Indust. & Busi. Management, Yung Ta Institute of Technology & Commerce
Pingtung, Taiwan
yswu@mail.ytit.edu.tw
[4] Simpleact Incorporated
Taipei, Taiwan
lcliu2000@simpleact.com.tw

Abstract. In this paper we propose a color-markered augmented reality system with integrated 1-D barcode reader for product promotion purposes. The barcode on each product becomes the key used to access remote databases and retrieves designated marker pattern and 3D model data. In this way, the association between marker and 3D model data becomes decoupled from the application itself, and become linked to the barcode instead. So because of the addition of barcode reader, the number of available markers and 3D models vastly increase at runtime so now our system can be practically used for commercial advertisement and product promotions. But to truly make our system accepted by the populace, the traditional ARToolkit black-and-white marker patterns, which are not visually exciting enough, must be replaced with colorful markers. So we tested several colorful point stickers provided by convenience stores, such as 7-11, in Taiwan as alternative markers. Experiments were performed to test how these stickers can be used as colorful markers for our AR system, and our results show that even though not all colorful stickers can be used as markers in their current composition of background and foreground images, but certain stickers can already be used as stable markers without any modification. The results of our experiments are presented.

Keywords: ARToolkit, Augmented Reality, 1-D Barcode, Colored Marker, Product Promotion.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 200–209, 2010.

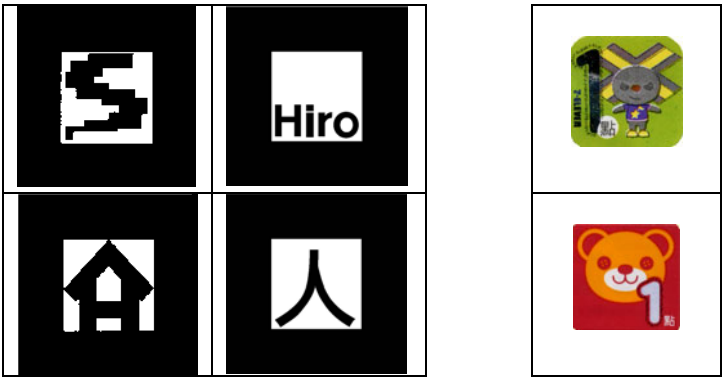
1 Introduction

Since its inception, the ARToolkit library has been widely used in various marker-based augmented reality applications[1-6]. But the GPL-licensed, non-commercial version is limited in that the association between AR marker and 3D models has to be pre-defined in the applications and can not be built at runtime, which is undesirable for non-commercial promotional advertisement applications. Another drawback is that it uses black-and-white marker which is also a hindrance to building AR applications that are geared toward commercial advertisement. An AR system for product advertisement should allow the manufacturers the flexibility to update their AR markers and 3D models at any time. We resolved the need to pre-define markers by integrating a 1-D barcode reader with ARToolkit. With this integration, the end-users can just scan and decode 1-D barcodes on the merchandise at runtime and our system will use the decoded numbers as key to retrieve its associated marker pattern and 3D model data from a remote location. And once the pattern and 3D model data are retrieved, the marker associated with the downloaded pattern will be located and its location and orientation will be used as guide to display the 3D model. This solution decouples static associations between markers and 3D models, and allows advertisers to promote their products by adding entertainment values through the use of AR by changing markers and 3D models whenever they wish.

However, the traditional black-and-white marker patterns present a problem: because they are not be visually appealing and the advertisers would not put black-and-white markers on their products with colorful packaging. So we need to find alternative markers that are colorful. The requirements for a replacement marker is that it must be square, of certain minimum size, its border should be enclosed with sufficient thickness, and that the image inside that border must not be rotational symmetric so that marker orientation can be properly detected. These properties are well illustrated by the traditional black-and-white marker patterns as shown below in Fig. 1(a).

During our search for alternative markers, we came upon idea of testing the colored point stickers distributed by convenience stores as possible markers. Stores, such as 7-11, FamilyMart, OKMart, pass out various stickers with difference patterns on them for product promotion purposes. These stickers are usually small, square and colorful, in Fig. 1(b), so may be a feasible alternative to traditional markers. But the drawbacks to using these stickers as markers are that the background may not be completed enclose the foreground image, or if enclosed, the enclosure may not be of sufficient thickness, We will show in our experiments that how this concern may not apply in the case of color markers.

In Taiwan, about 49% of married women as well as college students collect these point stickers provided by the convenience stores in order to exchange the points for special ordered toys as well price discounts on certain items when sufficient number of points have been collected. However, these colorfully designed point stickers do not serve any other function. As such, they could be ideal markers used to enhance the stores' product promotions with our marker AR system integrated with a 1-D barcode reader.



(a) (b)

Fig. 1. (a)Traditional Black-and-White ARToolkit Markers vs. (b) Convenience Store Point Stickers

2 AR-Based System with 1-D Barcode Reader

Our implementation of the barcode-based AR system includes a server of graphics data and a PC-based AR application software, as shown in Fig. 2. The graphics server is used mainly for storage and will respond to the AR application's requests of marker files and 3D graphics files that correspond to specific barcodes. Instead of using the standard templates provided by ARToolkit as markers, we propose the use of colored point stickers provided by the convenience stores in Taiwan as markers for the proposed system. The AR application software installs on a Windows-based PC computer with a webcam. The application uses the webcam to detect and decode a one-dimensional barcode, then using the decoded barcode as key to attempt to retrieve the marker and its associated 3D graphics files from the graphics server. The server will contact the database that serves that particular merchant's graphics data to see if the files exists and have been updated. If the merchant has already placed the data related to the barcode into the graphics database, then application should be able to download the marker pattern and 3D model files and store them locally. If, however, no data related to the barcode can be found on the graphics server, then the application would ask the user to present a valid barcode and attempts to decode it again. If successfully, the application should locate the marker and use its location to display the 3D graphics on top of the actual marker.

The main process flow of the AR application as divided into the following steps: (1) capture an image from the webcam and determine the existence of a one-dimensional barcode within the image; (2) decode the one-dimensional barcode; (3) send the decoded barcode number as an index to the graphics server and request the corresponding AR marker and 3D graphics files; (4) the AR application receive the delivered marker and 3D graphics file from the server; (5) search within the webcam image for the physical AR marker; and (6) calculate the transformation matrices according to coordinates of the AR marker and then display the 3D graphics on the printed AR marker using the transformation matrices. Fig. 3 shows the concept of the

proposed system by using a black-and-white ARToolkit marker on a commercial product. The detailed process flow of the AR application is shown in Fig. 4

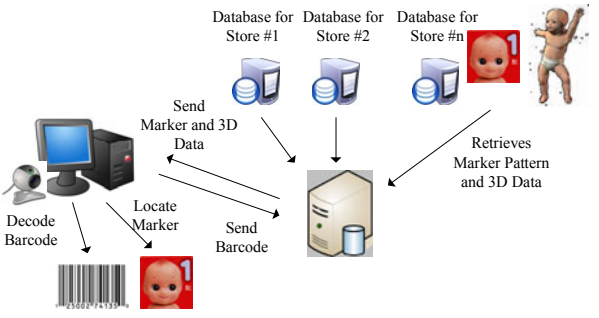


Fig. 2. Structure of barcode-based AR system

Fig. 3. The example of the AR Application Software

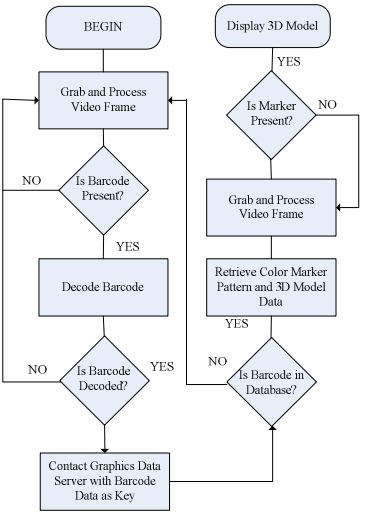


Fig. 4. Process Flowchart of the AR Application

In the stage of detection of one-dimensional barcode within a webcam-captured image, the fact that 1-D barcodes are straight vertical lines is used as feature for detection. In a given region, the gradient differences are calculated and summed for each pixel at 90, 0, 45, and 135 degrees orientations and differences between the sums are greater than a certain threshold value, then this region is judged to be part of a one-dimensional bar code. The algorithm is as follows:

For any given square (n x n) region extracted from the center of the screen, the color information are reduced to grayscale then thresholded to black-and-white, where the value of black pixels are set to '1', and the white pixels "0". We denote the value for any pixel at (x,y) position as bw(x,y). Then for each black-and-white pixel located at $(x, y) \in n \times n$ position, the counters C0, C45, C90, and C135, are used for 0, 45, 90 and 135 degrees, respectively. These counts are calculated as follows:

$$\text{if } \sum_{i=-1}^{1} \left| bw(x-n, y) - 1 \right| = 0, \text{ C0++}, \quad \text{if } \sum_{i=-1}^{1} \left| bw(x-i, y+i) - 1 \right| = 0, \text{ C45++},$$

$$\text{if } \sum_{i=-1}^{1} \left| bw(x, y-i) - 1 \right| = 0, \text{ C90++, and} \quad \text{if } \sum_{i=-1}^{1} \left| bw(x-i, y-i) - 1 \right| = 0, \text{ C135++}$$

and when these counts are obtained for the entire region, we calculate the possibility that this region is part of a barcode in this way:

if (|C90-C0| > |C135-C45| AND
||C90-C0|-|C135-C45||) > threshold2,
set region_in_barcode = TRUE

So once the threshold is achieved, we determine that a 1-D barcode exist within the view of the camera, a single frame is then capture and processed to decode the 1-d barcode and retrieves the 3D model for display on the physical colored marker.

However, since the viewing angles provided by ARToolkit are based upon the position and orientation of the marker, these angles may be limited by the type of merchandise on which the marker is printed. So in addition to using the AR marker's coordinates for viewing the 3D graphics, such as by rotating the marker to achieve rotation of the 3D model, or by moving the marker closer to the camera lens for zooming effect, our AR applications also has additional controls for the purposes of increasing viewing angles. We did this by inserting special effects that are marker-independent, that is, these controls allow rotations, zooming, and other effects without moving the marker. These additional controls are achieved by calculating the combined spatial transformation matrix and apply matrix multiplication to the coordinates of each triangle of the 3D model before drawing the triangle surface on the screen, and that each value within the transformation matrix is stored in memory. In this way, the 3D graphics can be quickly reset to its original position in order to facilitate a variety of interactive special effects. The order the multiplying the transformation matrices is defined as rotation first, followed by zooming, and finally the translation transformation matrix is applied. These transformation matrices, using homogeneous coordinates, are defined as follows [7]:

$$[X' \quad Y' \quad Z' \quad 1] = [X \quad Y \quad Z \quad 1] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ T_x & T_y & T_z & 1 \end{bmatrix}$$

Translation

$$[X' \quad Y' \quad Z' \quad 1] = [X \quad Y \quad Z \quad 1] \begin{bmatrix} Z_x & 0 & 0 & 0 \\ 0 & Z_y & 0 & 0 \\ 0 & 0 & Z_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Zoom

$$[X' \quad Y' \quad Z' \quad 1] = [X \quad Y \quad Z \quad 1] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & 0 \\ 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Rotation about X-axis

$$[X' \quad Y' \quad Z' \quad 1] = [X \quad Y \quad Z \quad 1] \begin{bmatrix} \cos\theta & 0 & -\sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ \sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Rotation about Y-axis

$$[X' \quad Y' \quad Z' \quad 1] = [X \quad Y \quad Z \quad 1] \begin{bmatrix} \cos\theta & \sin\theta & 0 & 0 \\ -\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Rotation about Z-axis

In addition to these transforms, we have also included a few surprise transforms that would add entertainment values to the system.

3 Experimental Setup and Results

In our experiment, we first verified the accuracy of the barcode decoder, and once we can be certain that 1-D barcodes can be decoded correctly, we then tested our barcode-based AR system with 14 colored stickers, of similar sizes, collected for various convenience stores, as show below in Fig. 5. The square stickers range in size from about 11/20" (1.4 cm) to about 5/8" (1.6cm) on each side. The sole circular sticker is about 1.6cm in diameter. The rectangular sticker is about 1.5cm by 1.9cm in size. And before testing, each color marker was separately trained in order to generate their pattern files. We found that the circular sticker could not be trained at all and so categorized it as unstable and unusable as marker. The rectangular sticker from 7-11 can also be trained by taking a square area starting from the upper left corner and ignore the rest. After training, we started the barcode verification phase of our test. For this test, we randomly chose one of the stickers and pasted it next to a valid EAN-13 barcode and presented it to the camera. Fig. 6 below shows that the barcode is decoded correctly and the marker pattern and its associated 3D model, which is a teapot, being retrieved from a remote database. This figure is deliberately shown in the original resolution of the webcam, which is low (640x480).

After the graphics data are retrieved, we then test whether the marker pattern can be recognized and located accurately, with the correct orientation. We found that most of the stickers passed this phase of the test, but in the case of certain markers, the denser parts of the barcode next to the marker can sometimes be mistaken for the actual marker, which similar to the problem illustrated by Fig. 10. In these few cases, we categorize these markers as unstable as rejected them from further tests. We then categorized all markers that passed this stage of testing as semi-stable..

Fig. 5. Color Stickers Used in the Experiment

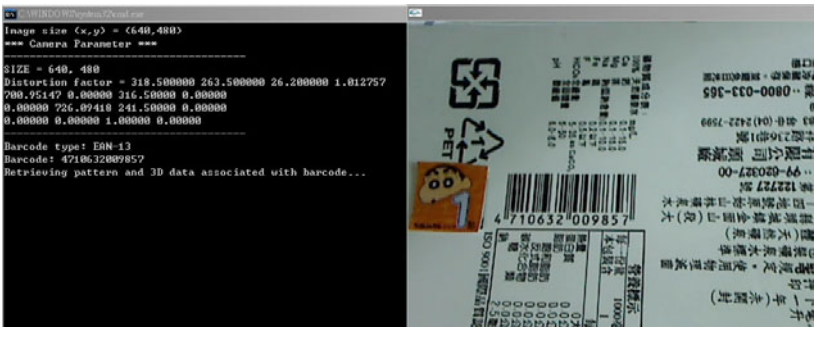


Fig. 6. Barcode Decoding Experiment

In the next stage of our experiment, we tested the further stability of each marker by using the same barcode, using the same pattern and 3D model data, with other markers. i.e. We placed each barcodes next to a marker not associated to it. If the original marker associated with the barcode is stable, then ARToolkit should not be able to locate the marker. However, ARToolkit falsely identify the replacement marker as the original marker. The marker whose pattern file does not result in any false detection is moved from the semi-stable category into the stable marker category. The same test is performed for every marker in the semi-stable category. Those markers that failed this stage of the test stay in the semi-stable category.

Fig. 7 below shows an example of multiple detections using the pattern file of a semi-stable marker, one true and one false. It is interesting to note here that even if two markers are similar in characteristics, the orientation of detected marker can be off by as much as 90 degrees as indicated by the tri-color axis.

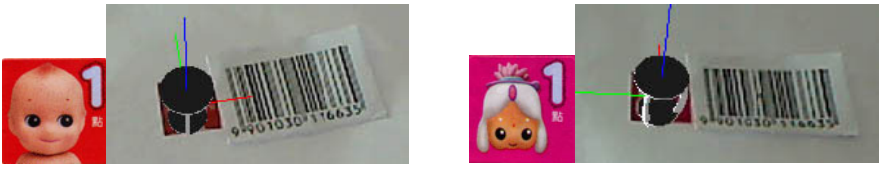


Fig. 7. Original Marker on the left and a Falsely Identified Marker on the Right

The final result is that out of the 14 stickers, 4 are unstable, 4 are semi-stable, and 6 are stable, or 3/7 of stickers may be used as markers. The results are shown in Table 1 below.

Table 1. Colored Stickers Used Markers in the Experiment Categorized

Stable	Semi-Stable	Unstable
		

Fig. 8 below shows the detection of a marker next to its associated barcode, and Fig. 9 below demonstrates the stability of stable markers. We placed two markers with similar colors and images next to it, and the location of the marker that matched the barcode was correctly detected.

Fig. 8. 3D Model Displayed on a Colored Marker

Fig. 9. Stable Markers in the Presence of Similar Markers

Next, we wish to test for conditions under which the stable markers would fail. We first tested for how far away from the camera the pattern of a stable marker can still be detected by measuring the size of the marker detected by the camera against its original size. We found that smallest detectable size is around 0.5cm x 0.5cm. If a stable colored marker is farther away the camera so that its size to the camera is smaller than 0.5cm on each side, then our system would fail to detect it. We performed a similar test in order to see if this result is valid for the standard black-and-white marker also, and found that we can achieve similar results with the black-and-white markers also.

Secondly, we want to test if the stable markers would still work if we replace the background from pure white to off white. For this test, we place a stable marker against backgrounds of several different colors, from off white to black. The result is that failure of detection occurred in all tested non-white backgrounds.

An interesting artifact was found during the test for stable markers: it is observed that if the colored marker is placed too close to the 1-D barcode, then during sudden movements ARToolkit can temporarily falsely identify the denser part of the barcode as the marker with the wrong orientation, as shown below in Fig. 10. But once the scene is stable again, the marker is correctly identified. Because of this, we suggest that the colored marker should be placed a small distance away from its associated barcode. This is so that the marker can be viewed with a camera at close range without the presence of the barcode.

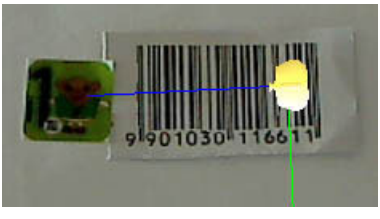


Fig. 10. False Detection during Fast Movements

Based on the results from these tests, it is postulated that in order to be used as stable colored markers for our ARToolkit-based system, the background of the marker should enclose most of the foreground image, but there does not seem to be a need for complete closure, as some of the stable markers have parts of their foreground images touching the edge of the sticker. It is also observed that the outside edges of stable markers could provide more stability by being jagged or rounded. The background should be textured rather than of a solid color, as shown below in Fig. 11. Because it is observed that, for colored markers, textured backgrounds can aid in location and identification by ARToolkit. Those colored markers with textured background are preferred over markers with solid-color background. Also, the foreground image should be made up of several contrasting colors that are distinct from the background colors, and of course, not be rotationally symmetric.

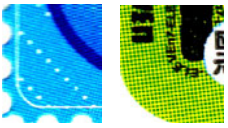


Fig. 11. Edges of Some of the Stable Markers

4 Conclusion

The barcode-based AR system proposed in this paper was successful in using a one-dimensional barcode as an index to download the marker and 3D graphics for display, and changed the way ARToolkit statically reads in 3D models into dynamically accessing 3D graphics files so that by simply updating the AR marker at remote databases, convenience stores can set the expiration date on a certain marketing or product promotion project, and not be dependent on ARToolkit's original method of using fixed marker and the 3D model. Future researches will explore the effect of different markers [8], maybe such as using the one-dimensional barcode not just as an index but as an AR marker itself. Although this approach will introduce some problems in the display of the 3D graphics and change the way ARToolkit recognizes markers, and allow even regular merchandise, without standard markers, can be used to experience AR-based advertising, and make it easier for the manufactures or convenience stores to add values to their products. In the future, we will add the adaptive binary threshold function during the marker training processing, and we expect this can make more colored markers stable in our system. Also, intelligent web agents may be implemented on the server side, in the future, to gather product information, such as size, colors, shape, and available barcode information. This information will hopeful be used to help the application side in locating colored markers against non-white backgrounds, and make decisions on how to display the 3D models.

References

1. Enhance Everyday Living using Virtual Information and Interactive Elements - Application of Augmented Reality, http://www.find.org.tw/find/home.aspx?page=news&id=5638
2. Augmented Reality has Become the New Wave and Walmart is Actively Participating, http://www.find.org.tw/find/home.aspx?page=news&id=5645
3. Augmented Reality-enabled Mobile Phone Browser-Layar, http://www.find.org.tw/find/home.aspx?page=news&id=5647
4. ARToolkit, http://www.hitl.washington.edu/artoolkit
5. Meiguins, B.S., Carmo, R., Almeida, L., Goncalves, A.S., Pinheiro, S.C.V., Garcia, M., Godinho, P.: Multidimensional Information Visualization using Augmented Reality. In: Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and its Applications, pp. 391–394 (2006)
6. Hu, Q., Liu, T., Yao, Y.: An Easy System of Spatial Points Collection Based on ARToolkit. In: Proceedings of 2009 WRI World Congress on Computer Science and Information Engineering, vol. 1, pp. 582–586 (2009)
7. Pulli, K., Aarnio, T., Miettinen, V., Roimela, K., Vaarala, J.: Mobile 3D Graphics with OpenGL ES and M3g. Morgan Kaufmann, San Francisco (2008)
8. Ababsa, F., Mallem, M.: A Robust Circular Fiducial Detection Technique and Real-time 3D Camera Tracking. Journal of Multimedia 3(4), 34–41 (2008)

Design and Implementation of e-Journal Review System Using Text-Mining Technology

Chun-Wei Tseng[1], Feng-Jung Liu[2],Wan-Chin Lu[1], and Shih-Hao Huang[1]

[1] Department of Information Management, Cheng Shiu University,
Kaohsiung County 833, Taiwan
cwtseng@csu.edu.tw, reye@csu.edu.tw, nuclear0304@hotmail.com
[2] Bachelor Program of Digital Contents and Technology Management,
Cheng Shiu University, Kaohsiung County 833, Taiwan
fjliu@csu.edu.tw

Abstract. With the advancement in information communication technology and the concept of the Web 2.0 framework, the e-learning becomes more popular and more diverse. In the last decade, the publication of journal papers has become a major channel for the outcome of researchers' studies. But, the current submission method of the international journal papers is mostly handled by traditional paper review process. As the submitted articles increase the burden of paper reviewing process gets heavy. It is not only time-consuming and laborious, but cannot find proper reviewers to review the specific manuscripts in some cases. Sometimes, the assignment of papers to specific reviewers is not objective enough. Thus, this study is based on the design concept of content management, CMS, to develop an online collaborative electronic journal paper review system. We adopt rich internet application, RIA, technology to web application development and apply text mining technologies to articles classification. We propose an assignment mechanism of paper reviews to achieve the automatic, fair and efficient paper reviewers' assignment. Expect the study will effectively simplify the complex assignment process and make the review work more objective and efficient.

Keywords: Content Management, Rich Internet Application, Text Mining.

1 Introduction

Along with the explosive growth of data, the shift from traditional document management to electronic file management makes the content management more complicated. [2,3] How to efficiently collect and maintain mass of contents? On the Web, functionality and content are mixed and matched to whatever extent the site owners can manage. It's a management issue. To mix and match content and functionality, you must be well organized. You must have a lot of structure in place that enables you to efficiently store and deliver your functionality where and when you need it. In short, you need a Content management system, CMS. CMS develops a new web site construction model which is a simple and efficient website management method. It can efficiently manage the web content and keep the advantages with decouple of content with design and portability.[1]

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 210–217, 2010.

Thus, in this study, we attempt to develop on-line electronic journal review system with the design concept of content management system. And we also utilize the RIA technology for web programming to improve the drawback with need to refresh the web pages as reading the pages and make users use the web applications like on the desktop computer. Additionally, we try to adopt text mining techniques for article classification and to develop automatic reviewer recommendation system to assist journal manager with the reviewer recommendation in order to simply the review process and to improve the improper reviewer with more objectivity and better efficiency.

2 Related Works

There are a number of standard classification techniques in literatures such as Rocchio algorithm[6,7], Naive Bayes algorithm [9], support vector machines[5,8], k Nearest Neighbor[10], and so on. These classification algorithms are mainly based on statistics theory and machine-learning methods. We briefly describe these typical classification algorithms as follows:

- Rocchio algorithm
  Rocchio algorithm was proposed by Rocchio, which is the best-known way of computing good class boundaries and used to compute the relation degree between the query and documents in information retrieval. Later in 1994, Hull renovated the Rocchio's formula applied to the document classification. On the basic concept of algorithm, it uses the training set to construct a prototype vector for each class and uses *centroids* to define the boundaries. The *centroid* of a class is computed as the vector average or center of mass of its member.

  In the Rocchio algorithm construction, it needs to determine a class and to train all of documents belonging to this class. The conducted vectors of these documents, which belong to this class, are represented with positive values; otherwise, they are represented with negative values. And then, sum up these vectors values to get the sum of vectors. The sum of vectors will act as the prototype vector and will be used to define the similarity of two vectors with the cosine function of the included angle between the two vectors.

  Easy to implement and the lower time complexity are the main advantages of Rocchio classification algorithm. However, it is seldom implemented to solve the real instances classification problems in classification systems.

- Naive Bayes
  The classier uses a probabilistic model of text. Although this model is a strong simplification of the true process by which text is generated, the hope is that it still captures most of the important characteristics.

- Support vector machines, SVM
  SVM is a powerful technique for data classification proposed by Boser, Guyon and Vapnik in 2002. It mainly solves the classification problem for two classes. In a high dimension space it searches a hyperplane to be the separation with the two classes and make the minimum of classification error.

– k Nearest Neighbor Classification, kNN Classification
Unlike Rocchio, k nearest neighbor or kNN classification determines the decision boundary locally. It is proposed originally by Cover and Hart in 1968, which mainly uses the vector space model to represent each documents' features. It requires no explicit training and can use the unprocessed training set directly in classification. While a new document is imported, it is assigned to the class in accordance with the features of k training data closest to the new document. The measure is to compute the similarity of the unclassified document with all of training documents. And then, the k closest training documents will be found. It is less efficient than other classification methods in classifying documents. In implementation of the proposed system, we used the well-known kNN algorithm as the document classification method.

3 Design of Reviewer Assignment Recommendation System

In this section, it mainly describes the design and data processing flows of the reviewer assignment recommendation system. In the proposed system there are four parts, including document collection, representation, classification and discrimination as shown Fig. 1. It will be described clearly as follows.

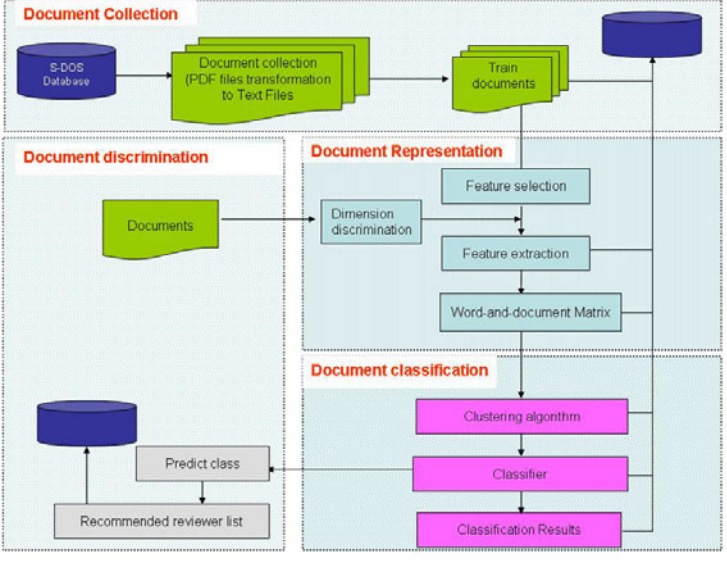


Fig. 1. System architecture of the reviewer assignment recommendation system

3.1 Document Collection

In document collection, it mainly collects some documents as the training data and transforms the PDF formatted files into text files using xpdf tool. We selected some journal papers from SDOS database. These data is used as training data for advanced document classification as shown Fig.2.

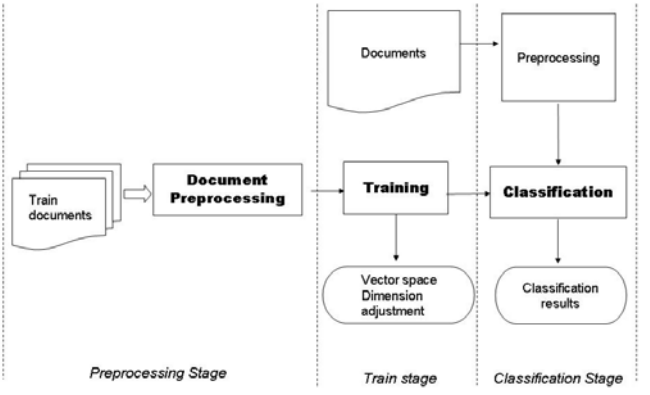


Fig. 2. Process flow of automatic document classification system

3.2 Document Representation

In the document representation part, the first step is feature selection. Next, feature extraction. At last, it is to construct the term-document matrix. First, on feature selection, in this study, we focus on the collections of English papers. English mainly use the blank to distinguish different words. After omitting the stop words as 'the', 'is', 'I', etc., if the word is new, never occurred previously, it will be stored into the word list; Otherwise, the count of the word will increase one. Additionally, we adopt Porter algorithm to solve the problem of same words. For example, the "gone" or "going" word is the different tense of the "go" word. As to the word-searching, we just compare these words in the abstract and keywords of the papers. If a keyword appears in the abstract of papers, it will be viewed as a feature value.

After feature extraction, we proceed to compute the TFIDF, Term Frequency Inverse Document Frequency, value [10] for each word and to sort by the weighted value. In order to reduce the dimension, we just select the top one hundred keywords. At last, with little impact of classification efficiency, we adopt the TFIDF formula as Eq. (1) to transform the documents into vector space representation.

$$TFIDF(t,d) = TF(t,d) * IDF(t) \tag{1}$$

where

$TF(t,d)$: The occurrence number of the term t in Document d;

$$IDF(t) = \log\left(\frac{|D|}{DF(t)}\right)$$

$DF(t)$:The number of papers which contain the term t.

$|D|$**:** The total number of terms in all documents

3.3 Document Classification

In this part, we try to adopt the Vector Space Model, VSM, to proceed to representation of document, which contains feature selection, feature extraction, TF-IDF weight calculation, the term-document matrix construction and so on. At the last, according to the trained model, we adopt k Nearest Neighbor, kNN, classification to conduct the classification results for professional area.

3.4 Document Discrimination

Same as the training documents, the testing documents should pass document preprocessing, feature selection and extraction, constructing the term-document matrix and, at last, according to the matrix and the *centroid* vectors of classes, judge which class it belongs to.

4 Implementation

In this section, it mainly explains our system implementation and related experimental considerations. As a result, it will recommend a list of reviewers for each paper in the proposed reviewer assignment recommendation system.

Paper99050501 (10).pdf
TitleApproximate fixed point theorems in Banach spaces with applications in game theory
TF(dish,d)=460
ITF(d)=0.8324

val	Approximate	xed	point	theorems	Banach	spaces	applications	appear	Journal	Mathematical	Analysis	Computer	Science	Cuza"	Carol	I
TF (t,d)	4	30	32	21	17	8	3	3	1	2	1	1	1	1	1	2
DF(t)	7	14	28	10	17	17	21	4	9	13	25	3	12	2	2	17
IDF(t)	2.65	2.35	2.05	2.50	2.26	2.26	2.17	2.89	2.54	2.38	2.10	3.02	2.42	3.19	3.19	2.26
TFIDF (t,d)	18.55	32.89	57.34	24.95	38.50	38.50	45.63	11.57	22.87	30.96	52.43	9.05	28.99	6.39	6.39	38.50

Paper99050501 (12).pdf
TitleCharacterization of nonsmooth quasiconvex and pseudoconvex functions
TF(dish,d)=275
ITF(d)=1.0558

val	nonsmooth	quasiconvex	pseudoconvex	Teacher	Training	Iran	Received	April	Available	online	September	Submitted	Mordukhovich Abstract	paper	c
TF (t,d)	5	12	8	1	1	1	1	1	1	1	1	1	1	4	
DF(t)	13	4	3	3	4	4	22	7	11	11	7	10	6	27	
IDF(t)	2.38	2.89	3.02	3.02	2.89	2.89	2.15	2.65	2.45	2.45	2.65	2.50	2.72	2.06	
TFIDF (t,d)	30.96	11.57	9.05	9.05	11.57	11.57	47.36	18.55	26.99	26.99	18.55	24.95	16.30	55.72	

Paper99050501 (13).pdf
TitleDirectional derivatives and subdi erential of convex fuzzy mappings and application in convex fuzzy programming
TF(dish,d)=484
ITF(d)=0.8103

val	Fuzzy	Systems	derivatives	subdi	erential	convex	fuzzy	mappings	application	Institute	Republic	China	Received	January	received	revised	July
TF (t,d)	44	40	10	9	22	72	142	33	5	1	1	2	1	1	1	1	2
DF(t)	3	5	8	3	3	28	6	9	9	7	2	5	22	6	4	4	5
IDF(t)	3.02	2.80	2.59	3.02	3.02	2.05	2.72	2.54	2.54	2.65	3.19	2.80	2.15	2.72	2.89	2.89	2.80
TFIDF (t,d)	9.05	13.98	20.74	9.05	9.05	57.34	16.30	22.87	22.87	18.55	6.39	13.98	47.36	16.30	11.57	11.57	13.98

Fig. 3. Partial processed result of document representation for three PDF papers

– Document collection

In this experiment, we collected one hundred journal papers from SDOS electronic library. And these papers' topics are filtered with the "optimization" keyword and are published between year 2000 and 2007. After that, these PDF papers will be transformed into text files using xpdf tool for advanced text mining.

– Document representation

We selected out thirty papers to proceed to feature selection, extraction, weight computation, the construction of a term-document matrix. The following is the partial processed result as shown in Fig.3.

– Document classification

After feature extraction and constructing the term-document matrix, it will proceed to the comparison of similarity for all documents vectors. In this study, we adopt the hyper graph cluster method. The detailed flow of document classification is shown as Fig.4.

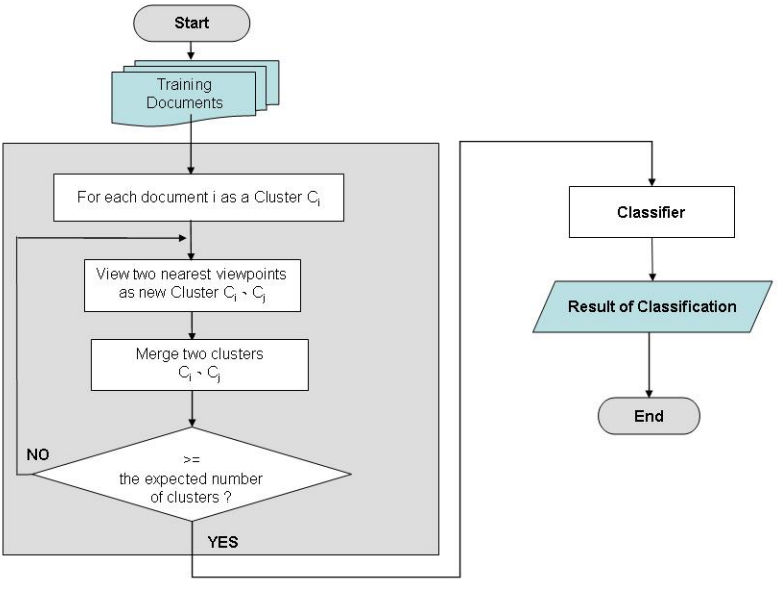


Fig. 4. It illustrates the detailed processing flow of hyper graph cluster

– Document discrimination

After completing the all of documents classification as described in section 4.3, we use the kNN classification to achieve the document discrimination for a new document. The detailed is shown as following. The default value of k is five. We compute the distance of the testing documents to all of documents and then select the five nearest points. Next, in 5-NN algorithm, we compute the number of classes of papers in the training data which the five nearest points locate to. At last, the testing document will be views as the document with the most class number.

– Automatic reviewer recommendation

After completing the all of documents classification into some expertise domains, the automatic reviewer recommendation system will recommend the top ten reviewers for each paper according to the following rules:

Case 1: The author who is one of the authors of the submitted article's references in the professional domain and appears with the most frequency number of the author's appearance.

Case 2: The author appears in the professional domain.

Case 3: If not satisfied with the above two, the author will be recommended as a reviewer who is the author of the most submitted papers in this professional domain.

After the reviewer list being recommended, the supervisor will assign the top three reviewers to proceed to the review job by E-mail notification. If the assignment is rejected or does not response in fourteen days, the new reviewers will be assigned until the review job completeness.

5 Conclusion and Future Works

In this paper, we constructed a paper reviewer assignment sub-system to help the e-journal editor's work more convenience and fairness. By using the text-mining technology, the editor can reduce time consuming in his position. We also suggest the system mechanism divided into four processing blocks, respectively, document collection, document representation, document classification and document discrimination. The result has showed our proposed system works well and increases the efficiency of editor's reviewer assignment job.

In future works, there are some of features still to be considered into our prototype. The first one is how to adopt the editor's experience to our assignment algorithm. Accumulating experiences may provide the system with efficient and accurate reviewers' assignment. We also find the ontology technology may be useful to get suitable keywords from the text of paper's body. Finally, the performance is a traditional issue in system implementation once we add full text of all selected paper into the system. How to reduce the time complexity of document classification is another important issue.

Acknowledgments. This work was supported by a grant from the National Science Council of Taiwan under Contract No. NSC 98-2622-E-230-010-CC3.

References

1. Boiko, B.: Content Management Bible, pp. 3–11. Wiley Publishing, Inc., Indianapolis (2002)
2. Sullivan, D.: Document Warehousing and Text Mining (2001) ISBN 0-471-39959-0
3. Robertson, J.: How to evaluate a content management system. KM Column, 1–7 (2002)
4. Garrett, J.J.: Ajax: A New Approach to Web Applications (2005)

5. Cervantes, J., Li, X., Yu, W., Li, K.: Support vector machine classification for large data sets via minimum enclosing ball clustering. Neurocomputing 71, 611–619 (2008)
6. Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice-Hall, Englewood Cliffs (1971)
7. Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In: ICML 1997 (1997)
8. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398. Springer, Heidelberg (1998)
9. McCallum, A., Nigam, K.: A comparison of event models for naïve bayes text classification. In: AAAI 1998 Workshop on Learning for Text Categorization (1998a)
10. Salton, G., Michael, J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)

Study on Architecture-Oriented Information Security Risk Assessment Model

Wei-Ming Ma

Information Management Department, Cheng-Shiu University
wma@csu.edu.tw

Abstract. In this study, we adopt the structure behavior coalescence methodology to construct an architecture-oriented information security risk assessment model (AOISRAM), which is integrated structure and behavior of the risk assessment model. AOISRAM solves many difficulties caused by the process-oriented approach in ISO 27001:2005 of information security risk assessment such as uneven distribution of resources, poor safety performance, and high risk. We find out the information security consultant, project manager are the key roles for the success of the risk assessment from structure behavior coalescence diagram. The feedback mechanism in the enterprise is essential to report and respond to the incidents for reducing the risk. This research achieves a beneficial model and knowledge for the information security risk assessment. This accomplishment may be valuable for the business and academic circles to follow and refer.

Keywords: Risk Management, Architecture-Oriented Risk Assessment Model, AOISRAM, Structure Behavior Coalescence.

1 Introduction

The information security incidents have most often been reported. The loss of enterprise operation is more and more serious because of information security incidents. There are more and more operation risks happening inside the enterprise because of such informational and electronic transformation. Consequently, the requirement to have an effective management framework of information security is urgent.

The international information security management standard, ISO 27001:2005, which includes personnel security, technology security, physical security and management security has been promulgated. When bringing in an information security management system (ISMS), a company usually encountered the process-oriented approach which treats the ISMS in system's structure view and behavior view separately. During the planning phase, separation of structure view and behavior view could cause many difficulties, such as uneven distribution of resources, poor safety performance and high risk.

Currently, most risk management models are categorized into the process-oriented approach. This research utilizes architecture-oriented modeling methodology so that structure view and behavior view are coalesced when decomposing the ISMS to obtain structural elements and behaviors deriving from interactions among these structure elements.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 218–226, 2010.

2 Background

The previous studies about information security, information security management system, and risk management model are described briefly.

2.1 Information Security

According to US Code Collection Title 44, the term “information security” means protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide[1][2][3]:

- Integrity, which means guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity;
- Confidentiality, which means preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information; and
- Availability, which means ensuring timely and reliable access to and use of information.

Information is the lifeblood of all organizations and can exist in many forms. It can be printed or written on paper, stored electronically, transmitted by mail or by electronic means, shown in films, or spoken in conversation. In today's competitive business environment, such information is constantly under threat from many sources. These can be internal, external, accidental, or malicious. With the increased use of new technology to store, transmit, and retrieve information, we have all opened ourselves up to increased numbers and types of threats.

There is a need to establish a comprehensive information security policy within all organizations. We need to ensure the confidentiality, integrity, and availability of both vital corporate information and customer information.

2.2 Information Security Management System

An Information Security Management System (ISMS) is a systematic approach to managing sensitive company information so that it remains secure. It encompasses people, processes and IT systems. BSI published a code of practice for these systems, which has now been adopted internationally as ISO/IEC 27001:2005(BSI, 1999; ISO, 2005). The Information Security Standard is published in the following parts:

- ISO/IEC 27001:2005 (formerly BS 7799-2:2002) Specification for Information Security Management
- ISO/IEC 27002:2005 (previously named ISO/IEC 17799:2005) Code of practice for Information Security Management.

ISO/IEC 27001:2005 specifies the requirements for establishing, implementing, operating, monitoring, reviewing, maintaining and improving a documented Information Security Management System within the context of the organization's overall business risks. It specifies requirements for the implementation of security controls customized to the needs of individual organizations or parts thereof. ISO/IEC 27001:2005 is designed to ensure the selection of adequate and proportionate security controls that protect information assets and give confidence to interested parties.

The ISO 27001 standard defines its 'process approach' as "The application of a system of processes within an organization, together with the identification and interactions of these processes, and their management". ISO 27001 is based on the methodology know as Plan-Do-Check-Act (PDCA) model, which must be structured to the every ISMS process [4]. This is an approach to developing, implementing and improving the effectiveness of an organization's ISMS. It describes in four key activities:

Plan - Establish the ISMS policy, objective, processes and procedures relevant to managing risk and improving information security to deliver results in accordance with an organization's overall policies and objectives. In this activity it needs to define requirements, assess risks, decide applicable controls, define scope of ISMS, define strategy of ISMS, define steps of risk assessment, define risk, select control objects and controls, prepare of statement of application.

Do - Implement and operate the ISMS policy, controls, processes and procedures, assessment and implement the threat risk plan, implement controls, the completion of the training implementation procedures, operational management, resource management, the implementation of procedures to detect and respond to security incidents.

Check - Monitor and review the ISMS. In this activity, it needs to assess and where applicable, measure process performance against ISMS policy, objectives and practical experience and report the result to management for review, implement and monitor procedures, review performance and acceptable level of risks periodically, construct internal the ISMS audit, and view the IMS records management activities and events affect the ISMS on a regular basis.

Act - Maintain and continuously improve the ISMS. It needs to take corrective and preventive actions which are based on the results of the internal ISMS audit and management review or other relevant information, achieve continual improvement of ISMS. It also needs to implement system enhancements and correct system activities, applications training courses, and related results of team communication to guarantee the achievement of system enhancement.

2.3 Risk Management Model

ISMS is a departure from the risk management through risk control measures to deal with the relative effectiveness of control enterprise information security risks and maintain a viable important mechanism for continuing operations to achieve the enterprise's operational requirements.

Risks are events that negatively impact the organization's ability to achieve their goals as far as the probability of their occurrence and the related consequences are concerned. Analyze risks means identifying and quantifying these events so that specific actions may be planned and developed [5][6][7].

Boehm studied a software risk management model. In the model risk management was divided in risk control and risk assessment [8]. Risk control included: risk monitoring, risk resolution, and risk management planning. Risk assessment was included: risk prioritization, risk analysis, risk identification as shown in Figure 1 [8][9].

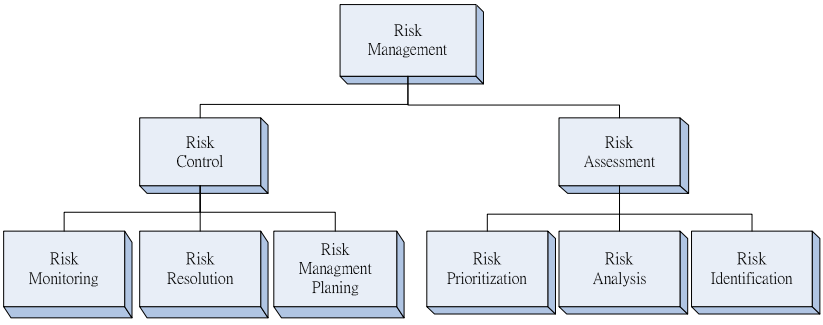


Fig. 1. Risk Management Steps (Redrawn from Boehm, 1991)

Objectives of Risk Analysis is to identify the main risks to information security in a systematic way, to ensure compliance of security management process with ISO 27001 standard, to present in a quantified way the events that may prevent the organization to achieve their goals, and to provide an overview of the aspects that need to be managed to assure compliance to the security policy. There are four risk treatment options [10][11]:

Risk reduction: if the risk is not acceptable, select appropriate safeguards, which are also called controls, and reduce the risk to an acceptable level.

Risk avoidance: if the risk is too large and it is uneconomical to reduce it, it may be prudent to altogether avoid the risk. This could be done by change of location, or avoiding using a risky technology.

Risk transfer: it may be viable to transfer the risk to other agency that is better at handling the risk. It may be an insurance company or a company whom the task could be outsourced. However, the responsibility may still remain with the organization.

Risk acceptance: Senior management may decide to knowingly accept the risk as it is too costly to take any action. This would be a conscious decision and management should be ready to consequence.

3 Architecture-Oriented Information Security Risk Assessment Model Design (AOISRAM)

We applied Structure Behavior Coalescence (SBC) methodology to design an architecture-oriented information security risk assessment model (AOISRAM) including: architecture hierarchy diagram, structure element diagram, structure element service diagram, structure element connection diagram, structure behavior coalescence diagram, and interaction flow diagram.

3.1 Architecture Hierarchy Diagram

Any management model can be illustrated by an architecture hierarchy diagram for the structure of a system's decomposition and combination [12][13]. Architecture hierarchy diagram can become easy to understand complex systems as shown in Figure 2.

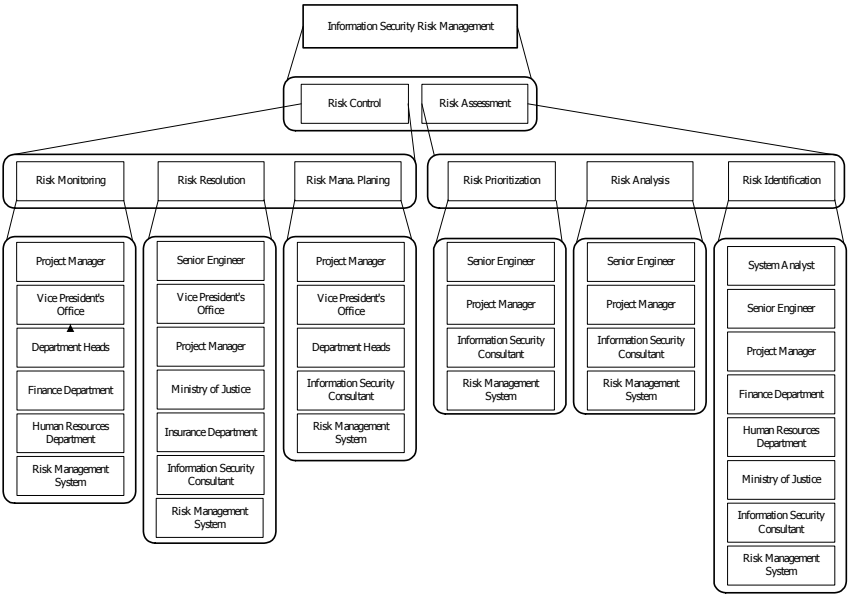


Fig. 2. Architecture Hierarchy Diagram of AOISRAM

This is required to construct the risk assessment model structure element diagram from a structural point of view. The structure elements of the risk assessment model are the basic elements, and they compose of the risk assessment structure. The necessary structure elements were analyzed from the information security risk assessment model. That is identified all builders and destroyers of the risk assessment model.

3.2 Structure Element Diagram

Collection of non-aggregated systems or structure elements of architecture hierarchy diagram become the structure element diagram. From Figure, we draw the risk assessment model structure element diagram which consists of eleven structure elements as shown in Figure 3.

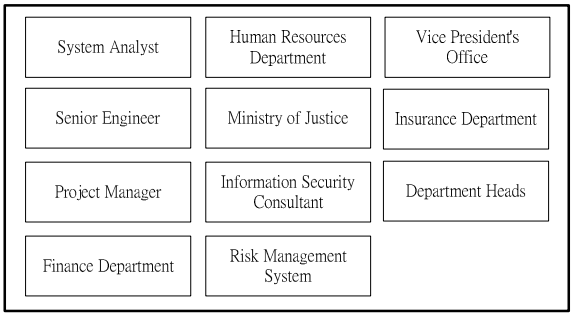


Fig. 3. Structure Element Diagram of AOISRAM

3.3 Structure Element Service Diagram

The structure elements provide many services through the interface or work content of the structure elements with input or output parameters is called a structure element service diagram [14][15]. Input parameter of the service is denoted by an arrow symbol directed to structure element. Output parameter of the service is denoted by an arrow symbol leave the structure element. Based on the collection of literature and sorted out the structure elements step by step, services of eleven structure elements ware obtained for the risk assessment model as shown in Figure 4.

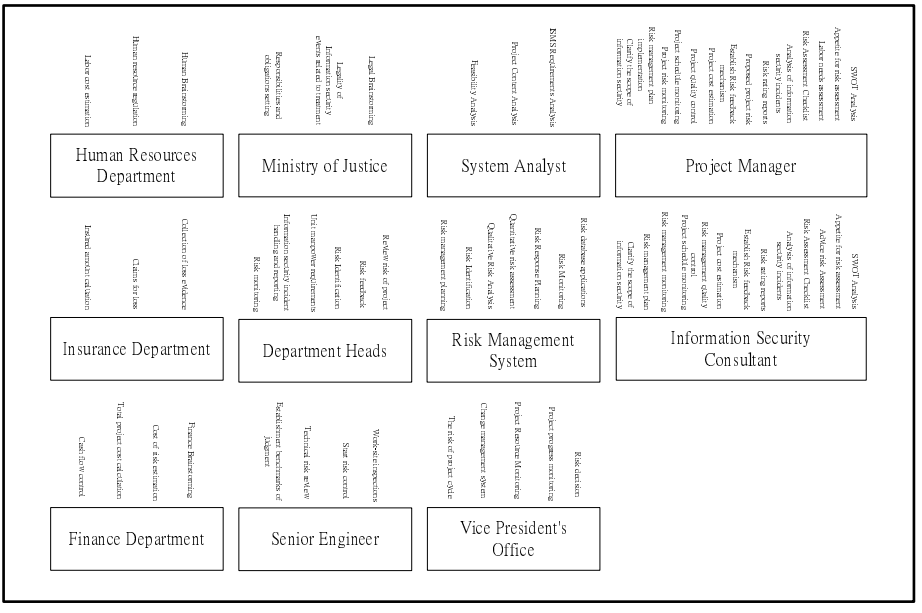


Fig. 4. Structure Element Service Diagram of AOISRAM

3.4 Structure Element Connection Diagram

A structure element connection diagram connect services between the various structure elements in accordance with its priorities. Rectangular frame is the system boundary, and the user or vice president is the external environment as shown in Figure 5.

3.5 Structure Behavior Coalescence Diagram

Behaviors are derived from interactions within structure elements or the outside environment for the risk assessment model. It becomes structure behavior coalescence diagram. The arrow represents to its behavior for service of supply and demand in each time occurred. The primary goal is to achieve an integrated model in which the structure and behavior are coalesced, preventing from a separation of structures and behaviors. From the structure element diagram and structure element service diagram, we further derive out eight behaviors of the risk assessment model as shown in Figure 6.

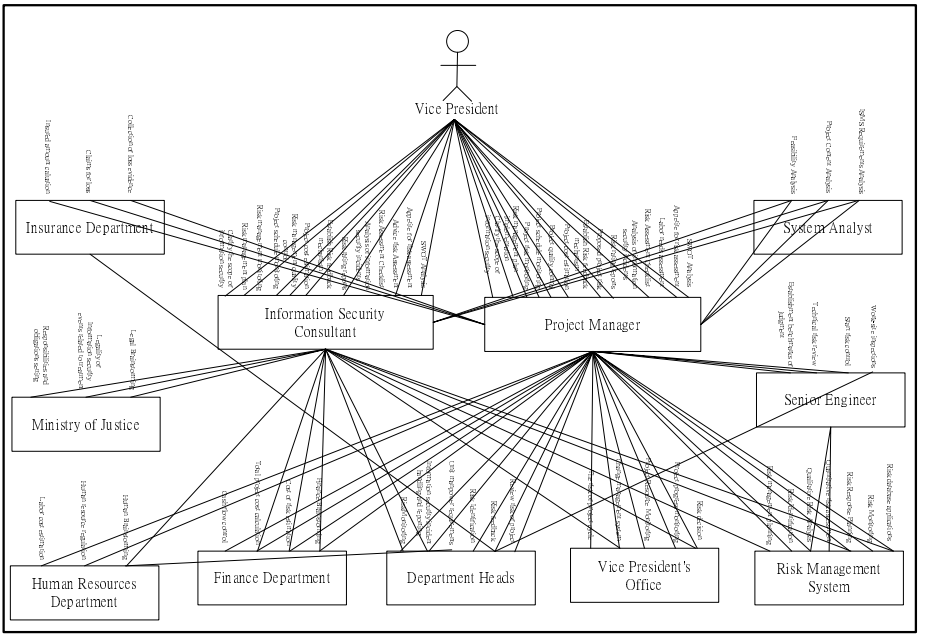


Fig. 5. Structure Element Connection Diagram of AOISRAM

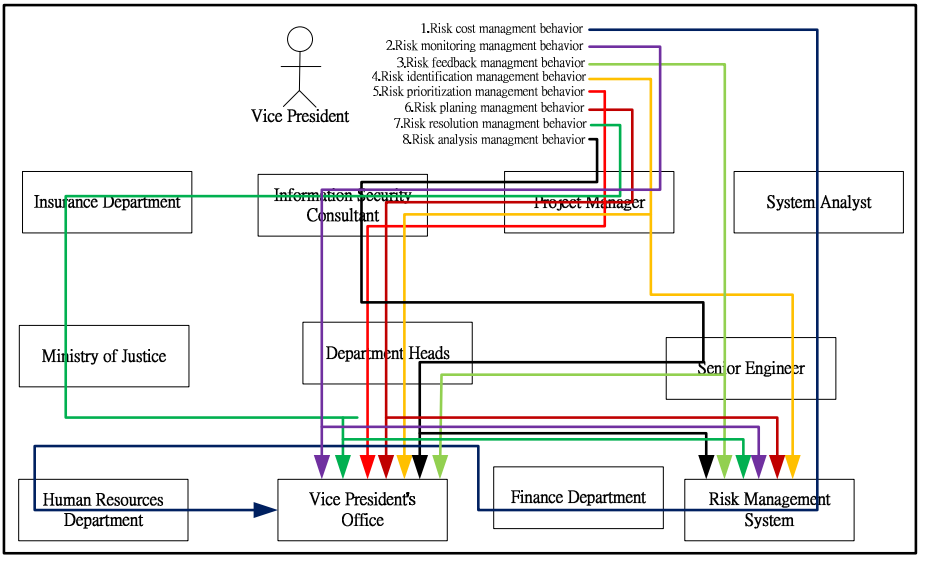


Fig. 6. Structure Behavior Coalescence Diagram of AOISRAM

3.6 Interaction Flow Diagram

Each behavior presented on the structure behavior coalescence diagram can be drawn as an interaction flow diagram. The construction of interaction flow diagram of the risk assessment model describes the outside environment and structure elements, and their interactions according to the time. There are eight interaction flow diagrams in total for the risk assessment model. We only show one interaction flow diagram here.

The sequence diagram exhibited in Figure 7 represents behavior of the risk assessment model. X-axis represents structure elements and the external environment in which information flow direction is from left to right. Y-axis represents the implementation of an interactive timeline from the top to the bottom in the time sequence.

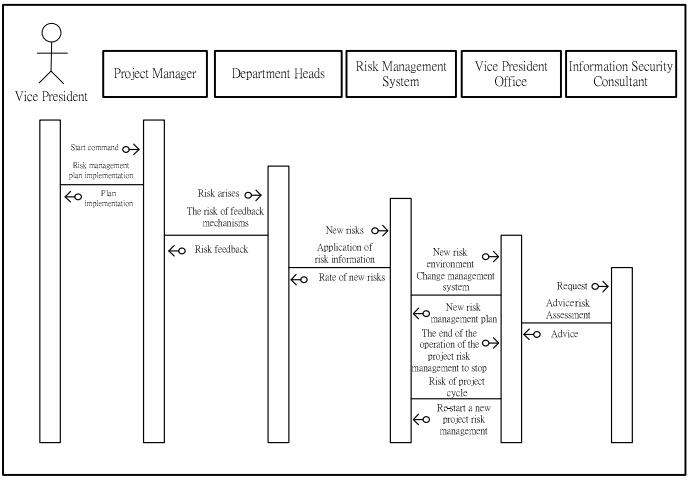


Fig.7. Interaction Flow Diagram for Risk Planning Management Behavior of AOISRAM

4 Conclusions and Recommendations

The information security incidents have been reported recently and the loss of enterprise operation becomes more and more serious problem. Risks are higher and higher for the enterprise because of information transformation and electronic commerce. Consequently, the requirement of an effective risk management framework of information security is urgently need.

In this study, we adopt the structure behavior coalescence methodology to construct an architecture-oriented information security risk assessment model (AOISRAM), which is integrated structure and behavior of the risk assessment model. AOISRAM solves many difficulties caused by the process-oriented approach in ISO 27001:2005 of information security risk assessment such as uneven distribution of resources, poor safety performance, and high risk.

This research achieves “what should be done?” and “how to progress?" in information security risk assessment model from architectural centric point of view to

avoid unreasonable design from the requirements. We find out the information security consultant, project manager are the key roles for the success of the risk assessment from structure behavior coalescence diagram. The feedback mechanism in the enterprise is essential to report and respond to the incidents for reducing the risk. This approach also improves quality of risk assessment, and provides important reference and guidelines to the risk assessment model.

This research provides a beneficial model and knowledge for the information security risk assessment. This accomplishment may be valuable for the business and academic circles to follow and refer. It is hoped that this research can offer a guideline for the information security risk assessment suitable for the enterprise and can be used as a reference for internal auditors and management.

References

1. International Organization for Standardization, ISO/IEC 27001 Information technology – Security techniques – Information security management systems – Requirements, ISO/IEC, http://www.iso.org/iso
2. Calder, A., Watkins, S.: IT Governance: A Manager's Guide to Data Security and ISO 27001 / ISO 27002. Kogan Page, London (2008)
3. Ma, W.-M.: Study of Consulting Service in Implementation of Information Security Management System. Journal of Global Business Operation and Management, 23–31 (2009)
4. Shewhart, W.A.: Economic Control of Quality of Manufactured Product. D. Van Nostrand Company, New York (1931)
5. Young, C.: Metrics and Methods for Security Risk Management. Syngress, Boston (2010)
6. Labodova, A.: Implementing integrated management systems using a risk analysis based approach. Journal of Cleaner Production 12, 571–580 (2004)
7. Poore, R.S.: Valuing Information Assets for Security Risk Management. Information Systems Security 9(4) (2000)
8. Boehm, B.W.: Software Risk Management: Principles and Practices. IEEE Software (1991)
9. Hung, M.-F.: Study on Architecture-Oriented Project Risk Management Model, Thesis, Information Management Department, National Sun Yat-sen University (2009)
10. Gilliam, D.P.: Security Risks: Management and Mitigation in the Software Life Cycle. In: Proceedings of the 13th IEEE International Workshops on Enabling Technologies, p. 3 (2004)
11. Yazar, Z.: A qualitative risk analysis and management tool – CRAMM, SANS Institute (2002)
12. Chao, W.S., Moore, J.M., Chang, C.S.: System Analysis and Design. Lambert, New York (2009)
13. Chao: Architect: System Analysis and Design—Using Software Architecture Model. Drmaster Culture Publishing, Taipei (2008)
14. Sweeney, R.: Achieving Service-Oriented Architecture: Applying an Enterprise Architecture Approach. Wiley, Hoboken (2010)
15. Lawler, J.P., Howell-Barber, H.: Service-Oriented Architecture: SOA Strategy, Methodology, and Technology. Auerbach Publications, New York (2007)

Constructing Problem-Based Learning Activities Using Self-assessment System

Feng-Jung Liu[1], Chun-Wei Tseng[2], and Wen-Chang Tseng[2]

[1] Bachelor Program of Digital Contents and Technology Management,
Cheng Shiu University, Kaohsiung County 833, Taiwan
fjliu@csu.edu.tw
[2] Department of Information Management, Cheng Shiu University,
Kaohsiung County 833, Taiwan
cwtseng@csu.edu.tw, lake75228@gmail.com

Abstract. Along with the information and communication technology getting more mature, the e-learning becomes more popular and more diverse. Many researches have pay much effort on exploiting the data-mining techniques to make user-learning more efficient. In this study, we mainly develop an experimented self-directed e-learning system, which tries to adopt adaptive testing based on the experts' knowledge and experiences to support problem-based learning activities. Within the item bank construction, we invite domain experts to assist the collection and creation of examination items and classification. Particularly, on the setting of item keywords, it is one of the most important processes for learners to easily discover related works as well as to easily share their collaborative learning activities. Additionally, after each evaluation, learners can not only follow the suggestions from the assessment system to find out the related materials which are collaboratively filtered by precursors' learning activities, but they also can easily contribute their learning modes in the same ways. We hope such collaborative self-assessment platform, which integrates the self-directed assessment system and the learning activity-based material recommendation system, make learners easier to share their learning experiences and then, improve the efficiency of self-directed learning.

Keywords: Collaborative learning, Problem-based learning, Assessment.

1 Introduction

Because of the advance of information and communication technology, ICT, E-learning becomes an important alternative way for learning. It provides learners with better alternative for learning without the limitation of time and space even it cannot substitute the whole traditional teachings. Essentially, the application of the Information technology on education is the auxiliary. That is, on education, the information technology just acts as a method or a tool in the process of learning. However, while human knowledge is accumulated to a huge mass of data, it will be aware that the self-directed learning style based on the problem-oriented policy will be raised again and will become one of adaptive learning modes in the knowledge-explosive era.[2,3].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 227–235, 2010.

Besides, information comes from different sources embedded with diverse formats in the form of metadata, making it troublesome for the computerized programming to create professional materials [3].

In this study, it mainly described an experimented self-directed e-learning system which tries to adopt adaptive testing based on the experts' knowledge and experiences into collaborative learning and tried to integrate with the learning resource recommendation system [4]. The design details will be described as follows.

2 Related Works

Along with the IT getting more mature and the rapid spread of web 2.0 concepts, the self-directed learning becomes more popular. And, the self-directed learning style based on the problem-oriented policy will be raised and will become one of adaptive learning modes in the knowledge-explosive era. With the technology development of the assessment analysis, the effect relationship among concepts can be constructed by analyzing the testing results.[5,6,7] The project's goal mainly focused on how to integrate the online evaluation system with the TjLRSE material recommendation system to support problem-based activities and to provide self-directed learners with adoptive course contents. With the evaluation system implementation, learners could understand their capabilities and efficiently learn their interested topics. Additionally, learners can efficiently improve their learning performance by learning of precursors' learning experiences. These related works are described for details as follows.

2.1 Self-directed Problem-Based Learning

Problem-based Learning (PBL) [1], pioneered at McMaster, is student-centered; it does not focus on the teacher who passes on information to the students, but on the learning process of the student. Furthermore, problems drive the learning process. That is, before students learn some knowledge they are given a problem. The problem is posed so that the students discover that they need to learn some new knowledge before they can solve the problem. So, students learn to deal with the problems they will be faced with in later professional practice.

PBL mainly utilized these important, associated and practical cases to provide the necessary learning resources and directions and made learners learn of knowledge and problem-solved skills. We followed the essence of PBL to design the self-assessment system and attempted to make learners easily join the problem-oriented learning activities during solving the questions.

2.2 TjLRSE, Tajen Learning Resource Search Engine

TjLRSE[4] is a content-searching engine in TAJEN University, Taiwan, which mainly provides search service on e-learning materials made by StreamAuthor[8] package. It not only services the learners with multimedia content query, but also provides learners with related keywords to discover related topic for advanced learning. As shown in the Fig. 1, the TjLRSE system consists of 4 parts, including web spider, search service, association rule and collaborative filtration. Web spider is

used to collect the index data, and search service mainly provides a web-based user interface. Association rule is mainly applied to find out the relations between keywords which are used by learners for searching the e-learning content. And, collaborative filtration is applied to automatically extract the correct keywords of each course by analyzing previous users' learning activities.

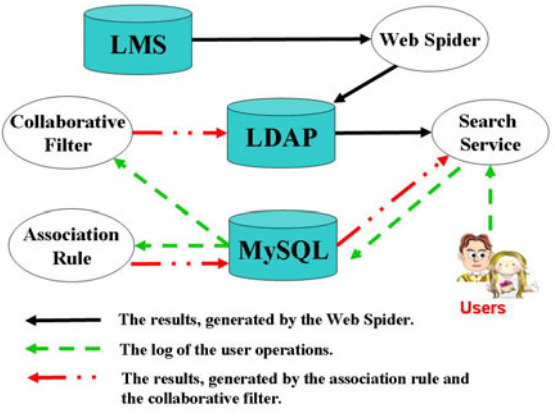


Fig. 1. Overview of the TjLRSE system

In implementing the material recommendation system, we adopted the techniques of the LDAP and the JAXB to reduce the complexity of context parsing. We used collaborative filtration technologies to conduct the feasible keywords of these contents. By mining these learners' activities to find the association among contents, the recommendation system will suggests learners with advanced learning material and related contents.

3 System Design

In the proposed problem-based self-directed learning system, it, besides, integrated the self-assessment system and TjLRSE material recommendation system. It also tried to lead learners to join the learning activities. In the material recommendation system, we successfully applied data-mining technology to advanced query services for learners.

In Fig.3, it described the self-assessment system supporting collaborative learning activities construction. In the system design, it consists of two main roles, teacher/expert and student. Student's main job is to join learning and testing activities. Students can use the self-evaluation system for on-line testing without limitation of time and space. After students finished the test, the system will suggest them with related materials according to their test results. These related materials are suggested by experts or teachers. In processing the suggested materials, teachers/experts also contribute their learning activities to material recommendation system. Different from the student's role, teachers are authorized to assign the reference material to each item.

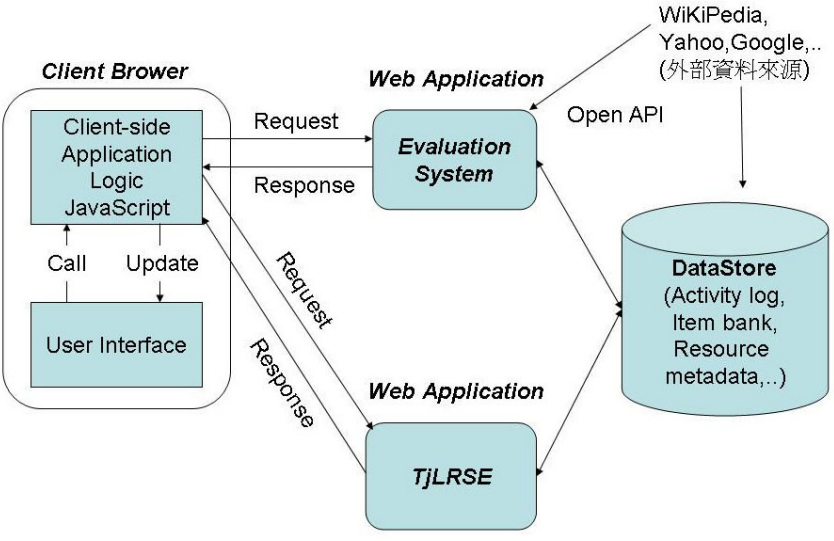


Fig. 2. Integration services of the self-assessment system and material recommendation system

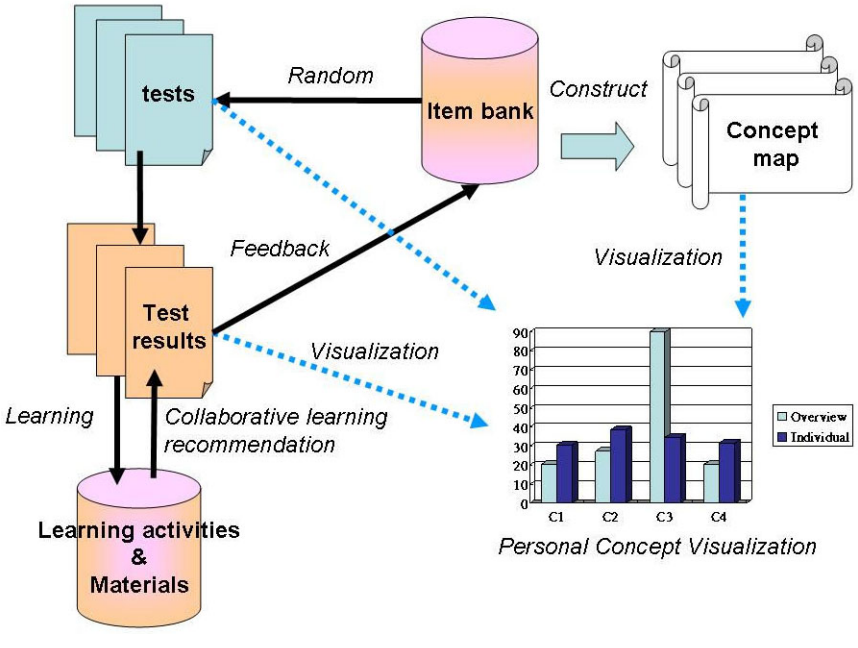


Fig. 3. The self-assessment system supporting learning activity construction

These associated materials will be provided as the URLs by material recommendation system. Students just clicked these hyperlinks and then these e-learning contents will be retrieved for them to study correctly and efficiently.

3.1 Item-Bank Construction

The basic unit of knowledge is the relation property between concept nodes as shown in Fig. 4. Further, these properties construct the knowledge network. Nevertheless, the same concept will mean differently while applied on different domains. Even in the same knowledge domain, different information will be retrieved while different topic is concentrated.

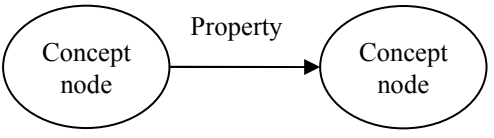


Fig. 4. Triple used as the construction of knowledge

An item consists of a question, an answer and many triples, which refer to related knowledge. Each item owns a concept independence coefficient to the concept triple. In intelligent test researches, the authors in [9] proposed an Evaluation Balance Table, EBT, to specify the key concepts of items. We made a little modification and different application of the EBT, named concept-associated item table. As show in Table 1, Q_i represents question i; C_j represents the topic concept; e_{ij} represents the weighted value of C_j related to Q_i. The e_{ij} value is ranged between 0(weak) and 5(strong). Each expert should consider the concept weight related to the test item in the item construction. With such concept-associated item table implementation, we not only use to check the similarity between items, but also evaluate the importance of each concept in the each item.

Table 1. Concept-associated item table. Q_i represents question i; C_j represents the topic concept; e_{ij} represents the degree of relation between Q_i and C_j.

Relationship Between Q_i and C_j		Concept C_j					
		C_1	C_2	…	C_k	…	C_n
Test Item Q_i	Q_1	e_{11}	e_{12}	…	e_{1k}	…	e_{1n}
	Q_2	e_{21}	e_{22}	…	e_{2k}	…	e_{2n}
	Q_3	e_{31}	e_{32}	…	e_{3k}	…	e_{3n}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Q_m	e_{m1}	e_{m2}	…	e_{mk}		e_{mn}

3.2 Setting Reference Material

The concept of Web 2.0 has led the creation of new learning mode, E-learning 2.0. In e-learning 2.0, the learning resources are not limited to the campus or classroom. Thus, the information overload problem of the explosive growth of web content is even more emphasized with the growing amount of text data in electronic form and the availability of the information. As shown in Fig. 5, the rich learning resources in the cyber net can offer teacher to explain the same concept. So, teachers can play a role of collaborative filtering and recommend their experienced learning resources about some concepts out of the rich Internet. In our implementing the setting of reference material, experts can easily select out the related content as the reference material for some units by using TjLRSE material recommendation system. Additionally, experts could filter the material with the preset keywords of items, but they can also find out the reference materials in their ways.

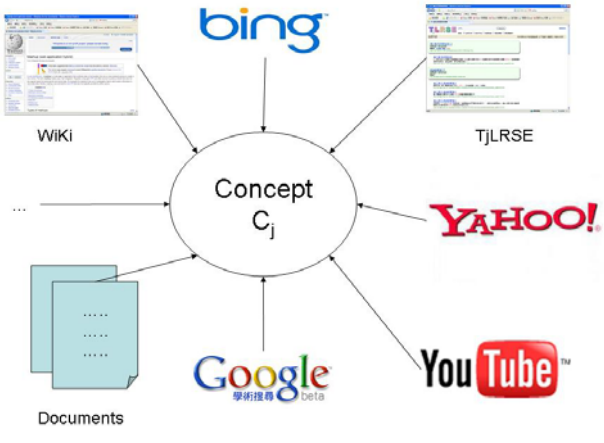


Fig. 5. Rich learning resourses

In implementation of TjLRSE, we kept at most ten terms, which are learned from learners' activities and set by experts, for each e-learning unit. Term frequency is also maintained for material recommendation. It is assumed that, in one unit if the term gets a higher frequency, then it will be better to be one of possible keywords for this unit.

4 Implementation

In implementing the self-assessment system, we mainly followed the item response theory [10], IRT, to evaluate the learners' ability. When a test starts, the assessment system will select the initial item out of the item bank. According to the item selection strategy of maximum information [11], the initial item selection should be not ever tested and should own maximum information property for testers. In the prototyping, we used the random selection instead of item selection implementation. After

Fig. 6. Teacher uses the "CPU" keyword to query and set the reference material for items

Fig. 7. It shows a student's test result, which doesn't provide the correct answer, but the related reference materials

answering each question, the ability of a learner will be automatically evaluated by the system. If the estimated ability of a learner does not meet the termination criterion yet, the next question will be provided automatically. Otherwise, the test will be terminated. The latest test result in textual form for each learner is reported by the self-assessment system and is displayed in graphical form implemented with exploiting Open Flash Chart [12]. The proposed system provides experts with a web

interface to query and set the reference materials for items as shown in Fig. 6. At the same time, these learning activities will be contributed to the collaborative filtering. After finishing the test, a learner could just follow the references, recommended by experts, to study the recommendation contents as shown in Fig. 7. The test results have been aggregated in order to provide the learners with a concept focused view. In Fig. 8, it offers learners a visualized concept overview for test results. It describes the comparison of tested items concepts with user's test result.

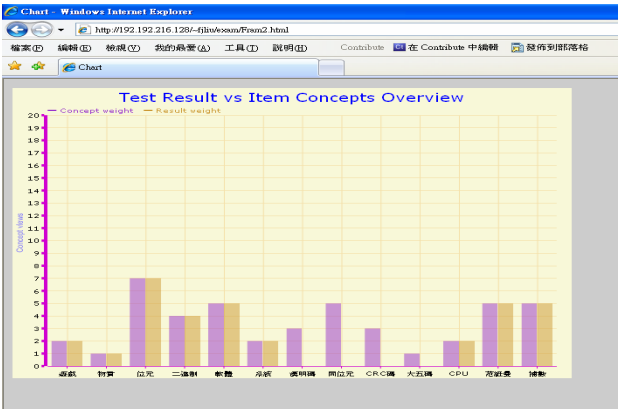


Fig. 8. Visualized item concept overview for test result. It compares user's test result with test items on concepts.

5 Conclusion and Future Works

In this paper, we adopted the self-evaluation system to lead the problem-oriented learning activities for collaborative learning. Such the e-learning system does not only invite domain experts to suggest reference materials for items, but also supports learners with e-learning activity-based material recommendation services. The experts can easily query out their interested topic materials using the TjLRSE recommendation system with keywords they input. These experts' query activity will be logged. By association rule implementation, it assists learners to achieve the integrity of course and the efficiency of learning.

The system prototype has been established but not completed. We hope the system not only provide on-line assessment and material recommendation, but also be easy for learners to join the learning activity in self-directed learning. After completing system's functionality evaluation, including item selection strategy, we will evaluate the performance of self-assessment learning system by questionnaire and by learning activities analysis. We believe that such the deployment will be helpful to achieve the better performance of studying and learning satisfaction for learners.

Acknowledgments. This work was supported by a grant from the National Science Council of Taiwan under Contract No. NSC98-2221-E-127-003.

References

1. Problem-based Learning, especially in the context of large classes, http://chemeng.mcmaster.ca/pbl/PBL.HTM
2. Masiello, I., Ramberg, R., Lonka, K.: Attitudes to the application of a Web-based learning system in a microbiology course. Computer & Education 45, 171–185 (2005)
3. Shih, B.-J., Shih, J.-L., Chen, R.-L.: Organizing learning materials through hierarchical topic maps: An illustration through Chinese herb medication. Journal of Computer Assisted Learning 23(6), 477–490 (2007)
4. Liu, F.-j., Shih, B.-J.: Learning Activity-based E-learning Material Recommendation System. In: IEEE International Symposium on Multimedia (ISM 2007), Taichung, Taiwan (December 2007)
5. Hwang, G.J., Hsiao, C.L., Tseng, C.R.: Computer-Assisted Approach to Diagnosing Student Learning Problem in Science Course. Journal of Information Science & Engineering 19(2), 229–248 (2003)
6. Tsai, C.-J., Tseng, S.S., Lin, C.-Y.: A Two-Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment. In: Alexandrov, V.N., Dongarra, J., Juliano, B.A., Renner, R.S., Tan, C.J.K. (eds.) ICCS 2001. LNCS, vol. 2074, pp. 429–438. Springer, Heidelberg (2001)
7. Liao, C.Y., Tseng, S.S., Weng, J.F.: An IRT-Based Approach to Obtaining Item-Aware Learning Achievement. In: The 23rd Workshop on Combinatorial Mathematics and Computation Theory, pp. 362–368 (2006)
8. CyberLink Stream Author, http://www.cyberlink.com/multi/products/main_7_ENU.html
9. Hsaio, H.W., Tseng, S.S., Hwang, G.J.: An evaluation model for the development of intelligent CAI system. In: Int. Conf. Computers Educ., Taiwan, China, pp. 337–339 (1993)
10. Hambleton, R.K., Swaminathan, H.: Item response theory: Principles and application. Klvwer Nijhoff, Boston (1985)
11. Lord, F.M.: Applications of item response theory to practical testing problems. Erlbaum, Hillsdale (1983)
12. The Open Flash Chart project home page, http://teethgrinder.co.uk/open-flash-chart/

Conducted EMI Analysis of a Three-Phase PWM Rectifier

Kexin Wei[1,2], Bin Liang[1,2], and Youjun Yue[2]

[1] School of Electrical Engineering & Automation, Tianjin University, Tianjin, 300072, China
[2] Tianjin Key Laboratory of Control Theory & Applications in Complicated Systems, Tianjin University of Technology, Tianjin, 300384, China
kxwei@tjut.edu.cn, tjliangbin@126.com

Abstract. Because of the strict electromagnetic interference (EMI) regulation, noises generated from high switching frequency converters need to be analyzed and suppressed by some methods. EMI noises in power converters are generally frequency related. This paper investigates the mechanisms of conducted EMI emissions associated with the typical three phases PWM rectifier system using frequency domain method. A simplified method for the calculation of Common-mode (DM) electromagnetic interference caused by the three-phase PWM rectifier is presented. Frequency domain noise current source has been represented first. The dominant high-frequency differential-mode current paths are identified later, the modeling principle of the paths has been explained at the same time, and this allows the noise spectrum to be predicted from knowledge of the component values. Theoretical predictions of EMI noises are verified by simulations in saber software through time domain and frequency domain analysis.

Keywords: PWM, EMI, modeling, differential-mode, frequency-domain.

1 Introduction

Main aims in modern power electronic systems are to deliver the power with maximum efficiency, minimum cost and weight in an integrated circuit [1]. Power electronics plays an important part in different industries when power processing is required such as in motor drives, cars and alternative energy systems. In power electronics, however, high du/dt and high di/dt are processed by fast switching to reduce loss which becomes primary sources of EMI.

Whereas, in the conventional design methodology, electromagnetic compatibility (EMC) issues are solved with “band-aid” methods by adding cumbersome and costly passive filters after a prototype is built. These traditional remedies usually have significant impact on the cost and the time-to-market for the products. To avoid such approaches after the development, it is necessary to take EMI into account at early design stage. How to predict EMI is thus becoming the major subject in recent power electronics researches. in recent years, the main EMI researches in power electronics focus on analysis of electromagnetic emissions by measurements, modeling and simulations [2-5].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 236–241, 2010.

This paper analyzed and explained the mechanisms giving rise to DM conducted interference in PWM rectifier systems. A simplified frequency domain model for noise prediction is presented here together with time domain simulations using the Saber package to determine the high-frequency current paths and sources. The Saber model, which includes high-frequency effects not normally considered by MATLAB or PSIM, proves to be an excellent tool for very detailed investigation of the high-frequency currents.

2 Noise Source Modeling

The topology of three-phase Voltage Source Rectifier (VSR) is shown in Figure.1. For simplified analysis, The proposed modeling method is based on the following basic assumptions: (1) The switching pattern is considered to be perfectly periodic. This is a pessimistic assumption, since the random changes in the period would decrease the EMI spectrum average [6]. (2) The power supplied in this system is three-phase smooth pure sine-wave source voltage;(3) The inductances used as unsaturated filters in net side are linear; (4) The switching device represented by equivalent loss resistor of the actual switch are in series with the ideal switch device[7].

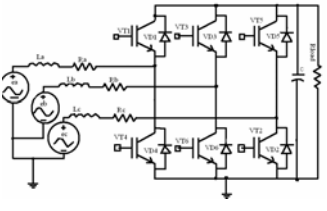


Fig. 1. The topology of the VSR

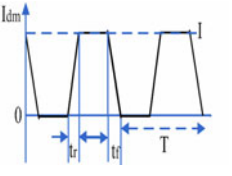


Fig. 2. Frequency domain circuit and noise current source representation

DM disturbances are considered to be caused by sudden changes (di/dt) of load current flowing through inductive paths and will be modeled by current sources. The switching noise source can be considered as a trapezoidal pulse train, Although the actual waveform will have different current rise and fall times, to simplify analysis, it is reasonable to assume the $t_f = t_r$. The frequency domain representation of the DM noise current can then be expressed in (1). [8]

$$I_{dm} = 2Id \frac{\sin(n\pi d)}{n\pi d)} \frac{\sin(n\pi_f / T)}{n\pi_f / T} \tag{1}$$

Where d is the duty cycle, T is the inverter switching period, I is the current amplitude, and n is the harmonic order.

3 The Equivalent Circuit of the DM EMI

According to the requirement of the GJB152A-97 EMI measurement standard, we can access the stable network impedance circuit (LISN)in circuit. The LISN which used in

test is shown in figure.2.The main effect of it are as follows: (1) Reducing the affect to the results of measurement which product by network impedance;(2) Isolating interference from the grid [9].Because the isolation effect of LISN, grid can be considered to be a power which just only have wave electromotive force and inside impedance. The equivalent circuit of the study subject is shown in figure.3.

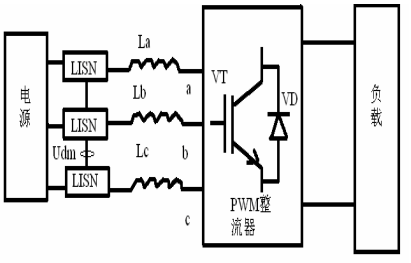


Fig. 3. Test pattern of DM EMI

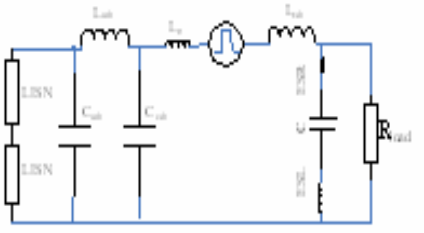


Fig. 4. Equivalent circuit for the PWM rectifier

For an accurate EMC model of the inverter it is necessary to take into account the HF parasitic paths. Fig. 4 shows the HF equivalent circuit for the 3-phase PWM rectifier.

The most important parasitic paths of this circuit are: the parasitic inductance of the emitter Le and the internal parasitic capacitances of the IGBT. The value of Le is taken from device datasheet and parasitic capacitances of IGBT are included in the IGBT Saber model. DM interference is mainly given rise to EMI current flow in transmission line. Under the high power case, transmission line is often long and thus the effect of it's parasitic parameters, Lcab and Ccab, should not be neglected. In this paper, we selected the classic π networks [12] for modeling the transmission line. However, Stray inductances of the connecting wires have very small values and affect principally the differential conducted emissions. For this study this inductance has been neglected because the length of the connecting wire is very small.

As for the differential-mode (DM) noise, it's source impedance is strongly influenced by the equivalent series resistance (ESR), and the equivalent series inductance (ESL) of the bulk capacitor [10]. In the HF range the equivalent circuit of the DC link capacitor consists of the series combination of capacitance, resistance and inductance, as show in fig.5.

When the frequency increases the impedance of the capacitor decreases linearly at rate of -20dB/decade. The impedance of the inductor increase until it equals that of the capacitor at the point of resonance. At corner frequency point the impedance is Rs. For higher frequencies the impedance of the inductor increases at rate of +20dB/decade. The impedance of the DC link capacitor bas a strong effect upon the differential conducted emissions [11]. And the values of them should be measured with an impedance analyzer.

On the basis of the data obtained through measuring by others [12-14] and with FFT analysis in saber simulation we can verified the previous numerical results of high-frequency parasitic parameters of the model. Thus we can get the high-frequency

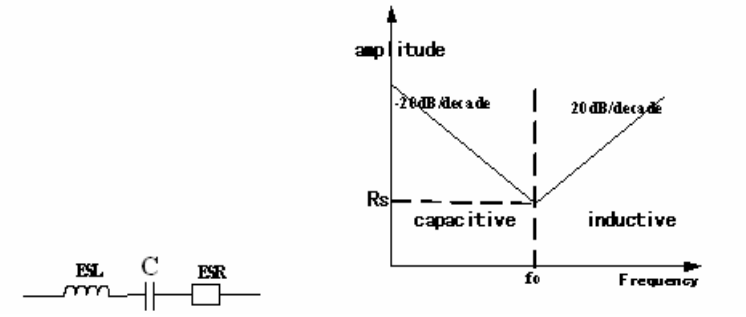


Fig. 5. High frequency model of DC link capacitor

equivalent model of the three-phase PWM rectifier as shown in figure.4. Now if we ignore the commutation model of rectifier bridges, and there are only two power tubes conducting at the same time, so LISNs can equivalent to be 100Ω in the DM channel.

4 The Simulation and Validation

Lacking of some instruments such as impedance analyzer, we get some relevant parameters according to the previous studies and analysis from others' literatures [12-14]. And the switching frequency is 20kHz.

Because we don't consider the commutation process of switch tube, IGBT element approximation is treated to be an trapezoidal ideal switch. Thus, PWM signal which provided to the main circuit is used without amplification, and the AC voltage source is considered to be ideal too. According to the configuration pattern in Figure.3, The simulation of DM interference voltage are get from each 50Ω balancing resistors. We can get the time-domain waveforms as in figure.6(a) by the saber simulation. This figure is the voltage on AC side, we can see that the voltage is in the same phase with current, but it should not be neglected that, the wave pattern is not the real sinusoidal wave, some time-variant harmonic components and unknown radio-frequency component exist in it as shown in fig 6(b). So, a further analysis to these components is needed to be done.

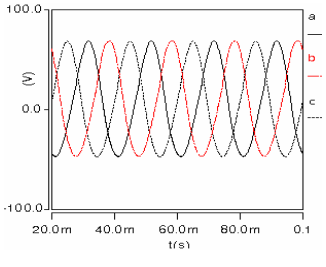


Fig. 6. (a) Time-domain voltage waveforms of AC side

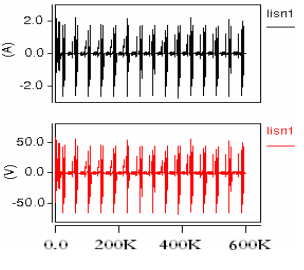


Fig. 6. (b) The frequency-domain analysis of the interfering voltage in each balancing resistors

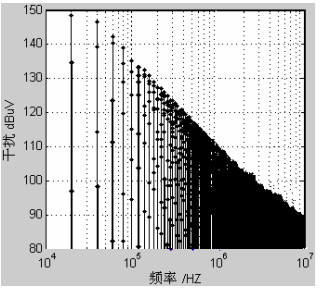


Fig. 7. Calculated interferential Spectrum

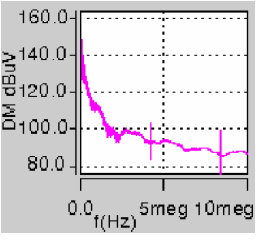


Fig. 8. Frequency domain simulation

As can be seen from the calculated diagram 7, the DM frequency spectrum spread over switching frequency point and it's multiples. The strongest EMI appear near the frequency point, the highest interference reach to 150dBuV, and as the frequency increase, the interference presents a downward trend rapidly. From the comparison in fingure.8, which is got after the FFT analysis for the time-domain waveforms, the results of theoretical analysis are in agreement with those of numerical simulation. And the accuracy of the model is good enough to predict the DM EMI, since the error of the numerical result is less than 3dBuV in high frequency range.

5 Conclusion

In this paper, we analyzed the mechanism of differential-mode interference produced on the three-phase rectifier. A simplified frequency domain model of three-phase PWM rectifier was proposed. Through taking advantage of the method, it gives immediately the exact operation point of the converter control variables. Although it is not possible to take into account all the parasitic elements in the propagation path, the accuracy of the predicted results is ensured and the established high-frequency model is verified by time-domain simulations. This research can provide reference for prediction of conducted EMI on the inverter system.

Acknowledgments

This project is from Tianjin Key Laboratory of Control Theory and Applications in Complicated Systems. And it is supported by the National Natural Science Fundation of China (50977063), the National High Technology Research and Development Program of China (863 program) (2008AA11A145), Key science and technology supporting project of Tianjin (09ZCKFGX01800), Special Found for Technology Innovation of Tianjin (08FDZDGX02000).

References

1. Zare, F.: EMI Issues in Modern Power Electronic Systems. In: Proceedings of the EMC Symposium, pp. 16–18 (2009)

2. Tarateeraseth, V., See, K.Y., Canavero, F.G.: Accurate Extraction of Noise Source Impedance of an SMPS under Operating Conditions. J. IEEE Trans. on Power Electronics 25(1), 111–117 (2010)
3. Foissac, M., Schanen, J.L., Vollaire, C.: Compact EMC model of power electronics converter for conducted EMC studies in embedded networks. J. Automotive Power Electronics (2009)
4. Ma, W.M., Zhao, Z.H., Meng, J., et al.: Precise methods for conducted EMI modeling, analysis, and prediction. J. Sci. China Ser. E Tech. Sci. 51(6), 641–655 (2008)
5. Akagi, H., Shimizu, T.: Attenuation of conducted EMI emissions from an inverter-driven motor. J. IEEE Trans. Power Electron 23(1), 282–290 (2008)
6. Implementing random space vector modulation with the ADMCF32X, Analog Devices, Norwood, MA, Applicat. Note ANF32X-54 (2000)
7. Zhang, X., Zhang, C.: PWM rectifier and control. M. Beijing Industry Press, Beijing (2003)
8. Lai, J.-S., Huang, X., Pepa, E.: Inverter EMI Modeling and Simulation Methodologies. In: Industrial Electronics Society, pp. 1533–1539 (2003)
9. Pei, X.J., Zhang, K., Kang, Y., Chen, J., et al.: Prediction of common mode conducted EMI in single phase PWM inverter. In: Power Electronics Specialists Conference, pp. 3060–3065 (2004)
10. Liu, Q., Wang, S., Wang, F., Baisden, C., Boroyevich, D.: EMI Suppression in Voltage Source Converters by Utilizing dc-link Decoupling Capacitors. J. IEEE Trans. Power Electron 22(4), 1417–1428 (2007)
11. Moreira, A.F., Lipo, T.A.: High-frequency modeling for cable and induction motor overvoltage studies in long cable drives. J. 38, 1297–1306 (2002)
12. Ran, L., Gokani, S., Clare, J.: Conducted electromagnetic emission in induction motor drive systems part 2: Frequency domain mode. J. IEEE Transactions on Power Electronics 13(4), 768–776 (1998)
13. Jiang, Y.: Research on conducted interference in DC power grid with three-phase inverter. Huazhong University of Science and Technology (2006)
14. Lai, J.-S., Huang, X., Chen, S.: EMI Characterization and Simulation With Parasitic Models for a Low-Voltage High-Current AC Motor Drive. J. IEEE Transactions on Industry Applications 40, 178–185 (2004)

Synchronization of Duffing-Holmes Oscillators Using Stable Neural Network Controller

Suwat Kuntanapreeda

Department of Mechanical and Aerospace Engineering
Faculty of Engineering
King Mongkut's University of Technology North Bangkok
Bangkok 10800, Thailand
suwat@kmutnb.ac.th

Abstract. This paper presents a neural network controller for synchronization of two Duffing-Holmes oscillators. A Duffing-Holmes oscillator is a chaotic system describing a dynamics of the forced vibration of a buckled elastic beam. The controller is a feedforward neural network trained to drive the first Duffing-Holmes oscillator so that its states converge to those of the other Duffing-Holmes. The training scheme is based on a model reference strategy with imposing stability conditions on the controller's parameters. The stability condition guarantees the convergence of the synchronization errors. Numerical simulations are conducted to illustrate the feasibility and effectiveness of the stable neural network controller.

Keywords: Neural network, Neural network control, Duffing-Holmes oscillators, Chaos control.

1 Introduction

A Duffing-Holmes oscillator is the nonlinear system that shows chaotic behavior. Chaotic systems are deterministic systems, but their behaviors are very sensitive to initial conditions and unpredictable. They possess strange oscillating motions which are not periodic. Well-known paradigms of the chaotic systems include the famous van der Pol oscillators, the Duffing-type oscillators, the generalized Lorenz equations, the Lur'e systems, and the Chua's circuits.

Control of the chaotic systems has received increasing attention since the pioneering work of Ott et al. [1] was published. The behavior of the chaotic systems is random-like, but the systems are actually deterministic. Their behavior is very sensitive to initial conditions and unpredictable. This makes the chaos control problem very challenging. Recently, many methods have been proposed for controlling of chaos such as backstepping design [2], sliding mode control [3], passivity-based control [4], neural network control [5], fuzzy logic control [6], linear matrix inequality (LMI) technique [7] and adaptive control [8].

This paper presents a neural network-based design for chaos synchronization of Doffing-Holmes oscillators. The two systems in synchronization are called master

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 242–251, 2010.

system and slave system, respectively. The controller is a multi-layer feedforward neural network trained to drive a slave system to synchronize with a master system. The training scheme is based on a model reference strategy with imposing stability conditions on the controller's parameters. The stability condition guarantees the convergence of the synchronization errors. Numerical simulations are provided to show the effectiveness of the proposed design.

2 Chaos Synchronizations

The two systems in synchronization are called master system and slave system, respectively. A master system and a slave system can be defined as

$$\dot{\mathbf{x}}_m(t) = \mathbf{f}(\mathbf{x}_m(t)) \tag{1}$$

and

$$\dot{\mathbf{x}}_s(t) = \mathbf{f}(\mathbf{x}_s(t)) + \mathbf{u}(t), \tag{2}$$

where $\mathbf{x}_m(t)$ and $\mathbf{x}_s(t)$ are the state vectors of the master and the slave systems, respectively, $\mathbf{f}(\cdot)$ is the nonlinear function vector, and $\mathbf{u}(t)$ is the control input. Let define the synchronization error vector as $\mathbf{e} = \mathbf{x}_s(t) - \mathbf{x}_m(t)$. Thus, the synchronization error system can be described by

$$\begin{aligned} \dot{\mathbf{e}} &= \dot{\mathbf{x}}_s(t) - \dot{\mathbf{x}}_m(t) \\ &= \mathbf{g}(\mathbf{x}_s(t), \mathbf{x}_m(t)) + \mathbf{u}(t), \end{aligned} \tag{3}$$

where $\mathbf{g}(\mathbf{x}_s(t), \mathbf{x}_m(t)) = \mathbf{f}(\mathbf{x}_s(t)) - \mathbf{f}(\mathbf{x}_m(t))$. Here, $\mathbf{u}(t)$ is the control signal designed such that the error vector converges to zero, resulting in two synchronized systems.

3 Duffing-Holmes Oscillators

The Duffing-Holmes equation describes the dynamics of the forced vibration of a buckled elastic beam when only one mode of vibration is considered [9]. It is a Duffing-type equation, which models a nonlinear oscillator with a cubic stiffness term to describe the hardening spring effect. However, the linear stiffness term of the Duffing-Holmes oscillator is negative.

The governing equation of the Duffing-Holmes oscillator can be written in the form

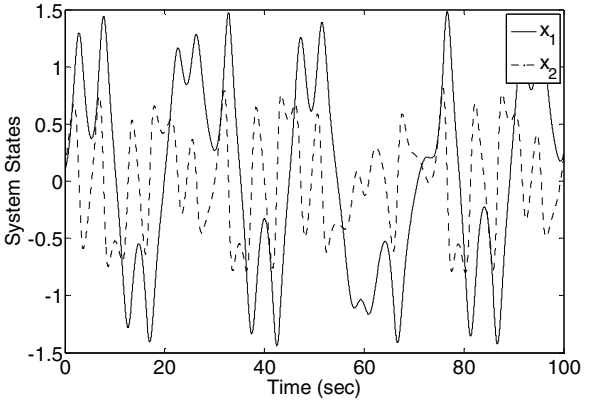
$$\ddot{x} + cv + k_0 x + k_1 x^3 = f\cos(\omega t) \tag{4}$$

where x is the oscillation displacement, $c > 0$ is the damping constant, $k_0 < 0$ is the linear stiffness constant, $k_1 > 0$ is the cubic stiffness constant, $f > 0$ is the amplitude of the excitation force, and $\omega > 0$ is the forcing frequency. When $f = 0$, the equation has three static equilibriums at $x = \pm\sqrt{-k_0/k_1}$ and $x = 0$. By defining the states of

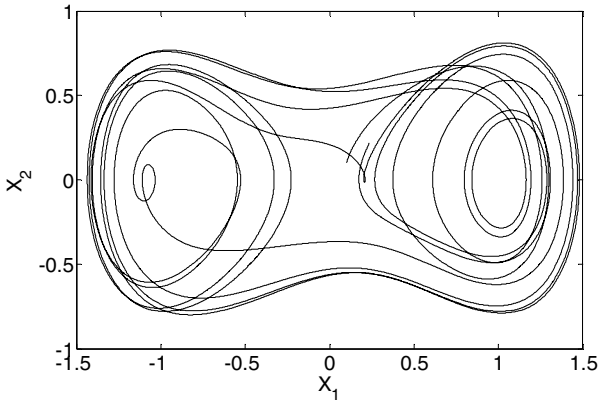
Eq.(4) as $x_1 = x$ and $x_2 = \dot{x}$, the Duffing-Holmes oscillator is represented by the state equations

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -k_0 x_1 - k_1 x_1^3 - c x_2 + f \cos(\omega t). \end{aligned} \tag{5}$$

An example of the chaotic behavior of Eq.(5) is shown in Fig. 1.



(a)



(b)

Fig. 1. Chaotic behavior of a Duffing-Holmes System ($k_0 = -1, k_1 = 1$, $c = 0.25$, $f = 0.3$, $\omega = 1$). The initial condition is $x_1(0) = x_2(0) = 0.1$. (a) State response and (b) Phase-space orbit

4 Synchronization of Duffing-Holmes Oscillators

According to Eq.(5), the master and the slave Duffing-Holmes oscillators can be defined as

$$\begin{aligned} \dot{x}_{m1} &= x_{m2} \\ \dot{x}_{m2} &= -k_0 x_{m1} - k_1 x_{m1}^3 - c x_{m2} + f\cos(\omega t), \end{aligned} \tag{6}$$

and

$$\begin{aligned} \dot{x}_{s1} &= x_{s2} \\ \dot{x}_{s2} &= -k_0 x_{s1} - k_1 x_{s1}^3 - c x_{s2} + f\cos(\omega t) + u \end{aligned} \tag{7}$$

where the lower scripts *m* and *s* stand for the master and the slave, respectively, and *u* is the control signal. By defining the synchronization errors as $e_1 = x_{s1} - x_{m1}$ and $e_2 = x_{s2} - x_{m2}$ and using Eq.(6) and Eq.(7), the synchronization error system can be written as

$$\begin{aligned} \dot{e}_1 &= e_2 \\ \dot{e}_2 &= -k_0 e_1 - c e_2 - k_1 (x_{s1}^3 - x_{m1}^3) + u. \end{aligned} \tag{8}$$

Here, the control signal *u* is designed such that the synchronization errors converge to zero, resulting in two synchronized systems.

5 Control Design Process

In this paper the control objective is to drive the synchronization errors to zero. Let express the control signal as $u = u_1 + u_2$, where the signal $u_1 = k_1(x_s^3 - x_m^3)$ is the nonlinear feedback term and the signal u_2 is generated by the neural network, which will be trained using the design process discussed later. By substituting $u = u_1 + u_2$, Eq.(8) becomes

$$\begin{aligned} \dot{e}_1 &= e_2 \\ \dot{e}_2 &= -k_0 e_1 - c e_2 + u_2. \end{aligned} \tag{9}$$

This error system will be used to train the controller.

The first step in the control design process is the training of a model network. Fig. 2 illustrates how the model network can be trained. Note that the model network must be trained prior the training of the controller network. The model network is trained to mimic the dynamics behaviour of Eq.(9). The inputs of the network comprise the synchronization errors, $[e_1(k), e_2(k)]$, and the control signal, $u_2(k)$ at the time step *k*. The outputs are the estimates of the synchronization errors at the next time step,

$$\mathbf{e}_{nn}(k+1) = \begin{Bmatrix} e_{nn1}(k+1) \\ e_{nn2}(k+1) \end{Bmatrix} = NN_{md}\left(\begin{Bmatrix} u(k) \\ e_1(k) \\ e_2(k) \end{Bmatrix}; (\mathbf{V}_{md}, \mathbf{a}_{md}, \mathbf{W}_{md}, \mathbf{b}_{md}) \right). \tag{10}$$

where the lower script *md* refers to the model network. By considering the differences between the synchronization errors and the model network's outputs,

$$\boldsymbol{\varepsilon}_{md} = \begin{Bmatrix} e_1 - e_{nn1} \\ e_2 - e_{nn2} \end{Bmatrix},$$

the parameters of the model network can be adjusted, for example, with the backpropagation algorithm.

Let the training index of performance be $J_1 = \boldsymbol{\varepsilon}_{md}^T \boldsymbol{\varepsilon}_{md}$. The gradient vector of the performance index can be expressed as

$$\frac{\partial J_1}{\partial \mathbf{A}_{md}} = \frac{\partial J_1}{\partial \mathbf{e}_{nn}} \frac{\partial \mathbf{e}_{nn}}{\partial \mathbf{A}_{md}},$$

where $\mathbf{A}_{md}$ is a set of the model network's parameters, $\mathbf{V}_{md}$, $\mathbf{W}_{md}$, $\mathbf{a}_{md}$ and $\mathbf{b}_{md}$. Note that, at this step, there is no controller network present. Once the model network has been trained, it is used in a procedure that adjusts the weights of the controller network.

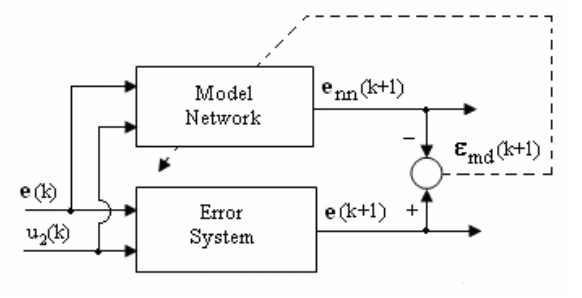


Fig. 2. Training of the model network

The next step is the training of the controller network which is another multi-layer feedforward network. The control design is suggested by the block diagram in Fig. 3. The design is based on a model reference neural network control strategy [10]. The input of the controller network is the synchronization errors at the time step *k*, $[e_1(k), e_2(k)]$. The control signal,

$$u_2(k) = NN_c\left(\begin{Bmatrix} e_1(k) \\ e_2(k) \end{Bmatrix}; \left(\mathbf{V}_c, \mathbf{a}_c, \mathbf{W}_c, \mathbf{b}_c\right)\right), \tag{11}$$

is the output of the controller network at the time step *k*. Here, the lower script *c* refers to the controller network. The controller network is trained to drive the slave system such that the differences between the synchronization errors and the outputs of a reference model,

$$\boldsymbol{\varepsilon}_c = \begin{Bmatrix} e_{re1} - e_1 \\ e_{re2} - e_2 \end{Bmatrix},$$

are minimized.

Let the training index of performance be $J_2 = \boldsymbol{\varepsilon}_c^T \boldsymbol{\varepsilon}_c$. The gradient vector of the performance index can be expressed as

$$\frac{\partial J_2}{\partial \mathbf{A}_c} = \frac{\partial J_2}{\partial u} \frac{\partial u}{\partial \mathbf{A}_c},$$

where $\mathbf{A}_c$ is a set of the controller network's parameters, $\mathbf{V}_c$, $\mathbf{W}_c$, $\mathbf{a}_c$ and $\mathbf{b}_c$. In order to calculate $\partial J_2 / \partial u$, the derivative $\partial \boldsymbol{\varepsilon}_c / \partial u$ is needed. However, this derivative is usually not available, but it can be estimated. In Fig. 3, the controller network's weights are adjusted by backpropagating the differences between the synchronization errors and the outputs of a reference model, $\boldsymbol{\varepsilon}_c$, through the model network along the path shown as a dashed line. The process that backpropagates the differences, $\boldsymbol{\varepsilon}_c$, through the model network from its output to its input is basically a technique for estimating the derivative $\partial \boldsymbol{\varepsilon}_c / \partial u$. In this step the model network's weights are frozen.

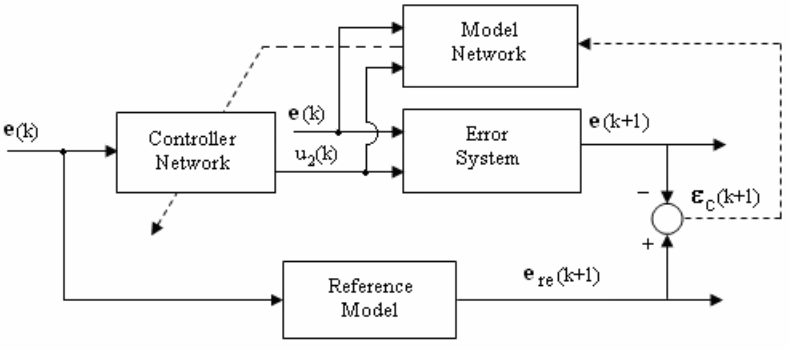


Fig. 3. Training of the controller network

In order to guarantee the convergence of the synchronization errors, the training algorithm used to adjust the parameters of the controller network is modified by imposing pre-derived stability conditions on the trained parameters. The concept was proposed in [11]. A derivation of the stability conditions is given below.

The controller network is assumed to be implemented as a single hidden layer feedforward network with a linear output layer and without bias vectors in both layers. Since $k_0 < 0$, the error system (9) is nonhermitian. To solve the nonhermitian problem, a direct weight matrix, $\mathbf{U}_c = [U_{ji}]$, connecting between the inputs and the outputs of the controller network is needed. Thus, the neural network control law (11) becomes

$$u_2 = \mathbf{U}_c^T \mathbf{e} + \mathbf{W}_c^T \boldsymbol{\sigma}\left(\mathbf{V}_c^T \mathbf{e}\right) = \mathbf{U}_c^T \mathbf{e} + \mathbf{W}_c^T \boldsymbol{\sigma}(\mathbf{z}), \tag{12}$$

where $\mathbf{e} = [e_1 \quad e_2]^T$ and $\mathbf{z} = \mathbf{V}_c^T \mathbf{e}$. Note that the learning parameters of the controller network consists of $\mathbf{U}_c$, $\mathbf{V}_c$, and $\mathbf{W}_c$. Substituting Eq. (12) into Eq.(9) results

$$\begin{aligned}\dot{\mathbf{e}} &= \mathbf{A}\mathbf{e} + \mathbf{B}\left(\mathbf{U}_c^T\mathbf{e} + \mathbf{W}_c^T\boldsymbol{\sigma}(\mathbf{z})\right) \\ &= \left(\mathbf{A} + \mathbf{B}\mathbf{U}_c^T\right)\mathbf{e} + \mathbf{B}\mathbf{W}_c^T\boldsymbol{\sigma}(\mathbf{z}))\end{aligned} \tag{13}$$

where $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -k_0 & -c \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Let $V(\mathbf{e}) = \mathbf{e}^T\mathbf{P}\mathbf{e}$ be a Lyapunov candidate, where $\mathbf{P} = \mathbf{P}^T$ is a positive definite matrix. The derivative of the Lyapunov candidate along any trajectory of Eq.(13) is

$$\begin{aligned}\dot{V} &= \dot{\mathbf{e}}^T\mathbf{P}\mathbf{e} + \mathbf{e}^T\mathbf{P}\dot{\mathbf{e}} \\ &= \left(\left(\mathbf{A} + \mathbf{B}\mathbf{U}_c^T\right)\mathbf{e} + \mathbf{B}\mathbf{W}_C^T\boldsymbol{\sigma}(\mathbf{z})\right)^T\mathbf{P}\mathbf{e} + \mathbf{e}^T\mathbf{P}\left(\left(\mathbf{A} + \mathbf{B}\mathbf{U}_c^T\right)\mathbf{e} + \mathbf{B}\mathbf{W}_C^T\boldsymbol{\sigma}(\mathbf{z})\right) \\ &= \mathbf{e}^T\left(\left(\mathbf{A} + \mathbf{B}\mathbf{U}_c^T\right)^T\mathbf{P} + \mathbf{P}\left(\mathbf{A} + \mathbf{B}\mathbf{U}_c^T\right)\right)\mathbf{e} + 2\mathbf{e}^T\mathbf{P}\mathbf{B}\mathbf{W}_C^T\boldsymbol{\sigma}(\mathbf{z}).\end{aligned}$$

Let assuming that there exists a vector $\mathbf{q}$ such that

$$\left(\mathbf{A} + \mathbf{B}\mathbf{U}_C^T\right)^T\mathbf{P} + \mathbf{P}\left(\mathbf{A} + \mathbf{B}\mathbf{U}_C^T\right) = -\mathbf{q}\mathbf{q}^{\mathbf{T}} - \mathbf{I} \tag{14}$$

and

$$\mathbf{P}\mathbf{B}\mathbf{W}_C^T + \sqrt{2}\mathbf{q} + \mathbf{V}_{\mathbf{C}} = \mathbf{0} \tag{15}$$

where $\mathbf{I}$ is a identity matrix. Then, it yields

$$\begin{aligned}\dot{V} &= -\mathbf{e}^T\mathbf{e} - \mathbf{e}^T\mathbf{q}\mathbf{q}^T\mathbf{e} - 2\mathbf{e}^T\mathbf{q}\sqrt{2}\boldsymbol{\sigma}(\mathbf{z}) - 2\mathbf{e}^T\mathbf{V}_c\boldsymbol{\sigma}(\mathbf{z}) \\ &= -\mathbf{e}^T\mathbf{e} - \left\|\mathbf{e}^T\mathbf{q} + \sqrt{2}\boldsymbol{\sigma}^T(\mathbf{z})\right\|^2 - 2\boldsymbol{\sigma}^T(\mathbf{z})\left(\mathbf{z} - \boldsymbol{\sigma}(\mathbf{z})\right).\end{aligned} \tag{16}$$

Where $\mathbf{z} = \mathbf{V}_c^T\mathbf{e}$. Choosing the activate functions to be hyperbolic tangents,

$$\boldsymbol{\sigma}(\mathbf{z}) = [\tanh(z_1) \quad \tanh(z_2) \cdots \tanh(z_p)]^T, \tag{17}$$

yields $\boldsymbol{\sigma}^T(\mathbf{z})\left(\mathbf{z} - \boldsymbol{\sigma}(\mathbf{z})\right) > 0$. Thus, Eq. (16) becomes $\dot{V} < 0$. Therefore, it can be concluded that the synchronization error vector, $\mathbf{e}$, converges to zero whenever the parameters of the controller network satisfy the stability conditions given by Eqs.(14) and (15). These stability conditions should be imposed in the training process of the controller network to guarantee the convergence. The reader is referred to [11] for a modified backpropagation algorithm that imposes stability conditions into the backpropagation algorithm. Note that the parameters adjusted in the training process compose of the controller network's weight matrices, $\mathbf{U}_c$, $\mathbf{V}_c$ and $\mathbf{W}_c$, the matrix $\mathbf{P}$, and the vector $\mathbf{q}$.

Once the controller network has been trained, it can be used as a feedback controller as shown in Fig. 4. Note that the model network and the reference model are not presented.

5 Simulation Results

The effectiveness of the proposed controller is verified via computer simulations. The parameters of the Duffing-Holmes oscillator are $k_0 = -1$, $k_1 = 1$, $c = 0.25$, $f = 0.3$

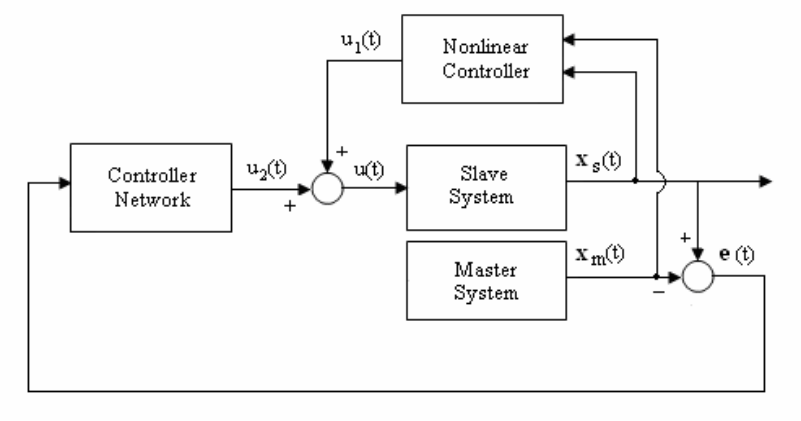


Fig. 4. Operational trained neural network control system

and $\omega = 1$. The reference model that is used for training the controller networks is selected as a second order linear system

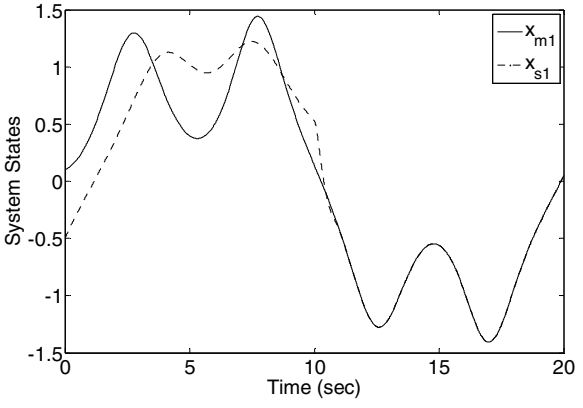
$$\begin{aligned} \dot{e}_{re1} &= e_{re2} \\ \dot{e}_{re2} &= -25e_{re1} - 8e_{re2} \end{aligned} \tag{18}$$

which has the damping ratio of 0.8 and the settling time of 1 second.

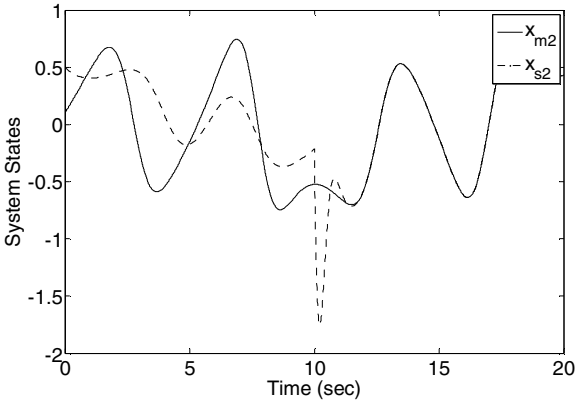
The model network is configured as a single hidden layer network; with three inputs to the input layer, followed by a ten-neuron hidden layer, and an output layer with two neurons. In short, it is said that the model network has a structure of 3-10-2. The activate functions of the hidden layer and the output layer are hyperbolic tangents. Three thousand input-output data points randomly sampled from the error system (9) are used to train the model network with the sampling time of 0.005 seconds. The training process follows the diagram in Fig. 3, using the backpropagation as the training rule. The final mean-square-value of the differences, ε_{md}, between the synchronization errors and the model network's outputs is 9.323×10^{-8}.

The structure of the controller network is 2-5-1 without bias vectors. The activate functions of the hidden layer and the output layer are hyperbolic tangents and linear, respectively. Another three thousand input-output data points randomly sampled from the reference model (18) is used to train the controller network. The sampling time of 0.005 seconds is also used. The training process follows the diagram in Fig. 4, using the backpropagation training rule with imposing the stability conditions (14) and (15). The final mean-square-value of the differences, ε_c, between the synchronization errors and the outputs of a reference model is 1.166×10^{-4}. Fig. 6 displays the state response and the synchronization errors of the controlled Duffing-Holmes oscillators. The controller is activated at t = 10 sec.

In Fig. 5, the results illustrate that the controller is able to drive the slave system to synchronize the master system as desired. The synchronization errors converge to zero within about 1 second after the controller was activated. The values of the trained parameters are shown in Table 1.



(a)



(b)

Fig. 5. System responses of the Duffing-Holmes systems. (a) States x_{m1}, x_{s1} and (b) States x_{m2}, x_{s2}.

Table 1. Parameters of the trained controller network

$\mathbf{V}_c^T$	$\mathbf{W}_c$	$\mathbf{U}_c$	$\mathbf{P}$	$\mathbf{q}^T$
$\begin{bmatrix} -0.6191 & 1.6276 \\ -0.5811 & 1.7463 \\ 0.4421 & -3.0212 \\ -0.7522 & 0.9854 \\ 0.8689 & -0.5688 \end{bmatrix}$	$\begin{bmatrix} -0.0428 \\ -0.0389 \\ 0.0272 \\ -0.0579 \\ 0.0784 \end{bmatrix}$	$\begin{bmatrix} -1.2129 \\ -0.0748 \end{bmatrix}$	$\begin{bmatrix} 11.5250 \\ 12.7321 \\ 12.7321 \\ 76.3145 \end{bmatrix}$	$\begin{bmatrix} 0.8228 & 1.1570 \\ 0.7615 & 0.8670 \\ -0.5575 & 0.6685 \\ 1.0533 & 2.4284 \\ -1.3206 & -3.8307 \end{bmatrix}$

6 Conclusions

In this paper, a neural network was proposed as an alternative solution for chaos synchronization of the Duffing-Holmes oscillators. The training scheme for the controller is based on a model reference strategy with imposing stability conditions on the controller's parameters. The stability conditions guarantee the convergence of the synchronization errors. Numerical simulations show that the trained neural network controller is able to drive the states of the slave systems to asymptotically synchronize the states of the master systems as desired.

References

1. Ott, E.F., Grebogi, C., Yorke, J.A.: Controlling Chaos. Phys. Rev. Lett. 64, 1196–1199 (1990)
2. Peng, C.-C., Chen, C.-L.: Robust Chaotic Control of Lorenz System by Backstepping Design. Chaos, Solitons and Fractals 37, 598–608 (2008)
3. Nazzal, J.M., Natsheh, A.N.: Chaos Control using Sliding-mode Theory. Chaos, Solitons and Fractals 33, 695–702 (2007)
4. Sangpet, T., Kuntanapreeda, S.: Output Feedback Control of Unified Chaotic Systems Based on Feedback Passivity. Int. Journal of Bifurcation and Chaos 20, 1519–1525 (2010)
5. Kuntanapreeda, S.: An Observer-based Neural Network Controller for Chaotic Lorenz System. In: Kang, L., Cai, Z., Yan, X., Liu, Y. (eds.) ISICA 2008. LNCS, vol. 5370, pp. 608–617. Springer, Heidelberg (2008)
6. Meda-Campana, J.A., Castillo-Toledo, B., Chen, G.: Synchronization of Chaotic Systems from a Fuzzy Regulation Approach. Fuzzy Sets and Systems 160, 2860–2875 (2009)
7. Kuntanapreeda, S.: Chaos synchronization of unified chaotic systems via LMI. Physics Letters A 373, 2837–2840 (2009)
8. Sangpet, T., Kuntanapreeda, S.: Adaptive Synchronization of Hyperchaotic Systems via Passivity Feedback Control with Time-varying Gains. Journal of Sound and Vibration 329, 2490–2496 (2010)
9. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, Heidelberg (1983)
10. White, D.A., Sofge, D.A.: Handbook of Intelligent Control: Neural, Fuzzy and Adaptive. Van Nostrand Reinhold, New York (1992)
11. Kuntanapreeda, S., Fullmer, R.R.: A training rule which guarantees finite-region stability for a class of closed-loop neural network control systems. IEEE Transactions on Neural Networks 7(3), 745–751 (1996)

Codes Base on Unambiguous Products

Ho Ngoc Vinh[1], Vu Thanh Nam[2], and Phan Trung Huy[2]

[1] Vinh Technical Teachers Trainning University, Nghean, Vietnam
hnvinh.skv@moet.edu.vn
[2] Hanoi University of Science and Technology, Hanoi, Vietnam
vtnam@cmc.com.vn, phanhuy@hn.vnn.vn

Abstract. In this paper, we propose the notion of +-unambiguous product which is expanded from unambiguous product and the definitions of alternative product, alternative code, even alternative code on a pair (X, Y) of languages. Some basic properties of +-unambiguous product, alternative codes and even alternative codes related to usual codes are given which show that these new codes can be considered as generations of codes. Necessary and sufficient conditions for alternative codes and even alternative codes are established. The independence of the claims in these conditions are proved. The existence of algorithms to decide whether a pair (X, Y) is an alternative or is an even alternative code, in case both components X, Y are regular, is shown.

Keywords: +-unambiguous product; alt-code; ealt-code; independency of conditions; generations of code.

1 Introduction

Unambiguity of the products on words, finite as well as infinite, brings many interesting properties and they seem to relate closely to codes. To enrich theory of codes, the unambiguous products in relationships with codes, automata, algebra ... areas are studied in many works. In [9] Pascal Weil used unambiguous automata, unambiguous relation semigroup as tools to establish a nice result showed the relation of groups of two codes Y, Z, where Z is finite, with groups in the wreath product $X = Y \circ Z$ of Y, Z. The notion of unambiguous product which had been introduced by M.P. Schützenberger [8] and has been studied extensively by J. E. Pin (see J. E. Pin [6]) in relation with theory of varieties of languages. J. E. Pin - P. Weil [7] established a result described an interest relationship between a variety of ordered monoids and the polynomial closure of the corresponding variety $\mathcal{V}$ of regular languages, by introduced a notion of unambiguous polynomial closure which is the closure under disjoint union and unambiguous marked product of the form $L_0a_1L_1 \ldots a_nL_n$, where the a_i's are letters and the L_i's are languages belong $\mathcal{V}$. In [5] P. T. Huy - D. L. Van established a result to express ω-regular languages of infinite words which is accepted by nonambiguous Büchi V-automata as disjoint finite union of a type of unambiguous products of languages and ω-languages whose syntactic monoids are in $\mathcal{V}$, where $\mathcal{V}$ is a variety of finite monoids closed under Schützenberger product. Codes with new products and codes with controlled shuffle are considered in [1], [2] respectively.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 252–262, 2010.

Considering unambiguity of a code by unambiguity of products of codewords in this code, in this paper, we introduce notion of +-unambiguous product of two languages which fills a middle gap between notions of unambiguous product of two languages and unambiguity of a code. Using the notion of +-unambiguous product as a restriction of the notion of unambiguous product we can define alternative codes (*alt-code*) and even alternative codes (*ealt-code*) and to establish some basic properties of +-unambiguous products and alt-codes, ealt-codes in relationship with usual codes. By Remark 3, alt-codes and ealt-codes can be considered as generations of usual codes and allow us to establish a classification of subclasses of ealt-codes. In this classification, codes are grouped in the smallest subclass, alt-codes are in the middle, and the largest is the whole class of all ealt-codes. A necessary and sufficient condition for a pair (X, Y) of languages is an ealt-code is given by Theorem 1 (*i*, *ii*) in a relation with +-unambiguous product. The main result (Theorem 2) gives a necessary and sufficient condition for a pair (X, Y) of languages to be an alt-code. The independence of the four claims in this condition is shown in Theorem 3. As consequences of results, the existence of algorithms to determine whether the pair (X, Y) of languages X, Y is an alt-code and is an ealt-codes, in case X, Y are regular, is presented and proved in Corollary 2.

At first we recall some notions and notations, for more details, we refer to [3, 4]. Let A be a finite alphabet, A^* is the free monoid generated by A with the unit ε. We denote by $|w|$ the length of a word w; $x \leq_p y$ (or $x <_p y$) means that x is a *prefix* of y (resp. x is a *proper prefix* of y). A set $X \subseteq A^*$ is called prefix if and only if $\forall x, y \in X$, $x \leq_p y$ or $y \leq_p x$ implies $x = y$. Let $X \subseteq A^+$. A word w admits at most one factorization on X (by $w_1, w_2, ..., w_n$) if we can write $w = w_1 w_2 \ldots w_n$, $w_i \in X$, $\forall 1 \leq i \leq n$. X is called a *code* if every word $w \in A^+$ admits at most one factorization on X . For $X, Y \subseteq A^*$, we set $Y^{-1}X = \{ w \in A^* \mid yw \in X, y \in Y \}$, $XY^{-1} = \{ w \in A^* \mid wy \in X, y \in Y \}$. Let us note that every prefix set is a code.

2 +-Unambiguous Product

Let A be an alphabet and $X, Y \subseteq A^+$. The product XY of X, Y is called unambiguous (shortly, (X, Y) is unambiguous) if $\forall x, x' \in X$, $\forall y, y' \in Y$: $xy = x'y' \Rightarrow x = x'$ and $y = y'$. Let us note that if X is a code then the product XX is unambiguous. But the unambiguity is far from the sufficient conditions of a code. For instance, the set $X = \{a, bbbb, abbbb\}$ is not a code but the product XX is unambiguous. Here we introduce a way to expand the notion of unambiguous product to the definition of +-unambiguous product which is directly closed to codes, as it is shown by Proposition 2 below. For any $X, Y \subseteq A^+$ we call the +-product of X, Y the set defined by

$$(XY)^+ = (XY) \cup (XY)^2 \cup (XY)^3 \cup \ldots$$

Given $X, Y \subseteq A^+$, we define an alternative product (alt-product for short) for the pair (X, Y) of languages X, Y as follows

Definition 1. *Let A be an alphabet and X, $Y \subseteq A^+$, $w \in A^+$. Then we say that*

(*i*) *w admits a factorization by alternative product on the pair* (X, Y) (*shortly, FAP on* (X, Y)), *if* $w = u_1 u_2 \ldots u_n$ *for some sequence* $u_1, u_2, ..., u_n$, $n \geq 2$, $\forall 1 \leq i \leq n$, $u_i \in A^+$, *such that* $u_i \in X$ *iff i is odd,* $u_i \in Y$ *iff i is even.*

(*ii*) *w admits an even factorization by alternative product on the pair* (X, Y) (*shortly, eFAP on* (X, Y)), *if w admits a FAP on* (X, Y) *and n is even.*

(*iii*) *w admits a factorization by alternative product on X, Y* (*shortly, FAP on X, Y*) *if w admits a FAP on* (X, Y) *or* (Y, X). *In this case we say that w admits an alternative factorization of words in X, Y.*

Example 1. Let $X = \{a, ba\}$ and $Y = \{b, aba\}$. Then, the word $w = babaaba$ admits two following FAP on X, Y:

$f_1 : (a).(b).(a).(b).(a).(aba)$, this is an eFAP on (X, Y) and
$f_2 : (aba).(ba).(aba)$, this is a FAP on (Y, X).

Example 2. Each binary file can be viewed as a stream of bits 0, 1, for instance:

$$w = 10101100001100110110000010111101$$

Set $X = \{1\}^+$, $Y = \{0\}^+$, then w is a FAP on X, Y. By concatenating a bit 1 at the beginning of w and a bit 0 at the end of w, we obtain the result w' is always presented as an eFAP on (X, Y). For concrete w above,

$$w' = 1w0 = 1101011000011001101100001011101 0.$$

This shows a common use of alternate code in binary files in computer. (Let us note that X, Y are not codes).

Definition 2. *Let* $X, Y \subseteq A^+$. *We say that the +-product* $(XY)^+$ *is +-unambiguous,* (*shortly,* (X, Y) *is +-unambiguous*) *if and only if:* $\forall m, n \geq 2, x_1, x_2, ..., x_n, x_1', x_2', ..., x_m' \in X, y_1, y_2, ..., y_n, y_1', y_2', ..., y_m' \in Y$. *if* $x_1y_1x_2y_2 ... x_ny_n = x_1'y_1'x_2'y_2' ... x_m'y_m'$ *then* $m = n, x_i = x_i', y_i = y_i', \forall i = 1, ..., n.$

Let us remark that, if (X, Y) is +-unambiguous then (X, Y) is unambiguous. This shows that +-unambiguous product is a special case of unambiguous product.

From Definition, it easily verified some obvious following properties of the +-unambiguous product.

Proposition 1. *Let* $X, Y \subseteq A^+$. *If the pair* (X, Y) *is +-unambiguous, then the pair* (Y, X) *is also +-unambiguous.*

Proof. Suppose the contrary that the pair (Y, X) is +-ambiguous. Then, there is a word w admitting two different FAP on (Y, X),

$$w = y_1x_1y_2x_2 ... y_nx_n = y_1'x_1'y_2'x_2' ... y_m'x_m' \text{ and } y_1 \neq y_1'.$$

Choose any $x \in X$, $y \in Y$ and put $w' = xwy$. Obviously w' admits two different FAP on (X, Y), which is a contradiction. Consequently, (Y, X) is +-unambiguous. □

The following property shows a closed relation between codes and +-unambiguous product.

Proposition 2. *Let* $X \subseteq A^+$. *Then X is a code if and only if the pair* (X, X) *is +-unambiguous.*

Proof. ($\Rightarrow$) By assumption, X is a code. Hence, if a word w can be factorized on X as $w = x_1x_2 ... x_{2k}$, then this factorization is unique. Obviously it is a FAP on (X, X). It shows that the pair (X, X) is +-unambiguous.

($\Leftarrow$) We prove that if (X, X) is +-unambiguous then X is a code. Suppose the contrary that, X is not a code. Then, there exists a word $w \in X^+$ admitting two different factorizations: $w = x_1x_2 \ldots x_n = x_1'x_2' \ldots x_m'$ and $x_1 \neq x_1'$. From $x_1x_2 \ldots x_n = x_1'x_2' \ldots x_m'$ it follows $w' = (x_1x_2 \ldots x_n)(x_1x_2 \ldots x_n) = (x_1'x_2' \ldots x_m')(x_1'x_2' \ldots x_m')$

This means that there exists a word w' admitting two different FAP on (X, X), which is a contradiction. □

3 Alternative Codes

Based on the concepts of +-unambiguous product and alternative product we now can define a new type of codes for a pair (X, Y) of languages as follows.

Definition 3. *Let A be an alphabet and X, $Y \subseteq A^+$. Then:*

(*i*) *The pair (X, Y) is called an alternative code* (*alt-code for short*) *if every word $w \in A^+$ admits at most one FAP on X, Y.*

(*ii*) *The pair (X, Y) is called even alternative code* (*ealt-code for short*) *if every word $w \in A^+$ admits at most one eFAP on (X, Y).*

Remark 1. A pair (X, Y) may not be an alt-code even if $Z = XY$ is a code. Indeed, consider the following example:

Example 3. Let $X = \{a, aa\}$, $Y = \{ab, b\}$. It is easily seen that $XY = \{aab, ab, aaab\}$ is a prefix code, the pair (X, Y) is an ealt-code, but it is not an alt-code, because the word $w = ababaab \in A^+$ admits two different FAP on X, Y: $w = (a).(b).(a).(b).(aa).(b) = (ab).(a).(b).(aa).(b)$

Remark 2. The following example shows that: there is a pair (X, Y) which is not an alt-code even if X, Y are codes, and there are X, Y which are not codes while the pair (X, Y) is an alt-code.

Example 4. a) Let $X = \{ab, ba\}$, $Y = \{a\}$. Obviously, X, Y are codes while pair (X, Y) is not an alt-code. Indeed, the word $w = abaaba \in A^+$ admits two FAP on X, Y :

$$w = (ab).(a).(ab).(a) = (a).(ba).(a).(ba)$$

b) Let $X = \{a^2, a^3\}$ and $Y = \{b^2, b^3\}$. It is easy to shown that they are not codes, but the pair (X, Y) is an alt-code.

The following result shows some relationships between +-unambiguous product, alt-codes and ealt-codes.

Theorem 1. *Let X, $Y \subseteq A^+$. Then*

(*i*) *The pair (X, Y) is an ealt-code if and only if the pair (X, Y) is +-unambiguous.*

(*ii*) *The pair (X, Y) is +-unambiguous if and only if $Z = XY$ is a code and (X, Y) is unambiguous.*

(*iii*) *If the pair (X, Y) is an alt-code then (X, Y) is +-unambiguous but the converse is not true.*

(*iv*) *If $X \cup Y$ is a code and $X \cap Y = \varnothing$ then the pair (X, Y) is an alt-code but the converse is not true.*

Proof. (*i*) It is obvious from the definition of +-unambiguous product and of ealt-codes. We prove by contradictions.

(*ii*) Firstly, assume that (X, Y) is +-unambiguous. Obviously (X, Y) is unambiguous. If $Z = XY$ is not a code, then there exists a word $w \in Z^+$ admitting two different factorizations on Z: $w = z_1 z_2 \ldots z_n = z_1' z_2' \ldots z_m'$, where $z_i = x_i y_i$, $z_j' = x_j' y_j'$, $x_i, x_j' \in X$, $y_i, y_j' \in Y$, $\forall i = 1, \ldots, n$, $\forall j = 1, \ldots, m$. This shows that w admits two different FAP on (X, Y). Consequently, the pair (X, Y) is +-ambiguous, a contradiction to assumption. Hence, $Z = XY$ is a code.

Conversely, suppose that $Z = XY$ is a code and (X, Y) is unambiguous. Then, from $x_1 y_1 x_2 y_2 \ldots x_n y_n = x_1' y_1' x_2' y_2' \ldots x_m' y_m'$, by putting $z_i = x_i y_i$, $z_j' = x_j' y_j'$, we get $z_1 z_2 \ldots z_n = z_1' z_2' \ldots z_m'$ where $z_i, z_j' \in Z$, $\forall i = 1, \ldots, n$, $\forall j = 1, \ldots, m$. Because Z is a code, hence $m = n$, $z_i = z_i'$ or $x_i y_i = x_i' y_i'$, $\forall i = 1, \ldots, n$. Since (X, Y) is unambiguous, from $x_i y_i = x_i' y_i'$, $\forall$ i = 1, ..., n, we get $x_i = x_i'$, $y_i = y_i'$. Thus (X, Y) is +-unambiguous.

(*iii*) Suppose that the pair (X, Y) is an alt-code. Then, by definition, it is easily seen that (X, Y) is an ealt-code. According to (*i*), (X, Y) is +-unambiguous.

The converse is not true. For example, $X = \{a\}$, $Y = \{a^2\}$ with any $a \in A^+$. Obviously, $Z = XY = \{a^3\}$ is a code. It is easily seen that (X, Y) is unambiguous.

According to (*ii*), (X, Y) is +-unambiguous. But from $a^3 = a.a^2 = a^2.a$ it follows that a^3 admits two different FAP on X, Y. Thus, the pair (X, Y) is not an alt-code.

(*iv*) Let $X \cup Y$ be a code and $X \cap Y = \varnothing$. By definition, it is easily seen that (X, Y) is an alt-code. Conversely, let $X = \{a, a^2\}$, $Y = \{b\}$, it is easy to show that $X \cup Y = \{a, a^2, b\}$ is not a code, but (X, Y) is an alternative code. □

Corollary 1. *Let A be an alphabet and X, Y* $\subseteq A^+$. *Then*

(*a*) *The pair* (X, Y) *is ealt-code if and only if the two following conditions are satisfied:*
  (*i*) $Z = XY$ *is a code.*
  (*ii*) (X, Y) *is unambiguous.*

(*b*) *The conditions* (*i*), (*ii*) *above are independent.*

Proof. (*a*) This is a direct consequence of Theorem 1 (*i*, *ii*).

(*b*) The prove this, we give three instances:

Firstly, consider $X = Y$, X is a code. Then XY is also a code and (X, Y) is unambiguous.

Secondly, consider $X = \{a, a^2\}$, $Y = \{b, ab\}$. Then, $XY = \{ab, a^2b, a^3b\}$ is a prefix code, but (X, Y) is ambiguous since $w = (a)(ab) = (a^2)(b)$.

Thirdly, Lets $X = \{a, a^2\}$, $Y = \{a\}$. It is easily seen that, $XY = \{a^2, a^3\}$ is not a code while (X, Y) is unambiguous. □

Remark 3. By Proposition 2 and Theorem 1, we can see that each code X can be considered as the alt-code (X, Y) where $Y = X$, each alt-code as a special case of ealt-codes, but the inverse cases are not true in general. Hence, we can establish a classification on three subclasses of ealt-codes which show that the smallest is the class of codes regarding each code X as an alt-codes of the form (X, X), the middle is the class of all alt-codes, and the largest is the class of ealt-codes.

In the case of traditional codes, we have following basic criteria of codes due to [3] in relation with injective morphisms.

Proposition 3. *Let $h : A^* \to B^*$ be a monoid morphism from A^* to B^* and let $X = h(A)$. Then, X is a code if and only if h is an injective.*

In the rest of this section, we establish a similar result for the case of alt-codes.

Let A be an alphabet and $X, Y \subseteq A^+$. Set $B = A \cup \{e, f\}$ where e, f are new letters not in A. Then, we define an *erasing morphism* $\varphi : B^* \to A^*$ given by :

(1) $\varphi(e) = \varphi(f) = \varepsilon$.
(2) $\varphi(a) = a$, $\forall a \in A$.

Set $S = \{ ewf, ewe, fwe, fwf \mid w \in A^+ \} \subseteq B^+$. On S, we define a product ".": for any $x = i_1 u j_1$, $y = i_2 v j_2$ in B^+,

$$x.y = \begin{cases} i_1 u v j_2 & \text{if } j_1 = i_2 \\ 0 & \text{if } j_1 \neq i_2 \end{cases}$$

where 0 is a new zero of this product, 0 not in B^+, 1 is the new unit of $S^\wedge = S \cup \{1, 0\}$. Then $S^\wedge$ is a monoid. Let $X, Y \subseteq A^+$, consider $U_{X,Y} = \{ exf, fye \mid x \in X, y \in Y \} \subset S$ and $V_{X,Y} = <U>$ is the submonoid of $S^\wedge$ generated by $U_{X,Y}$, then $V_{X,Y} = U^+_{X,Y}$.

Proposition 4. *The pair (X, Y) is an alt-code if and only if $\varphi \mid_{V_{X,Y}}$ is injective.*

Proof. ($\Rightarrow$) Assume that the pair (X, Y) is an alt-code, we verify that $\varphi \mid_{V_{X,Y}}$ is injective. Indeed, suppose a contrary that $\varphi|_{V_{X,Y}}$ is not injective. Then we can choose two different elements, for instance, $u = ex_1fy_1ex_2fy_2 \ldots \neq v = fy_1'ex_1'fy_2'ex_2' \ldots$ and $\varphi(u) = \varphi(v)$. This implies $x_1y_1x_2y_2 \ldots = y_1'x_1'y_2'x_2' \ldots$, its a contrary with assumption. For other cases, we also deduce contradiction.

($\Leftarrow$) Assume that the $\varphi|_{V_{X,Y}}$ is injective, we verify that the pair (X, Y) is an alt-code. Indeed, suppose a contrary that the pair (X, Y) is not an alt-code. There exist two different alternative factorizations of a word w by X, Y. For instance, consider the case $x_1y_1x_2y_2 \ldots = y_1'x_1'y_2'x_2' \ldots$ Consider two words $u = ex_1fy_1ex_2fy_2 \ldots$, $v = fy_1'ex_1'fy_2'ex_2' \ldots$ in $V_{X,Y}$. We have $\varphi(u) = x_1y_1x_2y_2 \ldots = y_1'x_1'y_2'x_2' \ldots = \varphi(v)$. Hence $\varphi(u) = \varphi(v)$. This a contrary with assumption that $\varphi|_{V_{X,Y}}$ is injective. For other cases, we also deduce contradiction □

The following is obvious hence we omit its proof.

Proposition 5. *Let A, B are non empty and disjoint alphabets. Set $C = A \cup B$. For arbitrary monoid morphism $h: C^* \to D^*$, set $h(A) = X$ and $h(B) = Y$, define U as the set of all words in C^+ which can be factorized alternatively by words in A, B, and V as the subset of U consisting all words of the form $a_1b_1a_2b_2 \ldots a_nb_n$, $n \geq 1$, $a_i \in A$, $b_i \in B$, $i = 1, \ldots, n$. Then*

(*i*) *The pair (X, Y) is an alt-code if and only if $h|_U$ is injective.*
(*ii*) *The pair (X, Y) is an ealt-code if and only if $h|_V$ is injective.*

4 Characterization for alt-codes

The following theorem gives a necessary and sufficient characterization for a given pair (X, Y) to be an alt-code which permits us to establish algorithms to test alt-codes with X, Y are regular.

Theorem 2. *Let $X, Y \subseteq A^+$. Then, (X, Y) is an alt-code if and only if the four following conditions are satisfied:*

(i) XY *is a code and* $X^{-1}X \cap YY^{-1} - \{\varepsilon\} = \varnothing$;

(ii) $Y^{-1}(XY)^+ \cap (XY)^+ = \varnothing$;

(iii) $(XY)^+X^{-1} \cap (XY)^+ = \varnothing$;

(iv) $(XY)^+ \cap (YX)^+ = \varnothing$.

Proof. ($\Rightarrow$) Assume that (X, Y) is an alt-code. We proceed by contradictions.
(i) Suppose that (i) does not hold. There are two cases:

Case 1: $X^{-1}X \cap YY^{-1} - \{\varepsilon\} = \varnothing$. Then, there exists words $u \neq \varepsilon$, $x_1, x_2 \in X$, $y_1, y_2 \in Y$ such as $x_1 = x_2u$, $uy_1 = y_2$, we get $x_1y_1 = x_2y_2$ with $x_1 \neq x_2$. Thus, the pair (X, Y) is not an alt-code, which is a contradiction.

Case 2: $Z = XY$ is not a code, it means that there exists a word w admitting two different factorizations on Z: $w = z_1z_2 \ldots z_n = z_1'z_2' \ldots z_m'$, $z_1 \neq z_1'$, such that

$$z_1 = x_1y_1,\ z_2 = x_2y_2, \ldots,\ z_n = x_ny_n, \text{ with } x_i \in X, y_i \in Y$$
$$z_1' = x_1'y_2',\ z_2' = x_2'y_2', \ldots, z_m' = x_m'y_m', \text{ with } x_i' \in X, y_i' \in Y$$

Obviously these are two different FAP of w on (X, Y). Thus the pair (X, Y) is not an alt-code, this also is a contradiction. Hence, (i) is true.

(ii) Assume that $Y^{-1}(XY)^+ \cap (XY)^+ = \varnothing$, then there exists $y \in Y$, $u \in Y^{-1}(XY)^+ \cap (XY)^+$ such that $u = z_1z_2 \ldots z_n$, $yu = z_1'z_2' \ldots z_m'$, $z_i, z_j' \in XY$, $z_i = x_iy_i$, $z_j' = x_j'y_j'$, $x_i, x_j' \in X$, $y_i, y_j' \in Y$, $\forall 1 \leq i \leq n, \forall 1 \leq j \leq m$.

Put $w = yu$, we get $w = z_1'z_2' \ldots z_m'$. Since $u = x_1y_1x_2y_2 \ldots x_ny_n$, $w = yx_1y_1x_2y_2 \ldots x_ny_n = x_1'y_1'x_2'y_2' \ldots x_m'y_m'$. This implies that (X, Y) is not an alt-code, this is a contradiction. Thus $Y^{-1}(XY)^+ \cap (XY)^+ = \varnothing$.

(iii) Assume that $(XY)^+X^{-1} \cap (XY)^+ = \varnothing$, then there exist $x \in X$, $u \in (XY)^+X^{-1} \cap (XY)^+$ such that $u = z_1z_2 \ldots z_n$ and $ux = z_1'z_2' \ldots z_m'$, $z_i, z_j' \in XY$, $z_i = x_iy_i$, $z_j' = x_j'y_j'$, $x_i, x_j' \in X$, $y_i, y_j' \in Y$, $\forall 1 \leq i \leq n, \forall 1 \leq j \leq m$.

Put $w = ux = z_1'z_2' \ldots z_m'$. Since $u = x_1y_1x_2y_2 \ldots x_ny_n$, $w = x_1y_1x_2y_2 \ldots x_ny_nx = x_1'y_1'x_2'y_2' \ldots x_m'y_m'$. This implies that (X, Y) is not an alt-code, this is a contradiction. Thus $(XY)^+X^{-1} \cap (XY)^+ = \varnothing$.

(iv) Suppose that $(XY)^+ \cap (YX)^+ = \varnothing$, then there exists a word w admitting two different factorizations on X, Y: $w = x_1y_1x_2y_2 \ldots x_ny_n = y_1'x_1'y_2'x_2' \ldots y_m'x_m'$. This shows that (X, Y) is not an alt-code, a contradiction to assumption. Hence $(XY)^+ \cap (YX)^+ = \varnothing$.

($\Leftarrow$) Assume that all conditions (*i*), (*ii*), (*iii*), (*iv*) are satisfied. Suppose on the contrary that (X, Y) is not an alt-code. Then there exists a word w admitting two different FAP on X, Y: $w = u_1u_2 \dots u_i = u_1'u_2' \dots u_j'$, $i, j \geq 1$, $u_1 \neq u_1'$. Depending on the parity of i, j, we need to consider the following cases:

Case 1: $u_1 \in X, u_1' \in X$.

+ i and j are even. If there is a word w admitting two different FAP on (X, Y): $w = x_1y_1x_2y_2 \dots x_ny_n = x_1'y_1'x_2'y_2' \dots x_m'y_m'$ then XY is not a code, a contradiction to (*i*).
+ i and j are odd. If there is a word w admitting two different FAP on (X, Y): $w = x_1y_1x_2y_2 \dots x_n = x_1'y_1'x_2'y_2' \dots x_m'$, we get $w' = x_1y_1x_2y_2 \dots x_ny = x_1'y_1'x_2'y_2' \dots x_m'y$, it follows that w' admits two factorizations on XY. This means that XY is not a code, a contradiction to (*i*).
+ i is even and j is odd. If there is a word w admitting two different FAP on (X, Y): $w = x_1y_1x_2y_2 \dots x_ny_n = x_1'y_1'x_2'y_2' \dots x_m'$. By setting $u = x_1'y_1'x_2'y_2' \dots x_{m-1}'y_{m-1}' \in (XY)^+$, we get $w = ux_m' \in (XY)^+$. Then $u \in (XY)^+X^{-1}$, a contradiction to (*iii*).
+ i is odd and j is even. If there is a word w admitting two different FAP on (X, Y): $w = x_1y_1x_2y_2 \dots x_n = x_1'y_1'x_2'y_2' \dots x_m'y_m'$. Putting $u = x_1y_1x_2y_2 \dots x_{n-1}y_{n-1} \in (XY)^+$, we get $w = ux_n \in (XY)^+$. Then $u \in (XY)^+X^{-1}$, a contradiction to (*iii*).

Case 2: $u_1 \in X, u_1' \in Y$.

+ i and j are even. If there is a word w admitting two different FAP on X, Y: $w = x_1y_1x_2y_2 \dots x_ny_n = y_1'x_1'y_2'x_2' \dots y_m'x_m' \in (XY)^+ \cap (YX)^+$. Then $(XY)^+ \cap (YX)^+ = \varnothing$, a contradiction to (*iv*).
+ i and j are odd. If there is a word w admitting two different FAP on X, Y: $w = x_1y_1x_2y_2 \dots x_n = y_1'x_1'y_2'x_2' \dots y_m'$. By setting $w' = w.w = x_1y_1 \dots x_ny_1'x_1' \dots y_m' = y_1'x_1' \dots y_m'x_1y_1 \dots x_n \in (XY)^+ \cap (YX)^+$ ones can deduce $(XY)^+ \cap (YX)^+ = \varnothing$, a contradiction to (*iv*).
+ i is even and j is odd. If there is a word w admitting two different FAP on X, Y: $w = x_1y_1x_2y_2 \dots x_ny_n = y_1'x_1'y_2'x_2' \dots y_m'$. Setting $u = x_1'y_2'x_2'y_3' \dots x_{m-1}'y_m' \in (XY)^+$, we get $w = y_1'u \in (XY)^+$. Then $u \in Y^{-1}(XY)^+$, a contradiction to (*ii*).
+ i is odd and j is even. If there is a word w admitting two different FAP on X, Y: $w = x_1y_1x_2y_2 \dots x_n = y_1'x_1'y_2'x_2' \dots y_m'x_m'$, we get: $w' = x_1y_1x_2y_2 \dots x_ny = y_1'x_1'y_2'x_2' \dots y_m'x_m'y$. Setting $u = x_1'y_2'x_2' \dots y_m'x_m'y \in (XY)^+$, we have $w' = y_1'u \in (XY)^+$. Hence $u \in Y^{-1}(XY)^+$, a contradiction to (*ii*).

Case 3: $u_1 \in Y, u_1' \in Y$.

+ i and j are even. If there is a word w admitting two different FAP on (Y, X): $w = y_1x_1y_2x_2 \dots y_nx_n = y_1'x_1'y_2'x_2' \dots y_m'x_m'$, we get $w' = xy_1x_1y_2x_2 \dots y_nx_ny = xy_1'x_1'y_2'x_2' \dots y_m'x_m'y$, this shows that w' admits two different factorizations on XY. Thus XY is not a code, a contradiction to (*i*).
+ i and j are odd. If there is a word w admitting two different FAP on (Y, X): $w = y_1x_1y_2x_2 \dots y_n = y_1'x_1'y_2'x_2' \dots y_m'$, we get $w' = xy_1x_1y_2x_2 \dots y_n = xy_1'x_1'y_2'x_2' \dots y_m'$. This shows that w' admits two factorizations on XY. Thus XY is not a code, a contradiction to (*i*).
+ i is even and j is odd. If there is a word w admitting two different FAP on (Y, X): $w = y_1x_1y_2x_2 \dots y_nx_n = y_1'x_1'y_2'x_2' \dots y_m'$, we get $w' = xy_1x_1y_2x_2 \dots y_nx_n = xy_1'x_1'y_2'x_2' \dots y_m'$. Setting $u = xy_1x_1y_2x_2 \dots x_{n-1}y_n \in (XY)^+$, then $w' = ux_n \in (XY)^+$. This implies that $u \in (XY)^+X^{-1}$, a contradiction to (*iii*).

+ i is odd and j is even. If there is a word w admitting two different FAP on (Y, X): $w = y_1x_1y_2x_2 \dots y_n = y_1'x_1'y_2'x_2' \dots y_m'x_m'$, we have $w' = xy_1x_1y_2x_2 \dots y_n = xy_1'x_1'y_2'x_2' \dots y_m'x_m'$. Setting $u = xy_1'x_1'y_2'x_2' \ x_{m-1}'y_m'$, then $w' = ux_m' \in (XY)^+$. Thus $u \in (XY)^+X^{-1}$, a contradiction to (*iii*).

Case 4: $u_1 \in Y, u_1' \in X$. This case is analogous with the Case 2.

The proof is completed. □

Theorem 3. *The conditions* (*i*), (*ii*), (*iii*) *and* (*iv*) *in Theorem 2 are independent.*

5 Algorithms to Test for alt-codes and ealt-codes

For the case X, Y are regular languages, the following result plays a fundamental role which helps us to establish algorithms to test for alt-codes and ealt-codes.

Corollary 2. *If X and Y are regular languages, then there exists an algorithm to decide if the pair* (X, Y) *is*

(*i*) *an alt-code.*

(*ii*) *an ealt-code.*

Proof. (*i*) Since there exists algorithms for testing the emptiness of regular languages, there exists an algorithm to test whether a given pair (X, Y) is an alt-code according to Theorem 2.

(*ii*) By assumption, $Z = XY$ is regular, due to a well-known result of Sardinas-Patterson (see [3]), there is an algorithm to test if Z is a code or not. Using the facts that both $X^{-1}X$, YY^{-1} are regular, and the product XY is unambiguous if and only if $(X^{-1}X) \cap (YY^{-1}) = \varnothing$, ones can build an algorithm to test if (X, Y) is unambiguous. Combining two these algorithms, from Theorem 1 (*i*, *ii*), there exists an algorithm to test whether the pair (X, Y) is an ealt-code. □

Algorithm to test for ealt-codes

B_1. Testing whether $Z = XY$ is a code by Sardinas-Patterson algorithm .

B_2. Testing $X^{-1}X \cap YY^{-1} - \{\varepsilon\} = \varnothing$.

Algorithm to test for alt-codes

B_1. Testing whether $Z = XY$ is a code by Sardinas-Patterson algorithm.

B_2. Testing $X^{-1}X \cap YY^{-1} - \{\varepsilon\} = \varnothing$.

B_3. Testing $Y^{-1}(XY)^+ \cap (XY)^+ = \varnothing$.

B_4. Testing $(XY)^+X^{-1} \cap (XY)^+ = \varnothing$.

B_5. Testing $(XY)^+ \cap (YX)^+ = \varnothing$.

Complexity of algorithms

1. Algorithm to test for ealt-codes

B_1. By assumption $Z = XY$ is regular on an finite alphabet A. We can find a sujective morphism $h : A^* \to M$, M is finite, such that h saturates Z and $\{\varepsilon\}$. By some direct verifications we deduce the fact: any language obtained from Z and $\{\varepsilon\}$ by taking a finite number of boolean operations, left and right quotients, is also saturated by h.

Hence, all U_i are saturated by h. That is $U_i = h^{-1}(K_i)$ for some $K_i \subseteq M$, for $i = 1, ..., n$. Since the number subsets of M is $2^{|M|}$, so the number of U_i's is not larger than $2^{|M|}$. Hence, this step has time complexity about $O(2^{|M|})$.

B_2. Since X, Y, $\{\varepsilon\}$ are regular, we can find a sujective morphism $h : A^* \to M$, M is finite, such that h saturates X, Y and $\{\varepsilon\}$. Then $X = h^{-1}(B)$, $Y = h^{-1}(C)$, $\{\varepsilon\} = h^{-1}(1_M)$. Hence using fact above, $L = X^{-1}X \cap YY^{-1} - \{\varepsilon\}$ is saturated by h and $L = h^{-1}(K)$, where $K = B^{-1}B \cap CC^{-1} - \{1_M\}$ which is easily computed by B, C on the product table of M. The time complexity to compute K is $O(|M|.|B| + |M|.|C|)$. Hence the time complexity for this step not exceeds $2|M|^2$.

Totally, this algorithm has the time complexity about $O(2^{|M|} + 2|M|^2)$, if we choose a h in common in two steps B_1, B_2.

Remark 4. Let us recall a well-know result that, if L is a regular language accepted by automaton $\mathcal{A}$ with the sizes m (by number of states of $\mathcal{A}$) then we can build a monoid M having a size about 2^m and a morphism $h: A^* \to$ M saturating L.

2. Algorithm to test for alt-codes

B_1, B_2. Since the steps (B_1),(B_2) in this algorithm are nothing but the steps in algorithm to test for ealt-code, hence at first, we consider the rest three steps (B_3, B_4, B_5)

B_3. If X, Y are accepted by give finite automata $\mathcal{A}_1$, $\mathcal{A}_2$ of sizes m, n (by number of states of $\mathcal{A}_1$, $\mathcal{A}_2$ respectively) then we can design an automaton of size about $m+n$ to accept XY, $(XY)^+$. By the Remark 4, we can construct a finite monoid M of size 2^{m+n} saturating $(XY)^+$ and a finite monoid M' of size $2^{m+n}.2^n = 2^{m+2n}$ saturating $Y^{-1}(XY)^+$. Therefore, the time complexity in the step (B_3) is about $O(2^{m+n}.2^{m+2n})$ or $O(2^{2m+3n})$.

B_4. It is similar to the case of $(XY)^+$ and $(XY)^+X^{-1}$ for the this step, it is about $O(2^{3m+2n})$.

B_5. We can see that, this step is about $O(2^{m+n}.2^{m+n})$.

Let us remark that, the time complexity to design a morphism saturating Z is about $k = 2^{m+n}$ and the number of steps to in Sardinas-Patterson procedure verifying whether Z is code, is about 2^k. Therefore the time complexity for the step (B_1) is about $O(2^k)$.

We can construct a monoid saturating both $X^{-1}X$ and YY^{-1} by the time complexity about $O(2^{m+n})$. Hence the time complexity for the step (B_2) is about $O(2^{m+n}.2^{m+n})$

6 Conclusion

In this paper, new types of codes: alt-codes and ealt- codes are introduced. An our result shows that these codes can be considered as extension forms of traditional codes. The notion of +-unambiguous product as a middle notion between notions of unambiguous product and of unambiguous product of codewords is introduced and studied. For the case of regular languages, two algorithms to test for ealt-codes and for alt-codes are obtained. As we seen, complexity for these algorithms are of power size. In next works, we hope can find some better algorithms and many interesting problems of codes may be extended to the cases of ealt-codes and alt-codes.

References

1. Anselmo, M.: Automates et codes zigzag. R.A.I.R.O. Theoretical Informatics and Applications 25(1), 49–66 (1991)
2. Ahmad, K.: Quelques problèmes de Mélanges Contrôlés. Thèse Docteur en Informatique, Université de Nice Sophia-Antipolis (2002)
3. Berstel, J., Perrin, D.: Theory of Codes. Academic Press Inc., New York (1985)
4. Eilenberg, S.: Automata, Languages and Machines, vol. B. Academic Press, New-York (1976)
5. Huy, P.T., Van, D.L.: On Non-Ambiguous Büchi V-automata. In: Proceedings of the Third Asian Mathematical Conference 2000, Diliman, Philippines, October 23-27, pp. 224–233. World Scientific, Singapore (April 2002) ISBN 981-02-4947-0
6. Pin, J.E.: Variété des Languages Infinis et variete de semigroupes. These Docteur d'Etat (1982)
7. Pin, J.E., Weil, P.: Polynomial closure and unambiguous products. Theory of Computing Systems 30, 383–422 (1997)
8. Schützenberger, M.P.: On a question concerning certain free submonoids. J. Combinatorial Theory, 422–437 (1966)
9. Weil, P.: Groups, codes and unambiguous automata. In: Mehlhorn, K. (ed.) STACS 1985. LNCS, vol. 182, pp. 351–362. Springer, Heidelberg (1985)

A Study on the Topology Control Method for Bluetooth Scatternet Formation

Chih-Min Yu

Dept. of Communication Engineering, Chung-Hua University
No. 707, Sec 2, Wufu Rd, Hsinchu, 30012, Taiwan, R.O.C
ycm@chu.edu.tw

Abstract. In this paper, a topology-configurable method for forming a Bluetooth scatternet is proposed. The heuristic method describes two mechanisms, the const-hop algorithm and the variant-hop algorithm. With a constant k parameter, the const-hop algorithm propagates k in its downstream direction to determine roots and constructs their associated subnets. With a constant k, a counter variable v, and a return variable r as parameters, the variant-hop algorithm generates appropriate roots locally and evenly configures the subnet size. A computer simulation shows that the proposed method achieves good network scalability and generates an efficient scatternet configuration for a Bluetooth multihop network.

Keywords: Ad-hoc networks, Bluetooth, sensor network, scatternet formation.

1 Introduction

Bluetooth is emerging as a potential technology for short-range wireless ad hoc networks. This technology enables the design of low power, low cost, and short-range radio which is embedded in existing portable devices. Initially, Bluetooth technology was designed as a cable replacement solution for portable and fixed electronic devices. Today, people tend to use a number of mobile devices, such as cellular phones, PDA's, digital cameras, laptop computers, etc. Consequently, there exists a strong demand for connecting these devices into networks. As a result, Bluetooth has become an ideal candidate for the construction of ad hoc personal area networks.

Many scatternet formation algorithms [1]-[7] have been proposed to construct a Bluetooth ad hoc network. Currently, most of the topology control algorithms partition their Bluetooth scatternet after collecting the complete information on the nodes [5]-[7]. In [5], a node which has complete knowledge of all the nodes is elected as the leader of the scatternet. This leader then partitions the entire scatternet topology via a predefined formula. In [6], a super master collects all node information, determines the role of each node, and shapes its topology into a line, bus, star or mesh by its corresponding parameters. With a traffic-dependent model, the configured Blueweb [7] uses a route master to collect complete topology information and reconfigure the scatternet into several subnets to improve routing performance.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 263–271, 2010.

To make the topology configurable without prior knowledge of the nodes, two scatternet formation algorithms have been proposed in this paper: the const-hop algorithm and the variant-hop algorithm. The const-hop algorithm uses a designated root to propagate a constant k and a counter limit $k1$ in its downstream direction to determine new roots locally, as well as to build their associated subnets. With this method, the subnet size can also be controlled by appropriately selecting a constant k. With a constant k, a counter variable v, and a return variable r as parameters, the variant-hop algorithm can appropriately select the new roots and evenly configure the size of the subnet.

The remainder of this paper is organized as follows: Section 2 describes the detailed operation of the const-hop scatternet formation algorithm; Section 3 presents the detailed operation of the variant-hop scatternet formation algorithm; Section 4 uses computer simulations to demonstrate the system performance of these two algorithms; and, finally, a conclusion is drawn in Section 5.

2 Const-Hop Scatternet Formation Algorithm

In order to make the topology configurable without collecting any information on the nodes in advance, a const-hop scatternet formation algorithm is herein proposed. In this algorithm, two parameters, including a constant k and a counter limit $k1$ are introduced to determine new roots, with each root responsible for configuring and managing its own subnet. In addition, the new roots are determined locally on a layer-by-layer basis in the downstream direction (out from the designated root) during the scatternet formation.

At the start, all nodes in a network are assumed to stay in the inquiry scan state. A particular node is given as the designated root to set a counter limit $k1=k$, where $k1$ is an integer variable and k is the constant. With these two parameters, the first root starts the inquiry cycle, pages up to 7 neighboring slaves, and forms its own piconet. Each slave then switches its role to master (called S/M node) to inquire and page one additional slave. After each S/M node connects to its slave, a role-exchange mechanism is executed to make the S/M node function a relay and the slave function a master. Then, these new masters decrease $k1$ by 1 and continue to propagate the two parameters in the downstream direction.

In this way, when the $(k1)th$ master is reached, $k1=0$. The master becomes a new root and the counter limit $k1$ is reset to k. The tree-shaped subnet of the designated root is created. Then, this new root asks its upstream masters to try to connect with one additional slave until its immediate upstream root is reached (the procedure is referred to as “return connection” in this paper). As a result, the tree-shaped subnet of the designated root is converted into the web-shaped subnet.

At the same time, the new root repeats the same procedure as that of the designated root to build its own subnet and propagates the two parameters to determine new roots. This procedure is continued until the leaf nodes are reached. All the leaf nodes will request their immediate upstream masters to conduct the return connection procedure until its immediate upstream root is reached and the whole scatternet is formed. Finally, each root manages its own web-shaped subnet.

Here, *k=2* in Figure 1 is used as an example to describe the const-hop scatternet formation process. Initially, the designated root, R1, inquires and pages slaves to form its piconets. Each slave then switches its role to master (called S/M node) to inquire and page one additional slave. After each S/M node connects to its slave, a role-exchange mechanism is executed to make the S/M node function a relay and the slave function a master. As a result, R1 connects with the first tier masters. There is a relay (slave/slave node) between R1 and its immediate downstream masters.

Then, the first tier masters decrease *k1* by 1 and continue to connect with their downstream masters. When the second tier masters are reached and the counter limit *k1=0*, these masters become new roots and reset *k1* to *k*. The tree-shaped subnet of the designated root is created. These new roots ask their upstream masters to start the return connection procedure and connect with one additional slave until its immediate upstream root, R1, is reached. The topology of the designated root is finished and it generates a web-shaped subnet.

At the same time, these new roots start to page new slaves and connect with their immediate downstream masters (leafs in this example), to build their own tree-shaped subnets. When the leaf masters are reached, these masters start the return connection procedure until their immediate upstream roots, R2, R3, R4 and R5, are reached, and the scatternet formation process is terminated. Finally, all roots have their corresponding web-shaped subnet, as shown in Figure 1.

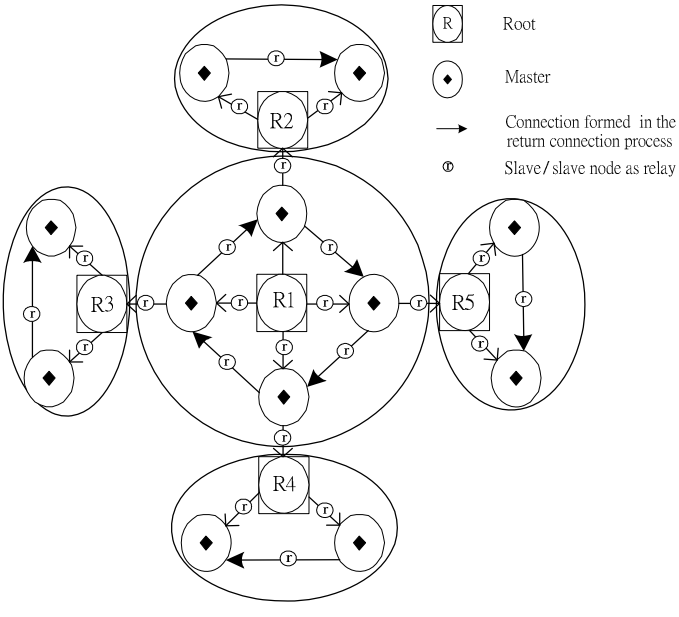


Fig. 1. The const-hop scatternet formation process

3 Variant-Hop Scatternet Formation Algorithm

To determine the appropriate new roots and evenly configure the size of the subnet, a variant-hop scatternet formation algorithm has also been proposed. Based on the constants *k* and *k1* as parameters, an additional variable *v* is introduced in the variant-hop algorithm to try to configure the size of the subnet. In addition, a return variable *r* and a root selection criterion are also added to determine the appropriate new roots locally.

Initially, the designated root sets a counter limit $k1=k*v$, where *k* is a constant and *v* is a counter variable ($v=1$, initially). The designated root inquires and pages up to 7 neighboring slaves, then forms its own piconet. Each slave then switches its role to master (called S/M node) to inquire and page one additional slave. After each S/M node connects to its slave, a role exchange is executed to make the S/M node function a relay and make the slave function a master. Then, these new masters decrease *k1* by 1 and continue to propagate the three parameters (*k1*, *k*, *v*) in the downstream direction.

In this way, when the *(k1)th* layer master is reached, $k1=0$ and the master becomes a new root. In addition, counter variable *v* is increased by 1 and counter limit *k1* is reset to *k*v*. Then, this new root immediately starts the return connection procedure and asks its upstream masters to try to connect with one additional piconet until its immediate upstream root is reached.

At the same time, the new root repeats the same procedure, as that of the designated root initially executed, to build its own subnet and propagates the three parameters to determine new roots locally. This procedure is iterated until the leaf nodes are reached. Then, the masters of all the leaf nodes set a return variable *r* ($r=1$, initially) and start the return connection procedure. The immediate upstream masters increase *r* by 1 until the immediate upstream roots are reached.

Finally, each immediate root uses a root selection criterion to decide whether it remains a root. The criterion is as follows. If variable *r* is less than or equal to *k1*/2 for all downstream paths, the root will change its role to a master and pass its downstream information to its immediate upstream root. Otherwise, the root will remain in its role as a root.

Here, $k=2$, $v=1$, and $r=1$ is used for illustration. Figure 2 shows the variant-hop scatternet formation process. Initially, the designated root, R1, connects with the first layer masters, as in the procedure described for the const-hop algorithm. There is a relay (slave/slave node) between R1 and its immediate downstream piconets. The first layer masters decrease *k1* ($k1=2$, initially) by 1 and continue to connect with their downstream masters. When the second layer masters are reached, the counter limit $k1=0$ and these masters become new roots.

Then, these new roots ask their upstream masters to conduct the return connection procedure until R1 is reached. At the same time, these new roots increase *v* by 1, set *k1* to $k*v=4$, and connect with their immediate downstream masters until the leaf nodes are reached. Since R3 is a leaf master itself, it will conduct the return connection, pass its own information to R1, and switch its role from a root to a

master. In addition, the downstream master of R4 is also a leaf node. It will start a return connection and set a return variable, r=1. Since r (r=1) is equal to $k1$/2, it will pass its downstream information to R1 and switch its role to master. At the same time, a new root, R5, is generated when $k1$ decreases to 0 and it connects with one master.

Finally, the leaf master of R5 sets r=1 and starts the return connection until R5 is reached. Since variable r (r=1) is less than $k1$/2 ($k1$/2=2), R5 sends its downstream information to R2. R2 retains its role as a root because it receives the information passed by R5. When the variant-hop scatternet formation process is terminated, both roots R1 and R2 have individual subnets, as shown in Figure 2. As seen, the total number of roots could be reduced from 5 to 2 using the root selection criterion. In order to simplify the illustration, relays (S/S node, used to interconnect masters among piconets) are not shown in Figure 2.

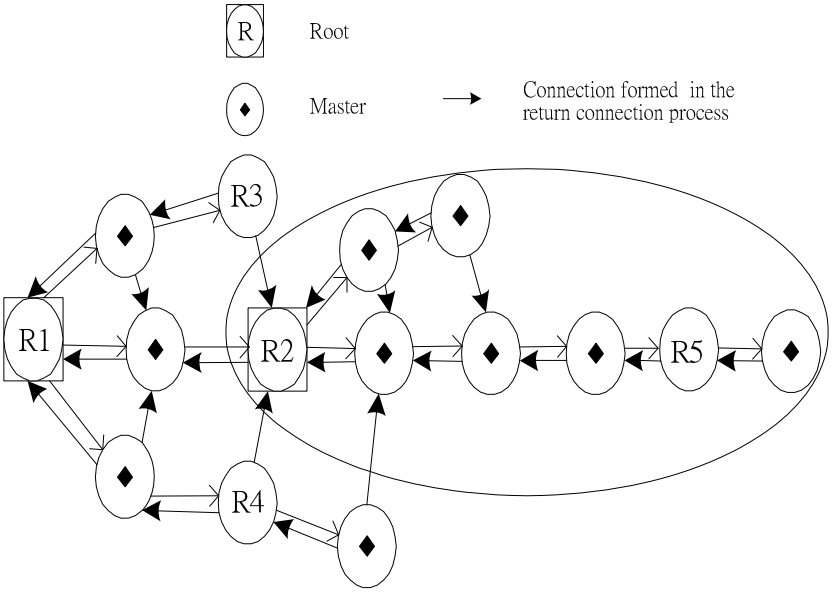


Fig. 2. The variant-hop scatternet formation process

4 Symtem Performance Simulation

A. Simulation Model and System Parameters

A simulation program was written to evaluate the system performance of the topology-configurable method. First, this study assumed that the Bluetooth nodes were uniformly located on a rectangular lattice, and that the number of neighboring nodes which could be reached by each node was between 2 and 4. The simulated node number ranged from 60, 70, 80,…, to 200. A set of performance metrics was

calculated by averaging over 100 randomly generated topologies for each simulated node number case. The parameters *k*, *v*, and *r* were used in a combinatory way for different simulated cases, and the simulation results were as follows.

B. Performance Results

Figure 3 shows the average number of roots for both the const-hop and the variant-hop algorithms with *k, v,* and *r* as parameters. As seen, the number of roots decreased as *k* increased, and that the *k*=4 case produced the smallest number of roots in terms of the largest average subnet size. There was a performance tradeoff for the value of *k*, the average number of roots, and the average subnet size.

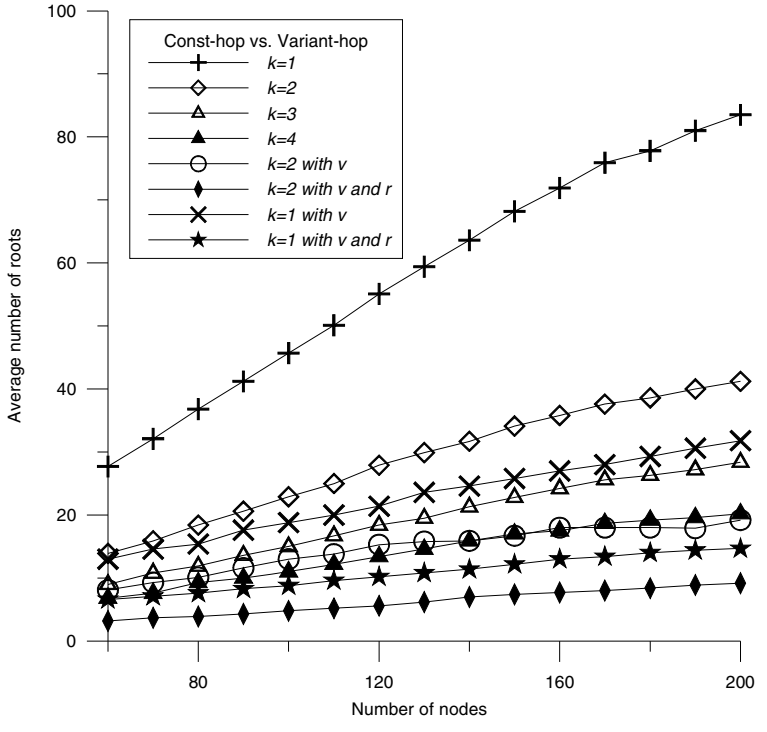


Fig. 3. Average number of roots in the scatternet

With both *k* and *v* as parameters, the root-number performance was improved and it almost achieved the performance of the *k*=4 case. In addition, using *k*, *v*, and *r* as parameters, the number of roots was reduced further. This result showed that adding the return variable *r* effectively reduced the number of roots.

Figure 4 shows the average number of piconets in a subnet for both the const-hop and the variant-hop algorithms. With *k* as a parameter, a higher average number of

piconets was produced in a subnet by increasing k, and the k=4 case generated the largest average number of piconets in a subnet. With both k and v as parameters, the k=2 case achieved a similar performance as the k=4 case when k was used as the only parameter. With k, v, and r as parameters, the k=2 case achieved the largest average number of piconets in a subnet. As shown in Figure 3 and Figure 4, this study reduced the number of roots effectively and generated a more efficient scatternet configuration by using k, v, and r as parameters in the variant-hop scatternet formation algorithm.

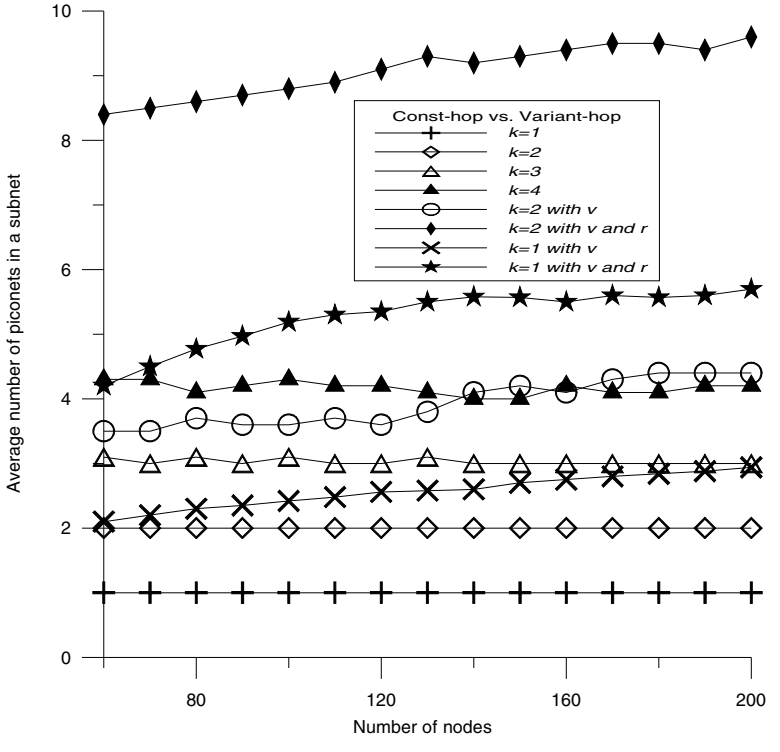


Fig. 4. Average number of piconets in a subnet

Figure 5 shows the average number of piconets in the subnet of the first root for both the const-hop and the variant-hop algorithms. With k as a parameter, the number of piconets in the subnet increased quickly as k increased, and the k=4 case produced the greatest performance deviation from Figure 4. The above performance deviation was effectively reduced in the other two cases (when k, v, or k, v, r were used as parameters) of the variant-hop algorithm.

With k as a single parameter, the const-hop algorithm made the topology controllable by selecting an appropriate k value, with each root managing its own subnet. However, the number of piconets of the first root grew more quickly than that of the other roots. Using both k and v as parameters, the variant-hop algorithm not

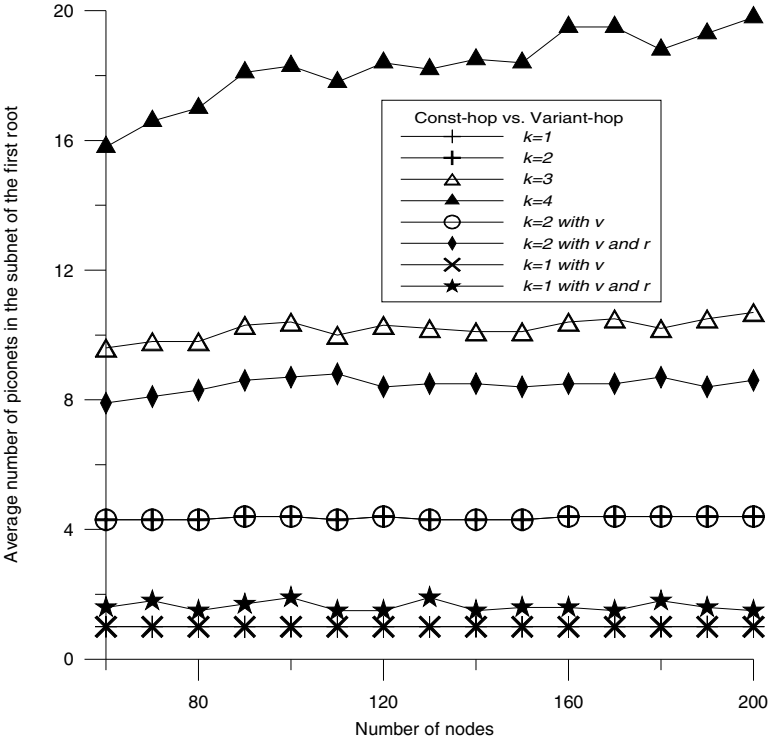


Fig. 5. Average number of piconets in the subnet of the first root

only achieved better network scalability than the single k parameter case, but it also reduced the subnet size of the first root. With k, v, and r as parameters in the variant-hop algorithm, the number of roots was reduced further than in the other two cases, and the subnet size of all roots was almost equal. As a result, the const-hop algorithm made the scatternet topology configurable, and the variant-hop algorithm determined appropriate new roots and generated an evenly distributed subnet configuration.

5 Conclusion

In this paper, two scatternet formation algorithms for configuring Bluetooth topology have been proposed: the const-hop algorithm and the variant-hop algorithm. Without any prior information on the nodes, the const-hop algorithm used a designated root to propagate a constant k and a counter limit $k1$ in its downstream direction to determine new roots and build their associated subnets. With this method, the subnet size could also be controlled by appropriately selecting a constant k. Based on k and $k1$ as parameters, the variant-hop algorithm added a counter variable v, a return variable r, and a root selection criterion to determine appropriate new roots and to generate an evenly distributed subnet configuration. The computer simulations showed that the proposed topology-configurable method achieved good network scalability and that it constructed various sizes of Bluetooth scatternet efficiently.

Acknowledgments

My heartfelt thanks to the support of the work by the National Science Council, Taiwan (NSC-99-22221-E-216-025).

References

1. Zaruba, G.V., Basagni, S., Chlamtac, I.: Bluetrees-scatternet formation to enable Bluetooth-based ad hoc networks. In: IEEE International Conference on Communications, vol. 1, pp. 273–277 (June 2001)
2. Petrioli, C., Basagni, S., Chlamtac, I.: Configuring BlueStars: Multihop scatternet formation for Bluetooth networks. IEEE Transaction on Computers 52(6), 779–790 (2003)
3. Cuomo, F., Melodia, T., Akyildiz, I.F.: Distributed self-healing and variable topology optimization algorithms for QoS provisioning in scatternets. IEEE Journal on Selected Areas in Communications 22(7), 1220–1236 (2004)
4. Persson, K., Manivannan, D., Singhal, M.: Bluetooth Scatternet Formation: Criteria, Models and Classification. In: First IEEE Consumer Communication and Networking Conference (2004)
5. Salonidis, T., Bhagwat, P., Tassiulas, L., LaMaire, R.: Distributed topology construction of Bluetooth personal area networks. IEEE Journal on Selected Area in Comm. 23, 633–643 (2005)
6. Chen, H., et al.: Controlling Network Topology in Forming Bluetooth Scatternet. IEICE Transactions on Communications E-88B(3), 943–949 (2005)
7. Yu, C.M., Huang, C.C.: On the Architecture Design and Performance Evaluation of A Configurable Blueweb Network. IEICE Transaction on Communication E-90B(5), 1104–1111 (2007)

A Study on the Global Configured Method of Blueweb Routing Protocol

Chih-Min Yu

Dept. of Communication Engineering, Chung-Hua University
No. 707, Sec 2, Wufu Rd , Hsinchu, 30012, Taiwan, R.O.C
ycm@chu.edu.tw

Abstract. Blueweb is a self-organizing Bluetooth-based multihop network equipped with a scatternet formation algorithm and a hybrid routing protocol. The hybrid routing protocol can be configured for a particular network through adjustment of a single parameter, the number of routing tier. In this paper, a global configured method is proposed to determine the desired configuration for Blueweb routing protocol. The global configured method is used in the route master and designs three blocks including the traffic generator, the query packet estimator, and the global tier decision blocks. The traffic generator block uses a uniform end-to-end traffic model in each master to generate the query packets for various N-tiers. The query packet estimator block measures the local and global query packets to compute the local query probability. The global tier decision block uses the parameter of local query probability to determine the proper number of routing tier. Computer simulation results show that this method can efficiently improve the routing performance and make the routing tiers configurable for a Blueweb routing protocol.

Keywords: Bluetooth, scatternet formation, hybrid routing protocol.

1 Introduction

Bluetree [1] is the first scatternet formation protocol for building a multihop Bluetooth ad hoc network. It adopts one or a few root nodes to start the formation of a scatternet. The resulting topology is tree-shaped and it uses master/slave nodes to serve as relays throughout the whole scatternet. Although its spanning tree architecture achieves a minimum number of connection links between any two nodes, its tree-shaped topology is not reliable under dynamic topological changes.

Based on the same assumption as Bluetree, Bluenet [2] sets up a scatternet in a distributed fashion and it shows that a mesh-like architecture achieves higher information-carrying capacity than a tree-shaped one. In BlueStars [3], each node initially executes an inquiry procedure in a distributed fashion to discover its neighboring devices, then a number of masters are selected based on the number of their neighbors, and finally a number of gateways are selected by these masters and a mesh-like scatternet is formed.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 272–281, 2010.

Another important issue in a Bluetooth multihop network design is routing protocol. Until now, a number of routing protocols have been proposed for Bluetooth multihop networks. In the proactive category [1], such as in the Bluetree, each master node maintains a routing table. The main problem here is the overhead in routing information exchanges, although little delay is involved in determining a route. In the reactive category [4][5], a flooding method is usually used to search for the optimal path from a source node to a destination node and this will incur a certain amount of delay. However, the reactive approach provides better network scalability. In [6], the performance of a hybrid routing protocol is presented for Bluetooth scatternets and it consumes small amount of storage, low routing overhead, and low route discovery latency. Nevertheless, the paper did not try to construct and optimize this hybrid routing protocol for Bluetooth scatternet to achieve its excellent routing performance. Blueweb [7] already propose a hybrid routing protocol in which we use the reactive approach globally in the router master and the proactive approach locally in the master to discover the optimal path for source routing. Nevertheless, these papers [6]-[7] did not try to optimize this hybrid routing protocol for Bluetooth scatternet to achieve its excellent routing performance.

To optimize the Blueweb routing protocol, a test-bed simulation scheme called the global configured method is proposed to determine its proper routing configuration. This method designs three blocks including the traffic generator to generate the query packets, the query packet estimator to compute the local query probability, and the global tier decision blocks to determine the proper number of routing tier.

The rest of this paper is organized as follows: In Section 2, we review the scatternet formation algorithm and the routing protocol of Blueweb. In Section 3, we describe the detailed operation of the global configured method. In Section 4, computer simulations are used to verify the routing performance improvement of the Blueweb network. Finally, a conclusion is drawn in Section 5.

2 A Review on Blueweb

2.1 Scatternet Formation Algorithm

The scatternet formation of Blueweb is executed in two phases. In the first phase, a coordinator called the route master initiates the scatternet formation procedure by paging up to 7 neighboring slave nodes, and forms the first piconet. The slave nodes then switch their roles to masters (called S/M nodes). Each S/M node only pages one additional neighboring slave node. After each S/M node connects to its slave, a role exchange mechanism is executed to make the S/M node function as a relay and make the slave node function as a master. Then the new master node begins to page up to 7 neighboring slave nodes. This procedure is iterated until the leaf nodes of the tree are reached and a tree-shaped topology is created.

In the second phase, a return connection mechanism is used to generate more connection paths among nodes and the tree-shaped topology is converted into a web-shaped topology. Figure 1 illustrates a simple Blueweb topology example.

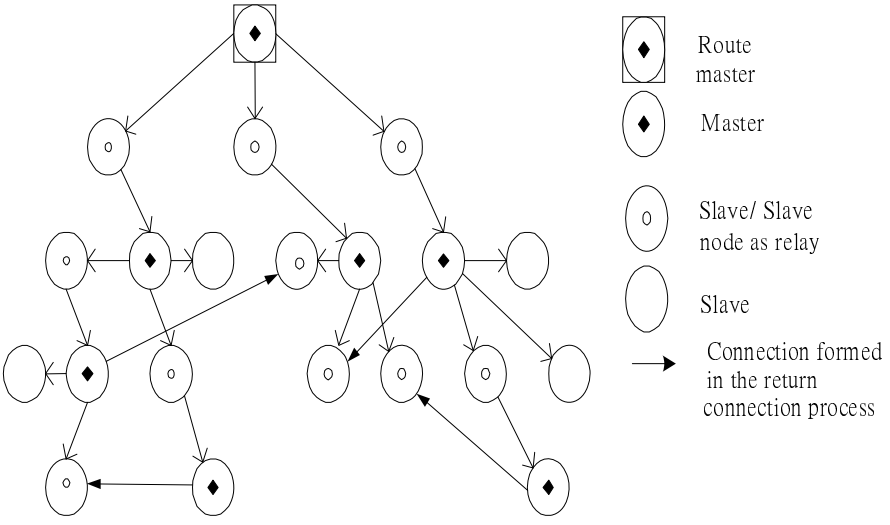


Fig. 1. An example of a connected Blueweb topology

2.2 Modified Source Routing Protocol

In the Blueweb scatternet formation period, some routing information can be exchanged among masters. In the first phase of scatternet formation, each master keeps a record of its directly connected upstream master. As a result, a query path can be easily formed by connecting all the masters in the upstream direction to the route master.

In the second phase of scatternet formation, each returning master will pass its own piconet information together with a list of its directly connected masters to the route master via its upstream masters. At the same time, each returning master including the route master will pass its own piconet information to its directly connected masters. Here, we define the directly connected neighboring piconets within its neighboring N tiers of a master as the N-tier piconets of the master. The associated N-tier piconet information will be stored in the master's ***N-tier piconet table***. In addition, those masters affected by the return connection mechanism will update their N-tier piconet table via relays. As a result, each master will keep its own piconet information and its N-tier piconet information. This information is used locally when a node inquires the master for a path to deliver packets.

After finishing the second phase of scatternet formation, the route master will have the routing information of all nodes and store it in a ***piconet list table***. This table contains a list of all the masters and their associated slaves. Meanwhile, the route master will compute the shortest path for any two-piconet pair using the all-pairs shortest path algorithm. This shortest path information is stored in a ***scatternet routing table*** and is used when any node inquires the route master for routing information to deliver packets.

Based on the routing information collected by all the masters including the route master, a modified source routing protocol is developed. This is a hybrid routing protocol and operates in two phases. In the first phase, an optimal path from source to destination is searched. In the second phase, the optimal path is used to transmit the packets.

3 Global Configured Method

In Blueweb, each master maintains its N-tier piconet information and the route master maintains the global routing information. In each master, the larger N-tier improves the routing performance but generates more routing overhead in terms of routing cost. The number of routing tier is leveraged by the hybrid routing protocol to improve the efficiency of a reactive route query mechanism. As a result, there is a trade-off between routing performance and cost through adjustment of the routing tier parameter.

During the scatternet formation phase, a global configured method is proposed in the route master to compute the global proper routing tier number initially. In the maintenance phase, each master maintains the proper number of routing tier.

3.1 Global Configured Method

Figure 2 shows the block diagram of the global configured method. For each iteration, the route master of Blueweb executes the traffic generator block first and generates a uniform end-to-end traffic for the whole scatternet. Secondly, the route master executes the query packet estimator block to probe the query packets and compute the local query probability. Thirdly, the improvement of local query probability is calculated in the global tier decision block to evaluate the performance of the new routing configuration. This algorithm is iterated until the proper number of tiers is determined.

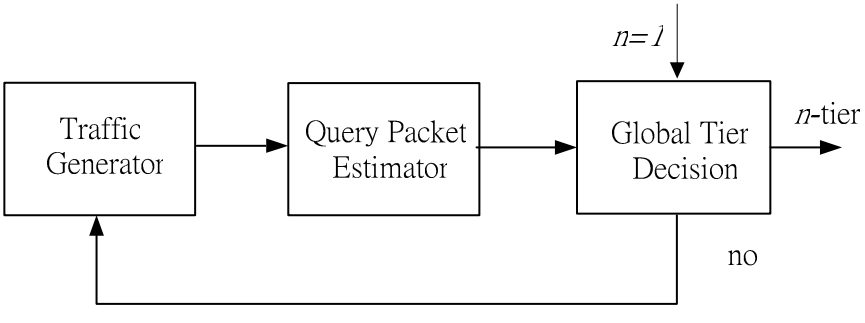


Fig. 2. Block diagram of the global configured method

3.2 Traffic Generator

In our simulation, scatternet topology was constructed with our Blueweb scatternet formation algorithm and the modified source routing protocol with the N-tier piconet

tables. Overall, we simulated topology with 40 nodes randomly distributed in a 40*40 m^2 geographical area.

With a uniform end-to-end traffic model, packets were generated in each node according to a Poisson arrival pattern. Here, we assumed a single packet with a length of 5 time slots was sent in each routing session. Each node queries its associated route master to acquire the routing path for packet transmission. Each route query packet and each route reply packet were assumed to last for only one time slot. Each node was provided a FIFO queue with a length of 160 packets. The source-destination pair in each routing session was randomly selected and packets were forwarded by using the modified source routing protocol.

3.3 Query Packet Estimator

We begin by examining the performance results of both the local and global query packets. This is because that each master is able to measure the amount of queries either from the local master or the route master. With a uniform end-to-end traffic model, each master generates the routing traffic to some randomly selected destinations and the amount of local and global query packets are recorded. The Blueweb 40-node example is simulated to demonstrate the traffic behavior of local and global query.

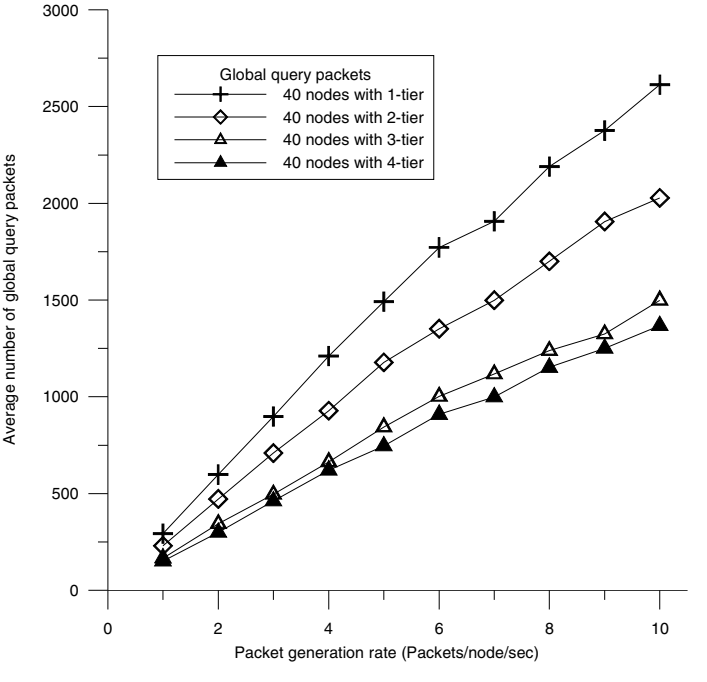


Fig. 3. The global query traffic of Blueweb

Figure 3 shows the average number of global query packets for each master in Blueweb. We observed that the amount of global query packets increases as the packet generation rate increases. In addition, the larger tier number produces the less packets of global query in each master. As a result, the 1-tier configuration achieves the largest amount of global query packets. This is because most of the queries generated for the destinations are out of its routing tiers.

On the other hand, Figure 4 shows the average number of local query packets for each master in Blueweb. From the performance results, the amount of local query packets increases as the packet generation rate increases. In addition, the higher tier in each node produces the more local query packets and the 4-tier generates the largest local query packets. This is because the increment of tiers will increase the frequency of local query.

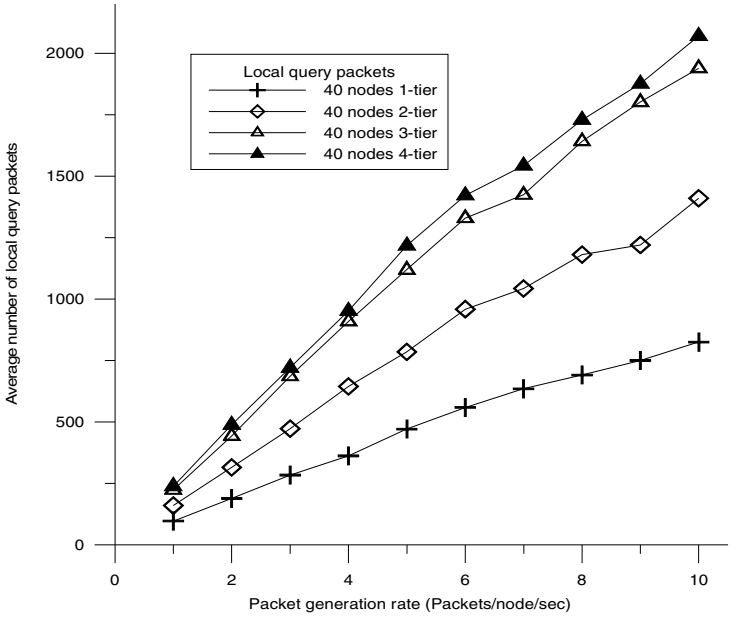


Fig. 4. The local query traffic of Blueweb

3.4 Global Tier decision

The probability of querying local master p_I is defined as the ratio of the total number of local query packets over the total number of query packets (including both the local and global query packets in Figure 3 and Figure 4). Figure 5 shows the performance improvement ratio of $p_{(I+1)} / p_I$ and the improvement ratio decreases as the number of tiers increases. The initial tier number I is set to 1 in this global tier decision block. The average performance improvement of 2-tier/1-tier is t_1 =1.7, 3-tier/2-tier is

t_2 =1.4, and 4-tier/3-tier is t_3 =1.07 as well as their standard deviation are below 0.03. Since the deviation is very small the average performance improvement ratio can be regarded as constant as the various packet generation rates.

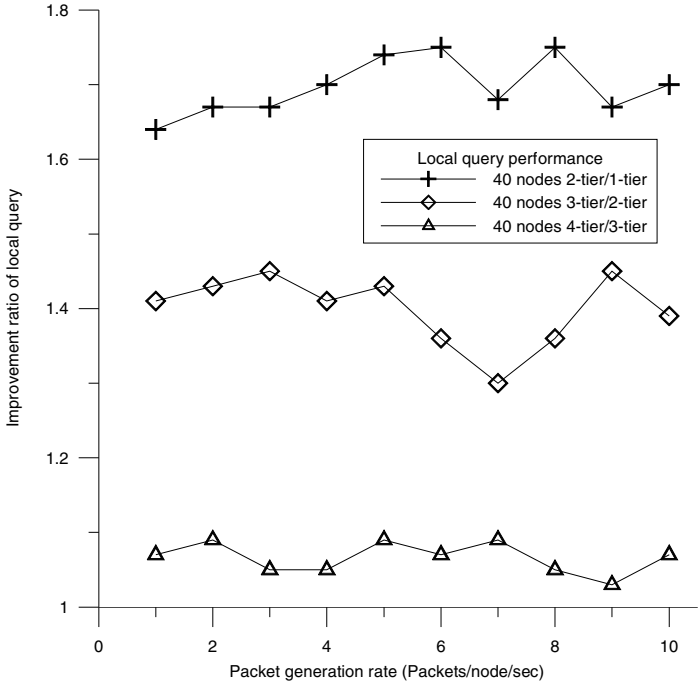


Fig. 5. The performance improvement ratio of local query probability

Based on the performance improvement ratio of query performance in Figure 5, the desired routing performance can be found out by the global configured method in the route master to select the proper number of tiers. A parameter t_I represents the average performance improvement ratio of the corresponding N-tier for various packet generation rates. The performance improvement threshold T=1,1 is a predefined system parameter for the decision algorithm in formation phase.

When the system performance improvement ratio t_I is below the predefined threshold T, the I will be selected as the desired number of routing tiers and the algorithm stops here. From the performance result of Figure 5, the I is determined to be 3 as the desired number of tiers since only its improvement ratio meets the decision criteria and the algorithm stops here for this Blueweb 40-node example.

After determining the proper configuration of Blueweb routing protocol, the route master passes the proper tier number n to its immediate downstream masters. Then, each master passes the n to its immediate downstream masters until the leaf masters are

reached. Finally, each master uses *n* as the init tier number and maintains its proper *n*-tier routing information.

4 Routing Performance Simulation

In this section, we simulate and evaluate the routing performance for Blueweb with the N-tier routing protocol. The performances are based on a uniform end-to-end traffic model to demonstrate the routing performance of Blueweb. A simulation program is written to evaluate the routing performance.

4.1 Simulation Model and System Parameters

In our simulation scenario, the scatternet topologies simulated were constructed by using the scatternet formation algorithms as described in Section 2. Overall, we simulated ten topologies each with 40 nodes randomly distributed in the same geographical area. Table 1 summarized the simulation parameters.

Table 1. The simulation parameters

Simulation time (seconds)	20
Number of nodes	40
N-tier in all masters	4
Traffic pattern	Poisson arrival
Scheduling scheme	Round robin
Routing protocols	Modified source routing
FIFO buffer size	400 packets
Source-destination pair	Randomly selected
Query or reply packet	1 time slot
Data packet (for each routing session)	5 time slots
Each routing session	1 data packet

4.2 Routing Performance

During the simulation, the packets of local and global query for the above scenarios are also computed. With the global configured method, the proper tier number is determined to be 3 for the Blueweb 40-node topology.

The average packet delay metric is defined as the average packet transmission time from the first transmitted bit at the source node to the last received bit at the destination node for every routing packet. In addition, our simulation adopts the Poisson arrival traffic pattern, the round robin scheduling algorithm, and the modified source routing protocol to evaluate this performance metric in a uniform end-to-end traffic model.

Figure 6 shows the average packet delay performance of Blueweb. The average packet delay increases as the packet generation rate increases. In addition, the larger number of tiers achieves better delay performance than the smaller number of tiers and

the 4-tier case generates the smallest average delay. However, the 3-tier case reduces the largest packet delay than the other cases. As a result, the global configured method works well to determine the desired configuration of Blueweb routing protocol.

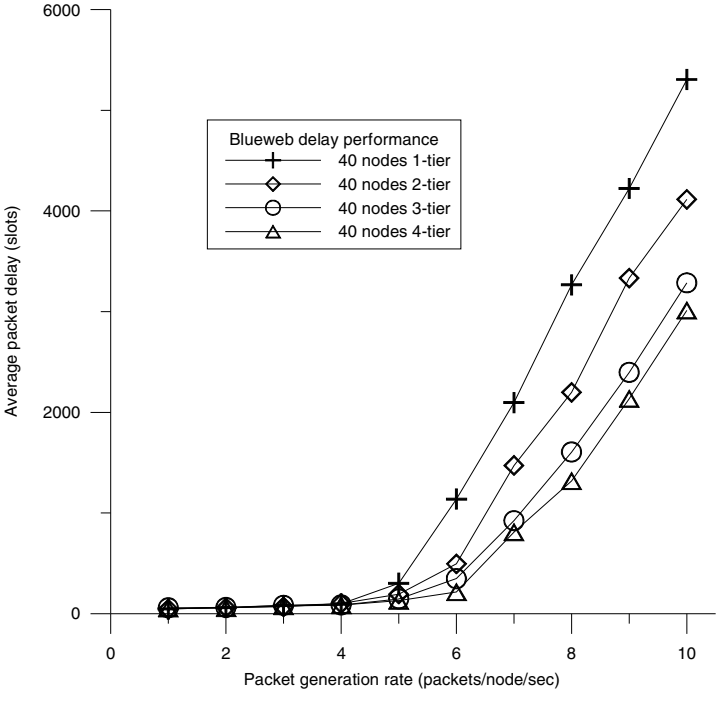


Fig. 6. Average packet delay of Blueweb 40-node example

5 Conclusion

In this paper, a test-bed simulation method called the global configured method is proposed to determine the desired configuration for Blueweb routing protocol. This method designs three blocks including the traffic generator to generate the query packets, the query packet estimator to compute the local query probability, and the global tier decision blocks to determine the proper number of routing tier. In addition, a Blueweb 40-node example is simulated to demonstrate the effectiveness of this global configured method. Finally, computer simulation results show that this method can efficiently improve the routing performance and make the routing tiers configurable for a Blueweb routing protocol.

Acknowledgments

My heartfelt thanks to the support of the work by the National Science Council, Taiwan (NSC-99-22221-E-216-025).

References

1. Zaruba, G.V., Basagni, S., Chlamtac, I.: Bluetrees-scatternet formation to enable Bluetooth-based ad hoc networks. In: IEEE International Conference on Communications, vol. 1, pp. 273–277 (June 2001)
2. Wang, Z., Thomas, R.J., Haas, Z.: Bluenet – A New Scatternet Formation Scheme. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences, pp. 779–787 (2001)
3. Petrioli, C., Basagni, S., Chlamtac, I.: Configuring BlueStars: Multihop Scatternet Formation for Bluetooth Networks. IEEE Transaction on Computers 52(6), 779–790 (2003)
4. Bhagwat, P., Segall, A.: A Routing Vector Method (RVM) for Routing in Bluetooth Scatternets. In: IEEE International Workshop on Mobile Multimedia Communications, pp. 375–379 (1999)
5. Prabhu, B.J., Chockalingam, A.: A Routing Protocol and Energy Efficient Techniques in Bluetooth Scatternets. In: IEEE International Conference on Communications ICC 2002, pp. 3336–3340 (2002)
6. Kapoor, R., Gerla, M.: A Zone Routing Protocol for Bluetooth Scatternets. IEEE Wireless Communications and Networking 3, 1459–1464 (2003)
7. Yu, C.-M., Lin, S.-J., Huang, C.-C.: On the Architecture and Performance of Blueweb: A Bluetooth-based Multihop Ad Hoc Network. IEICE Transaction on Communication E89-B(2), 482–489 (2006)

Energy Efficient Framework for Mobility Supported Smart IP-WSN

Md. Motaharul Islam, Nguyen Tien Dung,
Aymen Abdullah Al Saffar, Sang-Ho Na, and Eui-Nam Huh

Department of Computer Engineering, Kyung Hee University (Global Campus)
Youngin, Gyeonggi-do, South Korea
{motahar,ntiendung,aymen,shna,johnhuh}@khu.ac.kr

Abstract. Energy efficient Internet Protocol based smart Wireless Sensor Networks (IP-WSN) are gaining tremendous importance because of its broad range of commercial applications in health care, building & home automation, environmental monitoring, security & safety and industrial automation. In all of these applications mobility of sensor node with special attention to energy constraints is an indispensible part. Host based mobility management protocol is inherently unsuitable for energy inefficient IP-WSN. So network-based mobility management protocol can be an alternative to the mobility supported IP-WSN. In this regard, Proxy Mobile IPv6 has been standardized by the IETF NETLMM working group, and is starting to pay close attention among the telecommunication and Internet communities. In this paper we propose energy efficient Sensor Network based PMIPv6 protocol called Sensor Proxy Mobile IPv6 (SPMIPv6). We present SPMIPv6 architecture, respective message formats and analyze the energy dissipation and finally evaluate its performance.

Keywords: NETLMM, IP-WSN, IETF, 6LoWPAN, IEEE 802.15.4.

1 Introduction

Recently advancement in micro-electro-mechanical system and wireless communication have enabled the development of low cost, low power, multifunctional sensor nodes that are small in size and communicate in short distances [1]. A sensor network is a special type of communication network that is composed of a large number of sensor nodes that are densely deployed either inside the phenomena or very close to it [4]. The tiny sensor nodes consisting of sensing, data processing and communicating components are capable of holding IP stack [3]. That is why application of wireless sensor networks are now quite broad than the earlier. IP-WSN concept is being implemented in many sophisticated application from building and home automation to industrial manufacturing. By the introduction of adaptation layer over IEEE 802.15.4 Physical and Medium Access Control layer it becomes feasible to transmit IPv6 packet in IP-WSN [2]. Adaptation layer make usage of stateless compression technique to elide adaptation, network and transport layer header fields- compressing all the three layers down to a few bytes [3]. However IP-WSN introduces excessive signaling overhead due to its numerous tunneling over the air. Excessive signaling cost becomes a barrier for the real life implementation of low power IP-WSN.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 282–291, 2010.

Proxy Mobile IPv6 (PMIPv6), a network based localized mobility management protocol [5] provides mobility support to any IPv6 host within a restricted and topologically localized portion of the network and without requiring the host to participate in any mobility related signaling [9], [12]. In this paper we have emphasized different sort of mobility scenario with consideration of the energy consumption. In addition to that we have introduced the concept of PMIPv6 by modifying the functionality of its Mobile Access Gateway and Local Mobility Anchor to IP-WSN enabled gateway and anchor point. Then we propose the protocol architecture named Sensor Proxy Mobile IPv6 (SPMIPv6), its functional architecture, necessary message formats. Moreover we compare our network mobility model with MIPv6 model and finally evaluate performance of our proposed scheme.

The rest of the paper is organized as follows. Section 2 reviews the background related to PMIPv6 and IPv6 over low power wireless personal area network (6LoWPAN). Proposed Sensor PMIPv6 Protocol architecture, its mobility scenario, sequence diagram of message flow, message formats and operational architecture are depicted in section 3. Section 4 illustrates Performance evaluation and finally, section 5 concludes this paper.

2 Overview of PMIPv6 and 6LoWPAN

PMIPv6 is designed to provide network-based mobility management support to a Mobile Node (MN) in a topologically localized domain [13]. Therefore, an MN is exempt from participation in any mobility-related signalling, and the proxy mobility agent in the serving network performs mobility-related signalling on behalf of the MN. Once an MN enters its PMIPv6, the serving network assigns a unique home network prefix to each MN, and conceptually this prefix always follows the MN wherever it moves within a PMIPv6 domain. From the perspective of the MN, the entire PMIPv6 domain appears as its home network. The new principal functional entities of PMIPv6 are the mobile access gateway (MAG) and local mobility anchor (LMA). The MAG acts like an access router and LMA act as the mobility anchor point of the PMIPv6 domain [9], [12].

And 6LoWPAN [2] is a low power wireless personal area network working group at IETF. It defines an adaptation layer for sending IPv6 packets over IEEE 802.15.4. The goal of 6LoWPAN is to reduce size of IPv6 packets to make it fit in 127 bytes 802.15.4 frame. 6LoWPAN consist of a header compression scheme, fragmentation scheme and a method for forming IPv6 link local address on 802.15.4 network [15].

3 Proposed SPMIPv6 Scheme

3.1 Overview of SPMIPv6 Protocol

We propose SPMIPv6 protocol for network based localized mobility management protocol for IP-WSN. The SPMIPv6 architecture will consists of Sensor network based Localized Mobility Anchor (SLMA), Sensor network based Mobile Access Gateway (SMAG) [10], numerous fully functioned IPv6 header stack enabled sensor node. In this model SLMA will also incorporate the functionality of Authentication, Authorization,

and Accounting (AAA); we call it Sensor Authentication, Authorization, and Accounting (SAAA) service. The main role of SLMA is to maintain the reach ability to the sensor node's address while it moves around within the SPMIPv6 domain, and the SLMA includes a binding cache entry for each recently registered sensor node. The binding cache entry maintained at the SLMA is more specific than LMA in PMIPv6 with some additional fields such as sensor node identifier, the sensor node's home network prefix, and a flag bit indicating a sensor proxy registration. SMAG acts as an Edge Router. The main function of SMAG is to detect sensor nodes movement and initiate mobility related signaling with the sensor node's SLMA on behalf of the sensor node.

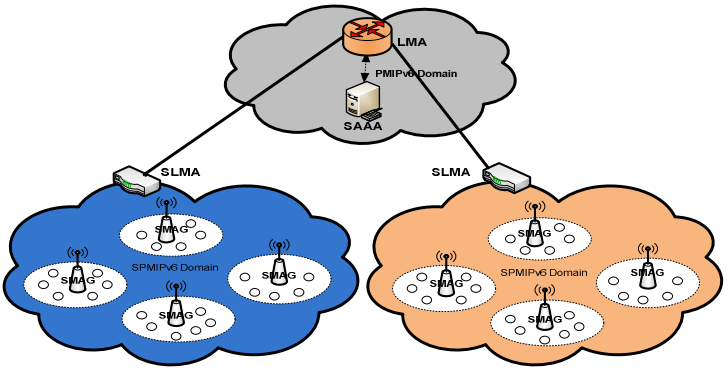


Fig. 1. Sensor Proxy Mobile IPv6 Architecture

The functionality of SLMA and SMAG in SPMIPv6 are different in many ways but similar in nature in comparison with LMA and MAG of PMIPv6. The major difference is both SLMA and SMAG works with low power 6LoWPAN sensor nodes. But both SLMA and SMAG deal with a plenty of sensor nodes. SLMA will act as a topological anchor point of all the SMAG. Inbuilt AAA functionality of SLMA helps the SMAG and sensor node to move the SPMIPv6 domain.

3.2 SPMIPv6 Mobility Scenario

We consider several different mobility scenarios.

Case-I: Movement of patient within the same SMAG of the SPMIPv6 domain
Case-II: Patient movement between different SMAGs of same SPMIPv6 domain
Case-III: Movement of patient between different SMAGs of different SPMIPv6
Case-IV: Movement of a SMAG-based PAN within the same SPMIPv6 domain
Case-V: Movement of a SMAG-based PAN between different SPMIPv6 domains
Case-VI: Patient monitoring in personal home environment

These scenarios are explained below.

Case-I: In this case, the mobilities of the patients will be handled by the appropriate SMAG, without the involvement of the SLMA. This, the simplest mobility scenario, arises frequently in hospital management: a patient can move within the PAN of a single branch of the hospital for purposes such as exercise and fresh air.

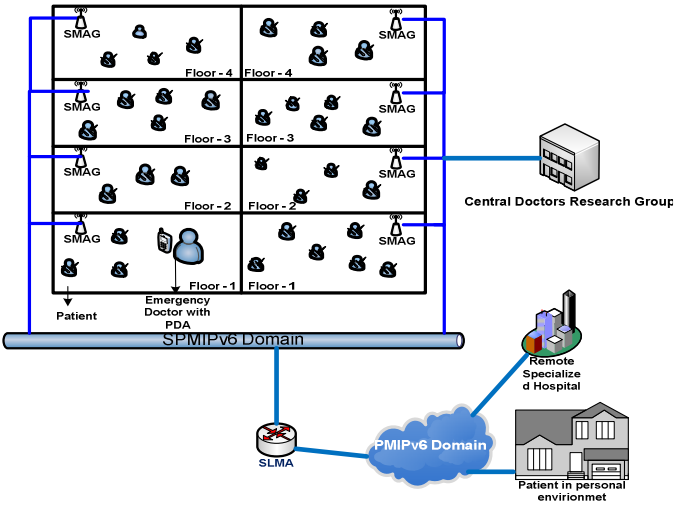


Fig. 2. SPMIPv6 based hospital management

Case-II: In this case, mobility will be handled by the appropriate SMAG with minimal initiative from the SLMA. The initial coordination will be performed by the SLMA alone; then the SMAG will oversee the remaining procedures. In our hospital management model, a patient can move from one PAN to another PAN in the same branch of the hospital.

Case-III: In this case, mobility is inter-domain, using the public PMIPv6 domain. The LMA, AAA, and SLMA will coordinate with one other. In our hospital management model, a patient can move on an emergency basis from one PAN of a hospital branch to a PAN of another branch of the same hospital.

Case-IV: In this case, mobility is based on the NEMO protocol, confined to the same domain. Only the SLMA and corresponding SMAGs will be involved. In our hospital management model, a patient with the whole set up can move from one PAN to another PAN.

Case-V: The final case also affords NEMO-based mobility, but is much different from case-IV. In our hospital management model, a patient can move on an emergency basis with its whole setup from one branch of a hospital to the more specialized branch of the same hospital.

Case-VI: Due to the increasing number of aging demographic group, we consider this case so that a patient can be monitor continuously from the patient's personal environment as discussed in our recent paper [19].

3.3 Proxy Binding Message Format for SPMIPv6

In the proposed proxy binding update and proxy binding acknowledgement message we have added a flag bit S. If S flag is set it indicates the SMIPv6 based operations. If S bit is not set then it will indicate other operations apart from SPMIPv6.The other flags indicate meaning as mentioned in [6], [7], [8], [9].

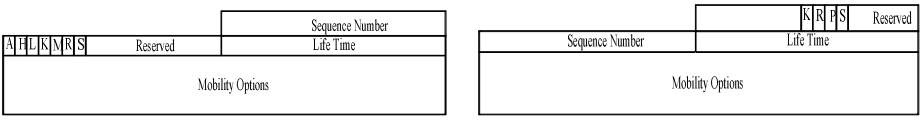


Fig. 3. SPMIPv6 PBU and PBA Message Formats

3.4 Architecture of the SPMIPv6

Figure 4 represents the functional architecture of SPMIPv6 which includes the functionality of SLMA, SMAG and Sensor node. It also depicts the interaction between the three entities. Since the sensor node is IP based so consists of all the layers including adaptation layer. Sensor node will be identified by 64 bits interface identifier. And it can easily generate its IPv6 address by combining interface identifier with network prefix provided by the corresponding Sensor Mobile Access Gateway. Here SMAG is full function device that support complete implementation of IPv6 protocol stack and sensor node is reduce function device that support minimum IPv6 protocol implementation.

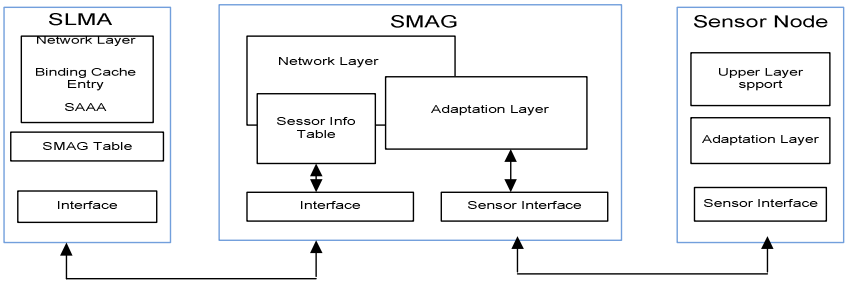


Fig. 4. Operational Architecture of SPMIPv6

4 Performance Evaluations

4.1 Network Model of IP-WSN

To evaluate the total signaling costs and mobility related cost, we compare our analytical model with MIPv6 and SPMIPv6. For analyzing signaling costs, we use a two-dimensional random walk model [11], [12], [14], [16] based on the properties of regular, absorbing Markov chains. Random walk mobility models are designed for dynamic location areas and are suitable for modelling user movement when mobility is generally confined to a limited geographical area. Such scenarios include homes, vehicles, hospitals, and departmental stores [17].

Table 1. System Parameter

Symbol	Description
BU	Binding Update Message
BA	Binding Acknowledgement Message
PBU	Proxy Binding Update Message
PBA	Proxy Binding Acknowledge Message
$D_{smag\text{-}slma}$	Distance between SMAG and SLMA
$D_{sn\text{-}smag}$	Distance between SN and SMAG
$M_{intra\text{-}pan}$	Intra PAN Mobility
$M_{inter\text{-}pan}$	Inter PAN Mobility
γ	Unit transmission cost in wireless link
σ	Unit transmission cost in wired link
RREQ	Router Request Message
RREP	Router Reply Message
C_{sd}	Sensor Mobility Cost
C_{bu}	Binding Update Cost

The total signaling cost of the proposed scheme based on MIPv6:

$$SC_{mipv6} = M_{intra\text{-}pan}.\ C_{sd} + M_{inter\text{-}pan}.(C_{sd} + C_{bu})$$

$$\text{Where, } C_{sd} = \gamma.\ (RREQ + RREP)\ D_{sn\text{-}smag}$$

$$C_{bu} = \sigma.(BU+BA)\ D_{sn\text{-}smag} + \gamma.(BU+BA)\ D_{smag\text{-}slma}$$

The total signaling cost of the proposed scheme based on SPMIPv6:

$$SC_{spmipv6} = M_{intra\text{-}pan}.\ C_{sd} + M_{inter\text{-}pan}.(C_{sd} + C_{bu})$$

$$\text{Where, } C_{sd} = \gamma.\ (\ RREQ + RREP\ )\ D_{sn\text{-}smag}$$
$$C_{bu} = \sigma.\ (PBU+PBA)\ D_{smag\text{-}slma}$$

4.2 Energy Consumption Model of IP-WSN

We consider IP-WSN with densely deployed IP sensing device. The network consists of two types of IP sensing device: Fully functional IP sensing device (IP-FFD) and reduced functional IP sensing device (IP-RFD). We have used model for the energy consumption per bit at the physical layer from [18], [20], [21], [22].

Table 2. Parameter values

Parameter	Value
No of IP-WSN Node (N)	25~120
Network Area (A)	120 x 120M
Node density (ρ)	0.00173~0.00833
Initial Energy	2 J
Transmit/Receive electronics (L_E)	50 nJ bit^{-1} m^2
Transmission Power	5.85 x 10^{-5} W
Number of SMAG	1-10
Transmission range (r)	25 m
Packet size	2KB

Here, E^{tx} and E^{rx} is the distance-independent amount of energy consumed by the transmitter and receiver electronics and the digital processing of each. Here α and β are path loss exponent ($2 < \alpha < 5$) and a constant [J/bit m2], r is a transmission range. L_{ctrl} is the length of control packets in bits, LE is the energy needed by the transmitter device to transmit or receive a packet, and T is the time period between two consecutive topological changes of the IP-WSN. $nffd_p(d_i)$ indicates the number of fully functional neighbouring node for a path p and range d.

$$E = E^{tx} + \beta * d^{\alpha} + E^{rx}$$

Since same type of transmitting and receiving device is concerned in IP-WSN

$$E^{tx} = E^{rx} = E_{dec}$$

$$E = 2 * E_{dec} + \beta * d^{\alpha}$$

$$C_i^{ctrl}(r) = [L_{ctrl} * \beta * r^{\alpha} + (n_i(r)+1) * L_{ctrl} * L_E]\frac{1}{T}$$

$$C_i^{\inf}(p) = [\sum_{i=1, j=2}^{N} (nffd_p(d_i)+1) * L_{data} * \beta d_{i,j}{}^{\alpha} + (n_p(d_i)+1) * L_{data}] * L_E$$

$$C_i^{total}(p) = \sum_{i=1}^{N} [C_i^{ctrl}(r)] + C_i^{\inf}(p)$$

So energy consumed by MIPv6 scheme can be calculated by the following mathematical derivation. Here E_i^{mipv6} indicates energy consumption by MIPv6 scheme.

$$E_i^{mipv6} = SC_i^{mipv6} * C_i^{total}(p)$$

So energy consumed by MIPv6 scheme can be calculated in the same way. Here $E_i^{spmipv6}$ indicates energy consumption by SPMIPv6 scheme.

$$E_i^{spmipv6} = SC_i^{spmipv6} * C_i^{total}(p)$$

The figure 5 depicts the energy consumption with respect to the IP-WSN node density in term of the MIPv6 and SPMIPv6. Energy consumption increases as the IP-WSN node density increase. Our proposed scheme increases the performance linearly with the comparison to MIPv6. And the energy consumption increases more rapidly as the density of IP-WSN node increases.

The figure 6 shows the energy dissipation with respect to number of IP-WSN source nodes used in both MIPv6 and PMIPv6. Energy dissipation increases almost exponentially as the no of source nodes increase. In this case, our proposed scheme dissipates much less energy with respect to MIPv6

The figure 7 shows the energy consumption with respect to payload. Energy consumption is linear in both MIPv6 and SPMIPv6. But due to fragmentation overhead energy consumption increased rapidly and then it represents the linear characteristics.

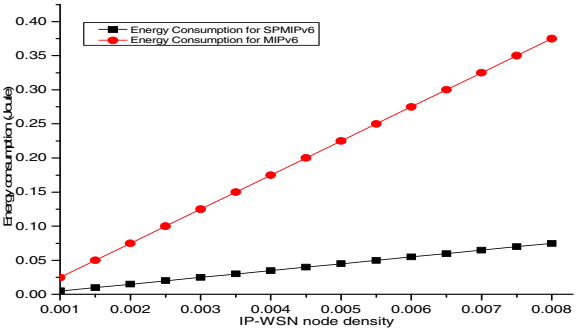


Fig. 5. Node Density vs. Energy Consumptions

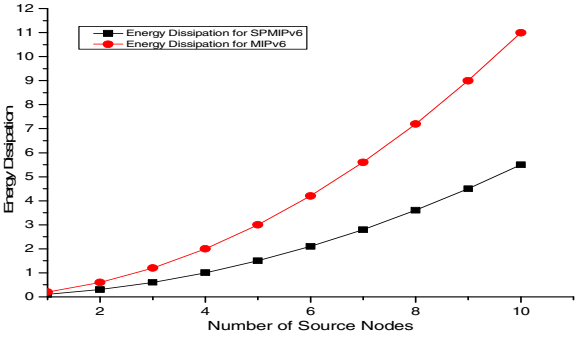


Fig. 6. Number of Source Nodes vs. Energy Dissipation

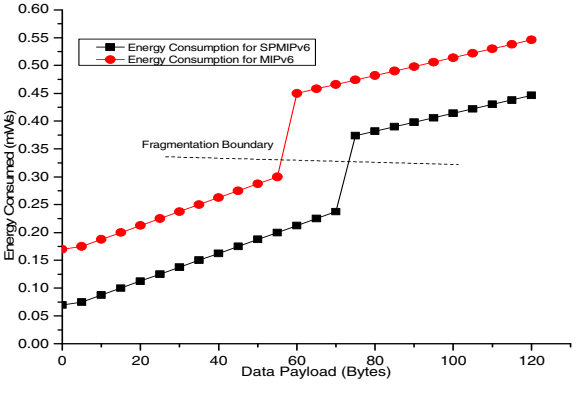


Fig. 7. Data payload vs. Energy consumption

5 Conclusions

Energy consumption for packet delivery for the individual tiny sensor node in IP-WSN is a big challenge to overcome. In IP-WSN, if the individual sensor node wants

to communicate with the gateway router then it generates huge air traffics and it deteriorates the performance at a large scale. IETF NETLMM working group has standardized network based localized mobility management protocol called PMIPv6. In this paper we propose a network based IP-WSN scheme based on the PMIPv6 called SPMIPv6 and further develop the architecture, corresponding message formats, analyzing energy consumptions and finally evaluate its performance. Analysis shows that the proposed scheme reduces the energy consumption than other scheme. In this paper we only focus IP-WSN of the same vendor and protocol stack. In future we will focus on the sensor network consisting of multi vendor and heterogeneous protocol stack.

Acknowledgements

"This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2010-0016959)".

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Communications Magazine 40(8), 102–114 (2002)
2. Kushalnagar, N., Montenegro, G., Schumacher, C.: IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): Overview, Assumptions, Problem Statement, and Goals, IETF RFC 4919 (2007)
3. Montenegro, G., Kushalnagar, N., Hui, J., Culler, D.: Transmission of IPv6 Packets over IEEE 802.15.4 Networks, IETF RFC 4944 (2007)
4. Akka, K., Younis, M.: A survey on routing protocols for wireless sensor networks. Ad Hoc Networks 3(3), 325–349 (2005)
5. Kempf, J.: Problem statement for Network-Based Localized Mobility Management (NETLMM), IETF RFC 4830 (2007)
6. Johnson, D., Perkins, C., et al.: Mobility Support in IPv6, IETF RFC 3775 (2004)
7. Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol, IETF RFC 3963 (2005)
8. Soliman, H., Castelluccia, C., El Malki, K., Bellier, L.: Hierarchical Mobile IPv6 Mobility Management (HMIPv6), IETF RFC 4140 (2005)
9. Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., Patil, B.: Proxy Mobile IPv6, IETF RFC 5213 (2008)
10. Chalmers, R.C., Almeroth, K.C.: A Mobility Gateway for Small-Device Networks. In: Second IEEE Annual Conference on Pervasive Computing and Communications, Washington DC, USA (June 2004)
11. Akyildiz, I.F., Lin, Y.B., Lai, W.R., Chen, R.J.: A new random walk model for PCS networks. IEEE Journal on Selected Area in Communication 18(7), 1254–1259 (2000)
12. Islam, M.M., Na, S.-H., Lee, S.-J., Huh, E.-N.: A novel scheme for PMIPv6 based Wireless Sensor Network. In: The 4th International Conference on Information Security and Assurance, Miyazaki, Japan, June 23-25 (2010)
13. Kim, M.-S., Lee, S.K.: A novel load balancing scheme for PMIPv6 based network. International Journal of Electronics and Communications, doi:10.1016/j.aeue.2009.03.003

14. Shidhu, B., Singh, H.: Location Management in Cellular Networks. World Academy of Science, Engineering and Technology 25, 314–319 (2007)
15. Kim, J.H., Hong, C.S., Shon, T.: A Lightweight NEMO Protocol to Support 6LoWPAN. ETRI Journal 30(5), 685–695 (2008)
16. Hasan, M., Akbar, A.H., Mukhtar, H., Kim, K.-H., Kim, D.-W.: A scheme to support mobility for IP based sensor networks. In: 3rd International ICST Conference on Scalable Information Systems, Vico Equense, Italy, June 4-6 (2008)
17. Shelby, Z., Bormann, C.: 6LoWPAN: The Wireless Embedded Internet. John Wiley & Sons Ltd., Chichester (2009)
18. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Trans. On Wireless Comm. 1(4), 660–670 (2002)
19. Hwang, S.-M., et al.: Multi-Modal Sensing Smart Spaces Embedded with WSN Based Image Camera. In: The 3rd International Conference on Pervasive Technologies Related to Assistive Environments, Samos, Greece, June 23-25 (2010)
20. Pathan, A.-S.K., Hong, C.S.: SERP: secure energy-efficient routing protocol for densely deployed wireless sensor networks. Ann. Telecommun. 63, 529–541 (2008), doi:10.1007/s 12243-008-0042-5
21. Razzaque, M.A., Hong, C.S.: Analysis of energy-tax for multipath routing in wireless sensor networks. Ann. Telecommun. 65, 117–127 (2010), doi:10.1007/s12243-009-0110-5
22. Singh, D., Lee, H.-J., Chung, W.-Y.: An Energy Consumption technique for global Healthcare Monitoring Applications. In: International Conference on Information Sciences, Seoul, Korea, November 24-26 (2009)

An Efficient Migration Framework for Mobile IPTV

Aymen Abdullah Alsaffar, Tien-Dung Nguyen,
Md. Motaharul Islam, Young-Rok Shin, and Eui-Nam Huh

Department of Computer Engineering
Kyung Hee University, Suwon, Korea
{aymen,ntiendung,shinyr,johnhuh}@khu.ac.kr,
motahar@icns.khu.ac.kr

Abstract. In this paper we present an efficient migration framework for mobile IPTV services. We present a secure migration for their devices when user migrates to receive MIPTV services for the first time, indoor and outdoor migration. In addition we use Conditional Access System (CAS) and Digital Right Management (DRM) to protect MIPTV service and digital content.

Keywords: MIPTV Services, CAS, DRM, Framework, Migration, Digital multimedia.

1 Introduction

In the progressive development of super-high speed broadband network and advancing of mobile devices capabilities, consumers are capable of connecting to the internet through Wifi Access Point or 3G to receive multimedia content and services. Mobile Internet Protocol TV (MIPTV) one of the services that are receiving tremendous demand by user in these days. MIPTV is technology that enables users to transmit and receive multimedia traffic including television signal, video, audio, text and graphic services through IP-based wired and wireless network [1]. MIPTV service providers are increasing their effort to provide diversity of services to user at home and to mobile devices while they are on the move. The recent development of STB that include wifi capability, make it possible for MIPTV provider to provide services [2]. However, wireless network environment has threats such as denial-of-service attack (DoS attack), replay attack, and man-in-the middle attack. Therefore, we proposed an efficient migration framework for mobile IPTV services to provide consumer with secure connection when migrating to MIPTV services. In addition, we explain the role of Conditional Access system (CAS) and Digital Right Management (DRM) in the MIPTV services. The rest of this paper is organized as follows. Section 2 related work. Section 3 proposed efficient migration framework for mobile IPTV services. Section 4 performance evaluation and lastly, section 5 the conclusion.

2 Related Work

2.1 CAS and DRM Role in MIPTV Security System

In present, multimedia such as video, data or voice have been digitalized and uploaded to the internet so consumer can have access to it anywhere anytime. One of this digital

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 292–301, 2010.

content that has a demand is known as MIPTV content. In order to protect unauthorized access of MIPTV services and digital content, Conditional Access System is used to ensure only subscribed member can receive services. And digital right management (DRM) will ensure the right of digital content is not been violated in anyway.

Conditional access system (CAS) is used to preventing non-subscriber from receiving the services [3]. There are three main functions: scrambling and descrambling, Entitlement Control Message (ECM), and Entitlement Management Message (EMM). There are two main concepts to be considered. The first one is entitlement which known as authorization. In scrambling, the information is transformed to make it unreadable to anyone except the one who possess away of descrambling it by a given key [3]. In this way, CAS will protect the business and the profit of charge broadcasting services provider. Several researchers proposed many techniques of CAS which are suitable to protect their business.

DRM is used to protect and manage the user right of digital content such as editing, copying or reproducing. The implementation of DRM will establish a protocol between user and service provider in how the content can be used. However, a variety of standardization efforts associated with certain aspects of DRM have been recently initiated including the Open Mobile Alliance (OMA), Open Digital Rights language (ODRL), MPEG-21, and Coral Consortium [12]. Yet, these efforts and the field of DRM itself are at an early stage of development, therefore a viable open DRM architecture has yet to emerge.

2.2 Authentication Mechanisms for User Mobile Devices

Here we will describe recently used user authentication mechanisms such as Kerberos, EAP-TLS and their advantages and disadvantages. Furthermore, we compare them to our proposed authentication mechanism.

2.2.1 Kerberos Authentication Mechanism

Kerberos is an authentication mechanism that is used in a distributed environment. It uses a third party authentication server that allow users and servers to trust each other and therefore securely establish communication. Kerberos works by encrypting data by using symmetric encryption for the authentication [4, 5].

For instant, when a user needs to access a service server (SS), he/she need two tickets to get authenticated to SS. A ticket granting ticket (TGT) received from authentication server (AS). The user will send TGT to ticket granting server (TGS) to prove identity. A user will receive second ticket from TGS and therefore access SS. The disadvantage is the user may have access to workstation and pretend to be someone else [5]. For its weakness in security issues, it is vulnerable against above mentioned security issues.

2.2.2 Authentication Mechanism for Anonymity and Privacy Assurance

The authentication mechanism uses Extensible Authentication Protocol Transport Layer Security (EAP-TLS) authentication and Symmetric key (PKI). It provide feature such as single sign on (SSO), privacy, and user anonymity as the content provider affiliated to the authentication server can use service without the need for a separate sign on process when the user get authenticated by Authentication, Authorization and Counting (AAA) server [4]. Using the services through anonymity

will secure the user anonymity and provide easy way to exchange session key to obtain secure data transportation between user and service provider. However, the overhead of client side certificate is it deadly weakness [4].

3 Proposed Migration Mechanism of Secure MIPTV Services

In this section, we provide mobile user with three methods of a fast and secure migration of their mobile devices to receive MIPTV services through STB with wifi access point or 3G. In the process of migrating user mobile devices we securely authenticating consumer to receive MIPTV services. Table 1 gives description of system parameters used in this scheme.

Table 1. The System Parameters

Notation	Description
ID_M	Mobile ID.
ID_{PW}	Mobile ID password.
Lic_M	Mobile generated license.
STB_N	Set-top box number.
STB_{ID}	Set-top box id.
STB_{PW}	Set-top box password.
STB_{Lic}	Set-top box generated license.
$Nonce_M$	Random number of user mobile.
$Nonce_{STB}$	Random number of user STB.
$Nonce_{SP}$	Random number of service provider.
D_{Nonce}	Random number of Kerberos server.
Nonce	Random number of Kerberos client.
E_{CT}	Digital content encrypted.

3.1 Initial Phase to Migrate Mobile Devices to STB through Wifi Access Point

In this initial phase we assume the user did not register to receive MIPTV services through their mobile devices. The user only has the services to be watch through STB at home. Therefore, we will use initial phase to register and authenticate user mobile for the first time through wifi access point to home STB which store user STB license that they receive from license provider when they applied for MIPTV services. By connecting their mobile devices to their STB through wifi access point or STB with built-in wifi technology [2] and providing their STB license, they are securely authenticated by service provider. See figure 1 for a brief explanation. This basic architecture was introduce in our previous work [13].

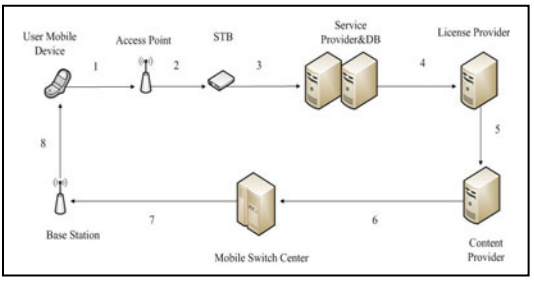


Fig. 1. Initial phase process to migrate mobile devices to STB through wifi access point

The following components of this system architecture are defined in this paper:

- User Mobile Devices: will have a USIM that contain information about the user such as identification, password, and user content license, etc.
- Wifi Access Point (AP): provide wifi access point between STB and users devices at home.
- Set-top box (STB): will have USIM that contains information about STB such as id, password, license and other information.
- Service Provider (SP): provide MIPTV services to user mobile devices base station or to home STB through wifi access point.
- Database (DB): Will store all consumers received information& update them.
- License Provider (LP): Generate a license for home STB or mobile devices user. The license will identify user and provide user with MIPTV services.
- Content Provider (CP): The content provider will prepare the requested digital content based in the agreed license between mobile user and content provider.

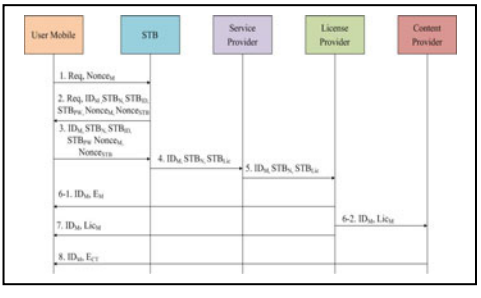


Fig. 2. Initial phase process of migrating mobile devices to STB though wifi access point

When the mobile user request MIPTV services through the user devices such as mobile device, the process is as follow (see figure 2):

Authentication Process: to watch MIPTV programs using mobile device, a request send through wifi access point to STB at home to migrate their devices to STB followed by mobile generated random number ($Nonce_M$). User STB request from mobile devices the following information such as $\{ID_M \mid STB_N \mid STB_{ID} \mid STB_{PW}\}$ followed by mobile and STB received random number ($Nonce_M$, $Nonce_{STB}$) to authenticate the user mobile device to access user STB at home.

- Mobile user send request $\{Req \mid Nonce_M\}$.
- STB request $\{ID_M \mid STB_N \mid STB_{ID} \mid STB_{PW} \mid Nonce_M \mid Nonce_{STB}\}$.
- Mobile user send $\{ID_M \mid STB_N \mid STB_{ID} \mid STB_{PW} \mid Nonce_M \mid Nonce_{STB}\}$

The STB received the requested information and compare it with the one stored in STB smart card (USIM). If it is valid, then transmit it to service provider. Otherwise, send an error message to mobile device through STB or base station.

Registration Process: When the STB complete mobile user authentication and migration to STB, it transmits $\{ID_M \mid STB_N \mid STB_{Lic}\}$ to service provider. The service

provider processes the registration and a request will be forward to database and to the license provider to generate a new user mobile license.

- $\{ID_M \mid STB_N \mid STB_{Lic}\}$.

License Generation Process: When a license provider receives $\{ID_M \mid STB_N \mid STB_{Lic}\}$ from service provider, it will generates another license for $\{Lic_M\}$ and forward one copy to content provider and user STB and then to user mobile.

- $\{ID_M \mid Lic_M\}$.

Content Transmit Process: The content provider receives a license from license provider for mobile device (Lic_M). Then prepare the digital content, encrypt it (E_{CT}) and transmit the requested program to user mobile through STB or base station.

3.2 Indoor Phase to Migrate Mobile Devices to STB through Wifi Access Point

In the indoor phase, we assume the user is at home watching program, but he/she want to suddenly go out but still want keep watching their program. The process of migrate devices to STB through wifi access point is less than before. We assume that the mobile user already have a mobile license in their mobile device.

Migrate to STB Process: Mobile user connect to STB through wifi access point. User send request to STB to watch the program through mobile device followed with mobile user random number ($Nonce_M$). The STB request $\{ID_M \mid Lic_M\}$ followed with random number ($Nonce_{STB}$). The mobile user forwards the requested information to STB and compares $\{ID_M \mid Lic_M\}$ and forward request to content provider.

- Mobile user send request $\{Req \mid Nonce_M\}$.
- STB request $\{ID_M \mid Lic_M \mid Nonce_M \mid Nonce_{STB}\}$.
- Mobile user send $\{ID_M \mid Lic_M \mid Nonce_M \mid Nonce_{STB}\}$.

Content Transmit Process: The content provider request a copy of user mobile information from user STB such as $\{ID_M \mid Lic_M\}$. If they are valid then prepare and encrypt content to provide services to user mobile. Therefore, the mobile user receives digital content and enjoy while they are indoor. If not valid, an error message will be send through STB or base station to user mobile.

3.3 Outdoor Phase to Migrate Mobile Devices through Base Station

In the outdoor phase, the MIPTV user want to receive the MIPTV services without connect their devices to STB at home. We assume that the mobile user already has a mobile license that allow user to receive MIPTV services through base station. When user comes home and still wants to watch MIPTV services privately, depending in how close they are to STB or base station they will receive MIPTV services. Therefore, users will have flexible methods to switch their devices indoor to outdoor and vice versa.

Authentication Process: Outdoor user request MIPTV services followed by mobile generated random number ($Nonce_M$) to service provider through base station. Service provider requests $\{ID_M \mid Lic_M\}$ followed by generated random number from user

mobile and compare them. If valid, content provider sends MIPTV services to user mobile devices. Otherwise mobile user receive error message through base station.

- Mobile user send request {Req | $Nonce_M$}.
- Service provider request {ID_M | Lic_M | $Nonce_M$ | $Nonce_{SP}$}.
- Mobile user send {ID_M | Lic_M | $Nonce_M$ | $Nonce_{SP}$}.

Content Transmit Process: The content provider receives a request from service provider includes {ID_M | Lic_M}. Then the contents are prepared and encrypt. Finally, transmit it to user mobile to be watching it.

4 Performance Evaluation

4.1 Security Evaluation

The system performance evaluation is based on security aspects, communication cost and handover latency. In the process of migrating mobile devices with other devices through wireless communication, a secure authentication and authorization is essential security issue. In the process of authenticating the user, a security threat might occur such as; Denial-of-service attack: occurs when the server is cheated by an attacker to update the false verification information for the next login phase [6]. Man-in-the middle: is a form of active eavesdropping where the attacker becomes the middle man in the communication between two users A and B [6].

Most of these security threats try to access user computers while users are communicating with each other. As a result, the privacy and security is compromised. A cryptographic Nonce is used between consumer mobile devices and STB to prevent threats. It is a pseudo-random or random number issued in an authentication protocol to make sure that old communication cannot be used in above mentioned security threats [8][9]. The nonces are different every time that 401 authentication challenge response code is presented, and client request has a unique number, therefore preventing other attacks from occurring [8].

4.2 Communication Cost Evaluation

To evaluate the cost of our proposed authentication mechanism, we will compare it with previously proposed authentications mechanism such as Kerberos and EAP-TLS. Therefore, we will calculate the number of message exchange between entities in each authentication mechanism to get the cost efficiency.

4.2.1 Kerberos Authentication Mechanism

First, we will calculate the number of exchange message between entities in order to compute the cost of authentication mechanism [11].

$$C_{aut\_Msg} = 2C_{M\_STB} + 2C_{M\_LP} + 4C_{M\_CP} \quad (1)$$

We will calculate the cost of exchanged messages between Mobile user and STB, Mobile user and License provider, and Mobile user and Content provider. We would like to mention that C_{M_LP} and C_{M_SP} have same value of 3.

4.2.2 Extensible Authentication Protocol Transport Layer Security Mechanism (EAP-TLS)

Here we will be using the same method that we used in previous mechanism in order to calculate the cost of exchange messages [10]. In EAP-TLS authentication mechanism the user mobile will request service from STB. The STB will request user mobile id to get authenticated to STB. The user mobile sends id to content provider and to license provider through STB or base station. The license provider receives the client certificate from user mobile. In return the user mobile will receive service certificate from license provider. Then user mobile send session key to license provider and the licenser provider send session key to user mobile. Finally user mobile receive broadcast key and session key from STB.

$$C_{aut\_Msg} = 3C_{M\_STB} + 1C_{M\_CP} + 5C_{M\_LP} \quad (2)$$

Here also we will calculate the cost of exchange message between Mobile user and STB, Mobile user and Content Provider and Mobile user and license provider.

4.2.3 Our Proposed Authentication Mechanism

In our proposed authentication mechanism we will calculate the cost of exchange messages between all entities from figure 2.

$$C_{aut\_Msg} = 3C_{M\_STB} + 1C_{STB\_SP} + 1C_{SP\_LP} + 1C_{LP\_CP} + 2C_{M\_LP} + 1C_{M\_CP} \quad (3)$$

In our proposed authentication mechanism, we will calculate the cost of exchange messages between Mobile user and STB, STB and Service provider, Service provider and License provider, License provider and Content provider, Mobile user and License provider and Mobile user and Content Provider (see figure 2).

Using the cost values we can calculate the authentication cost of Kerberos authentication mechanism, EAP-TLS authentication mechanism and proposed authentication mechanism by computing the partial costs of each step of the authentication mechanism. The execution of the authentication mechanism involves the exchange of messages between the entities. For example, the number of messages exchange between mobile users, STB and so on. The same calculation applies for other entities in other authentication mechanism (See table 2). Therefore, based in these values we compute authentication cost in all mentioned mechanisms [10].

Table 2. Authentication Cost Parameters

Symbol	Description	Value
C_{M_STB}	Mobile to STB	1
C_{STB_SP}	STB to Service Provider	2
C_{SP_LP}	Service Provider to License Provider.	3
C_{LP_CP}	License Provider to Content Provider	4
C_{M_LP}	Mobile to License Provider.	4.5
C_{M_CP}	Mobile to Content Provider	5

In table 3, we are comparing the authentication cost of exchanged messages between existing authentication mechanism and proposed one. Our proposed authentication mechanism exhibits greater performance in terms of authentication cost, compared to the Kerberos and EAP-TLS authentication procedure. This because

Table 3. Estimation of the Authentication Cost of Kerberos, EAP-TLS and Proposed Authentication Mechanism

Kerberos Mechanism	EAP-TLS Mechanism	Proposed Mechanism
$2C_{M\_STB} + 2C_{M\_LP} + 4C_{M\_CP} = 31$	$3C_{M\_STB} + 1C_{M\_CP} + 5C_{M\_LP} = 30.5$	$3C_{M\_STB} + 1C_{STB\_SP} + 1C_{SP\_LP} + 1C_{LP\_CP} + 2C_{M\_LP} + 1C_{M\_CP} = 26$

it includes less security operations and message that are exchanged between the involved entities, compared to Kerberos and EAP-TLS.

5 Numerical Results

Here, we will show the numerical results. First we will show the cost of authentication when we exchange message in the three authentication mechanism. Then we will show the impact of handover latency in Kerberos, EAP-TLS and Proposed authentication mechanism. We define latency as the time it takes a packet to travel from source to destination.

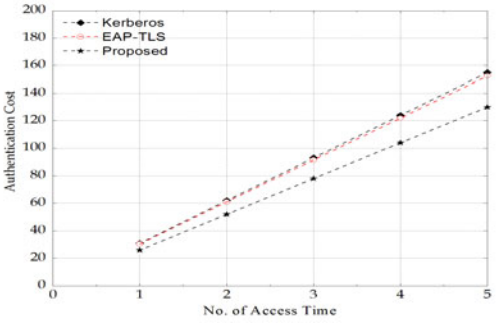


Fig. 3. Accumulated authentication cost vs. no. of time mobile user access the services

The figure 3 presents the accumulate authentication cost of the exchange messages Kerberos, EAP-TLS and Proposed authentication mechanism when user get authenticated to MIPTV services. We assume that the user is authenticated every time consumer gets access to restart MIPTV services after turn TV off. Therefore, we will use the number of access the user attempt and the cost of each authentication mechanism to show the different of authentication cost.

A side of reducing the authentication cost, our proposed authentication mechanism reduces the computational processing and energy cost at the level of mobile devices. For example, we will assume that number of access time the mobile user attempted was 5 times for proposed authentication mechanism and other authentication mechanism. The accumulated authentication cost for proposed authentication mechanism is 70 and for others 90. As a result our proposed authentication mechanism reduces the computational processing and energy cost by 70% compares to others authentication mechanism. Moreover, the reduced number of messages exchange optimizes the usage of the radio resources enhancing the efficiency of user authentication (see figure 3).

6 Conclusions

In this paper, we have presented an efficient migration framework for Mobile IPTV services. The main challenges facing the migration of MIPTV services from a STB or base station to mobile devices are providing mobile user with a fast, secure authentication and a convenient way of accessing, sharing, communicating and receiving MIPTV services. Our proposed mechanism eliminates the number of repeated authentication steps occurred in migration process. This will enhance the performance of user authentication, since proposed mechanism includes less number of security operation and message that are exchange in between entities in each authentication mechanism. We used cryptographic nonce between mobile devices and STB to prevent threats. A CAS used to protect and allow only charged subscriber to view the services. In addition, DRM was used to protect the digital content rights. In the numerical as a result, we compared our mechanism with other based on messages exchange and handover latency. In our proposed mechanism, cost efficiently and handover latency were less than other mechanism which enhances the service quality of real-time applications of mobile users.

Acknowledgment

This work was supported by the IT R&D program of KCC/MKE/KEIT [Contract number: Consolidation 5010-2010-0056-0058, Development of Open-IPTV Platform Technologies for IPTV Convergence Service and Content Sharing].

References

[1] IPTV, http://en.wikipedia.org/wiki/Mobile_IPTV
[2] Mobile IPTV Gets Real, http://www.dailyiptv.com/news/mobile-iptv-reality-030107/
[3] Conditional Access System, http://en.wikipedia.org/wiki/Conditional_access_system
[4] Park, J.H.: Subscriber Authentication Technology of AAA Mechanism for Mobile IPTV Service Offer. Spring Science and Business Media, LLC (2009)
[5] Kerberos (Protocol), http://en.wikipedia.org/wiki/Kerberos_protocol
[6] Denial-of-Service Attack, http://en.wikipedia.org/wiki/Denial-of-service_attack
[7] Man-in-the Middle Attack, http://en.wikipedia.org/wiki/Man_in_the_middle_attack
[8] Pries, R., Yu, W., Fu, X., Zhao, W.: A New Replay Attack against Anonymous Communication Networks
[9] Cryptographic Nonce, http://en.wikipedia.org/wiki/Cryptographic_nonce
[10] Cryptographic and Data Security, http://www.tcs.hut.fi/Studies/T-79.4501/2007AUT/lectures/lecture11.ppt

[11] Gouda, M., Haggag, M.: Enhanced Authentication Mechanism for Next Generation Networks. In: First International Conference on Computational Intelligence, Communication Systems and Networks. IEEE Computer Society, Los Alamitos (2009)
[12] Choi, J., Jung, S., Kim, Y., Yoo, M.: A fast and Efficient Handover Authentication Achieving Conditional Privacy in V2I Networks
[13] Alsaffar, A., Shin, Y., Nguyen, T., Huh, E.: Secure Migration of IPTV Services from a STB to Mobile Devices for Pay per View Video. In: 6th International Conference on Digital Content, Multimedia Technology and its Applications. IEEE Computer Society, Seoul (2010)

Auto-configuration Support for IPv4/IPv6 Translation in Smart Sensor Networks

Huan-wei Lin and Quincy Wu

Dept. of Computer Science and Information Engineering,
National Chi Nan University, Nantou, Taiwan
{s96321017,solomon}@ncnu.edu.tw

Abstract. In this paper, we give an overview of Smart Sensor Networks and illustrate the communication systems between smart sensor networks and IPv4 networks. In order to support the communication between smart sensor networks and current Internet, we introduce some existing IPv4/IPv6 translation mechanisms and point out their shortcomings such as scalability. This paper then proposes an enhancement on the IVI mechanism by adding the IPv6 auto-configuration feature. Simulation shows that the proposed approach works well with thousands of concurrent sessions.

1 Introduction to Smart Sensor Networks

Smart Sensor Networks are interconnected networks for delivering data from sensors to collectors.

There are some characteristics about Smart Sensor Networks [1]:

- Sensing and Measurement
- Integrated Communication
- Improved Interfaces and Decisions Support
- Advanced Components
- Advanced Control Methods

First of all, through intelligent devices, all behaviors can be detected in real time via sensing and measurement. Secondly, it contains an integrated two-way communication network which allows controllers to monitor all sensors in the system and even make some real-time responses through actuators. Therefore, advanced control methods would be designed to handle the whole smart sensor networks through integrated communications.

2 Scalability for Smart Sensor Networks

Scalability is critical for sensor devices in smart sensor networks. In the foreseeable future, there will be billions of sensor devices deployed in smart sensor networks.

Many network engineers propose to choose the Internet Protocol (IP) as a reliable technology to reach the sensors. Because IP-based sensors can connect to the IP-based networks directly without extra intermediate gateways, it is believed that IP is an elegant technology to interconnect new sensor networks with existing Internet seamlessly.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 302–310, 2010.

Since there may be billions of sensor devices, there will not be enough IPv4 addresses for large-scale deployment of sensor networks. Because the number of IPv4 addresses is restricted by its 32-bit addresses length, there can be at most four billion IPv4 addresses. The next generation Internet Protocol, IP Version 6 (IPv6), is the solution for this scenario. It uses 128 bits for its addresses. Also, the Unique Local Address is one of the reasons to use IPv6, which can be used to completely isolate a sensor network from the Internet.

Due to the above reasons, IPv6 is widely used as the protocol for sensors; thus, sensor devices are also considered as a potential killer application for IPv6. However, many existing networks only support the old IPv4 protocol. These IPv4 networks cannot establish direct communication with newly established IPv6 networks. To make a smooth transition from IPv4 to IPv6, it is important to develop a mechanism which allows both IPv4 and IPv6 networks to communicate with each other.

3 Existing Methods for IPv4/IPv6 Translation

Here are some existing solutions for IPv4/IPv6 translation. After research and investigate for these solutions, each of them has some problems to be solved.

3.1 SIIT (Stateless IP/ICMP Translation Algorithm)

IETF RFC 2765 refers to SIIT [2]. SIIT is mainly developed for packet header translation between IPv4 and IPv6. Standard of SIIT translator describes the way of translation between protocols. The translation way is to possibly include all header fields in IPv4 and IPv6, but not option header field in IPv4 and extension header field in IPv6. SIIT also includes bi-directional translation of ICMP message on both IPV4 and IPv6.

This algorithm can only be used as part of the solution for translation between IPv4 and IPv6. SIIT only mentions about the header filed translation. There is no address assignment and routing rules described in SIIT.

3.2 NAT-PT (Network Address Translation-Protocol Translation)

NAT-PT [3] allows IPv6 clients and applications to communicate with IPv4 clients and applications. An NAT-PT server is usually placed between an IPv4 network and an IPv6 network. Every NAT-PT server has some public IPv4 addresses. When connections across IPv4/IPv6 are needed, the NAT-PT server assigns an IPv4 address to an IPv6 client dynamically. In addition to the translation rules in SIIT, a NAT-PT server records the mapping between the IPv4 address and the IPv6 address for each connection to create a mapping table. Since the address mapping is not stateless, the translation will take more time on looking up table.

NAT-PT can be extended to NAPT-PT (Network Address Port Translation-Protocol Translation), which is also defined in RFC 2766. NAPT-PT can map many clients to the same IPv4 address, with different port numbers. This implies that the same IPv4 address can be used to support multiple connections, so that an IPv4 address is sufficient to represent many hosts.

Although NAT-PT is convenient in supporting IPv4/IPv6 translation, it is claimed to be harmful and should not be used in the future [4].

For example, one problem of NAT-PT is the lack of address mapping persistence. After a session is closed, even use the same application and same port to connect to NAT-PT server again, the NAT-PT server may map this connection to another IPv4 address different from last one. To applications which needs address retention, this behavior could cause them unable to work.

3.3 IVI

IVI [5] is similar to the NAT-PT. But it takes a subset of IPv6 address for IPv4 address to map for. So each IPv4 address can be mapped to an individual IPv6 address. SIIT is included in IVI to perform the translation and IVI DNS is also needed for interpretation between IPv4 and IPv6. And ALG [6] (Application Layer Gateway) is implemented for different applications.

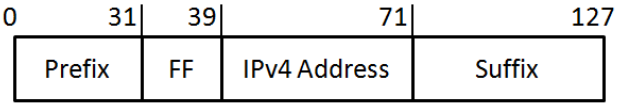


Fig. 1. IVI Address translation

In the address translation part, IVI uses the CERNET implementation to translate the address, as shown in Fig. 1. Bit 0 to bit 31 are the prefix, bit 32 to bit 39 are all ones to indicate that it is an IVI address. Bit 40 to bit 71 are IPv4 address, and the suffix are all zeros. By this implementation, it means IVI is stateless, but IVI abandons an important design of IPv6, the IPv6 auto-configuration. In IVI, it requires the address of an IPv6 client to be configured manually or by DHCPv6. And in this part, the address mapping is one-to-one, one IPv4 address can only be used by an IPv6 client.

Since one-to-one mapping does not solve the problem of lacking of IPv4 address, 1:N IVI [7] is developed. It turns to remap ports, which is described in stateless 1:N IVI. The IVI translator takes charge of port mapping, limit the number of port each client can use, see Fig. 2. It is obviously that with large amount of clients, the port for each client is not enough. For example, when a client is using Google Map, the connections can be dozens or up to a hundred.

Total available port : 65536
Number of Clients : 256
Number of port for each client –
65536 / 256 = 256

Fig. 2. Example of port restriction

An IPv6 client is assigned to No. 10,
and number of clients is 256.
The port numbers for this client can use are
10 + 256 * N, N = 0, 1...255.

Fig. 3. Example of port number assignment

Furthermore, the port number is also restricted, as shown in the example in Fig. 3. IVI assigns specific port number to the Client by the client's serial number, which the Client is assigned to when connects to IVI translator. It is less flexible to use the ports.

The above problem of IVI can be solved by using stateless 1:N IVI. Simply add a gateway between IVI translator and client for port mapping. This gateway receives the port range which IVI translator gives it, and remaps them into ports which clients are using. According to this method, an extra gateway must be added and configured for each client. It is not efficient to widely establish this mechanism.

The method described later in this document can solve the problems which IVI has, and keep it stateless. The most important part is, using IPv6 auto-configuration to assign IPv6 address.

4 SNAT (Stateless Network Address Translation)

To further improve IVI and overcome the inconvenience that IPv6 address has to be configured manually and port number restriction, in this section we shall propose a mechanism called SNAT to support auto-configuration. However, the translation remains stateless so that it can be scalable.

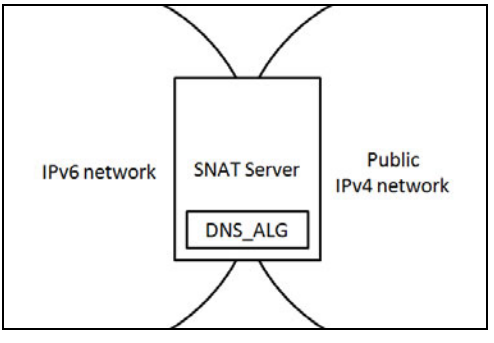


Fig. 4. Network components for translation

4.1 Network Environment

The SNAT server connects to public IPv4 network and IPv6 network. And have static IP address on both sides, details are shown in Fig. 4. Inside the SNAT server, DNS_ALG is implemented to deal with the packet. IPv6 Clients use IPv6 auto-configuration to get IP address.

4.2 Address and Port Translation Mechanism

- **IPv4 to IPv6 address translation**

Base on the address translation method described in NAT-PT, append the IPv4 address to an IPv6 prefix, see Fig. 5. The format is *prefix::<IPv4 address>*. The IPv6 prefix is given by the SNAT server.

- **IPv6 to IPv4 address translation**

Take the MAC address of IPv6 client which is 48 bit long, append the port number to it, see Fig. 6. Then translate the 64-bit data into 16-bit by using a 64 to 16 hash function. Create a hash table to save the 64-bit data and use the 16-bit data for index. The detail of hash table is shown in Table 1. If there is no collision happened, the mapping is unique.

IPv4 address = 163.22.2.1
IPv6 prefix = 2001:e10:6840:20::/64
Translated IPv6 address = 2001:e10:6840:20::163.22.2.1

Fig. 5. IPv4 to IPv6 address translation

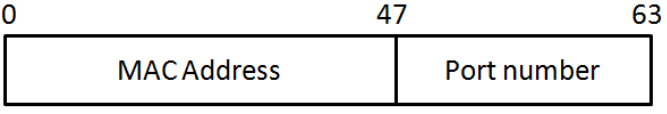


Fig. 6. IPv6 Client data format

Table 1. Hash table for port mapping

Table Index (16bit)	Table Content (64bit)
IPv4 port number	IPv6 Client MAC address + IPv6 Client port number

4.3 Network-layer Header Translation

IPv4 and IPv6 are different protocols, the format of header on Network-layer are different. So the header must be translated, through the method defined in SIIT. Details are shown in Table 2 and Table 3.

Table 2. IPv4 to IPv6 header translation

IPv4 Field	Translated to IPv6 Header
Version (0x4)	Version (0x6)
IHL	discarded
Type of Service	discarded
Total Length	Payload Length = Total Length – 20
Identification	discarded
Flags	discarded
Offset	discarded
Time to Live	Hop Limit
Protocol	Next Header
Header Checksum	discarded
Source Address	SNAT address mapping
Destination Address	SNAT address mapping
Options	discarded

Table 3. IPv6 to IPv4 header translation

IPv6 Field	Translated to IPv4 Header
Version (0x6)	Version (0x4)
Traffic Class	Discarded
Flow Label	Discarded
Payload Length	Total Length = Payload Length + 20
Next Header	Protocol
Hop Limit	TTL
Source Address	SNAT address mapping
Destination Address	SNAT address mapping
-	IHL = 5
-	Header Checksum recalculated

4.4 Details of Connection

- **IPv6 communicate with IPv6 and IPv4 communicate with IPv4**

  The SNAT server simply forwards the packet, acts like an ordinary router.

- **IPv6 communicate with IPv4**

See Fig. 7. The details are as the steps below.

a. At first, IPv6 address corresponds to the IPv4 Host is unknown to the IPv6 Client. In order to know it, IPv6 Client sends a DNS request (type AAAA) to the IPv6 DNS Server.

b. There is no such record for this IPv4 Host in IPv6 DNS Server. So the IPv6 DNS Server forwards this request to the SNAT server.

c. DNS_ALG in the SNAT server deals with the request, changes the request type from AAAA to A and sends it to the IPv4 DNS Server. The IPv4 DNS Server receives this request and returns the A record of IPv4 Host's IPv4 address to the SNAT server.

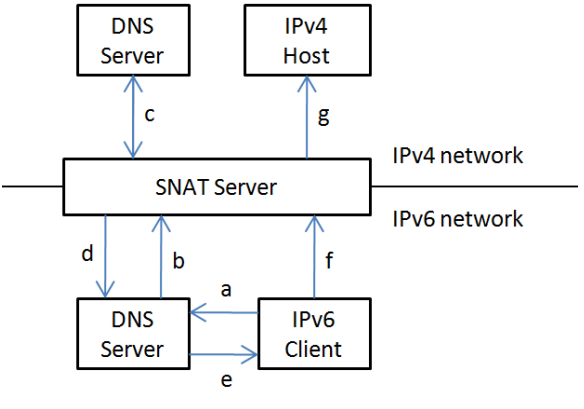


Fig. 7. IPv6 Client connect to IPv4 Host

d. The SNAT server translates the A record to correspond AAAA record and sends it back to IPv6 DNS Server.

e. The IPv6 DNS Server sends the AAAA record to IPv6 Client.

f. IPv6 Client connects to the SNAT server in order to reach IPv4 Host.

g. The connection from IPv6 to IPv4 is established.

- **IPv4 establish connection with IPv6**

See Fig. 8. The details are as the steps below.

a. IPv6 Host sends a STUN [8] request through the port which it wants to use.

b. The SNAT server then translates the packet and sends it to the STUN server with its IP address as source address.

c. STUN server then return packet with the IP address and port number this request used.

d. IPv6 Host receives the IPv4 public address and port number which STUN server tells it.

e. IPv6 Host uses the information received in step d to send a SRV [9] register to the DNS server in IPv4 network.

f. The SNAT server sends this register packet to the DNS server.

g. Now the IPv4 Client can ask DNS server for the name of IPv6 Host and get its IPv4 address and port number.

h. The IPv4 Client connects to the SNAT server, which the address and port number were registered on DNS Server.

i. The connection from IPv4 to IPv6 is established.

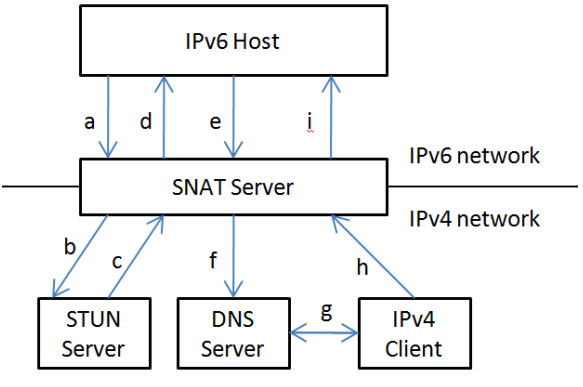


Fig. 8. IPv4 Client connect to IPv6 Host

5 Performance Evaluation

Since hash function is used for translation, there is possibility of collision to be happened. Here is a test for collisions when dealing with many connections.

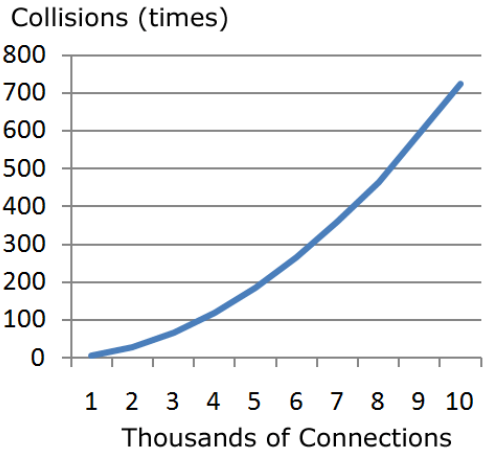


Fig. 9. Collisions when creating hash table

The test is based on 10 to 100 connections per client, and there are 100 clients, the result is shown in Fig. 9.

As the result, when number of total connections reaches ten thousand, collisions happened around seven hundred times. The collision rate is about 7%. In order to handle the collisions, when collision occurs, the SNAT server will add one to the current index and see if collision still happen. The SNAT server will continue doing this until there is no collision.

6 Conclusion

Smart Sensor Networks which uses IPv6 can communicate with IPv4 network through the SNAT server proposed in this paper. The mechanism increases the scalability of Smart Sensor Networks and can also be used for other IPv6 networks.

SNAT makes improvement based on NAT-PT, providing a mechanism to allow single IPv4 address shared by multiple IPv6 clients to communicate with IPv4 networks. Unlike the port restriction in IVI, each port of IPv6 clients behind an SNAT server can be mapped to an individual IPv4 port, and the number of ports used by each client is unlimited.

SNAT solves problems caused by previous solutions have, and makes the translation mechanism more integrated. The issue of security for IPv4/IPv6 translation is not mentioned in this document, which will certainly require further research.

Acknowledgment

This work is partly supported by National Science Council in Taiwan under grants NSC 99-2815-C-260-019-E and NSC 98-2218-E-029-004.

References

1. Pullins, S.: Key Technologies for a Modern Grid, October 10 (2006), http://www.smartgridnews.com/artman/publish/article_172.html
2. Nordmark, E.: Stateless IP/ICMP Translation Algorithm (SIIT), RFC 2765 (February 2000)
3. Tsirtsis, G., Srisuresh, P.: Network Address Translation - Protocol Translation (NAT-PT), RFC 2766 (February 2000)
4. Aoun, C., Davies, E.: Reasons to Move the Network Address Translator - Protocol Translator (NAT-PT) to Historic Status, RFC 4966 (July 2007)
5. Li, X., Bao, C., Chen, M., Zhang, H., Wu, J.: The CERNET IVI Translation Design and Deployment for the IPv4/IPv6 Coexistence and Transition, Internet Draft (Work in progress) (January 2010), http://tools.ietf.org/html/draft-xli-behave-ivi-07
6. Srisuresh, P., Tsirtsis, G., Akkiraju, P., Heffernan, A.: DNS extensions to Network Address Translators (DNS_ALG), RFC 2694 (September 1999)
7. Li, X., Bao, C., Chen, M., Zhang, H.: Address-sharing stateless double IVI, Internet Draft (Work in progress) (October 2009), http://tools.ietf.org/html/draft-xli-behave-divi-01
8. Rosenberg, J., Mahy, R., Matthews, P., Wing, D.: Session Traversal Utilities for NAT (STUN), RFC 5389 (October 2008)
9. Gulbrandsen, A., Vixie, P., Esibov, L.: A DNS RR for specifying the location of services (DNS SRV), RFC 2782 (February 2000)

Empirical Bayes Estimation of Customers' Guarantee Time Length of Loyalty

Hui-Hsin Huang

Department of Business Administration, Aletheia University, Taiwan
886-2- 2621-2121-5504
hoyasophia@gmail.com

Abstract. In this research we use soft computing methods and apply an empirical Bayes method to estimate the minimum customer alive duration. It is an important topic because the information of minimum length provides marketing decision maker to know the largest lower bound of which the customer will be alive. In this paper, we call this minimum duration the guarantee length of loyalty which means the value of each individual customer alive duration will be larger than or equal to this minimum length. This estimate can be used to help finding the best marketing timing for the extended of customer alive time. The model under consideration is based on a Bayes framework which is very flexible (general) so that many complicated factors that involve in marketing problem can be included in this model. In this research an asymptotic optimal empirical Bayes estimate will be derived. As the result, this model will be more practical in real situation.

Keywords: Bayes, duration, loyalty.

1 Introduction

During the Financial downturn period, the decline of consumption leads to recession of firms' profit. The marketing strategy decisions of corporations tend to become sparing. Therefore how to keep the loyal customer and to maintain the baseline revenue are important issues (Bolton, 1998; Reinartz & Kumar, 2000, 2003)[1][2][3]. Thus if we provide a model which estimate the minimum duration of customer patronage a brand or product, that is in this period of time customer doesn't switch to other brands, then the manager can control the marketing input to meet the expectation of profit output . Also, the corporation may consider tactics depending on different minimum duration of customer segments which produce different contribution to the firm.

Thus, the purpose of this paper is to use the stochastic model through soft computing methods to estimate the minimum alive duration of a customer of the brand which is called the guarantee length of loyalty (GLL). In this guarantee duration, the customer is active or alive which means in the duration length, the individual has not taken any business or patronage elsewhere. The minimum GLL of brand customer is an important information for the manager to make the marketing decision. In this paper, we consider

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 311–316, 2010.

a Bayes exponential model such that the prior of the exponential mean is very flexible so that the complicated situations that several untraceable factors are involved can the treated. Considering a square lose, we derive an empirical Bayes estimates which matches the Bayes estimate when sample size is large under the situation that prior is unknown.

The model of a customers' loyalty duration

We consider *X* as the loyal duration which is the total length of the customer using this brand. It is reasonable to consider *X* following a two-parameter exponential family with location and scale parameters:

$$f\left(x\middle|\lambda,\theta\right)=\frac{1}{\lambda}\exp\left(\frac{\left(x-\theta\right)}{\lambda}\right)I\left(x-\theta\right). \tag{1}$$

When $I(y)$ is the indicator function taking value 0 if $y \le 0$ and 1 otherwise. Since *X* $>\theta$ with probability 1, we call θ the guarantee length of loyalty (GLL). Now due to different demographic and individual characters among customers, θ would not be a constant. Therefore we consider θ as an unknown variable with pdf $g(\theta)$ .Here the $g(\theta)$ is unknown but satisfies some mild condition given by **(A1)** and **(A2).**

However for the scale parameter λ , we consider it as an unknown constant. To have information about θ , we may either estimate its mean or estimate its value when n samples are obtained. Since θ is random variable, our model is under a Bayes framework. According λ is assumed unknown and $g(\theta)$ is also unknown with some mild conditions, the GLL model includes a quite big family and can meet the particle application. The prior density $g(\cdot)$ satisfies the following two considerations:

(A1) $g(\cdot)$ is decreasing in $\theta > 0$ and $g(\theta) = 0$ for $\theta > b$. For some known value b, $0< b \le \theta$.

(A2) $g(\cdot)$ is (*r*-1) times differentiable, $g^{(r-1)}(\theta)$ is continuous on [0, *b*].

In (A1) assumption, b is the upper bound of customer alive. Beyond the upper bound, the probability of GLL is zero. The $g(\cdot)$ is a decreasing function which is indicated that the probability of GLL will be less by time increasing. In (A2) assumption, the probability density of GLL is a smooth curve.

According to $G(\theta)$ being satisfied (A1) and (A2)assumption, the Bayes exponential distribution is a huge family. In this Bayes framework, many complicated factors that involve in marketing problem can be included in this model.

In the area of Mathematical Statistics, some papers have studied this topic（Singn & Prasad, 1989; Prasad & Singh,1990）[4][5]. But in these literatures, the parameter λ of formula (1) is assumed known. This proposition is not practical especially in our study. Therefore we consider λ is an unknown parameter. Based on asymptotic optimal empirical Bayes method, this paper estimate the unknow piror distribution $G(\cdot)$ to create the estimating function and use the last sample to calculate θ . This paper

provide a best empirical Bayes estimators of θ in the square loss function. The asymptotic optimal empirical Bayes estimator $\hat{\theta}_G$ is indicated that there is a sequence $\{a_n\}$ if $\hat{\theta}_G$ is corresponding to a Bayes estimator of unkown prior distribution *G*, then

$$R\left(\hat{\theta}_{G^*}\right) - R\left(\hat{\theta}_G\right) = \mathrm{O}(a_n). \tag{2}$$

And $a_n \to 0$, $R(\hat{\theta}_G)$ is the Byes risk of $\hat{\theta}_G$.

2 Sampling

The samples we use are drawn from the customer database of credit card corporation. For the credit card marketing, if there is no record of consumption of the customer over six months, we call this customer the "death" one. According this, the datas we apply are the duration of customer transaction from the first time with this credit card to the end of death. This paper considers positive integer m and n. There are n stages in which we sample the data and in every stage, we draw m numbers samples. For example, X_{ij} is the *j* the sample in *i* stage. $X_{ij} = 7$ is *j* the customer drawn on *i* stage and he switches to other brand after using this brand seven months.

3 Estimate

According the Baye exponential mode, the parameters λ and *G* of θ are unknown. The θ is changing by *G* on different n stages. There is less limitation of *G* in this paper. Only, *G* must be satisfied the assumptions (A1) (A2). In the practice application, we can get the information of *G* depend on sampling.

Considering $K(\cdot)$ is a finite function on [0,1] which is satisfied:

$$\int_0^1 x^i k(x)dx = \begin{cases} 1 & if\ i = 0 \\ 0 & if\ i = 1,2,...,r-1 \end{cases}. \tag{3}$$

Where *r* is given in advance. In this paper we prefixed *r* = 3.

There are many ways to get the value of kernel function $K(\cdot)$.For example, if we give *r*=3, then $K(x) = a_0 + a_1 x + a_2 x^2$. Following the (3), we can get three simultaneous equations. By solving the equations, a_0, a_1, a_2 are obtained. And we consider $k_0 = \max_{0 \le x \le 1} |k(x)|$.

According (A1) assumption, the researcher can decide the domain although $g(\theta)$ is unknown. If the researcher think the variance of θ is large, the value of b can be considered larger. For example, in the durable product market, the duration of loyalty is longer, then we can create b value for five years.

Given

$$Y_i = \min(x_{i1}, x_{i2}, ..., x_{im})$$

and

$$W_i = \sum_{j=1}^{m} x_{ij} - mY_i$$

For the given values of n and *r*, we consider $h(n) = n^{\frac{-1}{(2r+1)}}$ in which *h* is a constant.

Given

$$\beta_n = \frac{(W_1 + W_2 + ...W_n)}{n(m-1)}$$

We propose the estimators as following:

$$\alpha_n(y) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(\frac{-m(y-Y_j)}{\beta_n}\right) I\left(y - Y_j\right). \tag{4}$$

$$f_n(y) = \frac{1}{nh(n)}\sum_{j=1}^{n} k\left(\frac{Y_j - y}{h(n)}\right). \tag{5}$$

$I(x)$ is an indicated function and *h* is the function of n. When $n \to \infty$, then $h \to 0$ ◦ If the total sample sizes are $m \times n$, we can use these numbers of samples to estimate the parameters or functions which are unknown in the models.

We present that $f_i(y|\theta_i, \lambda)$ is the function of Y_i in *i* the stage.

Then,

$$f_i(y|\theta_i, \lambda) = \frac{m}{\lambda}\exp\left(\frac{-m(y-\theta_i)}{\lambda}\right) I(y - \theta_i)$$

θ_i is varying by $G(\cdot)$.

Using the square error loss, the Bayes estimator $\varphi_G(y)$ is the posterior mean of θ .

$$\varphi_G(y) = y - \frac{\alpha_G(y|\lambda)}{f_G(y|\lambda)}$$

with

$$f_G(y|\lambda) = \frac{m}{\lambda}\int_0^{\infty} \exp\left(\frac{-m(y-\theta)}{\lambda}\right) I(y - \theta) dG(\theta). \tag{6}$$

And

$$\alpha_G(y|\lambda) = \int_0^{y} \exp\left(\frac{-m(y-\theta)}{\lambda}\right) d\, F_G(\theta|\lambda). \tag{7}$$

Because the $G(\theta)$ is unknown, we use equations (5) and (4) to estimate equations (6) and (7).

When the researcher draw *mn* samples, the values of $\alpha_n(y)$, $f_n(y)$ which are define in (4) and(5) can be calculated. The *mn* samples are called past data. They are conducted to estimate the unknown value of the model based on empirical Bayes method. Finally, it's needed to get *m* samples to become the samples of (*n*+1) stage for the estimation decision.

We consider the value $Y_{n+1} = y$ which is calculated from the samples in (*n*+1) stage. Thus, we propose the empirical Bayes estimator of θ is

$$\varphi_n^*(y) = \begin{cases} y - \left[\left(\dfrac{\alpha_n(y)}{f_n(y)} \vee 0 \right) \wedge y \right] & \text{if } 0 < y \le b \\ \varphi_n^*(b) & \text{if } y > b \end{cases} . \tag{8}$$

$$a \vee b = \max(a,b) \quad a \wedge b = \min(a,b)$$

Finally, we can obtain the estimator value of GLL from equation (8).

4 Data Analysis

The samples we use are drawn from the customer database of credit card corporation. For the credit card, if there is no record of consumption of the customer over six mounths, we call this customer "dead". According this, the data we apply are the duration of customer from the first transaction with this credit card to cancel the card. First, we discard the extreme value of duration and let t denote the total number of data. Secondly, we prefixed n and m ($n>m$) so that t is approximately close to $(n+1)m$ with $t \leqq (n+1)m$. There are n+1 subsamples and each subsample size is m. Third, we use the former *n* stages subsamples to estimate the unknown *G* and the last one stage subsamples to make the decision. Finally, use these GLL result to marketing application such as deciding the best promotion time.

5 Conclusion

The GLL estimation result can be compared to the customer demographic variables and make the cross analysis to explore the customer cluster rule. Also, we can combine the GLL value and customer lifetime value to calculate the minimum profit of the firm and distinguish the different between the actual performs and the estimation.

References

1. Bolton, N.R.: A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider The Role of Satisfaction. Marketing Scinece 17(1), 45–65 (1998)
2. Reinartz, J.W., Kumar, V.: On the Profitability of Long-life Customers in a Noncontractual Setting: An empirical Investigation and Implications for Marketing. J. Marketing 64(4), 7, 7–99 (2000)

3. Reinartz, J.W., Kumar, V.: The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. J. Marketing 67(1), 17–35 (2003)
4. Singh, R.S., Prasad, B.: Uniformly Strongly Consistent Prior-Distribution and Empirical Bayes Estimatiors with Asymoptotic Optimality and Rates in a Nonexponential Family. Sankhya 51(3), 334–342 (1989)
5. Prasad, B., Singh, R.S.: Estimation of Prior Distribution and Empirical Bayes Estimation in a Nonexponential Family. J. Statistic Planning Inference 24(1), 81–89 (1990)

An Integrated EPQ Model for Manufacturer's Replenishment Policies with Two Levels of Trade Credit Policy under Supplier Credits Linked to Ordering Quantity

Liang-Ho Chen[1,*], Jyh-Woei Chou[1], and Tien-Tsai Huang[2]

[1] Department of Industrial Engineering and Management, Tungnan University, ShenKeng, Taipei Hsien, 22202, Taiwan, R.O.C.
lhechen@mail.tnu.edu.tw
[2] Department of Industrial Management, Lunghwa University of Science and Technology, Toyuan, 333 Taiwan, R.O.C.

Abstract. In this paper, we not only develop an integrated EPQ model for manufacture's replenishment policies with two level of trade credit policy under supplier credits linked to ordering quantity, but also extend Teng and Chang [11] EPQ model more generalization in inventory system. We then provide the proper theoretical results to obtain the optimal solution. Finally, numerical examples are used to illustrate the proposed model and its optimal solution.

Keywords: Inventory; EPQ; Trade Credits; Permissible delay in payments.

1 Introduction

In the past decades, there were many researchers who have studied the inventory models with permissible delay in payments. As a matter of fact, in the business market, the supplier often provided a credit period to customers in order to stimulate the demand or decrease inventories of certain items. And the customers did not have to pay the supplier immediately after receiving the items, but instead, could delay their payment until the end of the allowed period. However, the supplier usually is willing to provide the retailer a permissible delay of payments if the retailer orders a large quantity.

In the literature, Goyal [3] first developed an economic order quantity (EOQ) model under the conditions of permissible delay in payments. Aggarwal and Jaggi [1] extended Goyal's model to the case of deterioration. Jamal et al. [8] analyzed Aggarwal and Jaggi's model to allow for shortages. Teng [9] amended Goyal's model by considering the difference between unit price and unit cost and established an easy analytical closed-form solution to the problem. Chung and Huang [2] proposed an economic production quantity (EPQ) inventory model for a retailer when the supplier offers a

* Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 317–326, 2010.

permissible delay in payments by assuming that the selling price is the same as the purchase cost. Huang [5] extended Goyal's model to develop an EOQ model in which supplier offers the retailer the permissible delay period M (*i.e.*, the supplier trade credit), and the retailer in turn provides the trade credit period N (with $N \leq M$) to its customers (*i.e.*, the retailer trade credit). Huang [6] incorporated both Chung and Huang [2] and Huang [5] to investigate the optimal retailer's replenishment decisions with two levels of trade credit policy in the EPQ framework. Jaggi et al. [7] incorporated the concept of credit-linked demand and developed an inventory model under two levels of trade credit policy to determine the optimal credit as well as replenishment policy jointly for the retailer. Ho et al. [4] formulated an integrated supplier-buyer inventory model with the assumption that the market demand is sensitive to the retail price and the supplier adopts a trade credit policy to determine the optimal pricing, shipment and payment policy. Teng and Chang [11] extended Hung [6] EPQ model and completed the shortcoming but also relax some dispensable assumptions on his model. They then established the theoretical results to obtain the optimal solution.

In this paper, an integrated EPQ model for manufacturer's replenishment policies with two levels of trade credit policy under supplier credits linked to ordering quantity is built. We then not only develop the proper theoretical results to obtain the optimal solution, but also extend Teng and Chang [11] EPQ model. Finally, numerical examples are used to illustrate the proposed model and its optimal solution.

2 Mathematical Formulation

The following notation are partially same as those in Teng and Chang 's [11] EPQ model.

D	the demand rate per year
P	the replenishment rate (*i.e.*, production rate) per year, $P \geq D$
A	the ordering (or set-up) cost per order (lot)
ρ	$1-\frac{D}{P} \geq 0$, the fraction of no production
c	the unit purchasing price
s	the unit selling price, $s \geq c$
h	the unit stock holding cost per item per year excluding interest charges
I_e	the interest earned per dollar per year
I_k	the interest charged per dollar in stocks per year by the supplier
Q_r	the minimum order quantity at which the delay in payments is permitted
T_r	the time interval that Q_r units are depleted to zero due to demand
M	the manufacturer's trade credit period offered by supplier in years
N	the customer's trade credit period offered by manufacturer in years
T	the cycle time in years
$TVC(T)$	the annual total relevant cost, which is a function of T
T^*	the optimal cycle time of $TVC(T)$
Q^*	the optimal lot size of $TVC(T)$

The total relevant cost consists of (a) cost of setup, (b) cost of purchased units, (c) cost of carrying inventory (excluding interest charges), (d) cost of interest charges for unsold items at the initial time or after the permissible delay *M*, and (e) interest earned from sales revenue during the permissible period.

In addition, the following assumptions are used throughout this paper.

(1) The replenishment rate (or production rate), *P*, is known and constant.
(2) Demand rate, *D*, is known and constant.
(3) Shortages are not allowed.
(4) If the order quantity is less than Q_r, then the payment for the items received must be made immediately. Otherwise, if the order quantity is greater than or equal to Q_r, then the delay in payments up to *M* is permitted. That is, If $Q \geq Q_r$, *i.e.* $T \geq Q_r / D = T_r$, the delayed payment is permitted. Otherwise, the delay in payments is not permitted.(i.e., setting $M = 0$).
(5) During the credit period the account is not settled, the manufacturer (or retailer) can accumulate revenue and earn interest after his/her customer pays for the amount of purchasing cost to the manufacturer (or retailer) until the end of the trade credit period offered by the supplier. That is, the manufacturer (or retailer) can accumulate revenue and earn interest during the period *N* to *M* with rate I_e under the condition of trade credit. And at the end of the permissible delay, the manufacturer (or retailer) pays off all units ordered, and starts paying for the interest charges on the items in stocks.
(6) The ending inventory is zero.
(7) Time horizon is infinite.

The annual total relevant cost consists of the following elements.

1. Annual ordering cost = A / T .
2. Annual stock holding cost = $hDT\rho / 2$.

According to the assumption (4), as well as the values of *N* and *M*, there are two cases for the manufacturer, case A: $N \leq M$ (see, figures 1 and 2) and case B: $M < N$ (see, figures 3 and 4), to occur in interest charged and interest earned per year.

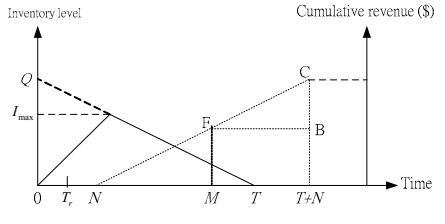


Fig. 1. Sub-case A-1: $T_r \leq T$ and $N \leq M \leq T + N$

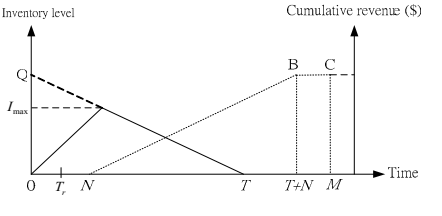


Fig. 2. Sub-case A-2: $T_r \leq T$ and $T + N < M$

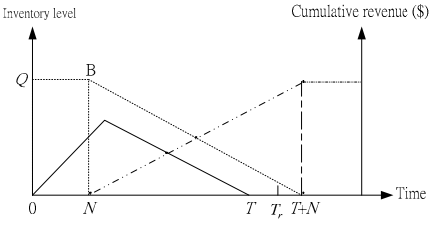


Fig. 3. Sub-case B-1: $T < T_r$

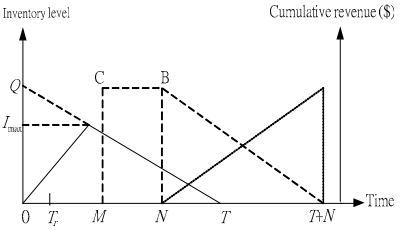


Fig. 4. Sub-case B-2: $T_r \le T$ and $M < N$

2.1 Case A: $N \le M$

In this case, we have following two sub-cases

(i) Sub-case A-1: $T_r \le T$ and $M \le T + N$

In this sub-case, the manufacturer buys all parts at time zero and must pay the purchasing cost at time *M*. Meanwhile, he/she starts to receive the money from his/her first customer at time *N*. Hence, the manufacturer pays off all units sold by *M* – *N* at time *M*, keeps the profits, and starts paying for the interest charges on the items sold after *M* – *N*. The graphical representation of this situation is shown in Figure 1. However, the manufacturer can not payoff the supplier by *M* because the supplier credit period *M* is shorter than the customer last payment time *T* + *N*. Hence, the manufacturer must finance all items sold after time *M* – *N* at an interest charged I_k per dollar per year. The interest charged per cycle is cI_k times the area of the triangle *BCF* shown in Figure 1. Therefore, the interest charged per year is given by

$$\frac{cI_k}{T}\{\frac{D[T+N-M]^2}{2}\}. \tag{1}$$

On the other hand, the manufacturer starts selling products at time 0, but getting the money at time *N*. Consequently, the manufacturer accumulates revenue in an account that earns I_e per dollar per year starting from *N* through *M*. Therefore, the interest earned per cycle is sI_e multiplied by the area of the triangle *NMF* as shown is Figure 1. Hence, the interest earned per year is given by

$$\frac{sI_e}{T}[\frac{D(M-N)^2}{2}]. \tag{2}$$

(ii) Sub-case A-2: $T_r \le T$ and $T + N \le M$

In this sub-case, the manufacturer receives the total revenue at time $T + N$, and is able to pay the supplier the total purchase cost at time *M*. Since the customer last payment time $T + N$ is shorter than the supplier credit period *M*, the manufacturer faces no interest

charged. The interest earned per cycle is sI_e multiplied by the area of the trapezoid on the interval [*N*, *M*] as shown in Figure 2. As a result, the interest earned per year is given by

$$\frac{sI_e}{T}[\frac{DT^2}{2} + DT(M-T-N)]. \tag{3}$$

Therefore, the annual total relevant cost for the manufacturer can be expressed as

$$TVC(T) = \begin{cases} TVC_1(T) & , Max\{T_r, M-N\} \le T, \\ TVC_2(T) & , T_r \le T \le M-N, \end{cases} \tag{4}$$

where

$$TVC_1(T) = \frac{A}{T} + \frac{hDT\rho}{2} + \frac{cI_k}{T}\{\frac{D[T+N-M]^2}{2}\} - \frac{sI_e}{T}[\frac{D(M-N)^2}{2}] \ . \tag{5}$$

$$TVC_2(T) = \frac{A}{T} + \frac{hDT\rho}{2} - \frac{sI_e}{T}[\frac{DT^2}{2} + DT(M-T-N)]. \tag{6}$$

2.2 Case B: $M \le N$

In this case, we consider two sub-cases:

(i) Sub-case B-1: $T < T_r$

Since $T_r > T$ implies the order quantity is less than Q_r, then the payment for the items received must be made immediately. Hence, the manufacturer must finance all items ordered at initial time 0 at an interest charged I_k per dollar per year, and start to payoff the loan after time *N*. Hence, the interest charged per cycle is cI_k multiplied by the area of the trapezoid on the interval [0, *T*+*N*], as shown in Figure 3. Therefore, the interest charged per year is given by

$$\frac{cI_k}{T}[NDT + \frac{DT^2}{2}]. \tag{7}$$

(ii) Sub-case B-2: $T_r \le T$

In this situation, the customer's trade credit period *N* is larger than or equal to the supplier credit period *M*. Consequently, there is no interest earned for the manufacturer. In addition, the manufacturer must finance all items ordered at time *M* at an interest charged I_k per dollar per year, and start to payoff the loan after time *N*. Hence, the interest charged per cycle is cI_k multiplied by the area of the trapezoid on the interval [*M*, *T*+*N*], as shown in Figure 4. Therefore, the interest charged per year is given by

$$\frac{cI_k}{T}[(N-M)DT + \frac{DT^2}{2}]. \tag{8}$$

Therefore, the annual total relevant cost for the manufacturer can be expressed as

$$TVC(T) = \begin{cases} TVC_3(T) & , T < T_r , \\ TVC_4(T) & , T_r \le T , \end{cases} \tag{9}$$

where

$$TVC_3(T) = \frac{A}{T} + \frac{hDT\rho}{2} + \frac{cI_k}{T}[\,DTN + \frac{DT^2}{2}]\,. \tag{10}$$

$$TVC_4(T) = \frac{A}{T} + \frac{hDT\rho}{2} + \frac{cI_k}{T}[\,\frac{DT^2}{2} + DT(N-M)\,]\,. \tag{11}$$

Remark 1. To ensure that the annual total relevant cost for the manufacturer in each sub-cases, (*i.e.*, the Equations (5), (6), (10) and (11)) are larger than or equal to zero, we have the following scenarios.

(1) $TVC_1(T) \ge 0$, if and only if

$D[cI_k(M-N)]^2 - 2A(h\rho + cI_k) - (h\rho + cI_k)(cI_k - sI_e)D(M-N)^2 \le 0$.

(2) $TVC_2(T) \ge 0$, if and only if $D[sI_e(M-N)]^2 - 2A(h\rho + sI_e) \le 0$.

(3) $TVC_3(T) \ge 0$, if and only if

$(h\rho + cI_k)(sI_e)D(M-N)^2 - 2A(h\rho + cI_k) - (h\rho)cI_kD(M-N)^2 \le 0$

(4) $TVC_4(T) \ge 0$, if and only if $D[cI_k(N-M)]^2 - [(h\rho + cI_k)](2A) \le 0$.

Remark 2. (Special case) If $T_r = 0$ which means that the supplier provides the permissible delay in payments without any condition on order quantity to the retailer, then our model will degenerate to the model proposed by Teng and Chang [11].

Optimal solution

To minimize the annual total relevant cost, taking the first-order and the second-order derivatives of $TVC_1(T)$, $TVC_2(T)$, $TVC_3(T)$, and $TVC_4(T)$ with respect to T, we obtain

$$\frac{d\,TVC_1(T)}{d\,T} = \frac{-[2A + D(cI_k - sI_e)(M-N)^2]}{2T^2} + \frac{(h\rho + cI_k)D}{2}, \tag{12}$$

$$\frac{d^2\,TVC_1(T)}{d\,T^2} = \frac{2A + D(cI_k - sI_e)(M-N)^2}{2T^3}, \tag{13}$$

$$\frac{d\,TVC_2(T)}{d\,T} = \frac{-A}{T^2} + \frac{(h\rho + sI_e)D}{2}, \tag{14}$$

$$\frac{d^2\, TVC_2(T)}{d\, T^2} = \frac{2A}{T^3} > 0, \tag{15}$$

$$\frac{d\, TVC_3(T)}{d\, T} = \frac{-A}{T^2} + \frac{(h\rho + cI_k)D}{2}, \tag{16}$$

$$\frac{d^2\, TVC_3(T)}{d\, T^2} = \frac{2A}{T^3} > 0, \tag{17}$$

$$\frac{d\, TVC_4(T)}{d\, T} = \frac{-A}{T^2} + \frac{(h\rho + cI_k)D}{2}, \tag{18}$$

and

$$\frac{d^2\, TVC_4(T)}{d\, T^2} = \frac{2A}{T^3} > 0\,. \tag{19}$$

For convenience, we set T_1^*, T_2^*, T_3^* and T_4^* be the optimal values for $TVC_1(T)$, $TVC_2(T)$, $TVC_3(T)$, and $TVC_4(T)$, respectively. That is, T_1^*, T_2^*, T_3^* and T_4^* satisfy the following equations:

$TVC_1(T_1^*) = \underset{k \le T}{Min} TVC_1(T)$, $TVC_2(T_2^*) = \underset{T_r \le T \le M-N}{Min} TVC_2(T)$,

$TVC_3(T_3^*) = \underset{T \le T_r}{Min} TVC_3(T)$, and

$TVC_4(T_4^*) = \underset{T_r \le T}{Min} TVC_4(T)$, respectively, where $k \equiv Max\{T_r, M-N\}$.

Theorem 1. When $N \le M$,

(A-1.1) If $\delta_1 = 2A + D(c\,I_k - s\,I_e)(M-N)^2 > 0$, then $TVC_1(T)$ is convex function for $T > 0$, hence, let $T_1 = \sqrt{\delta_1 / [(h\rho + cI_k)D]}$ satisfy $\frac{dTVC_1(T_1)}{dT} = 0$.

(1.a) If $Max\{T_r, M-N\} \le T_1$, then the optimal solution of $TVC_1(T)$ is $T_1^* = T_1$.

(1.b) If $T_1 < Max\{T_r, M-N\}$, then the optimal solution of $TVC_1(T)$ is $T_1^* = Max\{T_r, M-N\}$.

(A-1.2) If $\delta_1 = 2A + D(c\,I_k - s\,I_e)(M-N)^2 \le 0$, then $TVC_1(T)$ is concave and increasing function for $T > 0$, hence, there exits $T_1^{\Delta} > 0$ which satisfies $F(T_1^{\Delta}) = 0$, where

$F(T) = (h\rho + cI_k)DT^2 - 2\,c\,I_k D(M-N)T + \delta_1$.

(2.a) If $Max\{T_r, M-N\} \le T_1^{\Delta}$, then the optimal solution of $TVC_1(T)$ is $T_1^* = T_1^{\Delta}$.

(2.b) If $T_1^{\Delta} < Max\{T_r, M-N\}$, then the optimal solution of $TVC_1(T)$ is $T_1^* = Max\{T_r, M-N\}$.

(A-2) $TVC_2(T)$ is convex function for $T > 0$, hence, let $T_2 = \sqrt{2A/[(h\rho + sI_e)D]}$ satisfy $\frac{dTVC_2(T_2)}{dT} = 0$.

(a) If $T_r \le T_2 \le M - N$, then the optimal solution of $TVC_2(T)$ is $T_2^* = T_2$

(b)If $T_r \ge T_2$ or $T_2 \ge M - N$, then the optimal solution of $TVC_2(T)$ is $T_2^* = T_r$ or $T_2^* = M - N$.

Proof: Omitted.

Theorem 2. When $M \le N$,

(B-1) $TVC_3(T)$ is convex function for $T > 0$, hence, let $T_3 = \sqrt{2A/[(h\rho + cI_k)D]}$ satisfy

$$\frac{dTVC_3(T_3)}{dT} = 0.$$

(a) If $T_r > T_3$, then the optimal solution of $TVC_3(T)$ is $T_3^* = T_3$.

(b) If $T_r \le T_3$, then the optimal solution of $TVC_3(T)$ is $T_3^* = T_r$.

(B-2) $TVC_4(T)$ is convex function for $T > 0$, hence, let $T_4 = \sqrt{2A/[(h\rho + cI_k)D]}$ satisfy $\frac{dTVC_4(T_4)}{dT} = 0.$

(a) If $T_r > T_4$, then the optimal solution of $TVC_4(T)$ is $T_4^* = T_r$.

(b) If $T_r \le T_4$, then the optimal solution $TVC_4(T)$ is $T_4^* = T_4$.

Proof: Omitted.

3 Numerical Examples

To illustrate our proposed model, we take two examples in the following.

Example 1. We consider the values of the parameters as follows : A = \$150/order, D = 2500 units/year, c=\$50/unit, s=\$75/unit, h=\$15/unit/year, P=3000units/year, $I_k = 0.12$/\$/year, I_e=0.15/\$/year, $Q_r = \{150, 500\}$ units, $M = \{0.10, 0.15\}$ year, and $N = 0.12$ year.

From optimal solution procedures, the optimal replenishment interval, and optimal production quantity are shown in Table 1.

Table 1. Computational results with respect to different values of Q_r, M and N

M	Q_r	N = 0.12		
		T^*	Q^*	TVC^*
0.10	150	T_4^* =0.1188	Q_4^* =297.0	2824.9
	500	T_4^* =0.2000	Q_4^* =500.0	3175.0
0.15	150	T_1^* =0.1165	Q_1^* =291.1	2024.7
	500	T_1^* =0.2000	Q_1^* =500.0	2395.5

Example 2. We consider the values of the parameters as follows : A = \$200/order, D = 2500 units/year, s = \$90/unit, h = \$12/unit/year, P = 5000 units/year, $I_k = 0.15$ /\$/year, c = \$30/unit, I_e = {0.08,0.18}/\$/year, Q_r = {100, 300} units, $M = 0.05$ year, and $N = 0.08$ year. From Theorems 1 and 2, the optimal replenishment interval, and optimal production quantity can be shown in Table 2.

Table 2. Computational results with respect to different values of c, I_e, Q_r and M

c	I_e	Q_r	M = 0.05		
			T^*	Q^*	TVC^*
30	0.08	100	T_4^* =0.1155	Q_4^* =288.7	3801.6
		300	T_4^* =0.1200	Q_4^* =300.0	3084.2
	0.18	100	T_4^* =0.1155	Q_4^* =288.7	3801.6
		300	T_4^* =0.1200	Q_4^* =300.0	3084.2

4 Conclusion

In this paper, an integrated EPQ model with two level trade credit policies under supplier credits linked to the order quantity is developed. We then provide the theorems to find the optimal solutions for the provided models.

For the future study, we can consider both trade credits linked to order quantity for the manufacturer and customers. And we can also discuss the deterioration items in the models, the purchased, holding and deteriorated costs varied with time and /or quantity and so on.

References

1. Aggarwal, S.P., Jaggi, C.K.: Ordering policies of deteriorating items under permissible delay in payment. Journal of the Operational Research Society 46, 658–662 (1995)
2. Chung, K.J., Huang, Y.F.: The optimal cycle time for EPQ inventory model under permissible delay in payments. International Journal of Production Economics 84, 307–318 (2003)
3. Goyal, S.K.: Economic order quantity under conditions of permissible delay in payments. Journal of the Operational Research Society 36, 335–338 (1985)
4. Ho, C.H., Ouyang, L.Y., Su, C.H.: Optimal pricing, shipment and payment policy for an integrated supplier-buyer inventory model with two-part trade credit. European Journal of Operational Research 187, 496–510 (2008)
5. Huang, Y.F.: Optimal retailer's ordering policies in the EOQ model under trade credit financing. Journal of the Operational Research Society 54, 1011–1015 (2003)
6. Huang, Y.F.: Optimal retailer's replenishment decisions in the EPQ model under two levels of trade credit policy. European Journal of Operational Research 176, 1577–1591 (2007)
7. Jaggi, C.K., Goyal, S.K., Goel, S.K.: Retailer's optimal replenishment decisions with credit-linked demand under permissible delay in payments. European Journal of Operational Research 190, 130–135 (2008)
8. Jamal, A.M.M., Sarker, B.R., Wang, S.: An ordering policy for deteriorating items with allowable shortage and permissible delay in payment. Journal of the Operational Research Society 48, 826–833 (1997)
9. Teng, J.T.: On the economic order quantity under conditions of permissible delay in payments. Journal of the Operational Research Society 53, 915–918 (2002)
10. Teng, J.T., Goyal, S.K.: Optimal ordering policies for a retailer in a supply chain with up-stream and down-stream trade credits. Journal of the Operational Research Society 58, 1252–1255 (2007)
11. Teng, J.T., Chang, C.T.: Optimal manufacturer's replenishment policies in the EPQ model under two levels of trade credit policy. European Journal of Operational Research 195, 358–363 (2009)

Modeling a Dynamic Design System Using the Mahalanobis Taguchi System— Two-Step Optimal Algorithm

Tsung-Shin Hsu[3] and Ching-Lien Huang[1,2,*]

[1] Department of Industrial Engineering and Engineering Management
National Tsing Hua University
No. 101, Kuang Fu Road, Sec. 2, Hsinchu, Taiwan, R.O.C.
[2] Department of Industrial Management, Lunghwa University of Science and Technology
No. 300, Sec. 1, Wanshou Rd., Guishan, Taoyuan County 33306, Taiwan, R.O.C.
lynne.line@msa.hinet.net
[3] Department of Industrial Management, National Taiwan University of Science and Technology
No. 43, Sec.4, Keelung Rd., Taipei, 106, Taiwan, R.O.C.

Abstract. This work presents a novel algorithm, the Mahalanobis Taguchi System- Two Step Optimal algorithm (MTS-TSO), which combines the Mahalanobis Taguchi System (MTS) and Two-Step Optimal (TSO) algorithm for parameter selection of product design, and parameter adjustment under the dynamic service industry environments.

From the results of the confirm experiment, a service industry company is adopted to applies in the methodology, we find that the methodology of the MTS-TSO algorithm can easily solves pattern-recognition problems, and is computationally efficient for constructing a model of a system. The MTS-TSO algorithm is good at pattern-recognition and model construction of a dynamic service industry company system.

Keywords: Mahalanobis -Taguchi System (MTS), Data-Mining (DM), Two -steps optimal algorithm (TSO), Service Industry (SI).

1 Introduction

Searching for patterns and modeling via data-mining is typically done in static states. Notably, the Mahalanobis Taguchi System algorithm can effectively and efficiently overcome extraction problems for pattern recognition. The primary aim of the MTS algorithm is to accurately predict multidimensional attributes by constructing a global measure meter that combines Mahalanobis Distance (MD), Orthogonal Arrays (OA) and the Signal-to-Noise (SN) ratio. Studies using a dynamic environment for data-mining are scarce. The two-steps optimal (TSO) algorithm developed by Taguchi constructs a dynamic system to test attributes under dynamic environment. As the

* Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 327–332, 2010.

MTS can only handle pattern recognition problems in static and simple environments, this study proposes a novel method that integrates the TSO into the MTS to generate a new algorithm, the MTS-TSO algorithm. This proposed algorithm solves pattern-recognition and model-construction problems for service industry in a dynamic system.

A review of relevant literature reveals, Kaya, M., *et. al.*(2005) proposed A genetic algorithm (GA) and Fuzzy-based methods are used to construct association rules. Chiang, H. L.,*et. al.*(2005) submitted A Linear Correlation Discovering (LCD) method is utilized for pattern recognition, and Wang, C. Y., *et al.*(2005) presented a three-stage online association rule is utilized to mine context information and information patterns. Simultaneously, Daskalaki, S., *et. al.*(2003) submitted for the system lifecycle evaluation and estimation attributes.

Clearly, the MTS is a good and effective system for pattern searching. Das, P.,*et. al.*(2007) proposed a work shows that the MTS is applied for pattern recognition as follows, an MTS is used to resolve classification problems.

In the field of model construction, Kim, M. J. *et. al.* (2003) offered the GA algorithms have been utilized to generate a bankruptcy prediction model, and Ambrozic, T. (2003) presented an artificial neural network for predicting subsidence. From above mentions, the goal of this work is to solve the dynamic system construction problem. Thus, this study applied the MTS-TSO algorithm to the case of a service industry company to determine whether the model is good.

2 Model Construction

The MTS process; the process contains the following steps.

First, the process of the proposed algorithm establishes the MTS, which computes the Mahalanobis Distance (MD) from two data groups, transfers the MD values into observations y_{ij}, develops the *threshold* based on the minimum values of Type-I and Type-II errors, and to determines whether attributes belong to this group. Next, Using y_{ij} to calculate SN ratios for the attributes, and then selects these attributes to form a new pattern that represents the initial system. The algorithm is as follows.

The TSO algorithm is applied in the second stage. The TSO algorithm attempts to construct a service industry dynamic system, and construct the model, which is completed using by the following two steps. The first step determines whether β is significant. The second step adjusts the value of β. That is, the formula $Y_i = \beta M_i$ is generated and applied to a service trade company dynamic system as the rule of a diagnostic or prediction model. In summary, the algorithmic procedure has the following three steps.

Step 1. Construct a measure meter with the MTS process of a system

First, parameters, Mahalanobis Distance (MD) are computed. Equation 1 is the formula for standardized values:

$$MD = D_i^2 = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{k} a_{ij} y_{il} y_{jl} \ (l = 1 \ldots\ldots n) \tag{1}$$

The process is as follows.

1. Assess attributes from the normal data set.
2. Use the normal data set to calculate MD .

Step 2. Confirm the measure meters of the system

This stage generates some parameters, and *the-larger-the-best* (LBT rule) is adopted. Next, the important attributes of the normal data set are chosen based on *SN* ratios By Eq. 2, *d* is the number of an abnormal data set, $y_i^2 = MD$, and *i* = 1 to *d*.

$$\eta = -10\log_{10} {}^{1}\!/_{d} \left(\frac{1}{y_1^2} + \frac{1}{y_2^2} + ... + \frac{1}{y_d^2}\right) \tag{2}$$

The process at this stage is summarized as follows.

3. Calculate the MD
4. Compute *SN* ratios for the abnormal data set.
5. Determine the *threshold* value.

Step 3. Generate the dynamic model

The third step includes two process details for the TSO algorithm. The first step selects *SN* ratios from the second data group and maximizes these ratios. If the ratio is insignificant, the second step adjusts the value of β . The process is summarized as follows.

6. Confirm the measure meters to form a new pattern..
7. Construct a dynamic service trade company system.

Figure 1 presents the MTS-TSO algorithm.

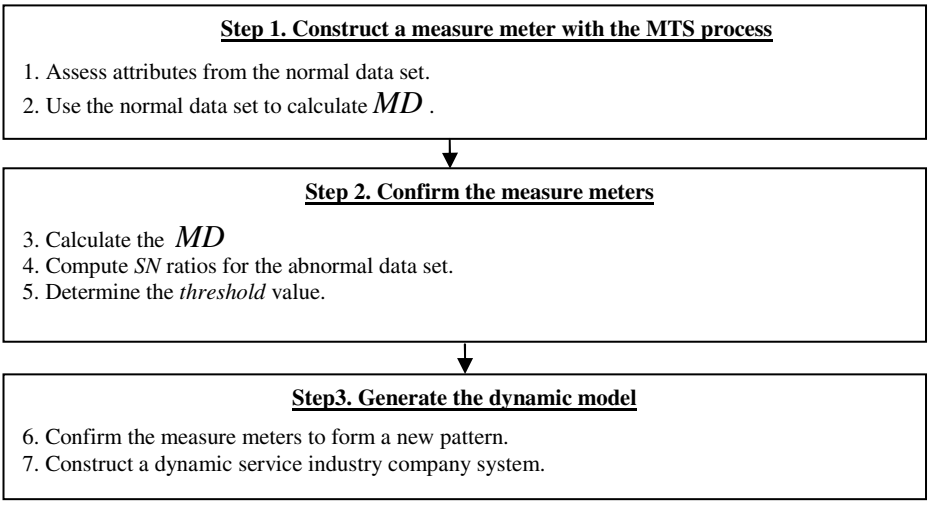


Fig. 1. Summary of the MTS-TSO algorithm

3 Case Application

The case company will be test by the performance measurement units, which has 30 work parameters for testing 32 work items. These attributes are coded from $y_1 - y_{30}$

The 23 data sets are selected from the normal data set and 9 data sets are extracted from the abnormal data set. The MTS-TSO algorithm is applied as follows.

Step 1. Build Measure meters of a system

First, the MD_s are calculated from the normal data set. The threshold is 0.7.

The threshold, 0.7, is determined to separate data sets to two groups. The value of the Type-I error is 0.2; that is, the correct rate is 80%.

Step 2. Confirm the measurement meters

At this stage, the calculated parameter is transformed into a SN ratio by using the L_{32} Orthogonal Array (OA) table. And the effective size of SN ratios is obtained.

Therefore, the attributes are reduced from 30 to 6 items, at the same time, the largest effect sizes among the SN ratios are chosen as follows. These attributes are v_3, v_5, v_{10}, v_{11}, v_{21}, and v_{22}.

Next, to confirm the reliability of model attributes, the expected value of the SN ratios is 1.06, which attains an optimal state. These data are then validated and verified. Thus, the confidence interval (CI), which confirms system reliability—also proposed by Taguchi—is derived using Eq.3.

$$CI_1 = \sqrt{F_{0.05,1,df\,2} \times Ve \times \left(\frac{1}{n_{eff}}\right)} \tag{3}$$

where $F_{\alpha,1,df}$ is the α value of F, α is an obvious level, $1-\alpha$ is a confidence interval, df is the item's of degree of freedom, v_e is the combination error item of variance, and n_{eff} is the number of efficient experimental times, which is shown as follows.

For confirmation reason, the second group is extracted and the experiment is conducted again; the 95% CI must remain within this range; that is, the new structure of the system model is valid.

4 Generate the Dynamic Model

4.1 To Reduce the Number of Attributes

At this stage, the second group is selected and the MTS process is applied again. Thus, the MD_s of second data group are computed (Table1).

Table 1. The MD_s of the second group

Sample	1	2	3	4	5	6	7	8	9	10
MD	0.92	0.93	0.68	0.94	0.80	0.84	0.65	0.80	0.9	0.82
Sample	11	12	13	14	15	16	17	18	19	20
MD	0.95	0.98	0.98	0.86	0.90	0.91	0.92	0.89	0.86	0.89
Sample	21	22	23	24	25	26	27	28	29	30
MD	0.69	0.8	0.92	0.85	0.78	0.79	0.88	0.86	0.87	0.8

Here the *threshold* value is 0.7, which is the smallest value and can be assessed by the type-I error, which is 0.20. The correct rate of the second data group is stay at 80%

4.2 The TSO Algorithm Is Used to Construct a Dynamic Model

In this stage, the TSO algorithm is utilized determine whether the dynamic model is in good condition. The processes are as follows.

Step 1. Maximize the *SN* ratios of the second data group.

The second data group is selected from a system.

To maximize the *SN* ratios of these data, attributes are chosen based on the largest *SN* ratio. These attributes are $v1$, $v2$, $v3$, $v4$, $v5$, and $v6$.

The β is employed to adjust the dynamic model and determine the β level, which is adjusted when it is not equal to 1. The processes are shown in the next step.

Step 2. Adjust the β values

According to the second data group, β is computed using the following formula for a manufacturing inspection dynamic system. Thus, the estimated value of the SN ratios is computed as -1, and listed as follows.

$$\hat{\eta} = \sum_{i=1}^{6} v_i - n\bar{\eta} \tag{4}$$

From Eq. (4), the value of the $\hat{\eta}$ is $\hat{\eta}$ = -1

The β value is 0.8, which is too low to reach 1. Thus, the attributes original value must be adjusted such that its β value is close to 1.

$$\hat{\beta} = \sum_{i=1}^{6} \beta_i - n\bar{\beta} \tag{5}$$

Next, from Eq. (5), the value of the β is $\hat{\beta}$ = 0.8

From the adjusting process, the SN ratios have been improved while β is adjusted to be close to 1.

5 Conclusion

The MTS is employed to process the pattern-recognition problem, and the TSO algorithm is utilized to formulate a dynamic service industry system.

Experimental results indicate that the MTS algorithm easily solves pattern-recognition problems, and is computationally efficient. Additionally, the TSO algorithm is a simple and efficient procedure for constructing a model of a service industry dynamic system. The MTS-TSO algorithm is good at pattern-recognition and model construction of dynamic service industry systems.

We conclude that the MTS-TSO algorithm can successfully be applied to accurately predict decisions in dynamic environments for resolving data mining problems.

References

1. Ambrozic, T., Turk, G.: Prediction of subsidence due to underground mining byarti.cialne neural networks. In: CG, vol. 29, pp. 627–637 (2003)
2. Chang, C.H., Ding, Z.K.: Categorical data visualization and clustering using subjective factors. In: DKE, vol. 53, pp. 243–262 (2005)
3. Chiang, H.L., Huang, C.E., Lim, E.P.: Linear correlation discovery in databases: a data mining approach. In: DKE, vol. 53, pp. 311–337 (2005)
4. Daskalaki, S., Kopanas, I., Goudara, M., Avouris, N.: Data mining for decision support on customer insolvency in telecommunications business. In: EJOR, vol. 145, pp. 239–255 (2003)
5. Das, P., Datta, S.: Exploring the effects of chemical composition in hot rolled steel product using Mahalanobis distance scale under Mahalanobis-Taguchi system. In: CMS, vol. 38, pp. 671–677 (2007)
6. Kaya, M., Alhajj, R.: Genetic algorithm based framework for mining fuzzy association rules. In: FSS, vol. 152, pp. 587–601 (2005)
7. Kim, M.J., Han, I.: The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. In: ESA, vol. 25, pp. 637–646 (2003)
8. Keim, D.A., Kriegel, H.P.: Using visualization to support data-mining of large existing databases. In: Lee, J.P., Grinstein, G.G. (eds.) Visualization-WS 1993. LNCS, vol. 871, pp. 210–229. Springer, Heidelberg (1994)
9. Nicholson, S.: The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. In: IPM, vol. 42, pp. 785–804 (2006)
10. Nasraoui, O., Rojas, C., Cardona, C.: A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation. In: CN, vol. 50, pp. 1488–1512 (2006)
11. Wang, C.Y., Tseng, S.S., Hong, T.P.: Flexible online association rule mining based on multidimensional pattern relations. In: IS, vol. 176, pp. 1752–1780 (2006)

A Fuzzy Model Applied on Assessing Operating Performance of Industrial PC

Tien-Tsai Huang, Su-Yi Huang, and Yi-Huei Chen

No. 300, Sec.1, Wanshou Rd., Gueishan Shiang, Taoyuan County 33306, Taiwan
Department of Industrial Management, Lunghwa University of Science and Technology
normanbb@mail.lhu.edu.tw

Abstract. Taiwan has maintained its competitiveness in the global industrial PC market due to its effective supply chain management. In the operational level, however, integration of corporate resources is required and the development of a growth and profit assessment model can result in the most value for the shareholders, the employees, the society and the nation. This research adopts the fuzzy logic approach to propose an assessment model for evaluating the performances of Taiwan's industrial PC companies. The strength of this approach transforms indistinct language and uncertain cogitation of human decision-making to quantitative data. Qualitative expert surveys and quantitative financial data are collected from Taiwan's industrial PC companies in 2008 to ensure the validity and objectiveness of model assessment. A total of 6 dimensions, including potentiality, capital structure, solvency, corporate performance, profitability, and cash flow, are elicited, and 20 assessment factors are identified to serve as the criteria for assessment.

Keywords: Fuzzy Approach, Industrial PC, Performance.

1 Introduction

The early applications of industrial PCs were limited to automatic production only. With the change of their life pattern and the advent of Internet along with the combination of information, communication, consumer electronics, optoelectronics and semiconductor, the applications of industrial PCs can cover cross-company, cross-area and multinational electronic data interchange. Therefore, industrial PC is defined as any kind of customized computer equipment produced with different types and patterns in different industries to meet customers' total solution in different industries for diverse demands.

Since one of industrial PCs' characteristics is their longer product lifetime, customers put emphasis on after-sales service, quality and product stability. Suppliers or manufacturers need to provide long, stable service, and longer turnover period of stock. Therefore, they are considerably different from consumer computers. Industrial PC industry is one that has high gross profits. In order to maintain profit making and operating growth, the existing manufacturers relocate or integrate related application resources, devoting themselves to discovering more unknown applications. In

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 333–341, 2010.

addition to current industrial automation and measurement market, life automation and retail market, computer integrating communication market, medical application market, commerical video game market, government military and aeronautic market, safety surveillance market and vehicle electronic market, the future applications will be wider and more diverse.

Industrial Personal Computer (IPC) had become famous when the stock price of "Firich Enterprises Co., Ltd." rose to NT$1,000 per share in June 2007. With the development of the global economy, the change of life pattern and the emphasis on leisure life, the stock prices of gambling, game and entertainment-concept companies are rising, and these sectors are closely related to the industrial PC industry. Tiawan's industrial PC indutsry takes the leading place in the related field owing to its competitive price, high system compatibility and coordination, long-term supply stability, strict quality control, complete component supply chain and strong inventory management. Furthermore, Taiwan's industrial PC manufacturers have R&D ability and can meet special industry requirements to customize products. In terms of cost reduction, international corporations regard Taiwan manufacturers as their optimal partners. The output value and future development of industrial PC industry should not be ignored. The market scale, or enterprise revenue, is important; actually, it is critical for money making. Only by continuously increasing profits can an enterprise sustain its operation and keep developing. In other words, enterprises must have good business operating accomplishments to maintain the sustained growth. The purpose of an enterprise is to make profits. The annual revenues become an important index to evaluate the enterprise's business operating accomplishments. Number speaks for itself. Profit making indicates good business operating accomplishments. Profitability becomes an important index of enterprise success.

Based on the rationale mentioned above, this research collects related data to make further analysis of industrial PC industry to achieve the following study purpose: analyze the current development and operating accomplishments of Taiwan's industrial PC industry; set up the fuzzy evaluation structure of business operating accomplishments for industrial PC industry; and report the results of empirical study to verify the operations of fuzzy evaluation structure mentioned above.

2 Preliminaries

The fuzzy number was proposed by Dubois & Prade in 1980. Traditional set is based on two-valued logic to describe things, where a relation of element x and set A can only belong to A or not belong to A. This is a concept of "Either This or That".

2.1 Fuzzy Number

Let $\tilde{A}$ be a fuzzy set on $\mu_{\tilde{A}}(x)$, then $\tilde{A} = (a,b,c,d), a < b < c < d$ is denoted as a trapezoidal fuzzy number (TFN) if its membership function is defined as follows:

$$\mu_{\tilde{A}}(x) = \begin{cases} \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & b \le x \le c, \\ \dfrac{d-x}{d-c}, & c \le x \le d, \\ 0, & otherwise. \end{cases} \tag{2.1}$$

2.2 Defuzzification

If C($\tilde{\mathrm{A}}$) is the value of fuzzy number $\tilde{\mathrm{A}}$ after defuzzification, then the following is the centroid method for defuzzification $\tilde{\mathrm{A}}$,

$$C(\tilde{A}) = \frac{\int x\mu_{\tilde{A}}(x)dx}{\int \mu_{\tilde{A}}(x)dx} \tag{2.2}$$

If $\tilde{A} = (a,b,c)$ is a TFN and $a \le b \le c$, $a,b,c \in R$, and then we have

$$C(\tilde{A}) = \frac{1}{3}(a+b+c)$$

2.3 Linguistic Variables

Linguistic variables are used to describe the fuzzy set by natural language under appointed domain, including four types of data: name, type, range and degree. When applying the fuzzy theory to measure subjective determination, the following two steps are taken:

1. Transform linguistry used by linguistic variables into fuzzy number.
2. Transform fuzzy number as crisp value via defuzzification.

3 Fuzzy Evaluation Approach

3.1 Fuzzy Evaluation Model

In constructing a fuzzy evaluation model, the following steps are taken:

1. Industrial PC manufacturers who participate in this evaluation are briefly referred to as "manufacturers under evaluation".
2. Decide the evaluation criteria (also called "evaluation dimension and factor").
3. Select the evaluation mode.

3.2 Manufacturers under Evaluation

Let m industrial PC manufacturers participate in the business operating evaluation, then $A = \{a_i\}, i = 1,2,...,m$, where *A* represents a set of decision-making case, a_i represents the *i* manufacturer under evaluation, which is the i^{th} decision-making case, and *m* represents the total number of decision-making case. Take the empirical case in Section 4 as an example, 22 industrial PC manufacturers are selected, so *m*=22.

3.3 Decision Evaluation Criteria

The criteria for deciding the evaluation structure must take the items and their relative importance into account regardless of one-phase evaluation factor or two-phase

evaluation factor and dimension. In other words, evaluation items and their relative weight should be put into the evaluation structure.

1. Form a review group
In order to effectively decide the evaluation criteria and their relative importance, it is necessary to found a review group. This research assumes *r* experts and scholars form an evaluation group and each member must have his/her properties related to profession, diversification and objectivity. Briefly, that is $B = \{b_k\}, k = 1,2,...,r$, where *B* represents the set of review group member, b_k represents the k^{th} reviewer, and *r* represents the total number of reviewers.

2. Decide the evaluation items
The way to decide the evaluation item is authorized to decide by the review group, and ensure the items of evaluation criterion; let $E = \{e_j\}, j = 1,2,...,n$, where E represents the set of evaluation criterion, e_j represents the j evaluation criterion, n represents the total number of evaluation criterion.

3.4 Relative Weight

In order to identify the relative importance of each factor e_j in E respectively, each component b_k in set *B* can use linguistic terms of weight (LTW) to assess the relative importance in each e_j, where LTW include important, very important and so on. This research adopts five LTW to identify each e_j, including Very Important (V), Important (I), Fair (F), Unimportant (U) and Very Unimportant (VU).

1. Linguistic weight triangular fuzzy number
r reviewers implement *t* times for the relative importance assessment of n-term evaluation criteria without memory, and related weight linguistic variables can be transformed into a corresponding triangular fuzzy number (TFN) according to the transform formula by Yu and Yao (2001). The transformation equation is as follows:

$$\tilde{L}_1 = (0,0,\frac{1}{v-1}), \tag{3.1}$$

$$\tilde{L}_u = (\frac{u-2}{v-1},\frac{u-1}{v-1},\frac{u}{v-1}), \tag{3.2}$$

$$\tilde{L}_v = (\frac{v-2}{v-1},1,1), \tag{3.3}$$

where $u = 2,3,...,v-1$, $\tilde{L}_u$ represents the *u* TFN of the corresponding linguistic variable, and *v* represents the total number of TFN corresponding to the LTW. Take the five LTW set in this research as an example, transform from [3.1] to [3.3]. The results are shown in Table 1.

Table 1. Comparison of the linguistic terms of weight and the triangular fuzzy numbers

LTW	Transformation Equation	TFNs
Very Unimportant (VU)	$\widetilde{L}_1$	(0,0,1/4)
Unimportant (U)	$\widetilde{L}_2$	(0,1/4,2/4)
Fair (F)	$\widetilde{L}_3$	(1/4,2/4,3/4)
Important (I)	$\widetilde{L}_4$	(2/4,3/4,1)
Very Important (V)	$\widetilde{L}_5$	(3/4,1,1)

2. Relative weight set

In order to avoid the extreme value of relative importance evaluation, this research adopts the median to determine the relative weight value. The detailed procedure is described as follows:

(1) Take the average value after summing three-time evaluation triangular fuzzy number for each b_k to all e_j, that is

$$\tilde{h}_{kj} = \frac{1}{3}\sum_{t=1}^{3} \tilde{g}_{kjt} \tag{3.4}$$

where $\tilde{h}_{kj}$ represents the TFN of the absolute weight average value that b_k evaluates the importance to all e_j.

(2) Obtain the absolute weight average value in the fuzzy sense after the defuzzification of $\tilde{h}_{kj}$ mentioned above is made. The formula is as shown below:

$$h_{kj} = C\left(\tilde{h}_{kj}\right) = \frac{1}{3}\left(\frac{1}{3}\sum_{t=1}^{3} \tilde{g}_{kjt}\right)$$

$$= \frac{1}{9}(x_{kj1} + x_{kj2} + x_{kj3} + y_{kj1} + y_{kj2} + y_{kj3} + z_{kj1} + z_{kj2} + z_{kj3}) \tag{3.5}$$

where h_{kj} represents the absolute weight average value evaluated by b_k to e_j in the fuzzy sense.

(3) Arrange all evaluated h_{kj} in each e_j in ascending order, and find out a median. Take j=3 as an example, find out a median of h_{k3} in a sequence of h_{k3}; the results of arranging $h_{13,}h_{23,}.....,h_{r3,}$ in ascending order are as follows:

$$h_{(1)3,}h_{(2)3,}.....,h_{(r)3} \tag{3.6}$$

where $h_{(1)3}$ represents the minimum value in h_{k3}, and $h_{(r)3}$ represents the maximum value in h_{k3}. Next, find out the median position of h_{k3} according to $\frac{r+1}{2}$ and then decide a median of h_{k3}. If *r* is an odd number, the median position of h_{k3} will be $h_{(\frac{r+1}{2})3}$. If *r* is an even number, a median of h_{k3} will be $\frac{\left[h_{(\frac{r}{2})3}+h_{(\frac{r+1}{2})3}\right]}{2}$.

(4) Decide the relative weight

If *r* is an even number, then we have the relative weight of e_j,

$$w_j = \frac{\frac{\left(h_{(\frac{r}{2})j}+h_{(\frac{r+1}{2})j}\right)}{2}}{\sum_{j=1}^{n}\frac{\left(h_{(\frac{r}{2})j}+h_{(\frac{r+1}{2})j}\right)}{2}} \tag{3.7}$$

where w_j represents the relative weight value of e_j in the fuzzy sense.

3.5 Evaluation of Business Operating Accomplishments

In order to understand business operating accomplishments by each manufacturer under evaluation, the evaluation can be made according to the related data, where the quantitative data refer to the numerical value in the annual financial statement. As for the qualitative data, the review group can evaluate the linguistic terms as mentioned above.

Qualitative Data: The assessment of the qualitative data can be handled following the steps to evaluate relative weight in the previous two sections. The detailed steps are as follows.

1. Each reviewer gives an evaluation linguistic variable as the qualitative evaluation criterion e_j. A set *P* will be used to represent all the assessment results collected. The four linguistic terms used to assess in this research include Excellent (EX), Very Good (VG), Good (G) and Poor (PR).

$$P = \{p_{kj}\}, k = 1,2,...,r,\ j = 1,2,...,n_1 \tag{3.8}$$

where p_{kj} represents a linguistic evaluation variable of b_j to e_j, while n_1 represents e_j in *n* terms, belongs to the qualitative evaluation criterion.

2. Transform each p_{kj} to the corresponding triangular fuzzy number $(\tilde{p}_{kj})$ in accordance with (3.1) to (3.3).

3. Average the triangular fuzzy number ($\tilde{f}_j$) of overall average evaluation by each reviewer after summing $\tilde{p}_{kj}$.

$$\tilde{f}_j = \frac{1}{r}\sum_{k=1}^{r} \tilde{p}_{kj}, j = 1,2,...,n_1 \tag{3.9}$$

4. Use the centroid in (2.5) for defuzzification to determine the overall average evaluation value in the fuzzy sense.

$$f_j = C(\tilde{f}_{kj}) = \frac{1}{3r}\sum_{k=1}^{r} \tilde{p}_{kj} \tag{3.10}$$

where f_j represents the overall average evaluation value that sums the qualitative evaluation item e_j by each reviewer in the fuzzy sense.

Quantitative Data. After collecting these qualitative data, the quantitative data can be handled according to the following steps:

1. Collect the quantitative data of evaluation criteria $e_j, j = 1,2,...,n-n_1$ in financial statements of each manufacturer under evaluation $a_i, i = 1,2,...,m$, and use statistic analysis to determine the average number (after (3.11)) and the standard deviation (such as (3.12)) formula to calculate the average value ($\bar{x}_j$) and determine that standard deviation (s_j) for each quantified value e_j .

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \tag{3.11}$$

$$s_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2} \tag{3.12}$$

where x_{ij} represents the j number of the quantified evaluation criterion for the i^{th} manufacturer, $\bar{x}_j$ represents the j average number of the quantified evaluation criterion for m manufacturers, s_j represents the j standard deviation of the quantified evaluation criterion for m manufacturers.

2. Because the error and probability of each estimate value $\bar{x}_j$ and each true value M_j can't be known, we take the reliability interval of $(1-\alpha)*100\%$ as the other way to estimate M_j . The estimation is as follows:

$$\left[\bar{x}_j - t_{m-1}(\alpha_{j1})\frac{s_j}{\sqrt{m}}, \bar{x}_j + t_{m-1}(\alpha_{j2})\frac{s_j}{\sqrt{m}}\right] \tag{3.13}$$

where $\alpha_{j1} + \alpha_{j2} = \alpha$, $m-1$ is the degree of freedom, and t_{m-1} is the t distribution of freedom $m-1$. Although the reliability interval can provide the error probability for M_j, a specified value must be used for the point estimate of M_j. Since M_j is an unknown parameter for the matrix and $\bar{x}_j$ is a point estimate, the estimation can be considered fuzzy if using $\bar{x}_j$ to estimate M_j. On the other hand, higher reliability can be considered subordination more approaching to 1 if $\left|\bar{x}_j - M_j\right|$ is smaller. On the contrary, the larger $\left|\bar{x}_j - M_j\right|$ error represents lower reliability, which can be considered subordination more approaching to 0. Therefore, the fuzzy number of $(1-\alpha)$ can be used for M_j calculation as shown in (3.14).

$$\tilde{x}_j = \left(\bar{x}_j - t_{m-1}(\alpha_{j1})\frac{s_j}{\sqrt{m}}, \bar{x}_j, \bar{x}_j + t_{m-1}(\alpha_{j2})\frac{s_j}{\sqrt{m}} \right) \tag{3.14}$$

where $\tilde{x}_j$ represents the average fuzzy evaluation value of the j quantitative factor for *m* manufacturers under evaluation.

3. Use certroid method in (2.5) to make defuzzification for (3.14) and determine the average evaluation value from fuzzy viewpoint.

$$x_j = C(\tilde{x}_j) = \frac{1}{3}\left[3\bar{x}_j + \frac{s_j}{\sqrt{m}}(t_{m-1}(\alpha_{j2}) - t_{m-1}(\alpha_{j1})) \right] \tag{3.15}$$

4. Use normal distribution formula to determine the Z value from each manufacturer under evaluation x_{ij}.

$$Z = \frac{x_{ij} - x_j}{s_j} \tag{3.16}$$

5. According to Table 2, the Z value mentioned above can be classified as Excellent, Good, Fair and Poor for four evaluation linguistic terms.

Table 2. Comparison of Z and evaluation value

Z value	Large number is excellent	Small number is Excellent
$Z \le -0.6745$	Poor	Excellent
$-0.6745 < Z \le 0$	Fair	Good
$0 < Z \le 0.6745$	Good	Fair
$0.6745 < Z$	Excellent	Poor

Refer to [3.1] to [3.3] and transform the linguistic terms evaluated above to corresponding triangular fuzzy number and proceed the defuzzification procedure.

3.6 Overall Business Operational Accomplishments

The overall business operational accomplishment of each industrial PC manufacturer can be obtained from [3.17].

$$D = \sum_{j=1}^{n} w_j f_j, j = 1,2,...,n \tag{3.17}$$

4 Conclusion

In a crucial competitive environment, the Industrial PC sectors encounter enormously difficult challenges. Hence, how to improve the performance and achieve the competitive advantage becomes an important issue. Performance can be viewed as the only way to assess the business operational accomplishments. However, it is not easy to evaluate the performance due to some subjective cognitive situations. In this paper, the fuzzy theory is applied to evaluate the performance in the IPC industry and both the quantitative and the qualitative data are considered in the model. It is more realistic in that the fuzzy model approaches the decision making behaviors by human beings and is more objective. The fuzzy model can also be employed on analogous industries of in the PC field.

References

1. Bonits, N., Keow, W.C.C., Richardson, S.: Intellectual Capital and Business Performance in Malaysian Industries. Journal of Intellectual Capital 1(l), 85–100 (2000)
2. Feroz, E.H., Kim, S., Raab, R.: Performance Measurement in Corporate Governance Managerial Performance in the Post-merger Period? Review of Accounting and Finance 4(3), 86–100 (2005)
3. Lumpkin, G.T., Dess, G.G.: Clarifying the Entrepreneurial Orientation Construct and Linking it to Performance. Academy of Management Review 21(1), 135–172 (1996)
4. Richard, Micholas: Production and Operations Management, 7th edn. Manufacturing and Services (1995)
5. Yu, M.M., Yao, J.S.: Modification of Concordance Analysis Based on Statistical Data and Ranking of Level 1–α Fuzzy Numbers. International Journal of Uncertainty Fuzziness and Knowledge–Based Systems 9(3), 313–340 (2001)

Inventory Models for Deteriorating Items with Variable Selling Price under Stock-Dependent Demand

Yen-Wen Wang[1], Chih-Te Yang[1,*], August Tsai[1], and Chiou-Ping Hsu[2]

[1] Department of Industrial Engineering & Management, Ching Yun University, Jung-Li 320, Taiwan
Tel.: +886 3 4581196, Fax: +886 3 4683298
ctyang@cyu.edu.tw
[2] Graduate Institute of Management Sciences, Tamkang University, Tamsui, Taipei 251, Taiwan

Abstract. In this paper, we consider the problem of determining the optimal replenishment policy for deteriorating items with variable selling prices under stock-dependent demand, in order to match realistic circumstances. The inventory problem of this study includes without shortages as well as complete backlogging models. A theoretical analysis of the existence and uniqueness of the optimal solutions without shortages and with complete backlogging is presented. Numerical examples of the parameters are also presented to illustrate these two models. Finally, we compare the optimal solutions without shortages to those with complete backlogging.

Keywords: Inventory, Selling price, Stock-dependent demand, Backlogging.

1 Introduction

It is important to control and maintain the inventories of deteriorating items in modern industrial organizations. In general, deterioration of an item is defined as the damage, spoilage, dryness, vaporization, etc., that result in decreased usefulness. IC chip, blood, alcohol, gasoline, and radioactive chemicals are examples of deteriorating items. Research on the inventory control of deteriorating items has been substantial and continues to expand. It began with Ghare and Schrader [1]; they were the first to present an economic order quantity model for an exponentially constant deteriorating inventory. Later, the assumption of a constant deteriorating rate was relaxed by Covert and Philip [2], who formulated the model with a variable deterioration rate of two-parameter Weibull distribution. This model was further generalized by Philip [3], who considered a three-parameter Weibull distribution. Shah [4] extended Philip's [3] model and considered that shortages were allowed. Goyal and Gunasekaran [5] developed an integrated production-inventory marking model for deteriorating items. Lin *et al.* [6] proposed an EOQ model for deteriorating items with time-varying demand and shortages. Finally, Goyal and Giri [7] presented a review of the literature

* Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 342–353, 2010.

on deteriorating inventories since the early 1990s. Following, there is a vast inventory literature on deteriorating items, the outline which can be found in review articles by Sarker *et al.* [8], Zhou and Lau [9], Yang [10], Dye *et al.* [11] and others.

Traditional inventory models developed for constant demand rate can be applied to both manufacturing and sales environments. However, the assumption of a constant demand for goods is not always applicable to real situations. It is usually observed in the supermarket that a display of consumer goods in large quantities attracts more customers and generates higher demand than those of fewer goods. Within the last twenty-five years, considerable attention has been given to situations in which the demand rate is dependent on the level of on-hand inventory. At the forefront of these studies are Gupta and Vrat [12], who were the first to develop inventory models for stock-dependent consumption rate. Later, Baker and Urban [13] established an economic order quantity model for a power-form inventory-level-dependent demand pattern. Mandal and Phaujdar [14] then developed an economic production quantity model for deteriorating items with a constant production rate and linearly stock-dependent demand. More recent studies in this area include those by Pal *et al.* [15], Padmanabhan and Vrat [16], Giri *et al.* [17], Ray and Chaudhuri [18], Ray *et al.* [19], and Pal *et al.* [20].

Since price has a direct impact on demand, pricing strategy is one of the major concerns of sellers/retailers who want to obtain maximum profits. Models with price-dependent demand also occupy a prominent place in the inventory literature. Cohen [21] jointly determined the optimal replenishment cycle and price for inventory that is subject to continuous decay over time at a constant rate. Wee [22] studied the joint pricing and replenishment policy for deteriorating inventory with a price elastic demand rate that declines over time. Abad [23] considered the dynamic pricing and lot- sizing problem of a perishable good under partial backlogging of demand. He assumed that the fraction of shortages backordered is variable and is a decreasing function of the waiting time. Wee [24, 25] extended Cohen's [21] model to develop a replenishment policy for deteriorating items with a price-dependent demand and Weibull distribution deterioration, and considered the absence or presence of a quantity discount separately. Abad [26] investigated joint pricing and lot-sizing under conditions of perishability, finite production and partial backlogging. Mukhopadhyay *et al.* [27, 28] reestablished Cohen's [21] model by taking the price elasticity of demand as normative and separately considering the time-proportional and two-parameter Weibull distribution deterioration rate. Chang *et al.* [29] introduced a deteriorating inventory model with a price-time dependent demand and partial backlogging. Recently, Dye [30] has proposed joint pricing and ordering policies for deteriorating inventory with price-dependent demand.

All of the inventory models for deteriorating items in the foregoing studies assumed that demand depends on either price or stock. In this paper, we consider the problem of determining the optimal replenishment policy for deteriorating items with variable selling prices under stock-dependent demand, in order to match realistic circumstances. The inventory problem of this study includes wthout shortages as well as complete backlogging models. A theoretical analysis of the existence and uniqueness of the optimal solutions without shortages and with complete backlogging is presented. Numerical examples of the parameters are also presented to illustrate these two models. Finally, we compare the optimal solutions without shortages to those with complete backlogging.

2 Assumptions and Notation

The mathematical model in this paper is developed on the basis of the following assumptions and notation:

1. The demand rate $D(t)$ at time t is
$$D(t) = \begin{cases} \alpha + \beta I(t), & I(t) > 0, \\ \alpha, & I(t) \le 0 \end{cases}$$
where α and β are positive constants, and $I(t)$ is the inventory level at time t.
2. The replenishment rate is infinite, and the lead time is zero.
3. Shortage is not allowed in Model I, whereas complete backlogging is permitted in Model II at a finite shortage cost C_2 per unit time.
4. The distribution of time until deterioration of the items follows an exponential distribution with a parameter of θ (constant rate of deterioration).
5. The unit cost C and the inventory carrying cost as fraction i, per unit time, are known and constant.
6. The variable selling price $S(t)$ at time t is:
$$S(t) = S_0 - \gamma D(t) = \begin{cases} S_0 - \gamma[\alpha + \beta I(t)], & I(t) > 0, \\ S_0 - \gamma\alpha, & I(t) \le 0, \end{cases}$$
where S_0, γ, α, β are positive constants and $I(t)$ is the inventory level at time t.
7. A denotes the ordering cost per order.
8. $P(T)$ and $P(t_1, T)$ represent the profits per unit time for Model I and Model II, respectively.
9. The inventory policy is a continuous review policy of EOQ type.

3 Model I: Without Shortages

The objective of the first model is to determine the optimal order quantity for items having a stock dependent selling rate and exponential decay with no shortages permitted. The inventory level decreases owing to the stock dependent demand rate as well as deterioration. Thus, the differential equation representing the inventory status is given by:

$$\frac{dI(t)}{dt} + \theta I(t) = -[\alpha + \beta I(t)], \ 0 \le t \le T, \tag{1}$$

with the boundary condition $I(T) = 0$. The solution of equation (1) is:

$$I(t) = \frac{\alpha}{\theta + \beta}\left[e^{(\theta+\beta)(T-t)} - 1\right], \ 0 \le t \le T. \tag{2}$$

The order quantity for each cycle can be written as:

$$Q = I(0) = \frac{\alpha}{\theta + \beta}\left[e^{(\theta+\beta)T} - 1\right]. \tag{3}$$

The material cost per cycle (denoted by MC) is:

$$MC = \frac{C\alpha}{\theta+\beta}\left[e^{(\theta+\beta)T} - 1\right]. \tag{4}$$

The inventory holding cost per cycle (denoted by HC) is given by:

$$HC = iC\int_0^T I(t)\,dt = \frac{iC\alpha}{(\theta+\beta)^2}\left[e^{(\theta+\beta)T} - (\theta+\beta)T - 1\right]. \tag{5}$$

The sales revenue per cycle (denoted by SR) is given by:

$$\begin{aligned} SR &= \int_0^T S(t)D(t)dt \\ &= (S_0 - \gamma\alpha)\alpha T + \left[\frac{\alpha^2\beta^2\gamma}{(\theta+\beta)^3} + \frac{S_0\alpha\beta - 2\alpha^2\beta\gamma}{(\theta+\beta)^2}\right]\left[e^{(\theta+\beta)T} - (\theta+\beta)T - 1\right] \\ &\quad - \frac{\alpha^2\beta^2\gamma}{2(\theta+\beta)^3}[e^{(\theta+\beta)T} - 1]^2 . \end{aligned} \tag{6}$$

Therefore, the total profit per unit time is given by:

$$\begin{aligned} P(T) &= \{\text{sales revenue per cycle-the ordering cost-the material cost} \\ &\quad \text{per cycle-inventory holding cost per cycle}\}/T \\ &= \frac{1}{T}(SR - A - MC - HC) \end{aligned}$$

Substituting Equations (4)-(6) into the above formula, we obtain :

$$\begin{aligned} P(T) &= \left[S_0\alpha - \gamma\alpha^2 - C\alpha + \frac{\gamma\alpha^2\beta^2}{(\theta+\beta)^2}\right] + \frac{\alpha}{(\theta+\beta)^3 T}\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma]\} \\ &\quad \left[e^{(\theta+\beta)T} - (\theta+\beta)T - 1\right] - \frac{\alpha^2\beta^2\gamma}{2(\theta+\beta)^3 T}e^{2(\theta+\beta)T} + \frac{1}{T}\left[\frac{\alpha^2\beta^2\gamma}{2(\theta+\beta)^3} - A\right]. \end{aligned} \tag{7}$$

Our main objective, precisely stated, is to find the optimal value of T in order to maximize the total profit per unit time, $P(T)$.

3.1 Results

The necessary condition for maximum profit per unit time in equation (7) is: $dP(T)/dT = 0$, which gives:

$$\begin{aligned} &\alpha\{(\theta+\beta)[S_0\beta - C(i+\beta+\theta)] - 2\theta\alpha\beta\gamma\}[\,(\theta+\beta)Te^{(\theta+\beta)T} - e^{(\theta+\beta)T} + 1] \\ &\quad - \frac{\gamma\alpha^2\beta^2}{2}\left[2(\theta+\beta)Te^{2(\theta+\beta)T} - e^{2(\theta+\beta)T} + 1\right] + A(\theta+\beta)^3 = 0 . \end{aligned} \tag{8}$$

It is not easy to find the closed-form solution of T from (8); however, we can show that the value of T which satisfies (8) not only exists, but also is unique under certain

conditions. To prove this, we first let $\Delta \equiv S_0\beta - C(i+\beta+\theta)] - 2\theta\alpha\beta\gamma$. We then obtain the following results:

Lemma 1: *If* $\Delta \le 0$, *then the solution of* T (*denoted by* T^*) *in Equation* (8) *not only exists, but also is unique.*

Proof: From Equation (8), we let $x = (\theta+\beta)T > 0$ and

$$F_1(x) = \alpha\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma\}[xe^x - e^x + 1] - \frac{\gamma\alpha^2\beta^2}{2}\left[2xe^{2x} - e^{2x} + 1\right] + A(\theta+\beta)^3 . \quad (9)$$

Taking the first order derivation of $F_1(x)$ with respective to x, we obtain:

$$\frac{dF_1(x)}{dx} = \alpha\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma\}xe^x - \gamma\alpha^2\beta^2(2xe^{2x}) .$$

If $(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma = \Delta \le 0$, then $F_1(x)$ is a strictly decreasing function with respect to x . Furthermore, we have $\lim_{x\to 0^+} F_1(x) = A(\theta+\beta)^3 > 0$ and $\lim_{x\to+\infty} F_1(x) = -\infty < 0$. Hence, by using the Intermediate Value Theorem, we can show that the solution of T in Equation (8) not only exists, but also is unique. This completes the proof of Lemma 1.

Theorem 1: *If* $\Delta \le 0$, *the total profit per unit time* $P(T)$ *is convex and reaches its global maximum at point* $T = T^*$.

Proof: Taking the second derivative of $P(T)$, and then finding the value of this function at point T^*, we obtain:

$$\left.\frac{d^2P(T)}{dT^2}\right|_{T=T^*} = \frac{\alpha\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma\}}{(\theta+\beta)T^*}e^{(\theta+\beta)T^*} - \frac{2\gamma\alpha^2\beta^2}{(\theta+\beta)T^*}e^{2(\theta+\beta)T^*} .$$

From Lemma 1, if $\Delta \le 0$, then it can be easily seen that total profit per unit time $P(T)$ is convex and T^* is the global maximum point of $P(T)$. This completes the proof of Theorem 1.

Next, by substituting $T = T^*$ into Equation (7), we get the maximum total profit per unit time, which is given by:

$$P_1^* \equiv P(T^*) = \alpha(S_0 - \gamma\alpha - C) + \frac{1}{(\theta+\beta)^2}\{L[e^{(\theta+\beta)T^*} - 1] - M[e^{2(\theta+\beta)T^*} - 1]\} . \quad (10)$$

3.2 Numerical Example

In order to illustrate the above solution procedure, we consider an inventory system with the data as follows: $A = 250$, $\alpha = 600$, $C = 5$, $i = 0.35$, $\beta = 0.2$, $\theta = 0.2$,

$S_0 = 15$ and $\gamma = 0.01$ in appropriate units. From the above theoretical results, the condition $\Delta = -0.78 < 0$ and then the optimal replenishment policy can easily be obtained as follows: the optimal length of the order cycle is $T^* = 0.4707$ and the optimal order quantity per cycle is $Q^* = 310.751$. The maximum total profit per unit time is $P(T^*) = 1382.51$.

4 Model II: Complete Backlogging

In this case, when the inventory is positive, the selling rate is stock dependent; whereas for negative inventory the demand (backlogging) rate is constant. Therefore, the inventory level decreases due to stock-dependent selling as well as deterioration during the period $[0, t_1]$. During the period $[t_1, T]$ demand is backlogged. The differential equations governing the inventory status are given by:

$$\frac{dI_1(t)}{dt} + \theta I(t) = -[\alpha + \beta I(t)], \ 0 \le t \le t_1, \tag{11}$$

$$\frac{dI_2(t)}{dt} = -\alpha, \ t_1 \le t \le T. \tag{12}$$

The solutions to the above differential equations after applying the boundary conditions $I_1(t_1) = 0$ and $I_2(t_1) = 0$, respectively, are:

$$I_1(t) = \frac{\alpha}{\theta + \beta}\left[e^{(\theta+\beta)(t_1 - t)} - 1\right], \ 0 \le t \le t_1, \tag{13}$$

and

$$I_2(t) = \alpha(t_1 - t), \ t_1 \le t \le T. \tag{14}$$

Let $t = 0$ in Equation (13) and we obtain the inventory level at the beginning of the cycle (maximum inventory level) as follows:

$$B \equiv I_1(0) = \frac{\alpha}{\theta + \beta}\left[e^{(\theta+\beta)t_1} - 1\right]. \tag{15}$$

On the other hand, let $t = T$ in Equation (14) and we obtain the maximum amount of demand backlogged per cycle as follows:

$$W \equiv -I_2(T) = \alpha(T - t_1). \tag{16}$$

From Equations (15) and (16), we obtain the order quantity, Q, as

$$Q = B + W = \frac{\alpha}{\theta + \beta}\left[e^{(\theta+\beta)t_1} - 1\right] + \alpha(T - t_1). \tag{17}$$

The material cost per cycle is:

$$MC = \frac{C\alpha}{\theta+\beta}\left[e^{(\theta+\beta)t_1} - (\theta+\beta)t_1 - 1\right] + C\alpha T \ . \tag{18}$$

The inventory holding cost per cycle is given by:

$$HC = iC\int_0^{t_1} I_1(t)\,dt \ = \frac{iC\alpha}{(\theta+\beta)^2}\left[e^{(\theta+\beta)t_1} - (\theta+\beta)t_1 - 1\right]. \tag{19}$$

The shortage cost per cycle due to backlog (which is denoted by SC) is given by:

$$SC = c_2\int_{t_1}^{T}\left[-I_2(t)\right]dt = \frac{c_2\alpha}{2}(T-t_1)^2 \ . \tag{20}$$

The sales revenue per cycle is:

$$\begin{aligned} SR &= \int_0^T S(t)D(t)dt = \int_0^{t_1}\{S_0 - \gamma[\alpha+\beta I(t)]\}[\alpha+\beta I(t)]dt + \int_{t_1}^{T}(S_0-\gamma\alpha)\alpha dt \\ &= (S_0-\gamma\alpha)\alpha T + \left[\frac{\alpha^2\beta^2\gamma}{(\theta+\beta)^3} + \frac{S_0\alpha\beta - 2\alpha^2\beta\gamma}{(\theta+\beta)^2}\right]\left[e^{(\theta+\beta)t_1} - (\theta+\beta)t_1 - 1\right] \\ &\quad - \frac{\alpha^2\beta^2\gamma}{2(\theta+\beta)^3}[e^{(\theta+\beta)t_1} - 1]^2 \ . \end{aligned} \tag{21}$$

Therefore, the total profit per unit time is given by:

$$\begin{aligned} P(t_1, T) &= \{\text{sales revenue per cycle-the ordering cost-the material cost per cycle} \\ &\quad \text{-inventory holding cost per cycle- shortage cost per cycle}\}/T \\ &= \frac{1}{T}(SR - A - MC - HC - SC) \end{aligned}$$

Substituting (18)-(21) into the above formula, we arrive at:

$$\begin{aligned} P(t_1,T) &= \left[S_0\alpha - \gamma\alpha^2 - C\alpha\right] + \frac{\alpha\{(\theta+\beta)[S_0\beta - C(i+\beta+\theta)] - 2\theta\alpha\beta\gamma\}}{(\theta+\beta)^3 T}\left[e^{(\theta+\beta)t_1} - (\theta+\beta)t_1\right. \\ &\quad \left. - 1\right] - \frac{A}{T} - \frac{\alpha^2\beta^2\gamma}{2(\theta+\beta)^3 T}\left[e^{2(\theta+\beta)t_1} - 2(\theta+\beta)t_1 - 1\right] - \frac{C_2\alpha}{2T}(T-t_1)^2 \ . \end{aligned} \tag{22}$$

4.1 Results

The necessary conditions for the total profit per unit time to be maximum are $\partial P(t_1, T)/\partial t_1 = 0$ and $\partial P(t_1, T)/\partial T = 0$, which give:

$$\begin{aligned} &\frac{\alpha\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma\}}{(\theta+\beta)^2 T}\left[e^{(\theta+\beta)t_1} - 1\right] - \frac{\alpha^2\beta^2\gamma}{(\theta+\beta)^2 T}\left[e^{2(\theta+\beta)t_1} - 1\right] \\ &+ C_2\alpha\left(1 - \frac{t_1}{T}\right) = 0 \ , \end{aligned} \tag{23}$$

and

$$\frac{-\alpha\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma\}}{(\theta+\beta)^3 T^2}\left[e^{(\theta+\beta)t_1} - (\theta+\beta)t_1 - 1\right] + \frac{A}{T^2}$$
$$+ \frac{\gamma\alpha^2\beta^2}{2(\theta+\beta)^3 T^2}\left[e^{2(\theta+\beta)t_1} - 2(\theta+\beta)t_1 - 1\right] - \frac{C_2\alpha}{2}(1 - \frac{t_1}{T^2}) = 0 . \quad (24)$$

For notational convenience, let $M = \gamma\alpha^2\beta^2 > 0$, $N = C_2\alpha > 0$ and $L = \alpha\{(\theta+\beta)[S_0\beta - C(i+\theta+\beta)] - 2\theta\alpha\beta\gamma\}$. Thus from Equations (23) and (24), we have:

$$T = t_1 + \frac{M(e^{2(\theta+\beta)t_1} - 1) - L(e^{(\theta+\beta)t_1} - 1)}{(\theta+\beta)^2 N}, \quad (25)$$

and

$$\frac{1}{(\theta+\beta)N}\{M[e^{2(\theta+\beta)t_2} - 1] - L[e^{(\theta+\beta)t_1} - 1]\}^2 + 2(\theta+\beta)t_1\{M[e^{2(\theta+\beta)t_1} - 1]$$
$$- L[e^{(\theta+\beta)t_1} - 1]\} - M[e^{2(\theta+\beta)t_1} - 2(\theta+\beta)t_1 - 1]$$
$$+ 2L[e^{(\theta+\beta)t_1} - (\theta+\beta)t_1 - 1] - 2A(\theta+\beta)^3 = 0 , \quad (26)$$

respectively.

Now we can present the remaining theorems.

Lemma 2: *If* $\Delta \le 0$, *then the solutions of* t_1 *and* T *in Equations* (25) *and* (26) (*denoted by* t_1^* *and* T^*) *not only exist, but also are unique.*

Proof: From Equation (26), we let $y = (\theta+\beta)t_1 > 0$ and

$$F_2(y) = \frac{1}{(\theta+\beta)N}\left[M(e^{2y} - 1) - L(e^y - 1)\right]^2 + 2y\left[M(e^{2y} - 1) - L(e^y - 1)\right] - M(e^{2y} - 2y - 1)$$
$$+ 2L(e^y - y - 1) - 2A(\theta+\beta)^3 . \quad (27)$$

Taking the first order derivative $F_2(y)$ with respect to y, we get:

$$\frac{dF_2(y)}{dx} = e^y(2Me^y - L)\left\{\frac{2}{(\theta+\beta)N}\left[M(e^{2y} - 1) - L(e^y - 1)\right] + 2y\right\} > 0 .$$

Thus, $F_2(y)$ is a strictly increasing function with respect to y. It can be shown that $F_2(0) = -2(\theta+\beta)^3 A < 0$ and $\lim_{x\to+\infty} F_2(y) = +\infty$. Therefore, by using the Intermediate Value Theorem, we show that there exists an unique t_1^* such that $F_2(t_1^*) = 0$, which means that t_1^* is the unique solution of Equation (26).

Once we obtain the optimal value t_1^*, then the optimal value T (denoted by T^*) can be derived from Equation (25), and is given by $T^* = t_1^* + \{M[e^{2(\theta+\beta)t_1^*} - 1] - L[e^{(\theta+\beta)t_1^*} - 1]\}/(\theta+\beta)^2 N$. This completes the proof of Lemma 2.

Theorem 2: *If* $\Delta \le 0$, *the total profit per unit time* $P(t_1,T)$ *is convex and reaches its global minimum at point* $(t_1,T)=(t_1^*,T^*)$.

Proof: Taking the second derivative of $P(t_1, T)$ with respect to t_1 and T, we have:

$$\left.\frac{\partial^2 P(t_1,T)}{\partial T^2}\right|_{(t_1^*,T^*)}=\frac{-N}{T^*}<0,$$

$$\left.\frac{\partial^2 P(t_1,T)}{\partial t_1 \partial T}\right|_{(t_1^*,T^*)}=\frac{N}{T^*},$$

$$\left.\frac{\partial^2 P(t_1,T)}{\partial {t_1}^2}\right|_{(t_1^*,T^*)}=\frac{-e^{(\theta+\beta)t_1^*}(2Me^{(\theta+\beta)t_1^*}-L)}{(\theta+\beta)T^*}-\frac{N}{T^*}<0, \tag{28}$$

and

$$\left.\frac{\partial^2 P(t_1,T)}{\partial {t_1}^2}\right|_{(t_1^*,T^*)}\times\left.\frac{\partial^2 P(t_1,T)}{\partial T^2}\right|_{(t_1^*,T^*)}-\left[\left.\frac{\partial P(t_1,T)}{\partial t_1 \partial T}\right|_{(t_1^*,T^*)}\right]^2$$

$$=\frac{Ne^{(\theta+\beta)t_1^*}}{(\theta+\beta){T^*}^2}(2Me^{(\theta+\beta)t_1^*}-L)>0. \tag{29}$$

Hence, from Lemma 2, it can be easily seen that (t_1^*,T^*) is the global minimum point of $P(t_1,T)$. This completes the proof of Theorem 2.

Finally, we substitute $t_1=t_1^*$ and $T=T^*$ into equation (22) to obtain the maximum total profit per unit time, which is given by:

$$P_2^* \equiv P(t_1^*,T^*)$$

$$=\alpha(S_0-\gamma\alpha-C)+\frac{1}{(\theta+\beta)^2}\left\{L[e^{(\theta+\beta)t_1^*}-1]-M[e^{2(\theta+\beta)t_1^*}-1]\right\}. \tag{30}$$

4.2 Numerical Example

The numerical example solved in the previous section is considered again with $C_2=3$ as an appropriate unit. In the same way, we check the condition: $\Delta=-0.78<0$. Thus, the optimal replenishment policy can easily be obtained as follows: the optimal length of the inventory interval with positive inventory is: $t_1^*=0.3321$, the optimal length of the order cycle is $T^*=0.7150$, and the optimal order quantity per cycle is $Q^*=422.866$. The maximum total profit per unit time is: $P(t_1^*,T^*)=1710.64$.

5 Comparison of Optimal Solutions

In this section, we compare the optimal solutions of Model I with those of Model II. For the sake of convenience, the values of the parameters are identical to those in Section 3.1 except for the shortage cost parameter of C_2. Three different values of C_2 are adopted: C_2 =3, 4, and 5. The computed results are shown in Table 1.

Table 1. Summary of the optimal solutions for different models

Models	Optimal length of order cycle	Optimal order quantity per cycle	Total profits per unit time
Model Ⅱ ($C_2 = 3$)	0.7150	422.866	1710.64
Model Ⅱ ($C_2 = 4$)	0.6636	413.754	1659.13
Model Ⅱ ($C_2 = 5$)	0.6299	395.270	1622.04
Model Ⅰ	0.4707	310.751	1382.51

From Table 1, it can be found that the optimal length of the order cycle, the order quantity per cycle, and the total profits per unit time of Model II are greater than those of Model I. Furthermore, in the case of Model II, the lower the shortage cost, the greater the optimal length of the order cycle, order quantity per cycle, and total profits per unit time.

6 Conclusion

In this study, we considered the problem of determining the optimal replenishment policy for deteriorating items with variable selling price under stock-dependent demand. We attempted to model situations in which the selling rate of deteriorating items depends on price and stock level with complete backlogging and without backlogging, and we showed, via theoretical analysis, both the existence and uniqueness of optimal solutions for both two models. Numerical examples of the parameters were also presented to illustrate these two models. Finally, we compared the optimal solutions without shortages to those with complete backlogging. Further research might incorporate the proposed model with other parameters, such as capital investment, which is required to improve storehouse equipment so as to effectively reduce the deteriorating rate of an item, partial backlogging, stochastic demand, a finite replenishment rate, and so forth.

References

1. Ghare, P.M., Schrader, G.H.: A model for exponentially decaying inventory system. International Journal of Production Research 21, 449–460 (1963)
2. Covert, R.P., Philip, G.C.: An EOQ model for items with Weibull distribution deterioration. AIIE Transaction 5, 323–326 (1973)

3. Philip, G.C.: A generalized EOQ model for items with Weibull distribution. AIIE Transaction 6, 159–162 (1974)
4. Shah, Y.K.: An order-level lot size inventory model for deteriorating items. AIIE Transaction 9, 108–112 (1977)
5. Goyal, S.K., Gunasekaran, A.: An integrated production-inventory- marketing model for deteriorating items. Computers & Industrial Engineering 28, 755–762 (1995)
6. Lin, C., Tan, B., Lee, W.C.: An EOQ model for deteriorating items with time-varying demand and shortages. International Journal of Systems Science 31, 391–400 (2000)
7. Goyal, S.K., Giri, B.C.: Recent trends in modeling of deteriorating inventory. European Journal of Operational Research 134, 1–16 (2001)
8. Sarker, B.R., Mukherjee, S., Balan, C.V.: An order-level lot size inventory model with inventory-level dependent demand and deterioration. International Journal of Production Economics 48, 227–236 (1997)
9. Zhou, Y.W., Lau, H.S.: An economic lot-size model for deteriorating items with lot-size dependent replenishment cost and time-varying demand. Applied Mathematical Modelling 24, 761–770 (2000)
10. Yang, H.L.: A comparison among various partial backlogging inventory lot-size models for deteriorating items on the basis of maximum profit. International Journal of Production Economics 96, 119–128 (2005)
11. Dye, C.Y., Chang, H.J., Teng, J.T.: A deteriorating inventory model with time-varying demand and shortage-dependent partial backlogging. European Journal of Operational Research 172, 417–429 (2006)
12. Gupta, R., Vrat, P.: Inventory model with multi-items under constraint systems for stock dependent consumption rate. Operations Research 24, 41–42 (1986)
13. Baker, R.C., Urban, T.L.: A deterministic inventory system with an inventory-level-dependent demand rate. Journal of the Operational Research Society 39, 823–831 (1988)
14. Mandal, B.N., Phaujdar, S.: An inventory model for deteriorating items and stock-dependent consumption rate. Journal of the Operational Research Society 40, 483–488 (1989)
15. Pal, S., Goswami, A., Chaudhuri, K.S.: A deterministic inventory model for deteriorating items with stock-dependent demand rate. International Journal of Production Economics 32, 291–299 (1993)
16. Padmanabhan, G., Vrat, P.: EOQ models for perishable items under stock dependent selling rate. European Journal of Operational Research 86, 281–292 (1995)
17. Giri, B.C., Pal, S., Goswami, A., Chaudhuri, K.S.: An inventory model for deteriorating items with stock-dependent demand rate. European Journal of Operational Research 95, 604–610 (1996)
18. Ray, J., Chaudhuri, K.S.: An EOQ model with stock-dependent demand, shortage, inflation and time discounting. International Journal of Production Economics 53, 171–180 (1997)
19. Ray, J., Goswami, A., Chaudhuri, K.S.: On an inventory model with two levels of storage and stock-dependent demand rate. International Journal of Systems Science 29, 249–254 (1998)
20. Pal, A.K., Bhunia, A.K., Mukherjee, R.N.: A marketing-oriented inventory model with three-component demand rate dependent on displayed stock level (DSL). Journal of the Operational Research Society 56, 113–118 (2005)
21. Cohen, M.A.: Joint pricing and ordering policy for exponentially decaying inventory with known demand. Naval Research Logistic Quarterly 24, 257–268 (1977)

22. Wee, H.M.: Joint pricing and replenishment policy for deteriorating inventory with declining market. International Journal of Production Economics 40, 163–171 (1995)
23. Abad, P.L.: Optimal pricing and lot sizing under conditions of perishability and partial backordering. Management Science 42, 1093–1104 (1996)
24. Wee, H.M.: A replenishment policy for items with a price-dependent demand and a varying rate of deterioration. Production Planning and Control 8, 494–499 (1997)
25. Wee, H.M.: Deteriorating inventory model with quantity discount, pricing and partial backordering. International Journal of Production Economics 59, 511–518 (1999)
26. Abad, P.L.: Optimal pricing and lot-sizing under conditions of perishability, finite production and partial backordering and lost sales. European Journal of Operational Research 144, 677–686 (2003)
27. Mukhopadhyay, S., Mukherjee, R.N., Chaudhuri, K.S.: Joint pricing and ordering policy for a deteriorating inventory. Computers and Industrial Engineering 47, 339–349 (2004)
28. Mukhopadhyay, S., Mukherjee, R.N., Chaudhuri, K.S.: An EOQ model with two-parameter Weibull distribution deterioration and price-dependent demand. International Journal of Mathematical Education in Science and Technology 36, 25–33 (2005)
29. Chang, H.H., Teng, J.T., Ouyang, L.Y., Dye, C.Y.: Retailer's optimal pricing and lot-sizing policies for deteriorating items with partial backlogging. European Journal of Operational Research 168, 51–64 (2006)
30. Dye, C.Y.: Joint pricing and ordering policy for deteriorating inventory with partial backlogging. Omega 35, 184–189 (2007)

Hierarchical IP Distribution Mechanism for VANET

Chiu-Ching Tuan, Jia-Ming Zhang, and Shu-Jun Chao

Graduate Institute of Computer and Communication Engineering
National Taipei University of Technology
Taipei, Taiwan
cctuan@ntut.edu.tw, homicide522@pchome.com.tw,
chao@dns.chsh.tpc.edu.tw

Abstract. Vehicular Ad Hoc Network (VANET) is characterized by high speed, unstable routing as well as no power concern, so high efficient addressing mechanism is required for vehicles to access to Internet or communicate with each other for some practical or emergent information. The existing methods of addressing for VANET usually do not cope with instant IP address assignment and recycling. This paper proposed an AP-based centralized IP distribution system using a hierarchical Dynamic Host Configuration Protocol (DHCP) mechanism to tackle efficient addressing and timely IP recycling for reuse, and the malfunctioning APs can be detected via hierarchical mechanism. The invalid IP addresses for parking vehicles can also be immediately recycled through the mechanism of periodical report performed by vehicle itself. The simulation results proved that it outperformed the existing addressing mechanism.

Keywords: VANET, IP recycling, hierarchical mechanism.

1 Introduction

Different from Mobile Ad Hoc Network (MANET), VANET is characterized by high speed, unstable routing and no power concern, the traditional strategies for MANET like routing protocol and addressing mechanism [1][2] cannot be directly employed to VANET due to the node moving at a high speed. There are still a number of similarities between VANET and MANET, each vehicle is equipped with a transceiver to transmit and receive information for certain applications, and they can exchange useful or emergent information through Access Points (AP) or communicate with other vehicles in ad hoc way. Instead of the traditional communication protocol of 802.11b referred to as Wi-Fi [3], 802.11p is the standard communication protocol employed by VANET, it is also called Dedicated Short Range Communication (DSRC) [4], which operates in the spectrum of 5.8 GHz and the communication range is up to 1000 m. Same as the vehicle, AP is also equipped with a transceiver for communication and connected to a Road Side Unit (RSU), and vehicle itself plays the role of On Board Unit (OBU), RSU and OBU are the two fundamental components operating in VANET using DSRC as the standard communication protocol.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 354–366, 2010.

Addressing is a critical issue for wired and wireless communications, DHCP is the common mechanism for IP address assignment, it can dynamically assign IP addresses to all hosts without the effort for administrator to handle configuration manually, and each assigned IP address is endowed with a lease, when the lease expires, the host must release its own IP address or send a request ahead of time to DHCP for expanding the lease. For VANET, each vehicle moves at a high speed, they must obtain IP addresses in time for instant information to execute specified behaviors. For instance, vehicle must know the situation of front road ahead of time to redirect its heading direction in order to prevent itself from the traffic jam or construction, the efficiency of IP assignment will have an indirect impact on the instant information retrieval because IP address must be assigned to vehicle prior to accessing to critical information.

This paper proposed a high efficient IP address distribution system using hierarchical DHCP mechanism for VANET, each AP is equipped with a DHCP server for IP address assignment, and APs are all connected to one central server; the duplicate IP addresses can be completely avoided via the assistance of central server. The proposed mechanism mainly focuses on IP addresses for IPv4 instead of IPv6 which comprises a number of difficulties of implementation and has not been utilized widely up to now.

This paper is organized as follows. Section 2 introduces the characteristics of existing addressing mechanisms for VNAET, section 3 depicts the proposed mechanism of addressing, section 4 demonstrates the simulation results, and the conclusion is in section 5.

2 Related Work

There are two forms of addressing for VANET, which are distributed and centralized mechanisms respectively, the former mainly relies on the vehicle itself to take charge of the job of IP address distribution without the assistance of RSU, and the latter is the well-known AP-based mechanism making use of AP with a RSU to distribute IP addresses.

The classic type of distributed addressing mechanism for VANET is Vehicular Address Configuration (VAC) [5], this paper presented a Leader-based IP distribution mechanism. Vehicles are characterized as two types, which are Normal car and Leader car respectively; Normal car stands for the normal vehicle without the ability to assign IP addresses, and Leader car with an on-board DHCP server takes charge of the task of IP address distribution, and its edge of communication range is concatenated with that of neighboring Leader cars one hop away to construct a Leader chain. Each Leader car has a *SCOPE* in which Normal car can own a unique IP address, and once the Normal car moves out of the *SCOPE* of the current Leader car, its IP address is no longer ensured a unique one and then the Normal car must perform the Duplicate Address Detection (DAD), requesting a new IP address from another Leader car. The scenario of VAC is depicted in Fig. 1. The length of *SCOPE* is decided according to the number of hops among Leader cars, so the lengths of *SCOPE* of Leader car A and B are both set to 1 considering the neighboring Leader cars one hop away.

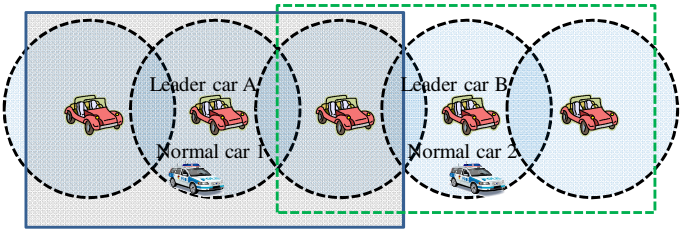


Fig. 1. The scenario of Vehicular Address Configuration (VAC)

In general, the Leader chain is difficult to maintain because of the variant relative speed of vehicles, if the distance between any two Leader cars exceeds *TH_max,* one Normal car between the two Leader cars will become a Leader car; on the contrary, if this distance shrinks, reaching *TH_min*, one of the two Leader cars will become normal.

Considering the aforementioned mechanisms performed by VAC, the IP address of Leader car or Normal car must be changed once the role changes due to the variant distance between any two Leader cars, which will increase the number of configuration. Furthermore, the problem with duplicate IP addresses still exists because of the delimited length of *SCOPE* in which unique IP address can be ensured, one Normal car like Normal car C or D exceeding the current *SCOPE* must be reconfigured, increasing the number of configuration as well. Periodical *HELLO* packet broadcasting with an interval of 800 ms performed by Leader cars is utilized to synchronize the IP address pools owned by Leader cars in order to ensure the unique IP addresses within their own *SCOPE*s. Normal car must listen to the *HELLO* packets for *T_start* time and then transmit an IP address request to the nearest Leader car for configuration; considering the *T_start* time, interval of each broadcasted *HELLO* packet and the competition among Normal cars sending requests simultaneously to the same Leader car, the consumed time for configuration may be considerable, which cannot accomplish instant configuration.

Another classic type of addressing mechanism for VANET is Centralized Address Configuration (CAC) [6], it presented a strategy of employing a centralized DHCP server which can provide unique IP addresses to all vehicles in the urban. Each RSU is equipped with an AP and connected to the centralized DHCP server to deal with configuration requests from vehicles. The mainly function of RSU for CAC is to relay configuration messages between vehicles and central DHCP server, all IP addresses are supplied by one source namely the central DHCP server, which can ensure that vehicles will not be configured with duplicate IP addresses, so the behaviors of DAD and reconfiguration in VAC are no longer needed, vehicle can possess an IP address for a long time, immune to frequent intermittent access to desired information, and vehicle can ask different RSU for extending the lease of IP address.

The scenario of CAC is depicted in Fig. 2. Although CAC has solved the problems with duplicate IP addresses and reconfiguration, the concern of efficiency of IP address configuration may still exist. Because of the RSUs deployed near the hot spots such as shopping malls or gas stations in the urban, a large number of vehicles could send IP address configuration requests to a number of RSUs or even one RSU for configuration at the same time, in such situation, a severe competition for IP address configuration will occur.

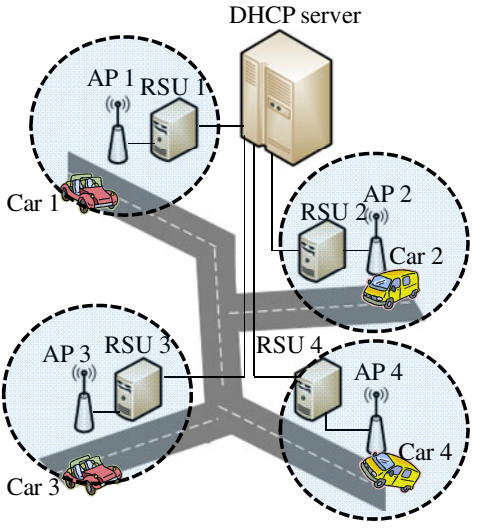


Fig. 2. The scenario of Centralized Address Configuration (CAC)

3 Hierarchical IP Distribution Mechanism

This paper proposed an addressing mechanism named Hierarchical IP Distribution (HID) in order to reach the objectives of unique IP address generation and high efficient configuration, and the invalid IP address can be immediately recycled without the concern of lease. The fundamental concept is to endow each RSU with the ability to directly assign IP addresses, namely on-board DHCP server is employed; commonly, AP is combined with an RSU to receive and transmit data, and AP is also one type of RSU, the terms "AP" and "RSU" can be used interchangeably, so the term "AP" will be used in HID. All APs are connected to one central server named Balance Server (BS) of which function is to distribute and synchronize the IP address pools managed by APs themselves. Instead of relaying messages performed in CAC, each AP in HID can directly deal with DHCP requests, namely AP has its own scheduler to configure all vehicles passing through its communication range. The fundamental architecture for HID is depicted in Fig. 3.

There is a hierarchical relationship which is set initially among all APs to assist them in assigning and recycling IP addresses and detecting malfunctioning APs through specific message-exchanging. The hierarchical relationship among APs in Fig. 3 is showed in table 1; current AP transmits periodical messages to higher level AP and receives that from lower level AP with an interval of 60 seconds for data-updating. With this periodical message-exchanging among APs, AP can know whether the lower level AP has malfunctioned or not, if so, it can immediately tackle the unassigned IP addresses of this malfunctioning AP for fault-tolerance.

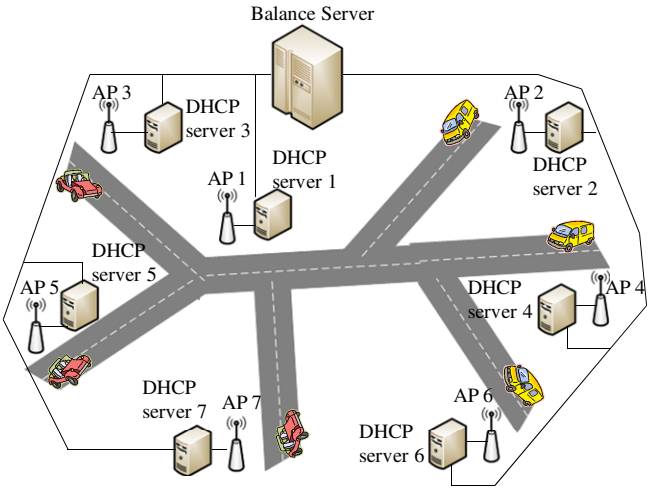


Fig. 3. The fundamental architecture for Hierarchical IP Distribution (HID)

Table 1. Hierarchical Relationship

Current AP	Higher Level AP	Lower Level AP
AP 1	BS	AP 3
AP 2	BS	AP 4
AP 3	AP 1	AP 5
AP 4	AP 2	AP 6
AP 5	AP 3	AP 7
AP 6	AP 4	—
AP 7	AP 5	—

3.1 The Method of AP Deployment

Each AP must be moderately deployed in order to perform the IP address distribution well, so the real traffic statistics recorded by Traffic Engineering Office in Taipei (TEOT) [8] is employed as a reference to deploy APs. There are two types of APs in this scenario, which are Border AP (BA) and Inner AP (IA) respectively, BA periodically broadcasts specific beacons which vehicles can snoop to decide to release its IP address to the current BA; vehicle can ask either BA or IA for configuration if it has no IP address when entering this urban. BA is just deployed in the entrance like a highway connected to the urban, and IA is deployed near the hot spot such as shopping mall or supermarket. The concept for IP deployment is showed in Fig. 4.

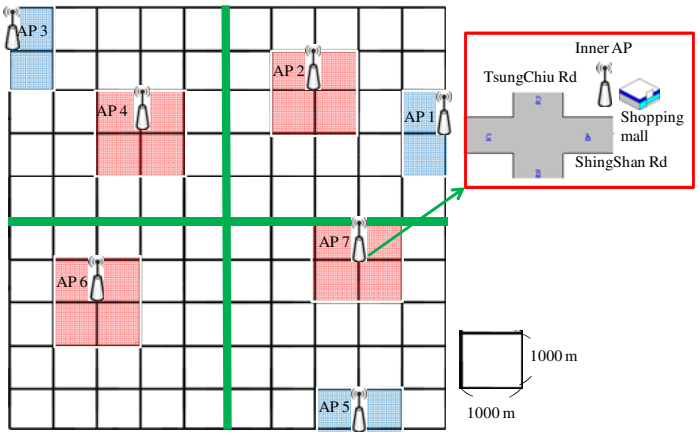


Fig. 4. The concept of AP deployment

100 small squares of which size is 1000 × 1000 m are combined into an larger area referred to as an urban, and each cross stands for an intersection. AP 1, 3, and 5 are BAs while AP 2, 4, 6 and 7 are IAs. In order to deploy APs averagely, the urban is divided into four sections, an IA and a BA will be deployed in one section except for the section without a BA in the third quadrant because there is inevitably a highway connected to a section according to TEOT. Each AP has a buffer for recycling IP addresses released from vehicles, and the size of buffer is set according to the differential traffic statistics for the location of each AP, which is also recorded by TEOT. Vehicle can extend the lease of IP address via any APs within this urban. The buffer size for BA and IA has to be designed carefully for efficient IP address assignment. Fig. 5 is the example of designing buffer size.

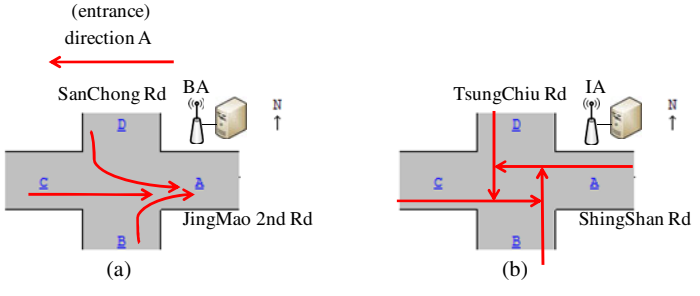


Fig. 5. The method of designing the buffer size

Due to the differential functionality between BA and IA, the calculation in designing buffer size must be also different. The responsibility of BA is to recycle the IP addresses of leaving vehicles, so the buffer size of BA employed in the entrance to the urban is set according to the sum of traffic for three directions which are B, C and

D in Fig. 5 (a), while that of IA installed inside the urban is set according to the sum of traffic for four directions which are A, B, C and D in Fig. 5 (b).

3.2 Available IP Address Segment Division

The private IP address for Class B ranging between 172.16.0.0 and 172.31.255.255 is employed as the example for available IP address segment division in this paper, its format is depicted in Fig. 6.

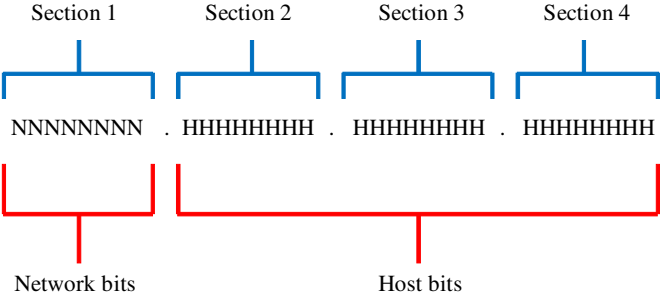


Fig. 6. The format of IP address for Class B

Section 1 is constant, and the subsequent three sections can be altered. A is method proposed for BS to divide the private IP address for Class B into a number of available IP address segments namely the IP address pools to all APs and itself; these segments absolutely do not overlap others through the algorithm of available IP address division performed by BS initially. Three equations for available IP address division are showed as follows.

$$SSA = \frac{SOS}{NOA + 1} \tag{1}$$

$$HEAD = AN \times SSA + IV \tag{2}$$

$$TAIL = (AN + 1) \times SSA + IV \tag{3}$$

SSA is the size of section 2 for each AP, *SOS* is the total size of section 2 for Class B, *NOA* is the number of deployed APs plus 1 which means BS itself; *HEAD* and *TAIL* are the minimum and maximum of section 2 for each AP respectively, *AN* is AP number ranging between 1 and the number of deployed AP, and *IV* is the initial value of section 2 which is 16 for Class B. For instance, 7 APs are deployed, *SOS* will be 31-16+1=16, *SSA* will be 16／(7+1) = 2, *HEAD* and *TAIL* for AP 7 will be 7×2+16=30 and (7+1)×2-1+16=31 respectively; the minimum and maximum of section 3 are constantly set to 0 and 255 respectively, while those of section 4 are constantly set to 1 and 254 respectively, so the available IP segment for AP 7 ranges between 172.30.0.1 and 172.31.255.254, of which retained IP addresses are eliminated and the first IP address 172.30.0.1 will be the IP address of AP 7 itself. With such IP address pool division, vehicle moving within the urban can has a unique IP address.

3.3 The Detail Operation of HID

In Fig. 7, Arrow (A) means that AP 1 puts the IP address released by Car 1 to its buffer which is designed as a circular queue, arrow (B) means that AP 1 directly assigns an available IP address to Car 2, and arrow (C) means that AP 1 takes one IP address to assign to Car 3 once it has run out of its available IP addresses. BS manages a form named IP Active Time Form (IATF), which keeps track of the status, report time and owner of each IP address in this scenario. The IATF is showed in table 2.

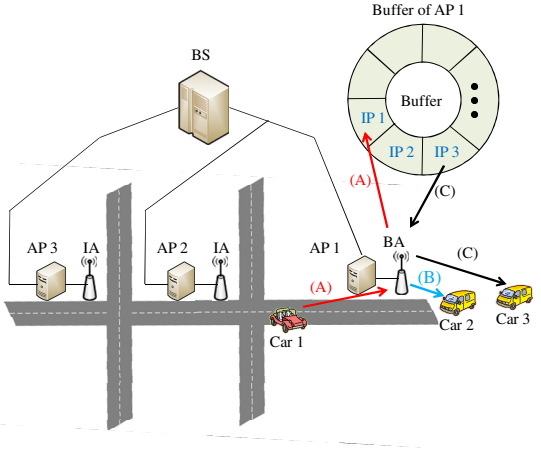


Fig. 7. The detail operation of HID

Table 2. IP Active Time Form

Address	Status	Report time	Owner
172.16.0.1	*available*	—	BS
172.16.0.2	*assigned*	10 : 05 : 31	—
172.16.0.3	*assigned*	10 : 01 : 34	—
172.16.0.4	*recycled*	—	AP 4
172.16.0.5	*assigned*	10 : 08 : 00	—
172.30.255.253	*assigned*	09 : 59 : 28	—
172.31.255.254	*recycled*	—	AP 2

After IP addresses are distributed by BS using the method of available IP address segment division, the values of status column for all IP addresses are all marked *"available"*, the values of report time column for that are all marked "—", and the values of owner column for that are marked the specific AP or BS itself.

The values of status column and owner column are marked *"recycled"* and the recycling AP or BS itself respectively when this IP address has been recycled; the values of active time column and status column will be marked a time value and *"assigned"* respectively if this IP address has been assigned to a vehicle. Same as VAC, each vehicle periodically broadcasts *HELLO* packet with an interval of 800 ms.

With the *HELLO* packet, vehicle can periodically transmit a life update packet for its IP address with an interval of 30 seconds to other ones in proximity in ad hoc way, this packet will be eventually relayed by AP to BS for updating the value of report time column with the current time value. BS periodically checks the IATF with an interval of 300 seconds, if the time value of active time column for an assigned IP address has not been updated for more than 60 seconds, it will be recycled by BS itself, namely the values of status column, report time column and owner column for this IP address will be marked *"recycled"*, "—" and BS respectively. Through this periodical reporting mechanism executed by vehicles, the invalid IP addresses of parking vehicles can be immediately recycled for reuse without considering the lease, which can tackle the problem with enormous IP address requests more than current available IP addresses in the urban at rush hour.

4 Simulation Results

The simulation environment is built based on Eclipse 3.4.2 using JAVA to evaluate the performance of HID. All simulation parameters are showed in table 3.

Table 3. Simulation Parameters

Parameter name	Value
Simulation area	10 km × 10 km
Number of deployed RSUs (NDR)	3, 7
Vehicles for each round (VER)	10-20, 20-30, 30-40
Velocity of vehicle (VV)	50, 70, 90, 110, 130 km/hr
Parking rate for vehicle	40 %, 60 %, 80 %
IP category	Class B
Maximum sensing area of AP	1 km
Maximum sensing area of vehicle	1 km
MAC protocol	IEEE 802.11p
Data rate	54 Mbps
Size of packet	352 bytes
Type of data transmission	Constant Bit Rate (CBR)
Movement model	Manhattan mobility model
Number of rounds (NR)	5000, 7500, 10000, 12500, 15000

Two equations are proposed to evaluate the rate of successful configuration which are Highest Traffic Assignment Rate (HTAR) and Average Assignment Rate (AAR)

in equation (4) and (5) respectively; the former considers the AP with the highest traffic, while the latter considers all APs. A_{ass} is the number of successfully assigned IP addresses by the AP with the highest traffic, A_{ava} is the number of all available IP addresses for the AP with the highest traffic, and n is the number of deployed APs. Total Configuration Time (TCT) and Average Configuration Time (ACT) in equation (6) and (7) are used to evaluate the configuration time for total vehicle and single vehicle respectively; n is the number of deployed APs, m is the number of configured vehicles, $T^{i,j}_{conf}$ is the configuration time of vehicle j successfully configured by AP i, and C_i is the number of all successfully configured vehicles by AP i.

$$HTAR = \frac{A_{ass}}{A_{ava}} \times 100\% \tag{4}$$

$$AAR = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{A^i_{ass}}{A^i_{ava}}\right) \tag{5}$$

$$TCT = \sum_{i=1}^{n}\sum_{j=1}^{m} T^{i,j}_{conf} \tag{6}$$

$$ACT = \frac{1}{\sum_{i=1}^{n} C_i}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} T^{i,j}_{conf}\right) \tag{7}$$

With a constant velocity of 90 km/hr for each vehicle, Fig. 8 and Fig.9 prove that HID outperforms CAC in $HTAR_i$ and AAR with a high rate of successful configuration. With 15,000 rounds executed, Fig. 10 and Fig. 11 shows that curves of $HTAR$ and AAR for CAC descends with higher velocity of vehicle, while that of HID remain smooth, more stable than CAC.

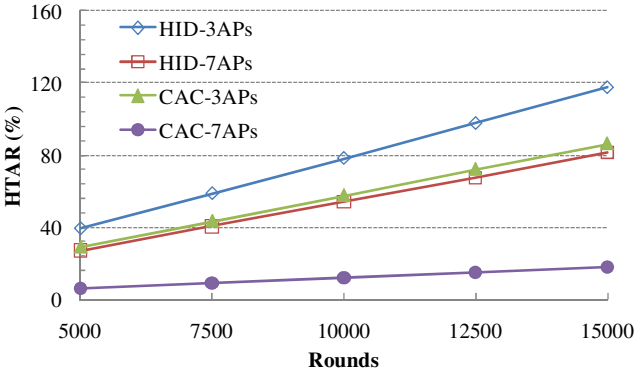


Fig. 8. HTAR with high traffic focused on round (VER = 30-40, VV = 90 km/hr)

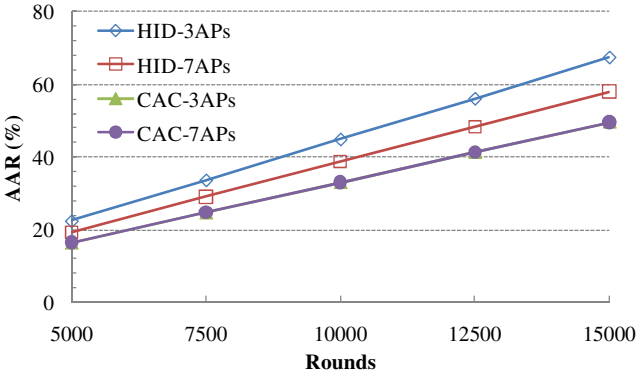


Fig. 9. AAR with high traffic focused on round (VER = 30-40, VV = 90 km/hr)

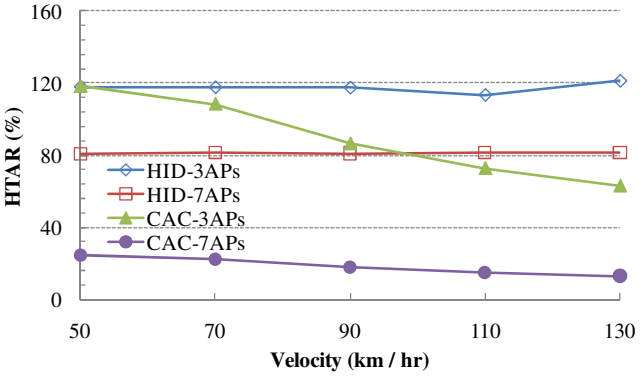


Fig. 10. HATR with high traffic focused on velocity (VER = 30-40, NR = 15000)

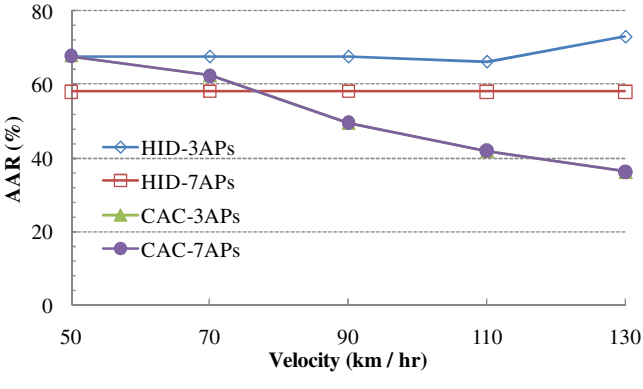


Fig. 11. AAR with high traffic focused on velocity (VER = 30-40, NR = 15000)

Fig. 12 and Fig. 13 show the results of *TCT* with constant velocity of 90 km/hr and *ACT* with constant velocity of 90 km/hr and 15,000 rounds executed respectively, which both prove that configuration time for HID is less than that for CAC.

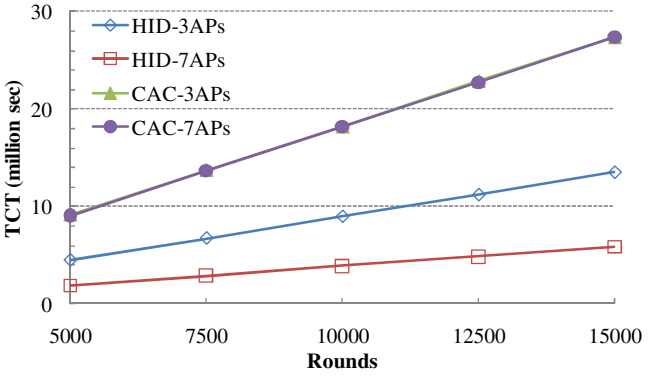


Fig. 12. TCT with high traffic focused on round (VER = 30-40, VV = 90 km/hr)

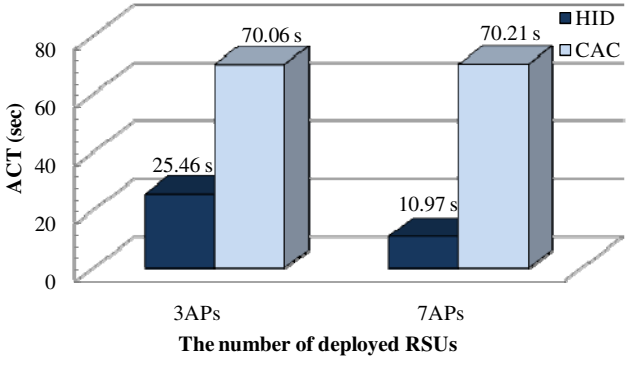


Fig. 13. ACT with high traffic (VER = 30-40, VV = 90 km/hr, NR = 15000)

5 Conclusion

Each AP endowed with a hierarchical function of DHCP server in HID can directly assign and recycle IP addresses, increasing the efficiency of IP assignment and decreasing the configuration time. With central BS which takes charge of IATF, the invalid IP addresses of parking vehicles can be detected and recycled immediately via periodical reporting mechanism. The results proved that HID significantly outperformed CAC in terms of IP assignment and configuration time. In the future, this system will be further modified for IPv6 and the fault-tolerance of the centralized mechanism will be discussed in detail.

References

1. Indrasinghe, S., Pereira, R., Mokhtar, H.: Host Address Auto Configuration for Mobile Ad Hoc Networks. In: The Proceedings of IEEE WCNC, pp. 1504–1511 (March 2005)
2. Haggerty, J., Pereira, R., Indrasinghe, S.: Conflict Free Address Allocation Mechanism for Mobile Ad Hoc Networks. In: 21st International Conference on Advanced Information Networking and Application Workshops, vol. 1, pp. 852–857 (May 2007)
3. WiFi Alliance, http://www.wi-fi.org/
4. Dedicated Short Range Communication (DSRC) Home, http://www.leearmstrong.com/DSRC/DSRCHomeset.htm
5. Das, S., Fazio, M., Gerla, M., Matematica, D., Palazzi, C.E.: Facilitating Real-time Applications in VANETs through Fast Address Auto configuration. In: 4th IEEE Consumer Communications and Networking Conference, pp. 981–985 (January 2007)
6. Mohandas, B.K., Liscano, R.: IP Address Configuration in VANET using Centralized DHCP. In: 33rd IEEE Local Computer Networks Conference, pp. 603–608 (October 2008)
7. Armitage, G., Branch, P.A., Pavlicic, A.M.: The Role of DHCP and RADIUS in Lawful Interception. CAIA Technical Report 040105A (January 2004)
8. Traffic Engineering Office. Taipei City Government, http://www.bote.taipei.gov.tw/main.asp
9. Zhao, J., Cao, G.: Vehicle-assisted Data Delivery in Vehicular Ad Hoc Networks. IEEE Transactions on Vehicular Technology 57(3), 1910–1922 (2008)
10. Arnold, T., Lloyd, W., Zhao, J., Cao, G.: IP Address Passing for VANETs. In: Proceedings of the 6th Annual IEEE International Conference on Pervasive Computing and Communications, pp. 70–79 (April 2008)
11. Chen, K.H., Dow, C.R., Lee, Y.S.: A Reliable Grid-based Routing Protocol for Vehicular Ad Hoc Networks. In: Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems, China, pp. 383–388 (October 2008)
12. Subramanian, A.P., Deshpande, P., Gao, J., Das, S.R.: Drive-by Localization of Roadside WiFi Networks. In: Proceedings of IEEE Communications Society Subject Matter, USA, pp. 1391–1399 (July 2008)
13. Xiao, B., Yu, B., Gao, C.: Detection and Localization of Sybil Nodes in VANETs. In: Proceedings of DIWANS, USA, pp. 101–108 (September 2006)
14. Khakbaz, S., Fathy, M.: Adding Reliability of Broadcast Methods in Vehicular Ad hoc Networks. In: Proceedings of The 2nd International Conference on Next Generation Mobile Applications, Services, and Technologies, pp. 385–388 (November 2008)
15. Mohandas, B.K., Naik, K., Goel, N.: A Service Discovery Approach for Vehicular Ad-hoc Networks. In: Proceedings of Asia Pacific Services Computing Conference, pp. 1590–1594 (October 2008)

VisMusic: Music Visualization with Interactive Browsing

Jia-Lien Hsu and Wei-Hsien Chien

Department of Computer Science and Information Engineering
Fu Jen Catholic University, Sinjhuang City, Taipei County 24205, Taiwan, R.O.C.
alien@csie.fju.edu.tw

Abstract. In this paper, we develop a system that provides a new way to browse streaming music on mobile device. Our system visualizes music information and social interaction history for users. The main design goals of our system are twofold. First, we provide a timeline-based way to capture music information at a glance. Second, we keep social interaction history and show by visualization to release exploration. Finally, we implement our prototype and perform evaluation to show the feasibility and effectiveness.

Keywords: visualization, streaming music, mobile device.

1 Introduction

With large amount of streaming music application on mobile device, users get the information only by word. There are few system designs with visualization which show the social interaction history on the user interface. Thus, it is a challenging task to visualize music information on mobile device with limit-size screen. The expression way of information and design of user interface may be crucial to browsing and listening to music.

In this paper, we design a visualization system to visualize music information and social interaction history for users. The system we design, named *VisMusic*, is an interactive visualization tool that browses an album archive and a song archive on mobile device. Our system also provides a streaming service on mobile device, in which users could listen to music on the Internet. Users may see the information via our visualization that will help user select the music they want. We have three distinguishing features of major contribution in this paper. First, the music information visualization helps users to comprehend music at a glance. Second, the visualization of social interaction history help users to identify which music is the most popular. Third, we design it with the gesture control, in which user can use fingers or stylus to slide those albums. Thus, it will help users to handle our system more easily.

The rest of the paper is organized as follows. We introduce some related work on information visualization in Section 2. In Section 3, we propose our design and method of music information visualization. In Section 4, we show the experiment results. Finally, we give a conclusion and future work in the Section 5.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 367–375, 2010.

2 Related Work

Some research on the information visualization problems currently focus on visualizing the content of information and interactive with users. In addition, there are fewer applications on mobile device designs with visualization. We show some information visualization research and mobile device application as follows.

Indratmo, Vassileva, and Gutwin design *iBlogVis* helping people to explore blog's content, which is different from traditional blog viewer [5]. Users get overview of a blog via *iBlogVis* that can quickly browse the key points on content and comments. That is the way to save users time and get what users want. However, when there have too many entries of blog, which would so difficult to compare with the tag on the top.

Viégas and Smith develop a visualization tool let users know the authors contribute to the post over a period [16]. For *Newsgroup Crowds*, users see the population of author in a particular newsgroup. For Authorlines, that is showing the authors activity in the newsgroup. Through the past posts of authors, we could know how they can be trusted.

Crampes, Villerd, Emery, and Ranwez design an automatic playlist composition tool, which make use of the expertise of DJs and users personalization to generate playlist [1]. MBox use music landscape to be visual indexing and use the music landscape to show the playlist path.

Laurier, Sordo and Herrera create a mood visualization tool, named *Mood Cloud* [9][10]. This application is a real-time system which provides automatic music mood prediction from audio content and browse music by mood.

Vincent Castaignet and Frédéric Vavrille create an interactive webradio, named *Musicovery*, which help user searching, finding and playing music [13]. This application is a webradio with visualization that user can filter music by mood, year, genre, and dance.

We compare the user interface of KKBOX, ezPeer, Last.fm, Musicovery, and Pandora Radio, respectively [3][8][12][13][14]. All of them are mobile device applications. They provide an online streaming music service for users. Users can express their preference for the music, but they can not get preference rate and hit rate from all users. The iTunes user interface is a mobile device music player, which is offer an offline service [7]. So, this application does not have social network service. Users may rate the song by themselves, but can not know other users behavior. However, above-mentioned applications do not have visualization in their user interface. We could see they show the information just a listing. In addition, we could not get music ID3 tag and social interactive history on the applications.

Compared to previous work on visualization and mobile device application, our object is in different way. First, we would like to create a clean and pure look in our interface. Second, we design our system with information visualization for mobile device. Third, we show ID3 tag and social interaction history on the user interface.

3 System Design and Method

In this section, we describe our system design and method as follows.

3.1 Design Goal

The main design goal of our system is a visualization of music information to help user exploring music. Our system design follows the information visualization principles,"Overview first, zoom and filter, then detail on demand" [15][17]. Since the main platform of VisMusic is target on mobile device. We also apply gesture control with finger and stylus in our application.

For human being, seeing a graphic is more sensitive than reading. For this reason, we hope visualization could help user reduce the search time. VisMusic visualize not only the information of music, but also the history of user interaction.

Our system used another user's history interaction as the way of social recommendation. In the end, we use their preference history to be a playlist.

3.2 The System Architecture

As shown in Fig. 1, we describe our system architecture in both server-side and client-side. The server will describe in Section 3.3, as well as the client GUI in Section 3.4.

3.3 System Design (Server)

Our system's music objects are store in the server. In order to listening music online with streaming music, we used songs in the format of MP3. There is not much memory in the mobile device. We store our analysis of the collection in the XML files. That can reduce the memory when we run the system. The PHP file is send and access the data between Flash and XML. Because of Flash just can download the XML file, that can not save the data to server. Therefore, we used it to connect two parts.

Our system keeps a user account, listening history and listening preference as user profile. We illustrate user profile module as follows.

User Profile

Our user profile contains the username, password, listening history, listening preference, and playlist. We will keep log as listening history, which remains the song when user listening the song and have not to consider the duplicate problem. The listening history is access in XML file with song URL. The user profile that we keep listening history log is contains username, password, and listening history. We also keep preference log, which remains the song when user express personal preference. The log of preference will consider the duplicate problem. The listening preference is access in XML file with song URL. However, if users dislike that song, they can take off the song log. In the end, user can use their log be a playlist that is base on listening preference

XML: Music Metadata

Fig. 3 shows the tree structure of XML file with our analysis of the collection. As shown in Fig. 3, we store hit rate data and preference rate data in every songs separately. The music ID3 tag that store in music and XML is search form internet in advance. We put the ID3 tag under the <tag> like the tree structure level 3 and 4 (see Fig. 3). The <file> tag of level 3 is store song duration and duration in seconds.

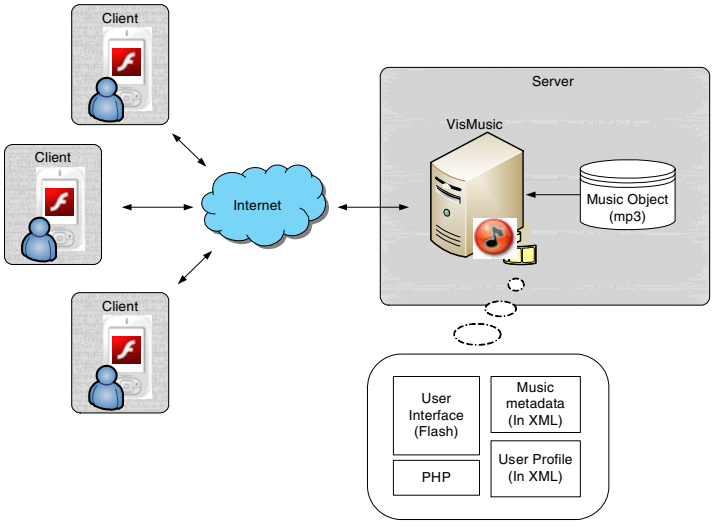


Fig. 1. The system architecture of *VisMusic*

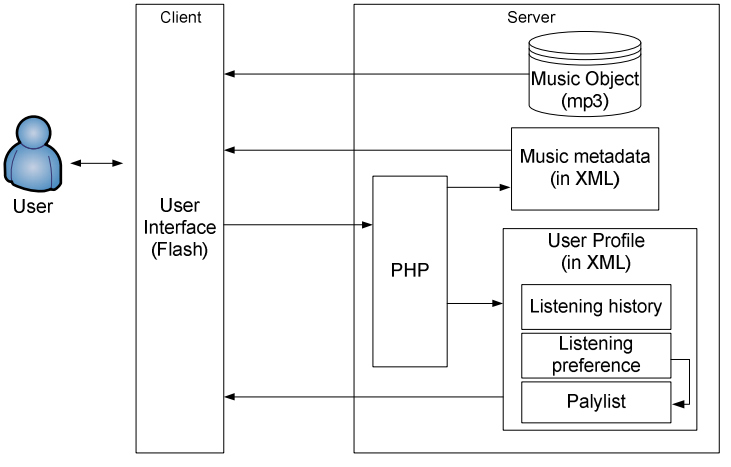


Fig. 2. The system procedure

3.4 Visualization and Interaction (client)

We create the visualization module with Adobe Flash. Users have to create an account or login first when they want to start the system. Our system display album entries according to publish time in chronological order. The VisMusic visual items in main frame arrange along timeline (see Fig. 4). The whole frame consists of two parts that are separate by timeline. The part above entries timeline visualizes the album information that includes tags, and songs line. The tags show album title, artist, and

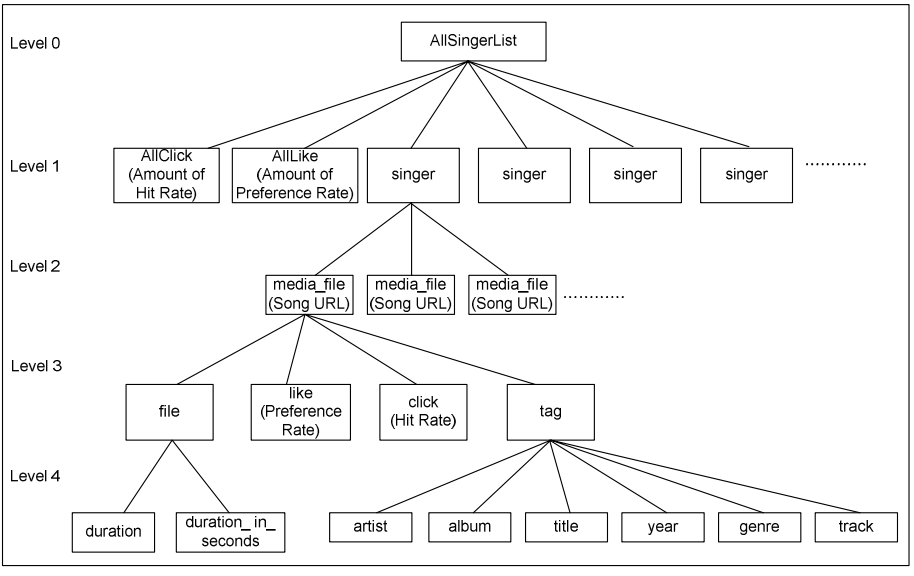


Fig. 3. The tree structure of XML file with music metadata

publish time. An album cover represents an album entry. The entry is connecting with album frames as Fig. 5 shows. The length of a songs line between album entry and timeline is representing the number of album songs, which is combining by ellipse.

The part below timeline visualizes the social interaction history [6] result in an album. The length of a line represent hit rate of an album. The dimension of a heart displays the preference rate of an album. There are two kind of status here. If the preference rate value cast up is positive it will be a pink heart; otherwise, it will be a brown broken heart. The left legend panel is a button that illustrates the function and how to control. It is a pop panel when user moves cursor over the see-through button. Fig. 4 shows the frame without legend panel.

The VisMusic visual items in album frame was arrange along track-line (see Fig. 5). Album frames consists of two parts, which were separate by track-line. The part above track-line visualizes the songs information that includes tags, entries and duration line. The tags are showing track and title. A song entry is represented by a CD shape. The entry is connecting an mp3 player panel (see Fig. 6), which can play the song that user select. The duration lines represented the duration of a song, which can compare with another songs let user know which one is short or long.

The part below track-line visualizes the social interaction history result in a song. The hit rate line and preference rate heart have similar meanings as main frame. The length of a line represent hit rate of a song. The dimension of a heart displays the preference rate of a song. The left pop panel is illustrates about album frame.

The mp3 player panel (see Fig. 6) is a pop panel when users select a song that will appear at once. We show the song ID3 tag on the panel. There have album, artist, title, publish year, and genre. The panel's top have two hearts, witch are preference button. The red heart represents like and the brown broken heart represents dislike. We will keep the log when user presses the button. This operation is influence listening preference and playlist. It also influences the social interaction history.

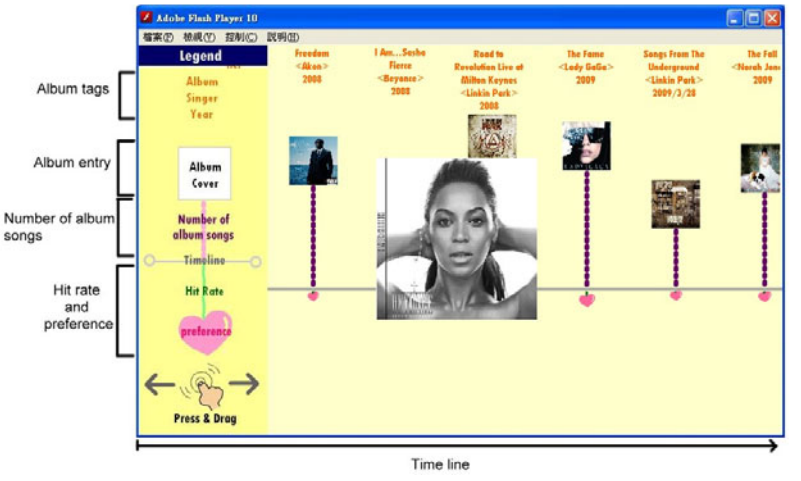


Fig. 4. The main frame of VisMusic with legends

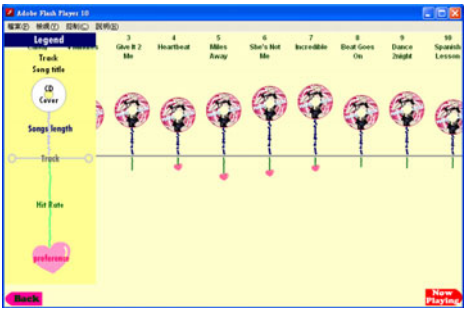


Fig. 5. The album frame

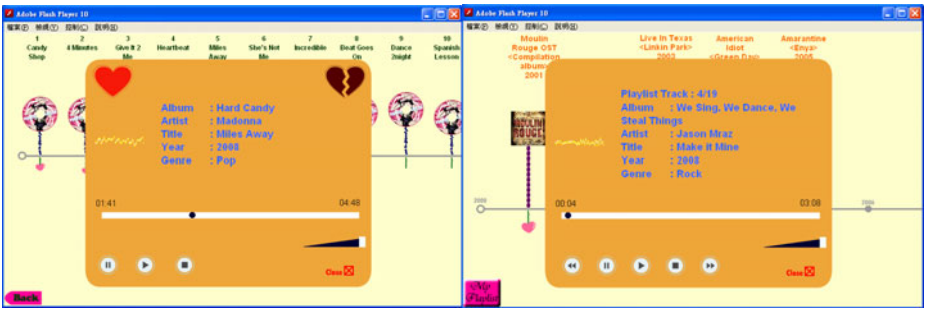


Fig. 6. MP3 player panel and Playlist MP3 player panel

We design a "My Playlist" button in the main frame. It is a pop panel with mp3 player (see Fig. 6). Users can listen to all the songs they like at that panel. The playlist is base on listening preference.

Our visualization tool: VisMusic starts from an overview of albums archive (see Fig. 4). We show album by album and display all albums entry along the timeline. Than user can zoom and filter by album entries. It displays all songs entry along the track- line (see Fig. 5). Finally, users can see and listen to song content through mp3 player panel by clicking on a song entry (see Fig. 6).

4 Evaluation

Our system offers a new way to browse music information and social interaction history, and there were few music visualization systems for comparison. Therefore, we focus on requesting feedback for the visualization design and assessing the effectiveness of our approach. The subjective measured is the user satisfaction with our system. We show the details of experiments as follow.

4.1 Participant

We ask twenty users (ten males and ten females) to use our system. Subject age ranges from twenty to thirty years old. There are fifteen participants having computer science background and the other five participants have not.

4.2 Experiment Set-Up

VisMusic is implemented as a mobile device application with Flash player. The music dataset in our system consists of about two hundred songs in the format of mp3. The genres of songs include popular, R&B, rock, vocal pop, alternative rock, and jazz. The languages of songs are in English and the release year is ranging from 2001 to 2009.

4.3 Procedure and Results

First, we give users a brief introduction and ask them to use our system. Users are required to report the *familiar time* in questionnaire. The *familiar time* is roughly the elapsed time from "starting application" to "listen the first song". The average time of all participants needed is 3.7 minuets. When users finish the evaluation tasks, we ask six questions for each user. The score ranges from 1 (Very disappointed) to 5 (Perfect).

As summarized in Table 1, most of users give high rating score of six questions. Still, there are few users giving us low rating score. In the following, we further ask these users the reasons why they do not like our system design.

As shown in Table 1, users thought our system is easy for use. According to the user profile and the results of question 3, we concluded the way of showing the "social interaction history" really has influence on users. More than half users select the song with highest hit rate and listen to the performance at first time. There is one

Table 1. Results of the questionnaire

No.	Questionnaire items \ rating score	1	2	3	4	5
1.	Is our system easy to use?	0	0	6	6	8
2.	What do you think for about the look of our user interface?	0	3	3	10	4
3.	If the system have music you never listen, will you select the music to listen because of another user's recommendation?	0	1	4	8	7
4.	Is our visualization helping you browse music more easily and find out the music more quickly?	0	0	7	10	3
5.	Is the display way of hit rate easy to understand?	0	1	9	5	5
6.	Is the display way of preference rate easy to understand?	0	1	7	8	4

user give low rating score of question 3, since the user (user #11) says, "social interaction history have a little influence for him," and the user likes to listening music album by album.

Regarding the question 5 shown in Table 1, there is one user give low rating score, since the user like to listening music by following the order of track in album. Therefore, the user will not pay attention to the hit rate.

Regarding the question 6 shown in Table 1, there is one user give low rating score. According to the user feedback, the user says, "we show the amount of preference value which only indicates summarized results of preference information and the summarized result is not enough to realize for her. The user suggested our system should have showed both the *like value* and *dislike value* on the user interface.

In summery, we may say our system works properly and fast to understand. Most of users express high satisfaction with our system. Thus, we may say our visualization is successfully for users.

5 Conclusion and Future Work

We discuss the design and implementation of VisMusic, which is an interactive visualization application for helping user exploration of music. By the mobile device service, we provide an online streaming music service to users. And, users could exploration music via VisMusic. From the visualization, besides the album title and the publish time of an entry, we figured the length of hit rate and the heart size of performance rate on the entry are influence the decision. The usability of our system is evaluated using subjective performance measure. The results of our evaluation show that users satisfaction is high.

Some future work is described as follows. We use social interaction history as recommendation way that is not enough. We would like to use methods of data mining to analysis user profile for designing more recommendation mechanism. Thus, our system would learn the user behavior from user profile and automatically generate

playlists for users. Moreover, some users would like to manipulate playlists ad hoc. Some users would like to show the personal preference. These are guidelines to revise and improve the user interface of our system.

Acknowledgments. This research was supported by Fu Jen Catholic University with Project No. 409831044042, and sponsored by the National Science Council under Contract No. NSC-98-2622-E-030-001-CC3.

References

1. Crampes, M., Villerd, J., Emery, A., Ranwez, S.: Automatic Playlist Composition in a Dynamic Music Landscape. In: Proceedings of the International Workshop on Semantically Aware Document Processing and Indexing (2007)
2. Cunningham, S.J., Nichols, D.M.: Exploring Social Music Behavior: an Investigation of Music Selection at Parties. In: Proceedings of the 10th International Society for Music Information Conference, Kobe, Japan (2009)
3. ezPeer, http://web.ezpeer.com/
4. Harris, J., Kamvar, S.: We Feel Fine (2005), http://www.wefeelfine.org/
5. Indratmo, J.V., Gutwin, C.: Exploring Blog Archives with Interactive Visualization. In: Proceedings of the ACM Conference on Advanced Visual Interfaces (2008)
6. Indratmo, Vassileva, J.: Social Interaction History: A Framework for Supporting Exploration of Social Information Spaces. In: Proceedings of the IEEE International Conference on Computational Science and Engineering, Vancouver, Canada, vol. 4, pp. 538–545 (2009)
7. Apple-iTunes, http://www.apple.com/itunes/
8. KKBOX, http://tw.kkbox.com/index.html
9. Laurier, C., Herrera, P.: Mood Cloud: A Realtime Music Mood Visualization Tool. In: Proceedings of the 2008 Computers in Music Modeling and Retrieval Conference, Copenhagen, Denmark, pp. 163–167 (2008)
10. Laurier, C., Sordo, M., Herrera, P.: Mood Cloud 2.0: Music Mood Browsing Based on Social Networks. In: Proceedings of the 10th International Society for Music Information Conference, Kobe, Japan (2009)
11. Lott, J., Schall, D., Peters, K.: ActionScript 3.0 Cookbook: Solutions for Adobe Flash Platform and Adobe Flex Application Developers. O'Reilly Media, Sebastopol (October 2006)
12. Last.fm, http://www.last.fm/
13. Musicovery: interactive webRadio, http://musicovery.com/
14. Pandora Internet Radio, http://www.pandora.com
15. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proceedings of the IEEE Symposium on Visual Languages, Boulder, Colorado, pp. 336–343 (1996)
16. Viégas, F.B., Smith, M.: Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS 2004) - Track 4, vol. 4 (2004)
17. Ware, C.: Information Visualization: Perception for Design, 2nd edn. Morgan Kaufmann, San Francisco (2004)

The Study of Plagiarism Detection for Object-Oriented Programming Language

Jong-Yih Kuo and Wei-Ting Wang

Department of Computer Science and Information Engineering
National Taipei University of Technology
jyku@ntut.edu.tw

Abstract. The purpose of this thesis is to study the plagiarism detection method aiming at C++ programs. We proposed the corresponding preventive measures by summarizing the common types of plagiarism attack through observation and statistical analysis. Using text analysis, structure analysis, and variable analysis would prevent misjudgment. Finally, we implemented the CPD(C++ Plagiarism Detection) system and compare it with other existing systems, and the experimental result shows us that our system can detect more kinds of plagiarism attacks than other existing systems.

Keywords: Programming Language, Plagiarism.

1 Introduction

In the era of electronic information, we can get a lot of information through the internet. In the internet, we often find similar documents. Therefore, students can search information throughout the Internet in order to hand out assignments. In this way, there would be similar parts in students' assignments. To avoid these issues, we have to compare the student's assignment and find out these similar parts. It would be a hard work for human eye.

There are number of researches [7][8][9] which focused on detecting plagiarism, using different methods to search the plagiarized part, and the accuracy should be the same as manual detection, which would be our direction of research. Some typical plagiarism patterns [6] are listed as follows:

1. Adding or changing comments.
2. Adding redundant spaces or lines.
3. Adding useless variables or functions.
4. Replacing function's name or variable's name.
5. Changing the order of process block's location.
6. Replacing control structure.

The 1, 2, 3, 4 plagiarism attack methods are more common and can be easy to detect. The 5, 6 plagiarism attack methods are used by more understanding of the semantics of program and can be difficult to detect. According to these problems, we would analyze programs by three phases - text, structure and variable.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 376–386, 2010.

The remainder of this paper is organized as follows. Section 2 briefly introduces the methods we used. Section 3 describes the system architecture and implementation of the CPD (C++ Plagiarism Detection) system. In the case Study, compare it with manual computing method, JPlag and Sid. The final Section is our conclusion.

2 Plagiarism Detection Methods

2.1 Text Analysis

Our approach applied Winnowing algorithm [4] for text analysis. There have four important variables in this algorithm. t means length of string. k means length of gram. w means the size of window. g means the distance between grams. The gram is to cut the string by block, and each gram differs by g length. k-gram means every continuing k characters in the string is a block. After cutting the program string into grams, all grams are processed into a sequence of hash value as its fingerprint. Dividing the sequence of windows by the number of hash values in one window is defined by user.

The restriction of algorithm's variable: (1) the length of gram (k) should be greater than zero and smaller than length of string. If the value of k is too smaller, the similarity will be too high, and it will lose its meaning. (2) The number of characters (g) between grams should greater than zero and smaller than k+1. The value of g will be 1 in general. But it might possibly increase the number of characters between grams in order to decrease computing time for analysis longer string. If value of g is greater than k, it will cause some information lost. (3) The length of window (w) should greater than k-1 and smaller than k+1. User can set the value depending on the size of document. Algorithm flow chart is in Fig. 1.

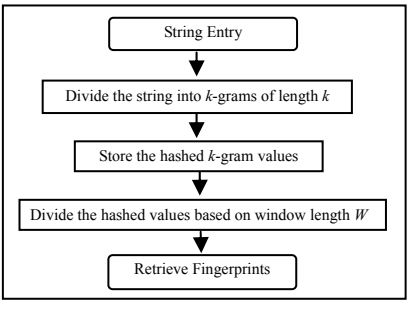


Fig. 1. Converting string to fingerprint flow chart

1. String processing: Remove punctuation or space and change capital letters to lowercase letters.
2. String dividing: Divide the program string into grams by user defined k.
3. Hash computing: Apply a hash function to a sequence of grams to produce a sequence of hash values.
4. Producing windows: Put a sequence of hash value into the windows to produce a sequence of the windows.

5. Producing fingerprints: Take out the smallest value in the window and record its position. It convenient to figure out whether it takes out the same fingerprint.

6. Comparing fingerprints: Compare fingerprints of two document to compute there similarity.

For example, the text: I love Pony Pony love me. Setting variable k = 5, w = 4, g = 1, after string processing is iloveponyponyloveme. After divided string is:

ilove lovep ovepo vepon epony
ponyp onypo nypon ypony ponyl
onylo nylov ylove lovem oveme

After hash computing are 23 4 7 6 9 20 18 17 2 15 14 21 12 0 19. Producing windows are (23, 4, 7, 6) (4, 7, 6, 9) (7, 6, 9, 20) (6, 9, 20, 18) (9, 20, 18, 17) (20, 18, 17, 2) (18, 17, 2, 15) (17, 2, 15, 14) (2, 15, 14, 21) (15, 14, 21, 12) (14, 21, 12, 0) (21, 12, 0, 19). Producing fingerprints are [4, 1] [6, 3] [9, 4] [2, 8] [12, 10] [0, 13]. After analyzing the document F1 and F2, we get the fp_1 and fp_2 respectively. Comparing method is taking fp_1 as the base comparing to fp_2. Then take the number of match to divide by number of fingerprints in the fp_1. After that, fp_2 is taken as the base comparing to fp_1. Applying the equation (10), we will get the text similarity. As Fig. 2 is the matching algorithm.

$$Sim(fp1, fP2) = \frac{Match(fP1, fP2), Match(fP2, fP1)}{2} \quad (1)$$

```
Match(fP1, fP2){
    int match = 0;
    for(int i = 0 ; i < fP1.size() ; i++)        {
        for( int j = k ; j < fP2.size() ; j++)
            if(fP1.value(i) == fP2. value(j))
                if((fP1. value(i+1) == fP2. value(j+1)) &&
                (fP1.position(i+1)- fP1.position(i)== fP1.position(j+1)-
                fP1.position(j))) {
                        match++;
                        break;
                }
    }
    return match / fP1.size()*fP2.size();
}
```

Fig. 2. Comparing algorithm

2.2 Variable Analysis

Donaldson used statistical methods [3] on the program plagiarism detection. But this method doesn't have the concept of layered structure, so it can not show enough properties of the structure. Variable Analysis is divided into two phases: the first phase is information collection. The second phase is information analysis.

2.2.1 Information Collection

The method records three kind information for each variable: (1) Basic information including data type and name. (2) Statistical similarity including the type of control structure. (3) The appeared level.

2.2.2 Information Analysis

Table 1 and Table 2 are information of two variables. The statistical similarity is 60.0%, and the formation similarity is 100.0%, so the average similarity is 80.0%.

Table 1. Count's information

Variable name	Data type	statistical similarity	formation similarity
count	int	Assign(0)	((
		for(0)	(
		dec(1)	((i
		for(0)	((
		inc(1)	(

Table 2. Time's information

Variable name	Data type	statistical similarity	formation similarity
time	double	Assign(1)	((
		for(0)	(
		dec(1)	((i
		for(1)	((
		inc(1)	(

2.3 Structure Analysis

2.3.1 DCS Tree

This study extended our previous work [5] to process the structure analysis. A program structure is converted a tree structure by the algebraic expression F(p) as shown Table 3. And this expression is used for text comparison to get a similarity.

Table 3. Symbols and corresponding descriptions

signal	description
S	seq
F	if
W	while
R	r
X	Another algebraic expression
L	λ, denotes null.
o	include
c	continue
+	Alternative operator
(	Layer(j)+1, denotes a move to child layer
)	Layer(j)–1, denotes a move to parent layer

2.3.2 Function Comparison

Text and structure analysis might cause a misjudgment for just change of the program's block location. The function comparison method can fix this problem. The process is to compare all the parameters and to apply text analysis according to the function range. Each highest result is taken to compute the average. If the result of the function comparison is higher than the text analysis, the text analysis result is replaced with the function comparison result.

2.3.3 Class Comparison

This method records three types of information for each variable as Fig. 3: (1) the data type and name of class attribute. (2) information about class member function. (3) information about class inheritance.

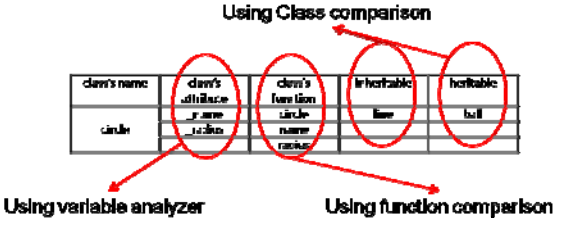


Fig. 3. Class information

3 Case Study

3.1 Case Description

Different users have different writing style of program. Changing the variable name, function locations would make plagiarism detection misjudge. So this research uses three kinds of analysis to prevent the misjudgment.

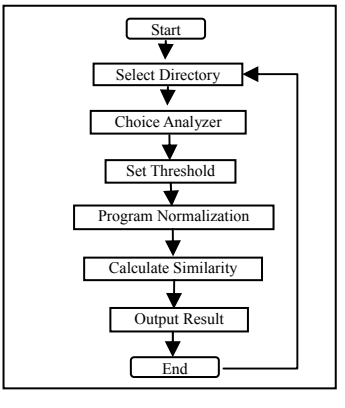


Fig. 4. System flow chart

3.2 System Flow

Fig. 4 describes our CPD systems flow. First the user must choose the program's document for comparison. Next, choose the analysis method that the user needs, and set the analyzer's threshold. After analyzing, the system would show the result of all comparison programs on the user interface. If the user wants to detect other programs, he need to repeat the steps above.

3.3 System Architecture

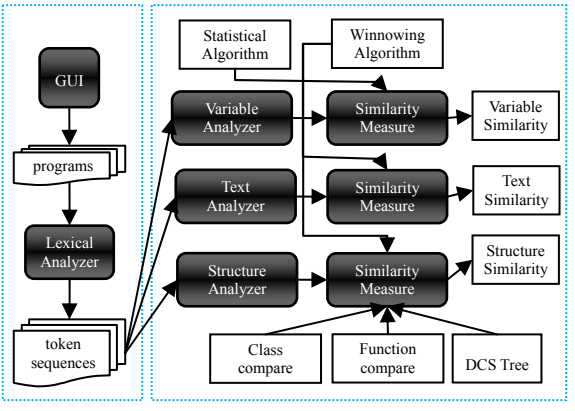


Fig. 5. System architecture

Fig. 5 is the system architecture. The user sets related parameter through GUI, then the system can compare multi programs immediately. The major modules are introduced as follows:

(1) GUI: The user can choose the path of program, choose the analyzer, and set the corresponding threshold.
(2) Lexical analyzer: Normalize the program, derive the program to a sequence of tokens, and set some attributes.
(3) Text Analyzer: For text analysis and comparison.
(4) Structure Analyzer: For Structure analysis and comparison.
(5) Variable Analyzer: Variable analysis and comparison.

3.4 System Implement

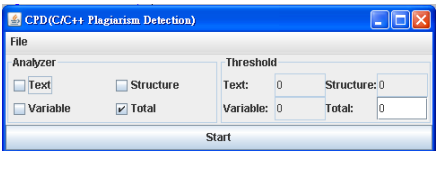


Fig. 6. Initial of user interface

Fig. 6 is the initial dialog of system. File menu is for choosing the path of the program. The user can choose Analyzer for analysis the program. There are three kinds of analyzers - text, structure, and variable. The total checkbox represents the text, structure, and variable analysis. The user can set the similarity threshold. When finished the above steps, the analyzing process is stared by pressing the start button.

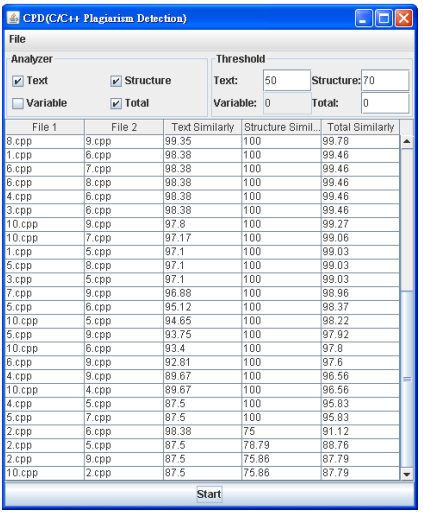


Fig. 7. The result of execution

The result of the analysis is shown in Fig. 7. The first column represents the first document of comparison combination. The second column represents the second document of comparison combination. The third column is the result of text analysis. The fourth column is the result of the text analysis. The fifth column is the result of the average. The result of analysis is sorted by average. If the result of analysis does not fit the threshold, it would not show on the GUI.

3.5 Experiment Design

The purpose of sample design is described as follow:

1. Change the function location.
2. Change the variable name.
3. Add more no use function or variable.
4. Declare variable in a different way.
5. Add or change comment.
6. Normalization program.

Table 4 is the description of basic sample codes that include the program name and the corresponding plagiarism attack. Table 5 is the description of object-oriented property sample codes.

Table 4. Basic sample codes

programs	contents
1.cpp	the original program.
2.cpp	changes function location according to 1.cpp.
3.cpp	normalizes program according to 1.cpp.
4.cpp	changes variable name according to 1.cpp.
5.cpp	adds more variable according to 1.cpp.
6.cpp	declares variable in different way according to 1.cpp.
7.cpp	adds more no use function according to 1.cpp.
8.cpp	adds or changes comment according to 1.cpp.
9.cpp	adds other useful source codes.
10.cpp	changes code location according to 9.cpp.

Table 5. Object-oriented sample codes

programs	contents
1.cpp	is a inheritance program.
2.cpp	changes the variable and function's name, location according to 1.cpp.
3.cpp	is an abstract class function overriding.
4.cpp	changes name, location according to 3.cpp.
5.cpp	is operator overloading.
6.cpp	changes the variable and function's name, location according to 5.cpp.
7.cpp	is a Function template.
8.cpp	is a virtual Inheritance.
9.cpp	is a Member Function override.
10.cpp	is a function overloading.

3.6 Experimental Results

The Experiment is separated by two parts. The first part is to compare the result of analysis of basic programs between CPD and manual detection. The second part is to compare the result of analysis object-oriented programs between CPD and manual detections. First, we defined methods of manual computing plagiarism similarity. (1) Omits the commands which includes output (printf, cout), comment. (2) Normalization program. (3) Calculates similarity.

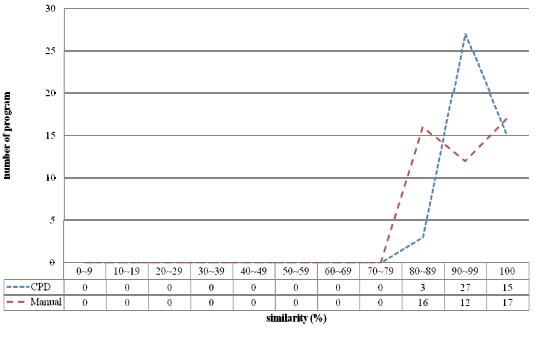


Fig. 8. Comparison of CPD with manual detection for basic programs

Using equation (2) to calculate similarity, *Sim(A,B)* represents the similarity for base program A. The result of *Sim(A,B)* is the match line. Line(A) represents number of program A's lines. Take the maximum similarity to be the final similarity.

$$Max\ (\frac{Sim\ (\ A\ ,B\ )}{Line\ (\ A\ )}, \frac{Sim\ (\ B\ ,A\ )}{Line\ (\ B\ )}) \tag{2}$$

3.6.1 Experiment 1

As Fig. 8 shown, we compare the result of manual detection with the result of CPD system. The CPD system can find 15 pairs of total match.

3.6.2 Experiment 2

As Fig. 9 shown, we compare the result of manual detection with the result of CPD system, the CPD system can find 15 matches.

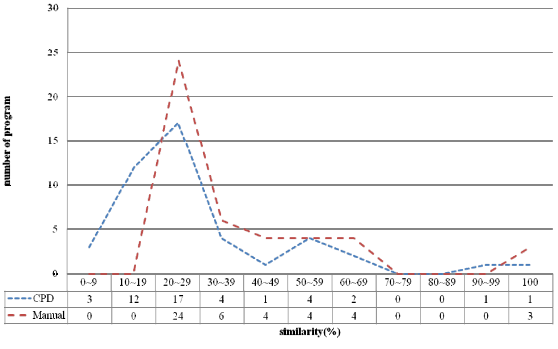


Fig. 9. Comparison of CPD with manual detection for object-oriented property programs

3.7 Related Research Comparison

The comparison of CPD system, Sid [2], JPlag [1] and manual detection is shown as Fig. 10. The results show that CPD system finds 15 pairs. The JPlag only detected 3 pairs, and Sid didn't detect anything.

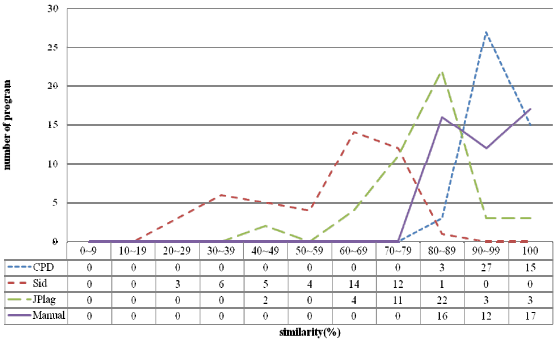
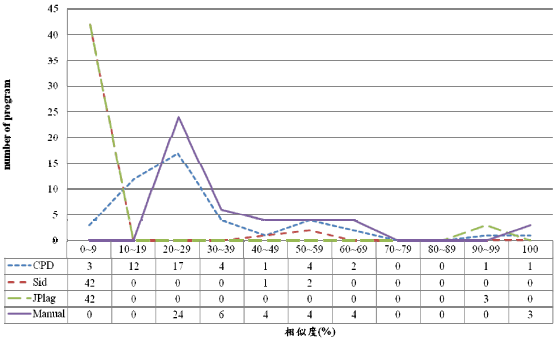


Fig. 10. Comparison of CPD with other detection methods for basic programs

Fig. 11 is the result of plagiarism detection for object-oriented property programs. The Sid and JPlag cannot find any pair of total match. From the two experiments, the CPD system can detect similar programs more accurately.



	0~9	10~19	20~29	30~39	40~49	50~59	60~69	70~79	80~89	90~99	100
CPD	3	12	17	4	1	4	2	0	0	1	1
Sid	42	0	0	0	1	2	0	0	0	0	0
JPlag	42	0	0	0	0	0	0	0	0	3	0
Mamal	0	0	24	6	4	4	4	0	0	0	3

Fig. 11. Comparison of related research

Fig. 12 is the condition with the same program code adding comments that would make Sid misjudge.

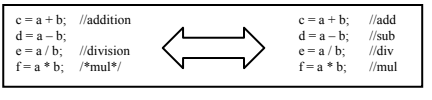


Fig. 12. Case 1

As Fig. 13 shown, Sid and JPlag cannot detect this kind of program. The left program omits the brackets. It will make Sid and JPlag misjudge.

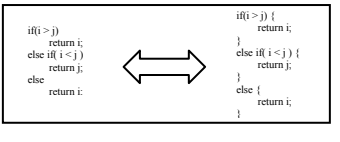


Fig. 13. Case 2

Fig. 14 show the program declares the variable in a different way. For the CPD system, variable analyzer categorizes these two programs into the same kind.

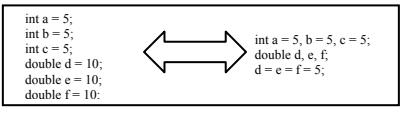


Fig. 14. Case 3

4 Conclusion

The proposed CPD system applied text, structure, and variable analysis to detect the plagiarism more effectiveness.

Acknowledgment

This work was supported by the National Science Council under grant number 98-2752-E-008-001- PAE and NSC- 98-2220- E- 027 -009.

References

[1] Malpohl, G.: JPlag: Detecting Software Plagiarism, http://www.ipd.uka.de:2222/index.html
[2] Chen, X., Francia, B., Mckinnon, B., Seker, A., Li, M.: SID: Plagiarism Detection, http://genome.math.uwaterloo.ca/SID/
[3] Donaldson, J.L., Lancaster, A., Sposato, P.H.: A Plagiarism Detection System. In: Proceedings of the Twelfth SIGCSE Technical Symposium on Computer Science Education (1981)
[4] Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: Local Algorithms for Document Fingerprinting. In: 2003 ACM SIGMOD International Conference on Management of Data (2003)
[5] Kuo, J.Y., Chu, L.: Intelligent Code Analyzer for Online Course Management System. In: Proceedings of the 3rd ACIS International Conference on Software Engineering Research (2005)
[6] Ji, J., Park, S., Woo, G., Cho, H.: Understanding the evolution process of program source for investigating software authorship and plagiarism. Digital Information Management (2007)
[7] Chow, T.W.S., Rahman, M.K.M.: Multilayer SOM With Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection. IEEE Transactions on Neural Networks 20, 1385–1402 (2009)
[8] Shen, Y., Li, S.C., Tian, C.G., Cheng, M.: Research on Anti-Plagiarism System and the Law of Plagiarism. Education Technology and Computer Science, 296–300 (2009)
[9] Yang, S., Wang, X.: A Visual Domain Recognition Method Based on Function Mode for Source Code Plagiarism. Intelligent Information Technology and Security Informatics, 580–584 (2010)

A Distributed Sleep Scheduling Algorithm with Range Adjustment for Wireless Sensor Networks

Kei-Chen Tung[1], Jonathan Chun-Hsien Lu[1], and Hsin-Hung Lin[2]

[1] Dept. of Computer Science and Information Engineering
Fu Jen Catholic University, Taipei, Taiwan
Distance_kk@hotmail.com, jonlu@csie.fju.edu.tw
[2] Graduate Institute of Applied Science and Engineering
Fu Jen Catholic University, Taipei, Taiwan
Jerome.lin@msa.hinet.net

Abstract. The network coverage for wireless sensor networks is an important issue because the information may not be considered useful when the total coverage drops below a certain required level. One way to maintain good coverage and extend the network lifetime is to schedule some sensor nodes to sleep between active cycles, as well as dynamically adjust the sensing ranges of active sensors. In this paper, we propose an algorithm to determine if a sensor node should sleep based on its residual energy and the size of the overlap area between the sensor and its neighbors. For those sensors that remain active, we use our algorithm to compute the sensing range of each active sensor such that the total coverage is maintained above a user-specified requirement for as long as possible. Simulation results show that our proposed scheme achieves a better performance in providing the user-required coverage and extending the system lifetime than the Coverage-Aware and Random-Sleep methods.

Keywords: wireless sensor network, sleep scheduling, network coverage, dynamic sensing range adjustment.

1 Introduction

Due to the fast development of embedded system and wireless communication technologies in recent years, wireless sensor network (WSN) has become one of the most important research areas. A Sensor network is a network system composed of multiple small and inexpensive devices deployed in a region to provide monitoring and communication capabilities for commercial and military applications including fire detection, asset tracking, habitat monitoring, and security surveillance [1-6]. In general, a wireless sensor network consists of a base station and many sensor nodes operating on small-sized batteries as shown in figure 1. When a sensor node detects an event within its coverage area, it would generate a report and send it in a data packet to the base station. The base station is responsible for collecting data from all the sensor nodes for further processing and decision making. A WSN may contain hundreds or even thousands of sensor nodes and it is desirable to operate these nodes as energy efficiently as possible. A sufficient number of sensor nodes must be deployed in order to provide a satisfactory performance.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 387–397, 2010.

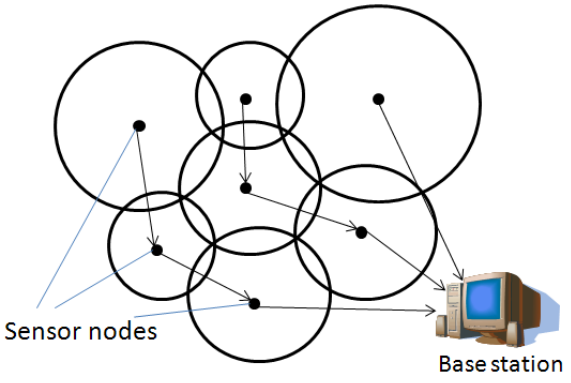


Fig. 1. An example of wireless sensor network

Some sensor nodes will stop functioning due to electronic breakdown or energy exhaustion after the network system has operated for a period of time. Dead sensors can result in holes in coverage or packet routing, which may lead to serious network degradation [7-10]. The system cannot perform its desirable function satisfactorily since the base station will not be able to collect sufficient information when the coverage is too low. One way to prevent this situation is to try to make the energy consumptions of different sensor nodes as even as possible by asking sensors with more residual energy to cover larger areas. In addition, sensor nodes in some system such as IEEE 802.15.4 ZigBee [11] can switch between active state and sleep state. A node that enters the sleep state will stay in an energy-saving mode by not participating in most activities. In this paper, we propose a mechanism to try to maintain the coverage level of a WSN above the user requirement for as long as possible. We first determine the set of the sensors in areas of high density that should enter the sleep state to conserve battery energy. For those sensor nodes that remain active, we would dynamically adjust their sensing ranges such that the total coverage can be maintained at or above the user requirement.

The rest of this paper is organized as follows: Section two lists the related work. We describe our sleep scheduling and sensing range adjustment algorithm in section three, while section four shows the experimental results and the performance comparisons. The conclusion is given in section five.

2 Related Work

Various mechanisms have been proposed to maintain a good coverage level for WSN. One approach is to use mobile sensors that are frequently trying to move to new target locations based on the virtual force interactions between sensors [12-15]. Each sensor is assumed to generate attractive or repulsive force on the other sensors. There will be a repulsive force to separate any two sensors if they are too close and an attractive force if they are far away. The idea is to constantly keep the sensors spread out in a balanced way. Each sensor then calculates its sensing range according to its relative residual energy and position to the other sensors in the neighborhood, where a sensor with

relatively higher amount of energy should cover a larger area. The major disadvantages of this type of approaches are its heavy computation and constant moving, which could require significant amount of energy [16]. Another set of approaches try to mover sensors to predefined regular positions such that they will be spread out perfectly evenly. ISOGRID [17] and CLP [18] both suggest the movement of sensors to the vertices of a hexagon such that the system forms a perfect cellular structure. This approach requires a minimum number of active sensors to occupy all the hexagon vertex positions, which may not be a problem at an early stage of system operation. However, there may not be enough active sensors to maintain a perfect structure as sensors gradually die out after the system has operated for a period of time.

Another class of mechanism is to do sleep scheduling [19-23], which is usually applied to networks of high density where sensors can alternate between the active and sleep modes from time to time to conserve energy. The active nodes in a neighborhood can cooperate with one another to maintain a good coverage. In the Random Sleep scheme [19], each sensor goes to sleep randomly with a probability set by the user. In the Distance-Based Sleep scheme [19], the sleep probability is based on the distance between the sensor and the base station, where a sensor node that is farther away will be put into the sleep state with higher probability. The Balanced-Energy Sleep scheme [20] tries to determine the sleep probability such that the maximal number of sensor nodes would consume energy at the same rate independent of the distance to the base station. In the Coverage-Aware Sleep Scheduling scheme [21], each sensor computes the total overlap area between itself and its active neighbors during each scheduling cycle. It then goes to sleep in the next cycle with a probability proportional to the size of the overlap area. The drawback of this method is that the computation may be too time-consuming to fit into the small computational power of a sensor when there are a large number of neighbors.

3 Distributed Sleep-Scheduling and Range Adjustment Algorithm

We hereby propose a mechanism called Distributed Sleep-scheduling and Range Adjustment (DSRA) that periodically determines the set of active nodes and sleep nodes for each cycle. Our system consists of a base station and many sensors nodes that do not move. We assume that every sensor node knows its own location by GPS or any other locationing mechanism. Every sensor node can adjust its own sensing range R_s ranging from 0 to R_{max}. A sensor is assumed to transmit its packets to the base station using multihop transmissions along the path of minimum number of hops if such path exists. The sensor is considered disconnected from the base station if there is no path between them. The sensors are assumed to be randomly deployed initially. A sensor node in sleep does nothing except waking up periodically to exchange control information with other neighboring nodes. The user is allowed to specify a coverage requirement C ranging from 0 to 100%, and our goal is to prolong the time period that the system coverage stays above this requirement for as long as possible. The DSRA algorithm consists of the three following stages:

A. Information Exchange

At the beginning of each cycle, every active node broadcasts Hello messages to find neighbor nodes within its transmission range. The Hello message contains the

following information: node coordinate, node id, residual energy, sensing range, type, state and S-value. Each active sensor node stores the received Hello messages in the *neighborhood information table* shown in table 1. Assuming its own location as the origin, each sensor can then identify the quadrant each of its neighbors resides in based on their coordinates.

Table 1. Neighborhood Information Table

Node ID	Coordinate	S-Value	Type	…	Quadrant	State
Node 1	(X_1,Y_1)	8.775	Interior	…	I	Active
Node 2	(X_2,Y_2)	10.615	Interior	…	II	Sleep
…	…	…	…	…	…	
Node n	(X_n,Y_n)	9.124	Boundary	…	IV	Active

Figure 2 shows an example where node *p* finds its neighbors *a, b, c, d,* and *e,* and the quadrant each resides in. A sensor node that has at least one active neighbor in each of the four quadrants is called an *interior node*; otherwise, it is called a *boundary node*. We do not allow a boundary node to sleep by setting its S-value to zero because its sleep may result in undesirably large coverage holes. An interior node, *g*, will calculate its S-value using the following formula:

$$\boldsymbol{S\_value} = \frac{\alpha}{g\text{'s energy}} + \beta \times (\text{total overlap areas in all four quadrants}), \qquad (1)$$

where α and β are two weight parameters. The formula takes into consideration both the sensor's own residual energy and the sum of the overlap areas between the sensor itself and the neighbors in all the quadrants. A sensor with large amount of energy should stay active, while a sensor with large coverage overlap between itself and its neighbors is a good candidate to sleep. The active sensor nodes will wait for some amount of time to account for network delay, and enters the next stage.

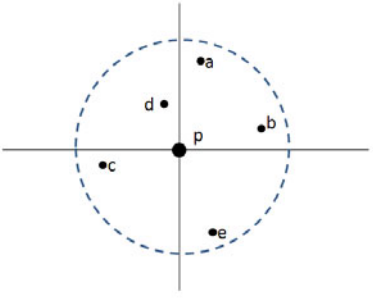


Fig. 2. Neighbors around an active sensor

B. Sleep Scheduling

After the calculation of the S-value, each active node will again broadcast its S-value to its neighbors for comparison. If a sensor node has the maximum S-value among its neighbors, it will broadcast its decision to sleep to its neighbors and switches to sleep. Otherwise, the sensor remains active and enters the next stage of sensing range adjustment.

C. Sensing Range Adjustment

After the sleep scheduling phase, every active sensor node is fully aware of the identities of all its active neighbors. The sensing range calculation is different for boundary node and interior node. A boundary node should adopt a larger sensing range to maintain good coverage. For each quadrant without any neighbor, a boundary node measures its distance to the physical network boundary. Let d be the maximum of these distances. Let $y = min\ (R_{max}, d)$. The actual sensing range to use will be $C * y$, where C is the coverage requirement the user specifies. The sensor will then notify all the neighbors of its sensing range. The interior nodes will wait until all the boundary nodes to finish their calculations and broadcast their sensing ranges. An interior node p first computes a temporary sensing range for each quadrant, and chooses the maximum of the four values as its final sensing range to use. The computation is as follows: If there is no boundary node in a quadrant, the sensing range for that quadrant will be set to $\left(D/\sqrt{2}\right)$, where D is the distance between the node and its nearest neighbor in that quadrant. Otherwise, let b be a boundary node in that quadrant and R_b be its sensing range. Let D_{pb} be the distance between the node p and b, and R_p be p's current sensing range, respectively. If R_b is greater than $(R_p + D_{pb})$, this means that node p's sensing area can be fully covered by node b. Therefore, node p will go to sleep and notify all the neighbors of its decision. Otherwise, the sensing range of p will be set to $(D_b - R_b)/\sqrt{2}$. After node p has calculated the sensing ranges for all the four quadrants, let R denote the maximum of these values. The new sensing range of node p will be equal to $f \times C \times R$, where f is the ratio of p's residual energy over the average residual energy of p's neighborhood.

4 Simulation Results

4.1 Simulation Environment

We built a simulator in MATLAB platform to evaluate the performance of our scheme. We randomly deployed 100 sensor nodes in a square area of size 200m× 200m. The initial energy of each sensor is 80J. Each packet is 512K bytes long. We use the following formulas to calculate the energy consumption [24]:

- Transmission energy consumption: $\boldsymbol{E_t = H + \theta \times packet\_size \times R_c^2}$
- Reception energy consumption : 0.25 J per packet
- Sensing energy consumption : $\boldsymbol{E_s = e_c + \gamma \times R_s^2}$

where γ and θ are both adjustment parameters. H and e_c denote the fixed costs of transmission and sensing, respectively. R_s and R_c represent the current communication range (maximum of 100m) and sensing range (maximum of 50m), respectively.

4.2 Simulation Results

In our simulation, we compared our method with two other methods: the Random Sleep (RS) [19] and Coverage-Aware (CA) [21] schemes on the metrics of coverage percentage, residual energy, the number of active nodes, and the total number of packet received by the base station. The RS scheme randomly selects sensors to sleep with a user-given probability, while the CA scheme puts sensor nodes to sleep in order to maximally reduce the overlap sensing area while trying to maintain 100% coverage. Both these two schemes use fixed sensing range, which is set to 20m in our experiment. The sleep probability in the RS scheme is set to 0.2 because that value produces the best experimental results. Figure 3 shows the comparison of the coverage percentages where all three schemes try to maintain a 100% coverage. We can see that the CA scheme does maintain full coverage in the early stage, but its lifetime is shorter than the others. Our DSRA can provide a coverage close to 100% with a longer lifetime because we try to make each sensor consume energy at about the same rate. The RS scheme also has a good lifetime, but the coverage percentage of RS is lower than the others because the sensors to sleep are chosen randomly. Figure 4 plots the number of active nodes in these three schemes. We can see that our DSRA scheme uses fewer active nodes to maintain the system coverage such that more nodes can preserve energy by switching to sleep and the system can operate longer.

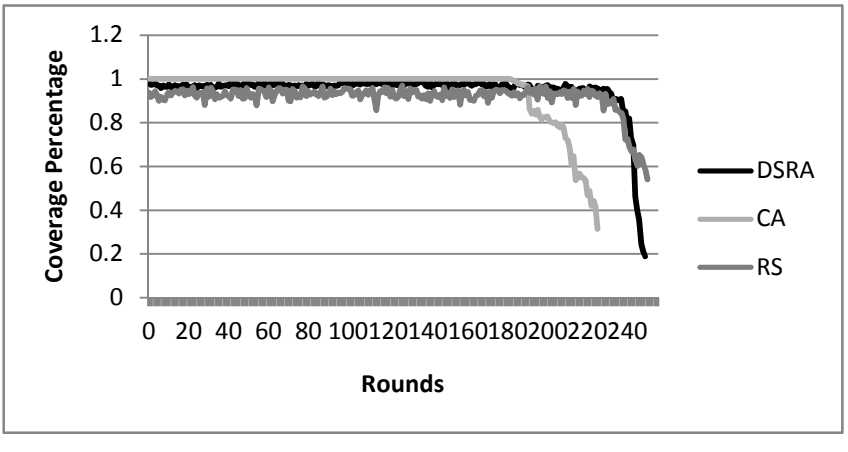


Fig. 3. Comparison of the coverage percentages

Figure 5 displays the accumulative number of packets successfully received by the base station. Both CA and our DSRA can receive almost all the packets since the coverage are about the same. However, the curve for CA flattens out after 200 rounds because most sensors are disconnected from the base station, while our DSRA still

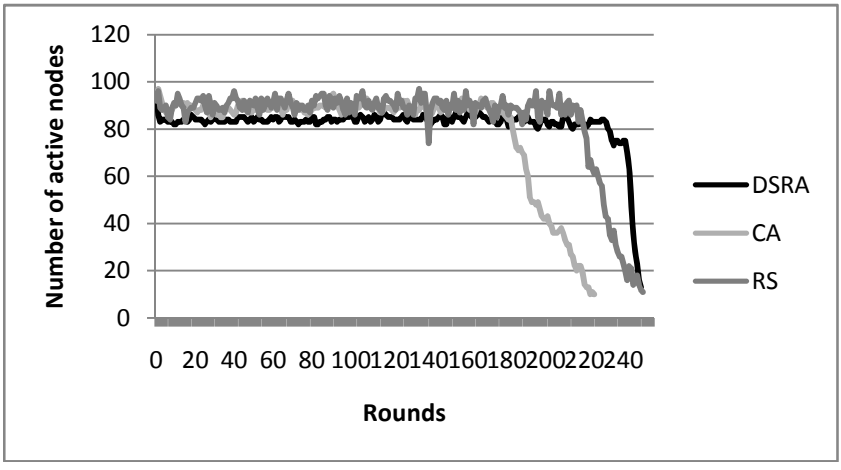


Fig. 4. Number of active nodes

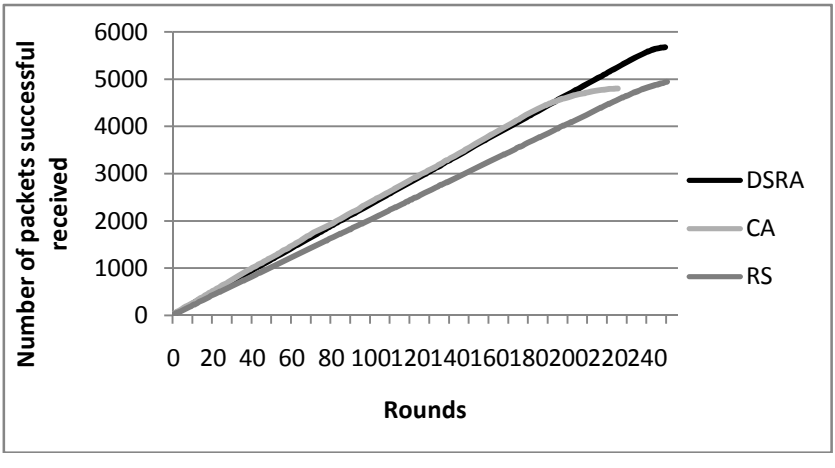


Fig. 5. Accumulative number of data packets successfully received by the base station

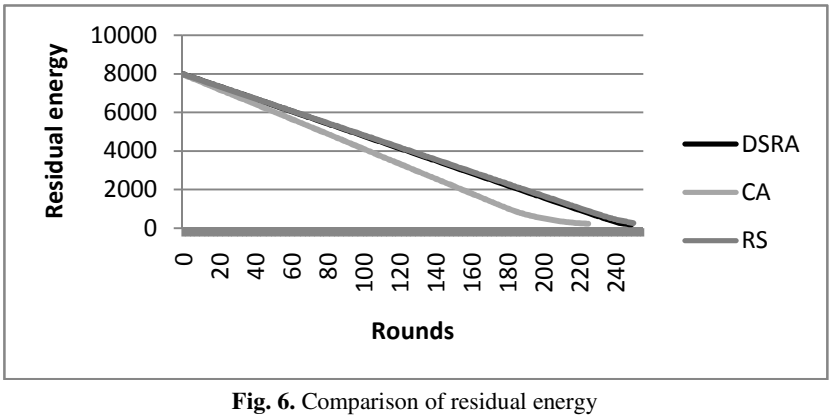


Fig. 6. Comparison of residual energy

keeps receiving packets. Figure 6 shows the residual energy in the system. CA consumes its energy faster than the others because it uses a larger sensing range and has fewer nodes in sleep. Although our DSRA uses fewer active nodes to operate in the network, it adjusts the sensing range of each active node effectively to provide a good coverage. Both the DSRA and RS consume energy at about the same rate.

Our DSRA allows the user to specify a minimum coverage requirement *C* while the others do not. Figure 7 plots the actual coverage achieved given different values of *C*. For the case of $C = 1$, the coverage percentage achieved is in the range of 0.95 to 0.99 that is a little bit below full coverage, while the system operates at a coverage above what the user specifies for all the other cases. In general, our DSRA can maintain the system coverage that satisfies the user requirement for as long as possible.

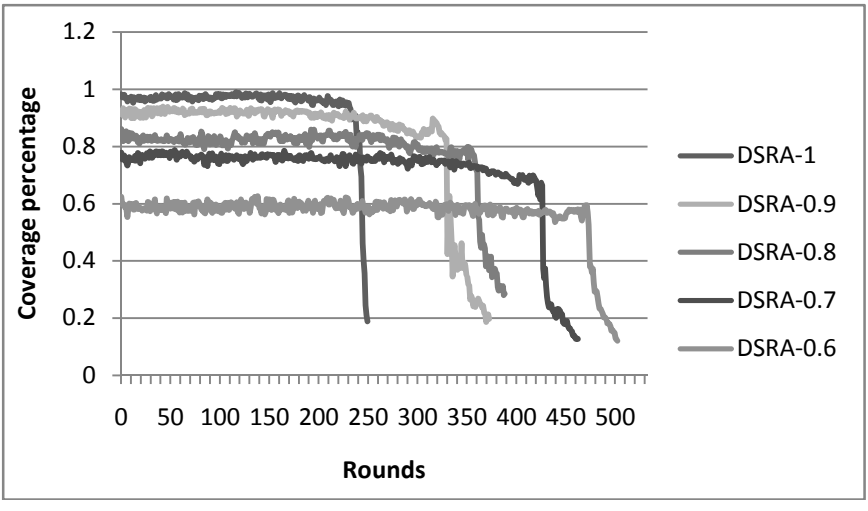


Fig. 7. Coverage percentages under different user requirements

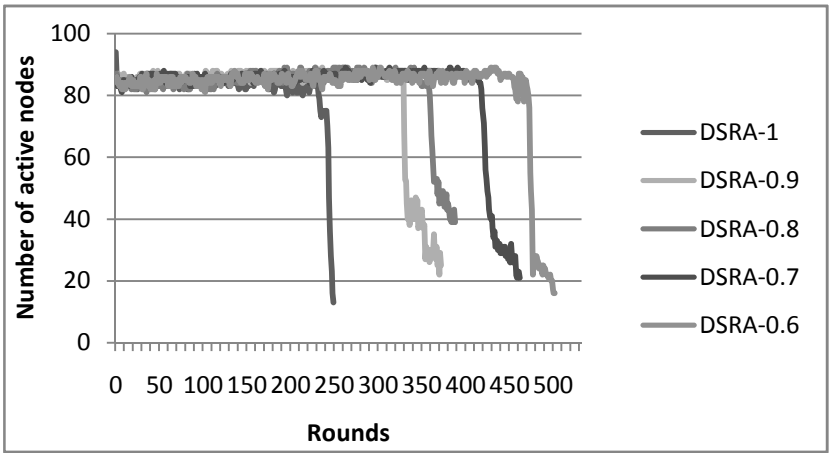


Fig. 8. Number of active nodes under different user requirements

Figure 8 shows the number of the active nodes given various values of *C*. It seems that the value of *C* does not affect the number of nodes in sleep. Figure 9 shows that the total number of packets received by the base station closely depends on the coverage percentage. When $C = 1$, the number of packets received by the base station is the largest initially, but the total number of packets is fewer than the others because the system lifetime is shorter. Figure 10 shows that the energy is consumed faster when the coverage requirement is higher.

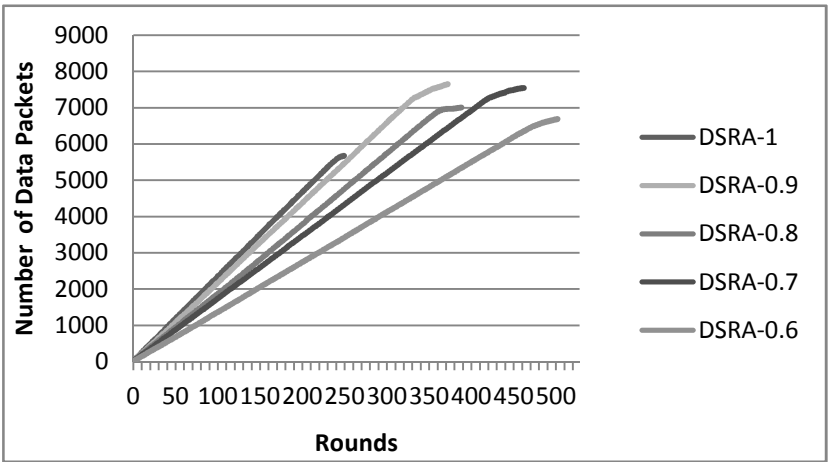


Fig. 9. Number of packets received under different user requirements

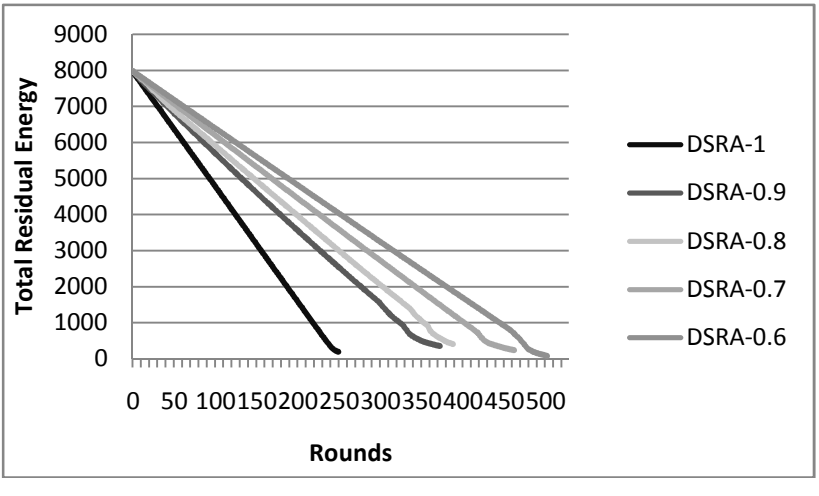


Fig. 10. Total residual energy under different user requirements

5 Conclusion

Wireless sensor network has been widely used in many important applications, and it is essential to keep a minimum coverage for the network to function properly. We proposed an efficient DSRA method to generate a total coverage at the desirable level the user requires. Some of the sensors in an area of high density will be selected to sleep to conserve energy. Each of the active sensors then sets its sensing range such that the total coverage by all the active sensors is maintained at the required level. The network is reconfigured periodically such that the sensors consume their energy at roughly the same rate. The simulation results showed that the DSRA scheme could achieve the required coverage for a long period of time, and the system lifetime can be extended longer than both the RS and CA schemes.

References

1. Akyildiz, I., et al.: A Survey on Sensor Networks. IEEE Communications Magazine 40(8) (2002)
2. Tian, D., Georganas, N.D.: A Coverage-preserving Node Scheduling Scheme for Large Wireless Sensor Networks. In: ADM International Workshop on Wireless Sensor Networks & Applications (2002)
3. Brooks, R.R., Ramanathan, P., Sayeed, A.M.: Distributed Target Classification and Tracking in Sensor Networks. Proceedings of IEEE 91(8) (August 2003)
4. Li, X., et al.: Coverage in Wireless Ad-hoc Sensor Networks. IEEE Transactions on Computers 52(6) (2003)
5. Huang, C.F., Tseng, Y.C.: The Coverage Problem in a Wireless Sensor Network. In: 2nd ACM International Conference on Wireless Sensor Networks and Applications, San Diego, CA, USA (2003)
6. Ye, F., Zhong, G., Cheng, J., Lu, S., Zhang, L.: PEAS: A Robust Energy Conserving Protocol for Long-lived Sensor Networks. In: 23rd International Conference on Distributed Computing Systems. IEEE Computer Society, Los Alamitos (2003)
7. Jiang, J., Dou, W.: A Coverage-preserving Density Control Algorithm for Wireless Sensor Networks. In: Nikolaidis, I., Barbeau, M., Kranakis, E. (eds.) ADHOC-NOW 2004. LNCS, vol. 3158, pp. 42–55. Springer, Heidelberg (2004)
8. Li, L., Sun, L., Ma, J., Chen, C.: A Receiver-based Opportunistic Forwarding Protocol for Mobile Sensor Networks. In: 28th International Conference on Distributed Computing Systems Workshops (June 2008)
9. Zabin, F., Misra, S., Woungang, I., Rashvand, H.F., Ma, N.W., Ahsan, A.: REEP: Data-Centric Energy-Efficient and Reliable Routing Protocol for Wireless Sensor Networks. IET Communications 2(8) (September 2008)
10. Wang, Y.H., Yu, C.Y., Chen, W.T., Wang, C.X.: An Average Energy Based Routing Protocol for Mobile Sink in Wireless Sensor Networks. In: First IEEE International Conference on Ubi-Media Computing (August 2008)
11. IEEE Standards, http://www.ieee.org
12. Zou, Y., Chakrabarty, K.: Sensor Deployment and Target Localization Based on Virtual Forces. In: IEEE INFOCOM (April 2003)
13. Younghwan, Y., Agrawal, D.P.: Mobile Sensor Relocation to Prolong the Lifetime of Wireless Sensor Networks. In: IEEE Vehicular Technology Conference (May 2008)

14. Li, S., Xu, C., Pan, W., Pan, Y.: Sensor Deployment Optimization for Detecting Maneuvering Targets. In: 8th International Conference on Information Fusion (July 2005)
15. Lin, C.H., Lu, C.H.: Efficient Relocation and Range Adjustment to Maintain Coverage in Wireless Sensor Networks. In: 5th Workshop on Wireless, Ad Hoc and Sensor Networks, Taiwan (2009)
16. Wang, G., Cao, G., Porta, T.L., Zhang, W.: Sensor Relocation in Mobile Sensor Networks. In: IEEE INFOCOM (2005)
17. Lam, M.L., Liu, Y.H.: ISOGRID: an Efficient Algorithm for Coverage Enhancement in Mobile Sensor Networks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (October 2006)
18. Wang, P.C., Hou, T.W., Yan, R.H.: Maintaining Coverage by Progressive Crystal-Lattice Permutation in Mobile Wireless Sensor Networks. In: International Conference on Systems and Networks Communications (October 2006)
19. Deng, J., Han, Y.S., Heinzelman, W.B., Varshney, P.K.: Scheduling Sleeping Nodes in High Density Cluster-based Sensor Networks. MONET Special Issue on Energy Constraints and Lifetime Performance in Wireless Sensor Networks (2004)
20. Deng, J., Han, Y.S., Heinzelman, W.B., Varshney, P.K.: Balanced-energy Sleep Scheduling Scheme for High Density Cluster-based Sensor Networks. In: 4th Workshop on ASWN (2004)
21. Shen, F., Liu, C., Zhang, J.: A Distributed Coverage-aware Sleep Scheduling Algorithm for Wireless Sensor Networks. In: 6th International Conference on Information Technology (2009)
22. Lu, G., Sadagopan, N., Krishnamachari, B., Goel, A.: Delay Efficient Sleep Scheduling in Wireless Sensor Networks. In: IEEE INFOCOM, Miami, FL (2005)
23. Balakrishnan, M., Johnson, E., Huang, H.: TEAN-Sleep for Distributed Sensor Networks: Introduction and α-Metrics Analysis. In: IEEE Military Communications Conference (2007)

An Innovative Routing Algorithm with Reinforcement Learning and Pattern Tree Adjustment for Wireless Sensor Networks

Chia-Yu Fan, Chien-Chang Hsu, and Wei-Yi Wang

Department of Computer Science and Information Engineering,
Fu-Jen Catholic University
510 Chung Cheng Rd., Hsinchuang, Taipei, Taiwan 242
cch@csie.fju.edu.tw, jhtsay96@csie.fju.edu.tw
matrixf@gmail.com, cch@csie.fju.edu.tw, wea97@csie.fju.edu.tw

Abstract. This paper proposes a new routing algorithm for wireless sensor network. The algorithm uses reinforcement learning and pattern tree adjustment to select the routing path for data transmission. The former uses Q value of each sensor node to reward or punish the node in the transmission path. The factor of Q value includes past transmission path, energy consuming, transmission reword to make the node intelligent. The latter then uses the Q value to real-time change the structure of the pattern tree to increase successful times of data transmission. The pattern tree is constructed according to the fusion history transmission data and fusion benefit. We use frequent pattern mining to build the fusion benefit pattern tree. The experimental results show that the algorithm can improve the data transmission rate by dynamic adjustment the transmission path.

Keywords: Wireless Sensor Networks, reinforcement learning, routing path, fusion benefit pattern tree.

1 Introduction

Wireless Sensor Networks (WSN) has been widely used in many application domains [2,5,8,11]. Peer-to-peer transmission protocol becomes the popular transmission method for innovative applications. Each node of the peer-to-peer WSN can only communicate with its neighboring nodes [16]. Many dynamic routing algorithms are proposed to transmit data for achieving energy savings. These algorithms are usually classified into three categories: clustered-based, tree-based, and mining-based [2,5,9,11,16]. Generally, the clustered-based method divides the sensor nodes into different clusters for data transmission. The cluster head collects data from its member nodes in the cluster and sends the aggregated data to the sink node. The member node of each cluster can only communicate with its cluster head. The head is used as the routers to the sink node. LEACH and HEED are the famous clustering energy-efficient algorithms [23]. Tree-based routing algorithm usually constructs a hierarchical tree to organize the sensor nodes into a structured topology for effective data transmission [8]. The leaf nodes of the structured tree are used to track, collect,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 398–405, 2010.

and transmit data [14]. It is suitable for mobile nodes of WSN. LFFT and DAB are the typical tree-based algorithms. The mining-based algorithm uses historical data to predict moving patterns of objects [8,12,17,21,22]. It uses the data mining algorithm to find the routing patterns for path predicting or selection [4,15,24]. STMP-Mine and MLOT are the newly algorithms for mining-based approach [16]. However, all of the above type of algorithm uses predefined or static structure to select the routing path [10]. They didn't consider the successful or fail transmission in the run-time routing. None of them integrate machine learning mechanism or intelligent technologies to support the path selection. How to design an intelligent path selection algorithm of WSN for run-time routing is an interesting research. Moreover, most routing algorithms focus on the energy-efficient data transmission. Seldom of them consider the fusion-efficient data transmission capability of each node in the path selection. It may influence the performance of data transmission as well as energy consumption in WSN.

This paper proposes an intelligent routing algorithm with reinforcement learning and pattern tree adjustment (IRARA) for WSN. It uses pattern mining method to construct a fusion benefit pattern tree based on fusion capability from past routing information. The routing path is selected from the fusion benefit pattern tree. It then uses reinforcement learning to adjust the pattern tree to record the routing path in the run-time environment. The Q value is used to reflect the successful or unsuccessful path prediction as well as the fusion capability. Q value records the data amount, fusion capability, transmission path, transmission success/fail times, and fusion capability energy consumption. The experimental results show that the integration of reinforcement learning and pattern tree adjustment can provide an excellent path selection in WSN.

The remainder of this paper is organized as follows. Section 2 introduces the system architecture. Section 3 is simulation. Section 4 concludes the work.

2 Proposed Method

Figure 1 shows the architecture of IRARA algorithm. It contains three main components: fusion benefit pattern tree construction, pattern tree adjustment, and path selection. Basically, the fusion benefit pattern tree uses frequent pattern mining algorithm [1,3,6,7,15,20] to discover the fusion benefit patterns from the past transmission information [13,19]. The fusion benefit is used as the support value for the pattern mining. It computes the data fusion benefit (FB_{nm}) of the path (n→m) in the cluster from the past transmission information.

$$FB_{nm} = \left\| S_{nm} * C_{nm} * F_{nm} \right\| \tag{1}$$

where S, C, and F are the stability of data transmission, fusion capability, and pattern frequency from the node n to m [18]. The longest fusion pattern is selected from pattern set. The fusion benefit pattern tree uses the longest fusion pattern from the pattern set as the backbone. The second longest pattern is then added to the fusion benefit pattern tree iteratively until no pattern is selected. Notably, the cyclic patterns will not be added to the fusion benefit pattern tree. Figure 2 shows the example of the fusion benefit pattern tree.

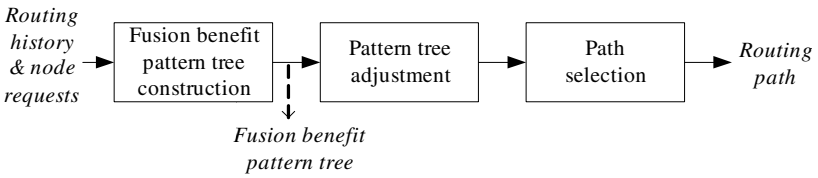


Fig. 1. System architecture

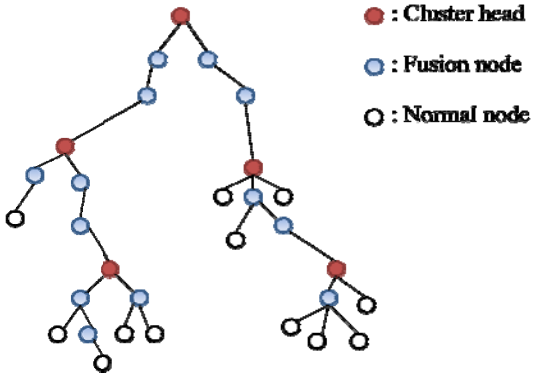


Fig. 2. Example of fusion benefit pattern tree

The pattern tree adjustment calculates the Q value of each node to adjust the tree structure of pattern tree for run-time routing. Q value is the reinforcement value to reflect the successful or fail path selection in run-time routing.

$$Q_n = \begin{cases} Q_{n-1} + \tau\gamma \times \frac{success}{visit}, if\ success \\ Q_{n-1} + \tau\gamma \times \frac{fail}{visit}, if\ fail \\ Q_0 = \frac{fc_{Max} + fc_{min}}{2} \end{cases} \quad (2)$$

where *n,* τ, γ, *success, fail*, *visit*, *fc, and* Q_0 are the time value, data transmission trajectory, transmission reward, success transmission times, fail transmission times, transmission times, fusion capability, and initial Q value, respectively. The Q value will be updated according to the last transmission path. The data transmission trajectory (τ) computes the variation of data transmission amount of the node.

$$\tau = \|\tau\,(1-\rho) + \Delta\,\tau\| \quad (3)$$

$$\Delta\,\tau = \|D_n - D_m\| \quad (4)$$

where D_n, D_m, and ρ are the cumulative amount of data transmission, average of data transmission, and energy consumption rate. Transmission reward (γ) will give reward or penalty if transmission is successful or fail.

$$\gamma = \pm \left\| \frac{1}{level/height} \times frequency \right\| \quad (5)$$

where $\pm$, *level, height,* and *frequency* are the success (+) or fail (-) transmission, node level, height of the pattern tree, and frequency of success transmission. The frequent pattern tree uses max heap tree rule to adjust the pattern tree based on the Q value. It means the Q value of the parent node is greater than the child node. If the parent node violates the rule, it will replace by its maximum child. If the Q value of node is less than or equal to zero, it marks as the failure node. The cluster head node will ask a best pattern from the pattern set to replace the failure node pattern in the fusion benefit tree. The best pattern means the maximum fusion benefit pattern without failure node in the path. The best pattern is retrieved by the sink node. If none of the pattern is found, it will delete the failure leaf node directly. If the failure node is not a leaf node, the Q value of the failure node is reset to zero.

Path selection selects a routing path from fusion benefit pattern tree. When a sensing request is received, it will traverses the fusion benefit pattern tree from root to the requested node for finding the maximum Q value path. The nodes in the path will be awaked during the requested time interval. The path selection process may fail if it cannot find the routing path or the transmission time window expired.

3 Simulations

We use NS2 to simulate 1000*1000 square meters network environment and deploy 100 nodes randomly. The initial energy and energy threshold of each node are 1 joule and 0.1 joule. The average energy consumption of data receiving and transmission are 0.003 joule and 0.0002 joule. The algorithm will work properly when the percentage of effective node number greater than 80%. Table 1 lists the related node information. The fusion benefit pattern tree construction uses frequent pattern mining algorithm to find the pattern set. The length of the longest pattern of each cluster is 3. So the depth of a cluster in the pattern tree is the same as the length of the longest pattern. Fig. 3 shows the partial display of the constructed fusion benefit pattern tree.

Table 1. Simulation of node information

Node#	Cluster#	Q value	Node#	Cluster#	Q value	Node#	Cluster#	Q value
0	1	0.8451	32	2	0.7546	89	3	0.4582
71	1	0.3222	46	2	0.5631	27	3	0.3677
10	1	0.2735	23	2	0.3838	93	3	0.4250
57	1	0.3723	65	2	0.4380	72	3	0.2820
15	1	0.2633	49	2	0.4649	39	3	0.3016
8	1	0.2552	80	2	0.4237	91	3	0.2520
61	1	0.3035	42	2	0.3053	7	3	0.2331
85	1	0.6075	13	2	0.3190	59	3	0.0106

The pattern tree adjustment then calculates the Q value of each node. Figure 4 shows the example of Q values computation of node #85. Notably, the awaked times

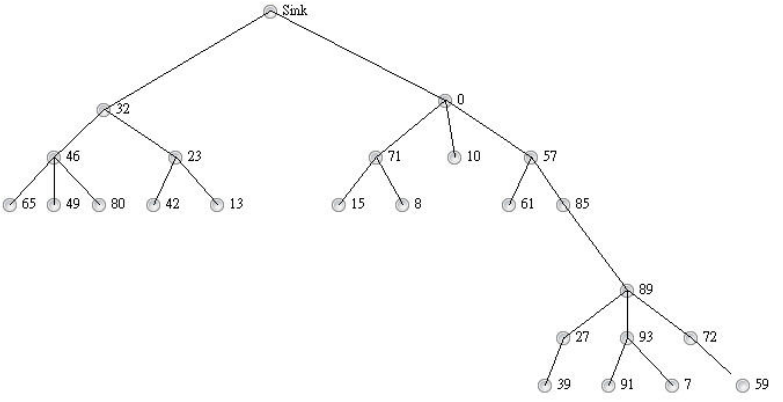


Fig. 3. Partial display of the fusion benefit pattern tree

$$\tau = 0.3706\ (1 - 0.0032) + 2.2746 = 2.2771$$

$$\Delta\tau = 2.652 - 0.3774 = 2.2746$$

$$\gamma = (-1) \times \frac{1}{3/9} \times 0.7 = -2.1$$

$$\|\gamma\| = \frac{-2.1}{10} = -0.21$$

$$Q7 = Q6 + 2.2771 \times (-0.21) \times 4/7 = 0.1617 + (-0.2733)$$

$$= -0.1116 < 0$$

Fig. 4. Example of node #85 Q value computation

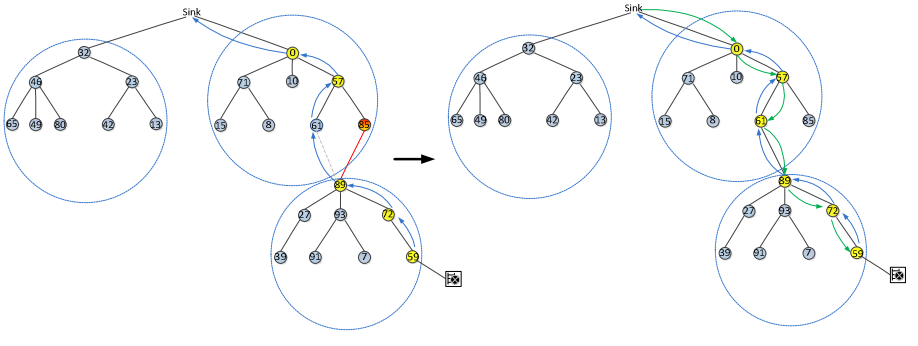


Fig. 5. Path pattern replacement of node #85

and Q value of node #85 are 7 and 0.6075. The Q value of node #85 is less than 0 and it marks as the failure node. Furthermore, the cluster head node #0 retrieves a path pattern from the pattern set, named node #61→node #89, to replace the node

#85→node #89 (Fig. 5). The fusion benefit pattern is adjusted according to the max heap rules. The path selection then selects the path with maximum Q value as the routing path (Fig. 6). Finally, the newly pattern is selected as the routing path: node 59→node 72→node 89→node 61→node 57→node 0→sink.

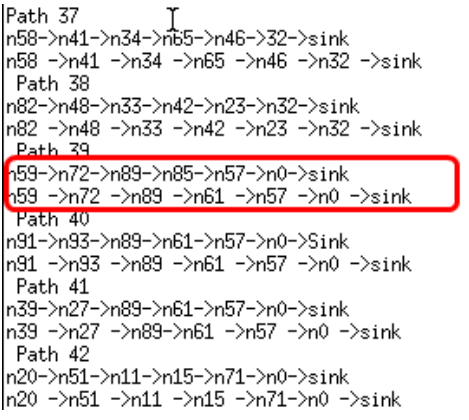


Fig. 6. Path selection process

4 Conclusion

The paper proposes an intelligent routing algorithm with reinforcement learning and pattern tree adjustment. The system uses frequent pattern mining method to construct a fusion benefit pattern tree. It uses the past fusion data and routing path to find the path patterns. The system uses a tree-based structure to store the frequent patterns. The system then uses reinforcement learning to real-time change the structure of the fusion benefit pattern tree. The Q value is used to record the data transmission trajectory and fusion capability of each node. If the Q vale less than zero, it is marked as the failure node. The failure node will be replaced from the discovered pattern set. Moreover, the reward and penalty value are used to the success and fail node transmission of Q value computation. The system also uses max-heap rule to adjust the node in each cluster. Finally, the system selects the path with maximum Q value as the transmission pattern for the requested sensor node. It traverses the fusion benefit pattern tree from root to the requested node. The node in the selected path will be awaked during the requested time window. The experimental results show that the system can provide an intelligent and innovative path selection method for conducting effective data fusion and energy saving.

Acknowledgments. This work is partly supported by National Science Council of ROC under grants NSC 99-2220-E-030-001 and NSC 99-2220-E-030-002.

References

1. Agrawal, R., Imiliemski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
2. Abbasi, A.A., Younis, M.: A Survey on Clustering Algorithms for Wireless Sensor Networks. Computer Communications 30(14-15), 2826–2841 (2007)
3. Cheng, Y., Ren, X.: Mining Moving Patterns Based on Frequent Patterns Growth in Sensor Networks. In: IEEE International Conference on Networking, Architecture, and Storage, pp. 133–138. IEEE Press, New York (2007)
4. Ci, S., Guizani, M., Sharif, H.: Adaptive Clustering in Wireless Sensor Networks by Mining Sensor Energy Data. Computer Communications 30(14-15), 2968–2975 (2007)
5. Chen, M.X., Wang, Y.D.: An Efficient Location Tracking Structure for Wireless Sensor Networks. J. Parallel and Distributed Computing 32(13-14), 1495–1504 (2009)
6. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation. In: ACM-SIGMOD International Conference on Management of Data, pp. 1–12 (2000)
7. Hong, T.P., Lin, C.W., Wu, Y.L.: Incrementally Fast Updated Frequent Pattern Trees. Expert Systems with Applications 34(4), 2424–2435 (2008)
8. Hung, C.C., Chang, C.W., Peng, W.C.: Mining Trajectory Profiles for Discovering User Communities. In: 1st Workshop on Location-Based Social Networks, Seattle, pp. 1–8 (2009)
9. Kung, H.T., Vlah, D.: Efficient Location Tracking Using Sensor Networks. Wireless Communications and Networking 3, 1954–1961 (2003)
10. Le, T., Sinha, P., Xuan, D.: Turning Heterogeneity into an Advantage in Wireless Ad-hoc Network Routing. Ad Hoc Networks 8(1), 108–118 (2010)
11. Lin, C.Y., Tseng, Y.C.: Structures for In-Network Moving Object Tracking in Wireless Sensor Networks. In: 1st International Conference on Broadband Networks, pp. 71–727 (2004)
12. Lin, K.W., Hsieh, M.H., Tseng, V.S.: A Novel Prediction-based Strategy for Object Tracking in Sensor Networks by Mining Seamless Temporal Movement Patterns. Expert Systems with Applications 37(4), 2799–2807 (2010)
13. Lin, L.J.: Self-Improving Reactive Agents based on Reinforcement Learning, Planning and Teaching. Machine Learning 8(3), 293–321 (1992)
14. Lin, C.Y., Peng, W.C., Tseng, Y.C.: Efficient In-Network Moving Object Tracking in Wireless Sensor Networks. IEEE Transactions on Mobile Computing 5(8), 1044–1056 (2006)
15. Tanbeer, S.K., Ahmed, C.F., Jeong, B.S., Lee, Y.K.: Efficient Mining of Association Rules from Wireless Sensor Networks. In: IEEE International Conference on Advanced Communication Technology, pp. 719–724. IEEE Press, New York (2009)
16. Tseng, V.S., Lu, E.H.C.: Energy-Efficient Real-Time Object Tracking in Multi-Level Sensor Networks by Mining and Predicting Movement Patterns. J. of Systems and Software 82(4), 697–706 (2009)
17. Tseng, V.S., Lin, K.W.: Energy Efficient Strategies for Object Tracking in Sensor Networks: a Data Mining Approach. J. of Systems and Software 80(10), 1678–1698 (2007)
18. Wang, W.Y.: An Intelligent Data Fusion Algorithm with Fusion Benefit Pattern Tree for Wireless Sensor Networks, Matster Thesis of Department of Computer Science and Information Engineering, Fu-Jen Catholic University. Taiwan (2010)

19. Watkins, C.J.C.H., Dayan, P.: Technical Note: Q learning. Machine Learning 8(3), 279–292 (1992)
20. Wu, B., Zhang, D., Lan, Q., Zheng, J.: An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure. In: IEEE International Conference on Convergence and Hybrid Information Technology, vol. 1, pp. 1099–1102. IEEE Press, New York (2008)
21. Xu, Y., Winter, J., Lee, W.C.: Prediction-based Strategies for Energy Saving in Object Tracking Sensor Networks. In: IEEE International Conference on Mobile Data Management, pp. 346–357. IEEE Press, New York (2004)
22. Yavaş, G., Katsaros, D., Ulusoy, O., Manolopoulos, Y.: A Data Mining Approach for Location Prediction in Mobile Environments. Data & Knowledge Engineering 54(2), 121–146 (2005)
23. Younis, O., Fahmy, S.: HEED: a Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad hoc Sensor Networks. IEEE Transactions on Mobile Computing 3(4), 366–379 (2004)
24. Yu, C.H., Peng, W.C., Lee, W.C.: Mining Community Structures in Peer-to-Peer Environments. In: 14th IEEE International Conference on Parallel and Distributed Systems, pp. 351–358. IEEE Press, New York (2008)

Swarm Intelligence for Cardinality-Constrained Portfolio Problems

Guang-Feng Deng and Woo-Tsong Lin

Department of Management Information Systems, National Chengchi University,
64, Sec. 2, Chihnan Rd., Wenshan Dist., Taipei 116, Taiwan ROC
95356502@nccu.edu.tw, lin@mis.nccu.edu.tw

Abstract. This work presents Particle Swarm Optimization (PSO), a collaborative population-based swarm intelligent algorithm for solving the cardinality constraints portfolio optimization problem (CCPO problem). To solve the CCPO problem, the proposed improved PSO increases exploration in the initial search steps and improves convergence speed in the final search steps. Numerical solutions are obtained for five analyses of weekly price data for the following indices for the period March, 1992 to September, 1997: Hang Seng 31 in Hong Kong, DAX 100 in Germany, FTSE 100 in UK, S&P 100 in USA and Nikkei 225 in Japan. The computational test results indicate that the proposed PSO outperformed basic PSO algorithm, genetic algorithm (GA), simulated annealing (SA), and tabu search (TS) in most cases.

Keywords: Particle swarm optimization, cardinality constrained portfolio optimization problem, Markowitz mean-variance model, nonlinear mixed quadratic programming problem, swarm intelligence.

1 Introduction

Portfolio optimization, which is the allocation of wealth among several assets, is an essential problem in modern risk management. Expected returns and risks are the most important parameters in portfolio optimization problems. Investors generally prefer to maximize returns and minimize risk. However, high returns generally involve increased risk.

The Markowitz mean –variance model, which is among the best models for solving the portfolio selection problem, can be described in terms of the mean return of the assets and the variance of return (risk) between these assets [1]. The basic model obtains the "efficient frontier", which is the portfolio of assets that achieves a predetermined level of expected return at minimal risk. For every level of desired mean return, this efficiency frontier indicates the best investment strategy.

From a practical perspective, portfolio selection problem consider many constraints of real-world investments, including trading limitations, portfolio size, etc. However, the basic Markowitz mean-variance model does not include cardinality constraints to ensure the investment in a given number of different assets, nor does it include bounding constraints to limit the funds invested in each asset. Although portfolio optimization using the standard Markowitz model is NP-hard, the solution to this

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 406–415, 2010.

problem with a sufficiently small number of variables can be solved by using quadratic programming. The problem becomes much more difficult if the number of variables is increased or if additional constraints such as cardinality constraints are introduced [2]. Such constraints formed nonlinear mixed integer programming problems, which are considerably more difficult to solve than the original problem. Exact solution methods are inadequate. Therefore, proposed heuristic solutions for the portfolio selection problem include evolutionary algorithms, tabu search (TS) simulated annealing (SA) and neural networks [2-5].

Particle swarm optimization (PSO), introduced by Kennedy and Eberhart in 1995, is based on a psychosocial model of social influence and social learning and has proven effective in many empirical studies [6]. This study proposes a novel application of PSO, a collaborative population-based meta-heuristic algorithm for the Markowitz mean-variance model, which includes cardinality and bounding constraints, to solve Cardinality Constrained Portfolio Optimization problems (CCPO problems). Due to the many variations of the original PSO, this study first investigated the performance of basic PSO in solving CCPO problems. The results showed that the constraints cause PSO to stagnate to the local optimum. Therefore, remedies are proposed to avoid stagnation in CCPO problems. The reflection strategy is to keep desired assets in portfolio in the search process. The replacement minimum hold strategy randomly adds assets with the minimum hold weight when assets needed to obtain a new solution are insufficient. The mutation strategy increases the search space.

The performance of the proposed PSO was compared with basic PSO and heuristic algorithms, including genetic algorithm (GA), simulated annealing (SA), and tabu search (TS). Performance was compared using five problems, involving 31-255 dimensions corresponding to weekly data for March, 1992 to September, 1997. The test data were obtained from the following indices: Hang Seng 31 in Hong Kong, DAX 100 in Germany, FTSE 100 in UK, S&P 100 in USA and Nikkei 225 in Japan. Results show that the proposed PSO is much more robust and effective than basic PSO algorithms in terms of tracing out the efficient frontier accurately. Compared to other heuristic algorithms, the proposed PSO obtained better solutions in most test problems.

Following this introduction, Section 2 presents the model formulation for the Cardinality constrained portfolio optimization problems, and Section 3 describes the application of PSO for solving this problem. The computational experiment in Section 4 evaluates the PSO model and experimental results. Section 5 presents conclusions and proposes future works.

2 Portfolio Optimization Problems and Particle Swarm Optimization

This section presents the standard Markowitz portfolio model and demonstrates an efficient frontier calculation. The cardinality constraints are then given for the Markowitz mean-variance model to be solved.

2.1 Portfolio Optimization Problems

The notation used in this analysis is based on Markowitz mean-variance model for solving the portfolio selection problem. Let N be the number of different assets, u_i be

the expected return of asset i (i=1,… ,N), $\sigma_{i,j}$ be the covariance between assets i and j (i=1,… , N; j=1, …, N), The decision variables x_i represent the proportion ($0 \leqq x_i \leqq 1$) of the portfolio invested in asset i (i=1,… , N) and a weighting parameter λ. Using this notation, the standard Markowitz mean-variance model for the portfolio selection problem can be presented as

$$Min \quad \lambda\left[\sum_{i=1}^{N}\sum_{j=1}^{N} x_i \cdot x_j \cdot \sigma_{ij}\right] - (1-\lambda)\left[\sum_{i=1}^{N} x_i \cdot u_i\right] \tag{1}$$

subject to

$$\sum_{i=1}^{N} x_i = 1 \tag{2}$$

$$0 \leq x_i \leq 1 \quad , \qquad i = 1,..., N \tag{3}$$

where $\lambda \in [0,1]$ is the risk aversion parameter. The case λ=0 represents the maximum expected return for the portfolio (disregarding risk), and the optimal portfolio is the single asset with the highest return. The case λ=1 represents the minimal total risk for the selected portfolio (disregarding return), and the optimal portfolio includes numerous assets. The two extremes λ=0 and λ=1, λ represent the tradeoff between risk and return. Equation (2) ensures that the sum of the proportions is 1. The equation $\sum_{i=1}^{N}\sum_{j=1}^{N} x_i \cdot x_j \cdot \sigma_{ij}$ obtains total variance (risk), which should be minimized and the equation $\sum_{i=1}^{N} x_i \cdot u_i$ obtains the total portfolio return, which should be maximized.

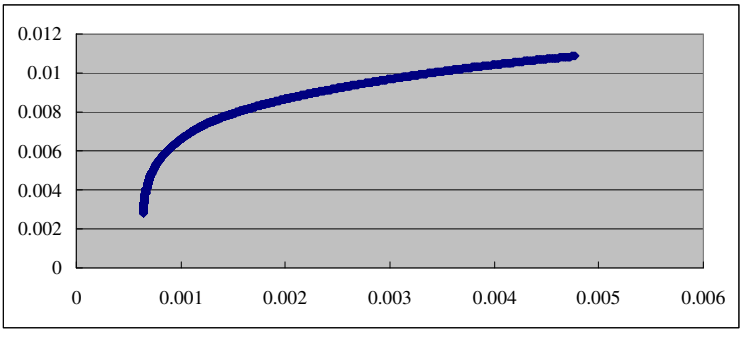


Fig. 1. Standard efficient frontier for Hang Seng 31 dateset

The portfolio selection problem is a multi-objective optimization problem, and all non-dominated solutions can be used to produce the efficient frontier. Figure 1 shows the efficient frontier as plotted by varying value λ corresponding to the benchmark Hang Seng 31. To calculate cardinality constraints for the Markowitz Optimal Model, this study used the model formulation presented in [3, 4]. In addition to the previously defined variables, let K be the desired number of assets in the portfolio, let ε_i be the minimum proportion of the portfolio allocated to asset i (i=1, …, N) if any of asset i is held, and let δ_i be the maximum proportion allocated to asset i (i=1,…, N) if any of asset

i is held, where $0 \leqq \varepsilon_i \leqq \delta_i \leqq 1$ (i=1,… , N). In practice ε_i represents a "min-buy" or "minimum transaction level" for asset i, and δ_i limits portfolio exposure to asset i. Zero-one decision variables are as follows:

$$z_i \begin{cases} 1 & \text{if any of asset } i \ (i=1...N) \text{ is held,} \\ 0 & \text{otherwise} \end{cases}$$

The cardinality constrained portfolio optimization problem is

$$Min \ \ \lambda \left[\sum_{i=1}^{N} \sum_{j=1}^{N} x_i \cdot x_j \cdot \sigma_{ij} \right] - (1-\lambda) \left[\sum_{i=1}^{N} x_i \cdot u_i \right] \tag{4}$$

subject to

$$\sum_{i=1}^{N} x_i = 1 \tag{5}$$

$$\sum_{i=1}^{N} z_i = K \tag{6}$$

$$\varepsilon_i z_i \leq x_i \leq \delta_i z_i \quad , \quad i = 1,..., N \tag{7}$$

$$z_i \in [0, 1] \quad , \quad i = 1,..., N \tag{8}$$

Equation (5) ensures that the sum of the proportions is 1, and Eq.(6) ensures that exactly K assets are held. Equation (7) ensures that if any of asset i is held (z_i =1) its proportion w_i must lie between ε_i and δ_i whilst if no asset i is held (z_i =0), its proportion w_i is zero. Equation (8) is the integrality constraint.

2.2 Particle Swarm Optimization

The PSO is a social population-based search algorithm of social influence that learns from its neighborhood. A PSO swarm resembles a population, and a particle resembles an individual. The PSO is initialized with a particle swarm, and each particle position represents a possible solution. The particles fly through the multidimensional search space by dynamically adjusting velocities according to its own experience and that of its neighbors [6, 7].

At each iteration t, the position $x_{i,j}^{t}$ of the ith particle is updated by a velocity $v_{i,j}^{t+1}$. The position is updated for the next iteration using

$$x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1} \tag{9}$$

where $x_{i,j}^{t}$ denotes the position of particle i in the dimension j in search space at time step t. The position of the particle is changed by adding a velocity $v_{i,j}^{t+1}$ to the current position. The velocity update rule is calculated as

$$v_{i,j}^{t+1} = v_{i,j}^{t} + c_1 \cdot r_1 \cdot \left(p_{i,j} - x_{i,j}^{t} \right) + c_2 \cdot r_2 \cdot \left(p_{g,j} - x_{i,j}^{t} \right) \tag{10}$$

The rule depends on three components: particle current velocity $v_{i,j}^{t}$, the best position and global best position. Where $v_{i,j}^{t}$ is the velocity of particle i in dimension j=1, …,n

at time step t. $p_{i,j}$ is the best position that the particle i has visited since the first time step. The global best position $p_{g,j}$ is the best position discovered by all particles found since the first time step.

The r_1 and r_2 are random values in the range [0, 1], sampled from a uniform distribution. These random values introduce a stochastic element to the algorithm. The c_1 and c_2 are positive acceleration coefficients used to scale the contribution of the cognitive and social components, respectively, and are also referred to as trust parameters, where c_1 expresses the confidence of a particle in itself, and c_2 expresses the confidence of a particle in its neighbors. Particles gain strength by cooperation and are most effective when c_1 and c_2 are well-balanced. Low values result in smooth particle trajectories, which allow particles to roam among different regions. High values cause particles to move abruptly from one region to another region.

To improve PSO convergence, [8 9 10] proposed a strategy for incorporating inertial weight w as a mechanism for controlling swarm exploitation and exploration by weighting the contribution of the previous velocity. The w control how much the memory of the previous flight direction influences the new velocity. For $w \geqq 1$, velocity increase over time, accelerates to maximum velocity, and the swarm diverges. Particle fails to change direction to move back towards promising areas. For $w \leqq 1$, particles decelerate until their velocity is zero. The velocity update with inertia is given as

$$v_{i,j}^{t+1} = w \cdot v_{i,j}^{t} + c_1 \cdot r_1 \cdot \left(p_{i,j} - x_{i,j}^{t} \right) + c_2 \cdot r_2 \cdot \left(p_{g,j} - x_{i,j}^{t} \right) \tag{11}$$

The authors in (11) also proposed maximum velocity v_{max}. The v_{max} was calculated as a fraction δ $(0 < \delta < 1)$ of the distance between the bounds of the search space as follows

$$v_{max,j} = \delta \left(x_{max,j} - x_{min,j} \right) \tag{12}$$

The PSO with constant inertia w and maximum velocity limitation $v_{\max}$ is referred to here as *Basic PSO*.

3 PSO for CCPO Problems

This section describes the proposed PSO approach to CCPO problems. Here, a particle position is formulated as a portfolio, and the position of particle i in dimension j is the proportion of capital to be invested in the jth asset.

This study found that PSO quickly stagnates to the local optimum in CCPO problems in order to satisfy the constraint on the desired number of assets K in the portfolio especially when CCPO problems consider high values for risk aversion parameter λ. This study therefore considered possible remedies for constraint satisfaction in the PSO to avoid stagnation in early phase.

3.1 Initialization

In PSO initialization phase, a size N swarm is randomly generated. The individuals in the swarm are randomly valued for each dimension between the bounds. Similarly, the velocity is initialized to zero in each dimension.

3.2 Constraints Satisfaction

New positions may leave the search space when updating particle positions in search process. In this case, the intuitive solution is setting the value of the new position to the boundary value for the asset of the portfolio. This causes rapid stagnation of the PSO to the local optimum. To attain better diversity to the search space, the reflection strategy suggested in Paterlini and Krink [11] is applied during the initial search phase. That is if the value of the new position leaves the domain of the search space, it is reflect back into the domain by

$$x_{i,j}^t = x_{i,j}^t + 2(x_j^l - x_{i,j}^t) \quad if\ x_{i,j}^t < x_j^l \tag{13}$$

$$x_{i,j}^t = x_{i,j}^t - 2(x_{i,j}^t - x_j^l) \quad if\ x_{i,j}^t > x_j^u \tag{14}$$

where x_j^u and x_j^l are the upper and lower bounds of each jth component, respectively. This method allows the particle to explore a larger search area and to escape from local minima, which improves solution quality. The reflection strategy terminates when no improvement is obtained after numerous iterations. Then the boundary value was set by

$$x_{i,j}^t = x_j^l, \quad if\ x_{i,j}^t < x_j^l. \qquad x_{i,j}^t = x_j^u \quad if\ x_{i,j}^t > x_j^u\,. \tag{15}$$

For handling the cardinality constraints, K is the desired number of assets in the portfolio. Given a set Q of K assets, Let K^{new} represent the number of assets after updating positions in portfolio (the numbers of the proportion w_i greater than 0). If $K^{new} < K$, then some assets must be added to Q; if $K^{new} > K$, then some assets must be removed from Q until $K^{new} = K$.

Considering the removal of assets in the case $K^{new} > K$. This study deletes the smallest assets. If $K^{new} < K$ assets, assets remaining to be added must be identified. This study randomly adds an asset $i \notin Q$ and assigns the minimum proportional value ε_i to the new asset.

According to Eq. (7), the value of $\boldsymbol{x_i}$ must also satisfy $0 \leqq \varepsilon_i \leqq x_i \leqq \delta_i \leqq 1$ for $i \in Q$. Let s_i represent the proportion of the new position belonging to Q. If $s_i < \varepsilon_i$, the minimum proportional value of ε_i replaces asset s_i. If $s_i > \varepsilon_i$, the proportional share of the free portfolio is calculated as follows :

$$x_i = \varepsilon_i + \frac{s_i}{\sum_{j \in Q, s_i > \varepsilon_i} s_i}(1 - \sum\nolimits_{j \in Q} \varepsilon_i) \tag{16}$$

This minimizes the proportional value of ε_i for the useless assets $i \in Q$ so that particles converge faster in the search process, especially in CCPO problems involving low values for risk aversion parameter λ.

3.3 Inertia Weight (w)

The inertia weight w controls how previous velocity affects present velocity. High w values emphasize exploration for the global search the optimal solution while low values emphasize the local search around the current search area. All population-based search techniques rely on global exploration and local exploitation to achieve good

performance. Generally, exploration should be most intensive in initial stages when the algorithm has very little knowledge about the search space, whereas later stages require additional exploitation requiring the algorithm to exploit information it has gained so far.

Since CCPO problems involve complex search space, the parameter w becomes vital in PSO algorithms. Therefore, the proposed PSO uses the time variant w for CCPO problems introduced by Shi and Eberhart [12]. The w is linearly reduced during the search process. Therefore, inertia values are initially large and decrease over time. Particles tend to explore in the initial search steps and tend to exploit as time increasingly. The w at time step t update is obtained by

$$w(t) = (w(0) - w(n_t))\frac{(n_t - t)}{n_t} + w(n_t) \tag{17}$$

where n_t is the maximum number of time steps required to execute the algorithm , $w(0)$ is the initial inertia weight, $w(n_t)$ is the final inertia weight, and $w(t)$ is the inertia at time step t. Usually $w(0) = 0.9$ and $w(n_t) = 0.4$ [16].

3.4 Acceleration Coefficients (c_1 and c_2)

If c_1 is larger than c_2, each particle has a stronger attraction to its own best position, and excessive wandering occurs. On the other hand, if c_2 exceeds c_1, particles are most attracted to the global best position, which causes them to rush towards the optima prematurely. The ratio between c_1 and c_2 coefficients is problem-dependent. Most applications use $c_1 = c_2$. Since c_1 and c_2 are usually static, their optimized values are found empirically.

To ensure a more global search during initial stages and a more local search during the final stages of CCPO problems, the proposed PSO adopts time variants c_1 and c_2 as demonstrated by Ratnaweera *et al.* [13]. Over time c_1 decreases linearly, and c_2 increases linearly. This strategy focuses on exploration in the early stages of optimization process by trusting itself, and encourages convergence to a good optimum near the end of the optimization process by trusting the best particle. The c_1 and c_2 at time step t update is given as

$$\begin{aligned} c_1(t) &= (c_{1,min} - c_{1,max})\frac{t}{n_t} + c_{1,max} \\ c_2(t) &= (c_{2,max} - c_{2,min})\frac{t}{n_t} + c_{2,min} \end{aligned} \tag{18}$$

where usually $c_{1,max} = c_{2,max} = 2.5$ and $c_{1,min} = c_{2,min} = 0.5$ [16].

3.5 Mutation

Similarly, to allow our proposed PSO algorithm to maximize diversity, mutation operator was used based on [12]. A similar mutation operator was used in the PSO for

multi-modal function optimization. Given a particle, a randomly chosen variable, say g_k, is mutated as given below:

$$g_k' \begin{cases} g_k + \Delta(t, \mathrm{UB} - g_k) & \text{if } flip = 0, \\ g_k + \Delta(t, g_k - \mathrm{LB}) & \text{if } flip = 1. \end{cases} \tag{19}$$

where *flip* denotes the random event of returning 0 or 1. The UB denotes the upper limit of the variable g_k while LB denotes the lower limit. The function Δ is defined as

$$\Delta(t, x) = x \cdot \left(1 - r^{\left(1 - \frac{t}{\max\_t}\right)^b} \right) \tag{20}$$

where r is a random number generated in the range [0, 1], max_t is the maximum number of iterations, and t is the iteration number. The parameter b determines the dependence of the mutation on the iteration number.

3.6 Termination

The algorithm terminates when no improvement occurs over repeated iterations.

4 Computational Experiments

To test the performance of the proposed PSO for CCPO problems, the computational experiments were performed. The experiment compared the performance of the proposed PSO with basic PSO, genetic algorithm (GA), simulated annealing (SA), and tabu search(TS) to CCPO problems.

4.1 Definition of Experiments

The proposed PSO searches for efficient frontiers by testing 50 different values for the risk aversion parameter λ in the cardinality-constrained Markowitz portfolio model. The experiment employed the five benchmark datasets used earlier in [3, 4] These data correspond to weekly price data from March, 1992 to September, 1997 for the following indices: Hang Seng 31 in Hong Kong, DAX 100 in Germany, FTSE 100 in UK, S&P 100 in USA and Nikkei 225 in Japan. The number N of different assets considered for each index was 31, 85, 89, 98 and 225, respectively. The sets of mean return of each asset, covariance between these assets and efficient frontier 2000 points are publicly available at http://people.brunel.ac.uk/mastjjb/jeb/orlib/ portinfo.html. The cardinality constraints used the values K =10, ε_i =0.01 and δ_i = 1 for problem formulation.

The criteria used to quantify the performance of proposed PSO for CCPO problem was accuracy. Accuracy refers to the quality of the solution obtained. This analysis used the standard deviation (risk) and return of the best solution for each λ to compare standard efficient frontiers and to measure percentage error respectively, and the lower value of standard deviation error and mean returns error was used as the percentage error associated with a portfolio. For example, let the pair (s_i, r_i) represent the standard

deviation(risk) and mean return of a point obtained by PSO. Additionally, let s_i^* be the standard deviation corresponding to r_i according to a linear interpolation in the standard efficient frontier. The standard deviation of return error e_i for any point (s_i, r_i) is defined as the value $100\,(s_i^* - s_i)/s_i^*$. Similarly, by using the return r_i^* corresponding to s_i according to a linear interpolation in the standard efficient frontier, mean return error η_i can be defined as the quantity $100(r_i - r_i^*)/r_i^*$. The error measure defined in [3] was calculated by averaging the minimums between the mean return errors e_i and the standard deviation of return errors η_i.

Numerous empirical studies show that the PSO is sensitive to control parameter choices such as inertia weight, acceleration coefficients and velocity clamping[14]. This study applied the following parameters suggested in the literature: For basic PSO, the value (w, c_1, c_2, v_{max}) set to (0.7298, 1.49618, 1.49618, 1) as suggested in [14]. In the proposed PSO, the value $c_{1,max}$=$c_{2,max}$ set 2.5, $c_{1,min}$=$c_{2,min}$=0.5, the $w_{initial}$ =0.9 linearly decreased w_{final} =0.4, and b=5 as suggested in [12, 13, 15, 16].The reflection strategy terminates when no improvement exceeds 20 iterations. Swarm size is set to 100 for all PSOs, and the algorithms terminate when no improvement occurs over 100 iterations. The average value of twenty-five trials for each test was recorded. Both PSO algorithms presented in this paper were coded in MATLAB language and tested on a Pentium M processor with 1.6 GHz CPU speed and 1 GB RAM machine, running Windows.

4.2 Computational Results

Table 1 shows the minimum mean percentage error of portfolio for the proposed PSO and basic PSO. The best minimum mean percentage error for each problem is in boldface. Clearly, the proposed PSO generally obtained a lower minimum mean percentage error than basic PSO did. To compare the proposed PSO with other heuristics, the same data sets were considered in the constrained portfolio problem. Table 1 also shows that comparable results were obtained genetic algorithm (GA), simulated annealing (SA), and tabu search(TS) for minimum mean percentage error. The results on GA, SA, and TS are from [3]. The proposed PSO was run for 1,000 iterations using 100 particles. The parameter settings were approximately the same number of solution searched by the heuristic with which we compare our results. The results for the proposed PSO were averaged over 25 trials. The minimum mean percentage error for each problem is given in boldface. The proposed PSO almost always obtained the best performance in most cases.

Table 1. Experimental results for CCOP problems

Problem name	assets(N)	GA	SA	TS	Basic PSO	Proposed PSO
Hang Seng	31	1.0974	1.0957	1.1217	1.1047	***1.0953***
DAX100	85	2.5424	2.9297	3.3049	2.9205	***2.5417***
FTSE100	89	1.1076	1.4623	1.6080	1.42781	***1.0628***
S&P100	98	1.9328	3.0696	3.3092	2.5554	***1.6890***
Nikkei	225	0.7961	***0.6732***	0.8975	0.96459	0.6870
Average		1.4953	1.8461	2.0483	1.7946	***1.4152***

5 Conclusion

This work developed an improved PSO for identifying the efficient frontier in portfolio optimization problems. The standard Markowitz mean-variance model was generalized to include cardinality and bounding constraints. Such constraints convert the portfolio selection problem into a mixed quadratic and integer programming problem, for which computationally efficient algorithms have not been developed. The proposed PSO was tested on CCPO problem set. The test results confirmed that incorporating the improved constraint handling increased PSO exploration efficiency in the initial search steps and increased convergence speed in the final search steps. Comparisons with basic PSO also showed that the proposed PSO is much more robust and effective, especially for low-risk investments. Solution comparisons showed that the proposed PSO outperformed genetic algorithm (GA), simulated annealing (SA), and tabu search (TS). Further research using Ant colony optimization to solve the CCPO problem is currently underway.

References

1. Markowitz, H.: Portfolio selection. Journal of Finance, 77–91 (1952)
2. Maringer, D., Kellerer, H.: Optimization of cardinality constrained portfolios with a hybrid local search algorithm. Or. Spectrum 25(4), 481–495 (2003)
3. Chang, T.J., Meade, N., Beasley, J.E., Sharaiha, T.M.: Heuristics for cardinality constrained portfolio optimisation. Computers & Operations Research 27(13), 1271–1302 (2000)
4. Fernandez, A., Gomez, S.: Portfolio selection using neural networks. Computers & Operations Research 34(4), 1177–1191 (2007)
5. Crama, Y., Schyns, M.: Simulated annealing for complex portfolio selection problems. European Journal of Operational Research 150(3), 546–571 (2003)
6. Kennedy, J., Eberhart, R.C., Shi, Y.: Swarm intelligence. The Morgan Kaufmann Series in Evolutionary Computation, San Francisco (2001)
7. Clerc, M.: Particle swarm optimization, London (2006)
8. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: Proceedings of IEEE World Congress on Computational Intelligence Evolutionary Computation (1998)
9. Fourie, P., Groenwold, A.: The particle swarm optimization algorithm in size and shape optimization. Structural and Multidisciplinary Optimization 23(4), 259–267 (2002)
10. Clerc, M., Kennedy, J.: The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation 6(1), 58–73 (2002)
11. Paterlini, S., Krink, T.: Differential evolution and particle swarm optimisation in partitional clustering. Computational Statistics and Data Analysis 50(5), 1220–1247 (2006)
12. Tripathi, P.K., Bandyopadhyay, S., Pal, S.K.: Multi-Objective Particle Swarm Optimization with time variant inertia and acceleration coefficients. Information Sciences 177(22), 5033–5049 (2007)
13. Ratnaweera, A., Halgamuge, S., Watson, H.: Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. IEEE Transactions on Evolutionary Computation 8(3), 240–255 (2004)
14. van den Bergh, F., Engelbrecht, A.P.: A study of particle swarm optimization particle trajectories. Information Sciences 176(8), 937–971 (2006)
15. Schutte, J., Groenwold, A.: A study of global optimization using particle swarms. Journal of Global Optimization 31(1), 93–108 (2005)
16. Engelbrecht, A.P.: Fundamentals of computational swarm intelligence. Wiley, Chichester (2005)

Immune Memory Mechanism Based on Cyclic Idiotypic Network

Chung-Ming Ou[1] and C.R. Ou[2]

[1] Department of Information Management, Kainan University, Luchu 338, Taiwan
cou077@mail.knu.edu.tw
[2] Department of Electrical Engineering, Hsiuping Institute of Technology, Taichung 412, Taiwan
crou@mail.hit.edu.tw

Abstract. Immune memory can be regarded as an equilibrium state of immune network system with nonlinear behavior. The rapid response of immune systems to the second-time antigen is owing to the stable structure of memory state forming by a closed loop of the idiotypic immune network. A dynamical system is proposed which explains how the memory state is formed and stabilized in the immune network.

1 Introduction

Immune memory mechanism can be explained through Jerne's immune network theory by investigation them as complex adaptive systems [1]. Many researches have been proceeded according to this Immune network theory [2]. Mathematical description of this theory is contributed by Perelson [3] and Parisi [4].

It is well-known that the introduction of a given amount of some antigen into a mammal's body will stimulate the production of antibodies directed against that antigen, if the antibody is with a high affinity for that antigen. The immune system of the animal has thus learned to produce high quantities of the antibody directed against that very antigen, which is called vaccinated. From the development of Jerne's network theory, we realize the following principles are needed. First is the need of interaction between various species of antibodies which plays an important role in immune regulation and memory. Secondly, the immune system is composed of a number of smaller network systems. Many idiotypic network models focus on the interactions between antibodies and antigens.

Immune memory mechanism can be modeled from immune network theory [4]. Once the foreign antigen is removed, the immune system will restore some information of such antigen by some memory mechanism. The effect of immune memory can contribute to the rapid response of the same type of antigen, so called the second immune response. The immune memory can be explained by the following network view point. Assuming that antibody Ab_1 is produced by modeling the stimulating antigen, the production of Ab_1 is increased in the presence of another type of antibody Ab_2. The population of T-helper cell TH_1 specified by Ab_1 is also increased. In this way, Ab_2 can be considered as some

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 416–425, 2010.

"internal image" of this antigen. This image will be remained after the antigen is removed. The interactions can be a long chain with length greater than 2.

Jerne's theory implicitly retains the concepts of network stability, which can be modeled through the nonlinear dynamical systems [5]. Clonal selection theory proposed by Burnet has explained the secretion of an antibody specific to an antigen. For more mathematical description of immune network theory, see [6]. However, lots of efforts are trying to faithfully express Jerne's idea. One philosphy is to reconsider the real immune system and clarify the roles of the lymphocyte and the antibody in the regulation through the nonlinear dynamical systems. Immune memory mechanism will be explained through this nonlinear dynamical system. On the other hand, Smith et al. [7] suggested that the immune memory is a member of the family of sparsely distributed memory; it derives associative and robust properties. The formation of immune memory is related to concentration levels of various immune cells during the primary immune and secondary immune responses. This concepts is the major idea throughout this paper. The existence of the immune memory implies then the stabilizing of the new steady state. The stability is a requirement of the homeostasis of the biological system. This study focuses on the immune memory strategy from the dynamical system's point of view.

The major goal of this paper is to study the memory mechanism based on nonlinear dynamical system of idiotypic immune network architecture proposed by [5]. We consider lymphocyte concentration in addition to antibody concentration for memory mechanism. Antigens bind populations of antibodies, which control the immune response and produce actions of differentiation and regulation.The arrangement of this paper is as follows. In section 2, some preliminary knowledge of immune memory is introduced. In section 3, dynamical behavior of idiotypic immune network is described. Stability analysis of immune network is given in section 4.

2 Preliminary Knowledge

Modeling an immune system is a major challenge for researchers. According to experimental findings, some theoretical considerations may help researchers select suitable set of interactions using mathematical knowledge. Firstly, it is well-known that immune system, if regarded as a network, is stable or quasi-stable [2] [8]. Therefore, equations describing dynamics of immune network must have stable solutions. Furthermore, the mathematical simulation of immune behavior must predict the 'switching' among states according to observed data. Based on Jerne's immune network theory, lymphocyte cells are communicating one another. The formation of such immune network systems is based on the following considerations: interactions between B-cells and T-cells (such as T-helper (T_H) cells and T-suppression (T_S) cells); interactions between Lymphocyte units which including antibody's communications.

2.1 Lymphocyte Unit

The immune system consists of the antibodies and lymphocytes, which include T cells and B-cells. The human immune system uses a large number of highly specific B- and T-cells to recognize antigens. Only B-cells secrete antibodies. Clonal selection theory explained the details of antibody secretion specific to an antigen where T-cells help regulating. The binding between antigen and specific lymphocytes trigger proliferation from immature lymphocytes to mature one and the secretion of antibodies. Antigen binds to receptors of the specific immature lymphocyte. The specific immature lymphocyte is selected by some antigen. Such binding triggers proliferations from both the T-cell and B-cell lymphocytes.

2.2 Idiotypic Immune Network

Idiotypic network theory implies that cells co-stimulate each other in a way that mimics the presence of the antigen [7]. Antibody and receptor of the lymphocytes can recognize each other. The epitope of antibody molecule is called idiotope. An epitope of antigen A_g is recognized by the antibody molecule Ab_1 and by the receptor molecule on the lymphocyte of LU_1. The antibody Ab_1 and the receptor of LU_1 have the idiotope which is recognized by antibody Ab_2 and the receptor on the lymphocyte of LU_2. On the other hand, the antibody Ab_1 and the receptor on the lymphocyte of LU_1 also recognize idiotopes on antibody Ab_n. Ab_n constitutes an internal image of the antigen A_g.Network forming by interactions between lymphocyte interactions. This Ab_n constitutes an internal image of the antigen Ag, see Fig. 1. The idiotypic network theory has been proposed as a central aspect of the associative properties of immune memory [7] [10].

2.3 Immune Memory

Immune system will react rapidly to the same antigens which had invaded human bodies. This phenomenon implies that immune system can "memorize" the formations of previously invaded antigens.

Immune memory mechanism is not fully understood so far; a number of mechanisms have been proposed. In vaccination, it is empirically known that an immune memory to a viral antigen is sustained more durable than the one to the non-viral antigen [11]. According to Smith et al. [7], at the end of an immune response, when the antigen is cleared, the B cell population decreases, leaving a persistent sub-population of memory cells. The newer view of memory cells is that they are not intrinsically longer-lived than virgin cells, but that memory depends instead on the persistence of antigen [12]. On the other hand, some researches, especially those related to immune network theory, imply that mechanisms of immune memory is formed by rather cyclic idiotypic networks than specific memory cells [5].

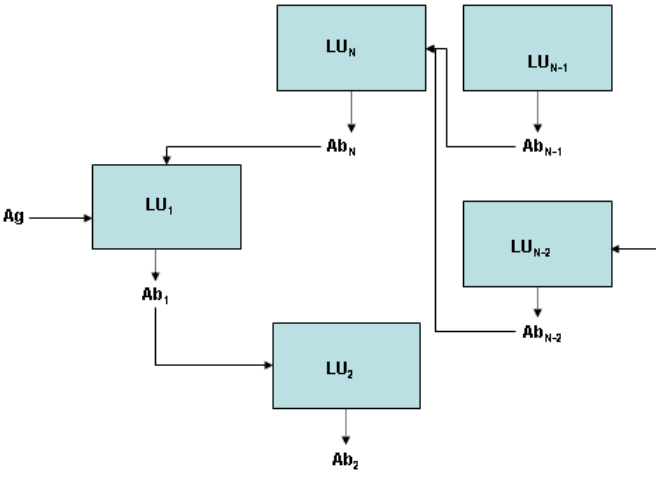


Fig. 1. Schematic diagram of the idiotypic network

3 Immune Memory and Dynamical Behaviors of Idiotypic Network

One major challenge for idiotypic networks is the modeling of the dynamical immune behaviors. Immune network theory tries to model network properties of lymphocytes in the absence of antigens; this is exactly the memory state of immune systems. We formulate the interactive behaviors among these lymphocytes by systems of ordinary differential equations; this investigation is based on the analysis of idiotypic network and the interactions within LUs.

3.1 BHS Model for Interactions within Lymphocyte Unit

We define the immune network first according to LU and some interaction matrix M as follows. M can be computed from the chemical structure of the antibodies.

Definition 1. *An immune network is defined by* $\langle \{LU_i\}_{i=1}^N, M \rangle$ *, where* $\{LU_i\}_{i=1}^N$ *is a directed graph of* N *lymphocyte units,* M *is an* $N \times N$ *interaction matrix with entries* $m_{i,j} \in [0,1]$ *and satisfies the following conditions:*

1. $m_{ik} > 0$ *represents* Ab_k *triggers the production of* Ab_i*;*
2. $m_{ik} < 0$*,represents* Ab_k *suppresses the production of* Ab_i *.*

In particular, $\langle \{LU_i\}_{i=1}^N, M \rangle$ *is cyclic, if*

- $m_{i+1i} \neq 0$ *for all* i*;*
- $m_{ij} = 0$*, otherwise.*

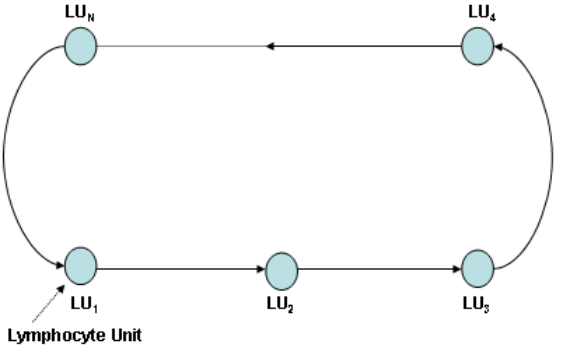


Fig. 2. Cyclic Immune Network

In fact, M plays the same role of the synaptic strength of the neural network. We define several states of the immune response as follows.

Definition 2. *Consider the immune system* $\langle \{LU_i\}_{i=1}^{N}, M \rangle$. *Three different equilibrium states are defined.* $\boldsymbol{H}$ *and* $\boldsymbol{S}$ *represent the concentrations of* T_H *cell and* T_S *cells, respectively. A dormant state is defined by* $(0,0) \in \boldsymbol{H} \times \boldsymbol{S}$. *A suppression state is defined as the equilibrium state with* $S > 0$; *a memory state is defined as the equilibrium state with* $S > 0$ *and* $H > 0$.

The existence of suppressed states could account for the immune tolerance. According to Sonoda [5], equations of the interaction of the lymphocytes can be described as

$$\frac{dH}{dt} = a_{HH} H\Phi(H, t_{HH}, n_{HH}) - a_{HS} H\Phi(S, t_{HS}, n_{HS}) + a_{HA} H\Phi(A, t_{HA}, n_{HA}) + \nu_H \Phi(A, t_{\nu H}, n_{\nu H}) - d_H H \quad (1)$$

$$\frac{dS}{dt} = a_{SH} S\Phi(H, t_{SH}, n_{SH}) - a_{SS} S\Phi(S, t_{SS}, n_{SS}) + \nu_S \Phi(A, t_{\nu S}, n_{\nu S}) - d_S S \quad (2)$$

H, S and A are variables representing the concentrations of T_H cells, T_S cells and antigens respectively. The coefficients a_{XY} represents the strength of the action of the Y cell to X cell. ν_X represents the concentration of immature-X cell. d_X represents the decay rate of X cell. Φ is a nonlinear function. For example, $\Phi(x,t,n) = \frac{x^n}{t^n + x^n}$. $\nu_X \Phi(A, t_{\nu X}, n_{\nu X})$ represent that an immature-lymphocyte is stimulated by antigens so as to proliferate to mature lymphocytes by the clonal selection theory. The antibody concentration B satisfies the following equation:

$$\frac{dB}{dt} = \nu_B \Phi(A, t_{\nu B}, n_{\nu B}) - d_B B \quad (3)$$

Equations (1)-(3) are called the BHS model which describes the dynamics within LUs.

3.2 Dynamics between Lymphocyte Units: Cyclic Idiotypic Network

BHS model is complex due to the numbers of LUs. However, this model can be greatly simplified by considering *cyclic* idiotypic networks, which can contribute to the immune memory mechanism. BHS model has its disadvantages. For example, it does not consider the interactions between antibodies and B-cells. Evidences show that idiotypic/anti-idiotypic interactions among antibodies and B-cells influence on shaping B-cells and T-cells repertoire in the newborn immune system [9].

Farmer et al. [10] introduce the matching coefficient m_{ij} to represent the strength of a matching between an epitope of i-th antibody and an idiotope of j-th antibody. m_{ij} can also be used for a matching between an epitope of i-th antibody and a paratope of j-th antigen. Now we concentrate on the dynamics of immune system based on cyclic idiotypic network $\langle \{LU_i\}_{i=1}^{N}, M \rangle$.

3.3 Immune Memory Described by Cyclic Idiotypic Network

The immune memory can be formed and preserved by a cyclic loop within the idiotypic network (Fig. 3). When an antigen is injected into the immune system, antibody stimulation propagates in LU_i successively. For all LUs forming a closed loop in the network, these active states are preserved and a *memory state* is asymptotically formed. Memory state is defined as an equilibrium state for dynamical systems of (1)-(2).

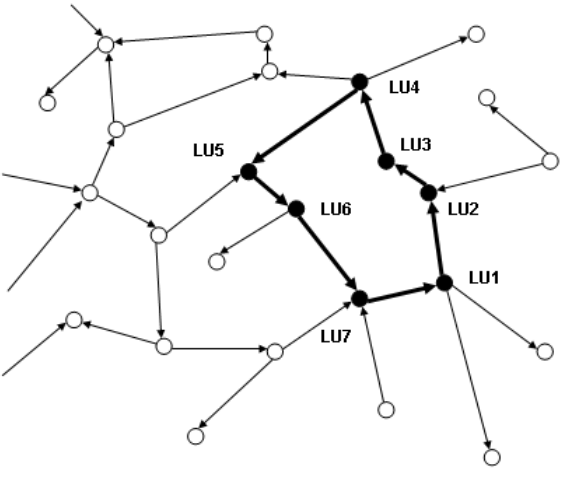


Fig. 3. Formation of the Memory State in the Cyclic Immune network ($N = 7$). For example, the arrow from LU_1 to LU_2 represents $m_{12} \neq 0$.

For T_H cell and T_S cell, it is suggested that both the clonal selection term and the decay term can be neglected [5]. Therefore, (1)-(2) can be deduced to the following equations for the active state.

$$\frac{dH_i}{dt} = [a_{HH}H_i - a_{HS}S_i + a_{HA}A_{i-i}]H_i \tag{4}$$

$$\frac{dS_i}{dt} = (a_{SH}H_i - a_{SS}S_i)S_i \tag{5}$$

There are three different cases for equilibrium states (besides dormant state):

- One equilibrium state, namely, dormant state.
- Two equilibrium state, one dormant state, the other is the memory state.
- Three equilibrium states, one dormant state, one is suppression state and the other one is the memory state

4 Analysis of Immune Network Systems

Once the immune regulation fails, the stability of immune system will be destroyed. This leads to multiple diseases. The stability of immune system depends mainly on the regulatory mechanism. Such mechanism in its essence is a self-regulatory, which can be characterized by immune networks without any central control mechanism. In details, every antibody has its own idiotypic determinants. Such idiotypic antibody may activate or suppress immune responses even the invaded antigens have been removed.

4.1 Formation of Immune Memory Based on Antigen Dynamics

Immune memory has been discussed in previous section based on the HS- and BHS-model. Now we provide another analysis of immune memory mechanism according to the dynamics of antigens. First we formerly define an active state of immune network as follows.

Definition 3. *The immune network* $\langle \{LU_i\}_{i=1}^N, M\rangle$ *is in an active state, if* $B_i, H_i, S_i > 0$*, for all* $i = 1, 2, \ldots N$*.*

When the immune network is in the active state, the B cell concentration remains (asymptotically) constant and the immune system can rapidly respond to the antigen. If the active state can be preserved for a "reasonable" time, such preserved state can be considered as a memory state.

According to this simplified model (4)-(5), we can describe the bistable behavior of T_H (T_S) cells, whose concentration can be a function of antigen concentration. When the antigen concentration is increasing, T_H will switch from a low active state into a high one. For suitable parameters, active states will act like Fig. 4. Two states are introduced, namely, low active state and high active state. When antigen invades, immune system switches from dormant state to low active state. The latter will switches furthermore to the high active state at some threshold value of antigen concentration $A = A_{tres}$. Accordingly, we have the following proposition for the immune system.

Proposition 1. *The immune system can reach high active state by the interactions of lymphocytes alone, without the interactions among antibodies*

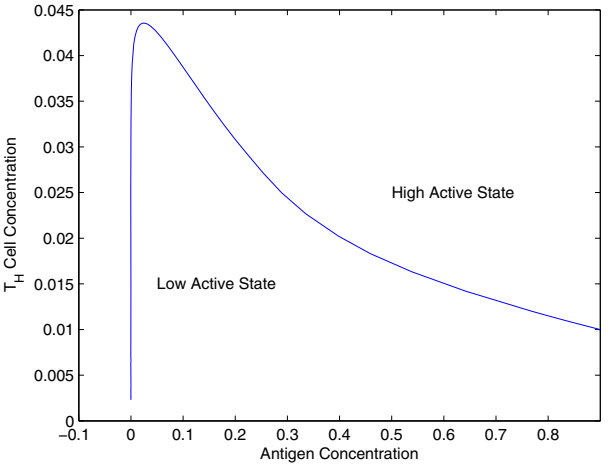


Fig. 4. Bistable Behavior of T_H cell concentration with respect to the antigen concentration A ($A_{tres} \approx 0.05$)

4.2 Dynamics of Antibody

After the antigen has been removed, the behavior of T_H cell and the B-cell in LU_i is as follows [5].

$$H_i \approx h \sum_{j=1}^{N} m_{ij} I_j + e_i \tag{6}$$

$$B_i \approx b_i \tag{7}$$

Where $e_i = 0$ represents LU_i is at the low active state, otherwise, high active state. Accordingly, we have the equation of antibody concentration in the active state. Now we have the equation of antibody dynamics in the cyclic immune network $\langle \{LU_i\}_{i=1}^{N}, M \rangle$.:

$$\frac{dI_i}{dt} = pI_{i-1}^2 + q_i I_{i-1} - rI_i I_{i+1} - rI_i I_{i-1} - sI_i, i = 1, \ldots N \tag{8}$$

4.3 Stability of Memory States

We now consider the immune memory mechanism from dynamics of antibody behavior. According to the first immune response, after the antigen has been removed, the immune system is under some stable states until the second (same) antigen stimulation.

The active state of LU_i is maintained by the cyclic immune network. The corresponding antibody I_i has two steady states,, namely, I_i^A and $I_i^D = 0$. In the memory state, concentrations of H and S are approximately constant, and the antibody concentration is around the active state $I_i^A \neq 0$, which shows the

existence of mature B cells. We will simulate this second immune response later. Now we theoretically give the following theorem.

Theorem 1. *Consider the cyclic idiptypic network* $\langle \{LU_i\}_{i=1}^N, M \rangle$.

1. *If all* LU_i *are in the high active states, then* I_i^A *are stable, for all* i.
2. *If any* LU_i *is not in the high active state, then* I_i^D *are stable, for all* i.

4.4 Simulations

In this section, we simulate the dynamics of (8) with $N = 3$ where each lymphocyte unit is at high active state. For this 3-cycle immune network, this simulation shows the second immune response of the same antigen is more rapidly than the first response, as the antibodies Ab_2 and Ab_3 generated in the second immune response are around 1.5 and 3 times than those in the first immune response. These antibodies become stable again soon after the second immune response ends, see Fig. 5.

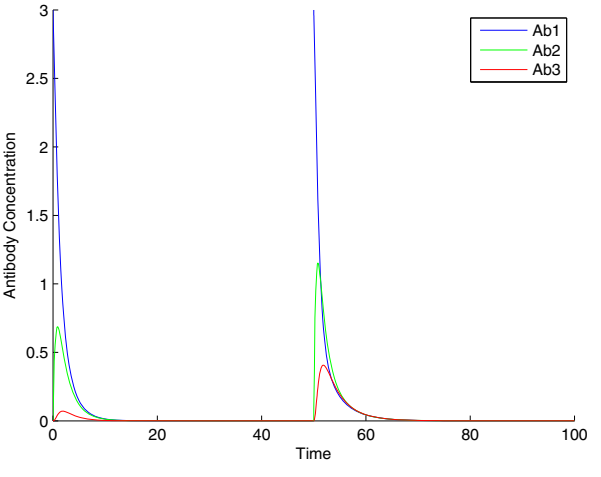


Fig. 5. Second Immune response to Antigen A_1 for (8) with $N = 3$. Memory state is reached around time=10.

5 Conclusions

We propose several dynamical systems of immune response mechanism based on the cyclic idiotypic network, namely, HS model, BHS model and Antibody-Antigen equation. According to the stability of each lymphocyte unit, the formation of immune memory can be deduced by close loop of the cell's interactions. Immune Memory mechanism can be interpreted by these models within active states.

Acknowledgment

The authors would like to thank support from National Science Council of Taiwan under the grant number NSC-98-2221-E-424-004.

References

1. Jerne, N.: Ann. Immunol. (Inst. Pasteur) 125C, 373 (1974)
2. Hoffman, G.W.: A theory of regulation and self-nonself discrimination in an immune network. Eor. J. Immunol. 5, 638–647 (1975)
3. Perelson, A.: Mathematical Approaches in Immunology. In: Andersson, S., et al. (eds.) Theory & Control of Dynamical Systems: Applications to Systems in Biology, pp. 200–230. World Scientific Publishing, Singapore (1992)
4. Parisi, G.: A Simple Model for the Immune Network. Proceedings of the National Academy of Sciences 87, 429–433 (1990)
5. Sonoda, T.: Formation and Stability of a Memory State in the Immune Network. J. Physical Society of Japan 61(4), 1408–1424 (1992)
6. Perelson, A. (ed.): Theoretical Immunology, Part One and Two, The Proceedings of the Theoretical Immunology Workshop. Addison-Wesley, Reading (1988)
7. Smith, D., Forrest, S., Perelson, A.: Immunological Memory is Associative, Artificial Immune Systems and Their Applications. In: 1999 The International Conference on the Multi-Agent Systems Workshop Notes, Kyoto, pp. 62–70 (1996)
8. Nielsen, K., White, R.: Nature 250, 235 (1974)
9. Seledtsov, V., Seledtsova, G.: A Pendular Mechanism for Maintenance of the Immune Memory. Medical Hypotheses 56(2), 160–162 (2001)
10. Farmer, J., Packard, N., Perelson, A.: THE Immune System, Adaptation, and Machine Learning. Physica 22D, 187–204 (1986)
11. Harada, K.: A Switching Memory Strategy in an Immune Network Model. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 519–525. Springer, Heidelberg (2004)
12. Matzinger, P.: Memories Are Made of This? Nature 369, 605–606 (1994)

Sensor Placement in Water Networks Using a Population-Based Ant Colony Optimization Algorithm

Konrad Diwold, Thomas Ruhnke, and Martin Middendorf

Parallel Computing and Complex Systems Group
Faculty of Mathematics and Computer Science
Universität Leipzig, Germany

Abstract. Water supply networks are of high economical importance. Accidental or deliberate contaminations pose a great threat to such networks and can have wide-ranging implications. Early warning systems can monitor the quality of the water-flow in such systems by means of sensors distributed in the network and report potential contaminations to minimize harm. Sensor placement in water networks usually addresses several objectives and depends on the specific network. Here a population-based ant colony optimization algorithm called — WSP-PACO — for sensor placement in water networks is proposed. The performance of the algorithm was tested on two realistic water networks under several test conditions. These solutions were compared to solutions of previous studies on these networks. The results suggest that WSP-PACO is highly suitable for solving the sensor placement problem in water networks.

Keywords: Water pollution, ant colony optimization, sensor placement, water networks.

1 Introduction

Industrial urbanisation has led to a need for concentrated high-volume water supply. Such demands cannot usually be met by natural water supplies such as rivers and wells but require elaborate artificial municipal water networks (MWNs). The highest threat to such networks are accidental or intentional contaminations as they can have a severe effect on human health as well as on the environment and thus national economies [18]. This has led to an increased focus on the design of early warning systems (EWS) for water supply networks that can detect and trace contaminations. Contamination detection is usually achieved by means of sensors. Therefore one important aspect of EWS is to find a suitable placement of sensors in the network which maximizes the potential detection and minimizes the time needed to detect and localize contamination, in order to minimize the impact to public health [19,3,8].

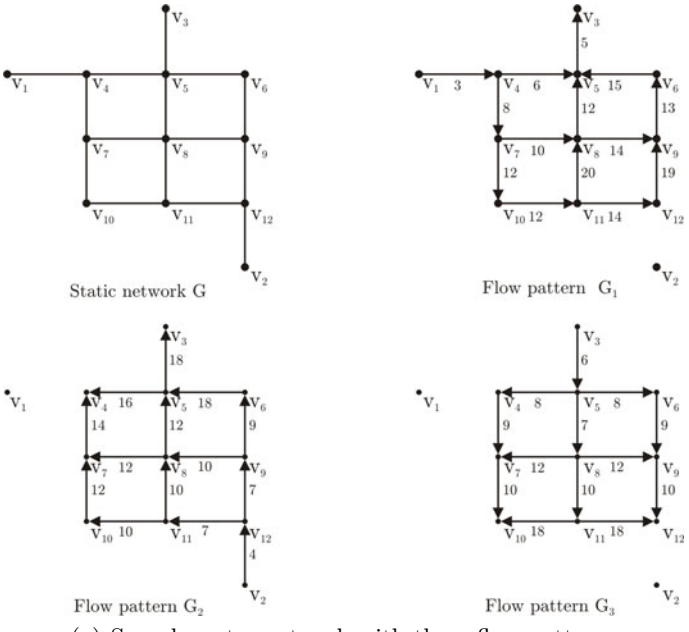
J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 426–437, 2010.

Detecting contaminations in MWNs is not an easy task. Usually only a limited amount of sensors are available, which can only cover a small part of the network. Supply nodes can also be associated with different amounts of consumption and may require prioritizing in terms of protection. Additionally MWNs are usually supplied by several independent water resources (i.e., water reservoirs) resulting in different flow patterns in the network which depend on the reservoirs currently in use [18]. Depending on how many objectives need to be optimized, sensor placement in MWNs pose a single or a multi-objective optimization problem.

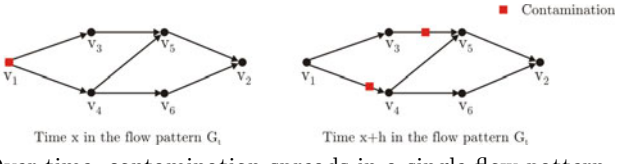
In the last years several algorithmic approaches have been proposed that tackle optimization problems related to the sensor placement problem. Among these problems are to find the minimum number of sensors required to detect contamination in a given MWN and how a fixed number of sensors should be distributed in a network in order to guarantee a good detection rate and time [15]. Potential sensor setups for MNWs can either be evaluated dependent on certain scenarios or independent from specific scenarios. When evaluated independently of a scenario, the optimization goal is to minimize the overall detection latency of any possible contamination in a MWN (e.g., [10,12,14]). In contrast scenario-dependent approaches evaluate sensor topologies according to their detection performance in specific contamination scenarios (e.g., [4,9,11]).

A wide range of methodologies have been used to solve sensor placement in MWNs such as predator–prey models [7], integer programming [2], and evolutionary computation [1]. Here we present a novel approach for the scenario-dependent sensor placement problem in MWNs with multiple objectives based on ant colony optimization. Ant colony optimization (ACO) [5] is a biologically inspired metaheuristic that mimics ant foraging behavior. Virtual agents (ants) construct solutions to a given problem by iteratively extending a partial solution. The extension choice is probabilistic and based on the local heuristics as well as on levels of virtual pheromone. Virtual pheromone is distributed on the elements of a solution after it has been established, with the level of pheromone depending on the relative fitness of the solution. In combination with pheromone evaporation this leads to a stronger concentration of pheromone on the elements of good solutions. As the optimization process proceeds the pheromone heuristic gains more and more influence (due to the increase in the amount of pheromone on good solutions) and will steer agents towards selecting elements that have been part of good solutions in the past. This leads to a convergence towards one single best solution. ACO is especially suited for discrete problems and has been used in many problem domains [6]. A population-based ACO (PACO) has also been introduced [13] to tackle multi-criteria optimization problems. As the sensor placement problem represents a discrete multi-criteria optimization problem in a discrete domain, the PACO methodology seems quite suited to tackle this problem.

This article is structured as follows. In Section 2 the scenario dependent municipal water networks sensor placement problem is described. A population-based ant colony optimization algorithm to solve this problem is introduced in Section 3. Experimental results and comparison to current algorithms are given in Section 4. Section 5 concludes the article.



(a) Sample water network with three flow patterns



(b) Over time, contamination spreads in a single flow pattern

Fig. 1. Example of a water network with different flow patterns and contamination spread after contamination induction

2 Modelling Sensor Placement in MWNs

2.1 Problem Representation

Water networks are usually modelled as graphs, with vertices corresponding to reservoirs, tanks, and supply nodes and edges constituting pipes and pumps. The static version of an MWN (i.e., the network without flow patterns) can thus be represented as a connected undirected graph $G = (V, E)$ consisting of a set of vertices V and a set of edges E where each edge $e \in E$ is of the form $e = \{v_i, v_j\}$ with v_i,$v_j \in V$, $i \neq j$. A flow pattern F for a water network $G = (V, E)$ assigns to each edge $e = \{v_i, v_j\} \in E$ a direction $d(e) \in \{(v_i, v_j), (v_j, v_i)\}$ and a value $w(e) > 0$ which denotes the time that the water needs to flow along in the assigned direction. Thus, if $d(e) = (v_i, v_j)$ then $w(e)$ denotes the time that water

needs to flow from v_i to v_j along e. In the following we consider changes of flow patterns over time. Thus, we consider sequences $F_1, F_2, \ldots, F_k$ of flows such that flow pattern F_i is applied to G from time t_i to time t_{i+1} for integer values t_i, $t_1 = 0 < t_1 < \ldots < t_{k+1} = T$. Such a sequence of flow patterns together with the time intervals is called a flow scenario. It is assumed that a flow pattern is repeated periodically after time T.

For contamination it is assumed that it can be induced at any vertex $v \in V$, and will spread through the water network according to the current flow pattern (see Figure 1b). A contamination event is a vertex where a contamination happens together with the time when this happens. A contamination scenario is a contamination event together wit a flow scenario. Sensors can be placed on vertices and it is assumed that if a contamination reaches a vertex that is equipped with a sensor the contamination will always be detected.

2.2 Objective Functions

As pointed out earlier there are different objective functions that can be used to evaluate a sensor placement in an MWN [15]. How good a placement meets certain objective is usually evaluated via its performance for a set of contamination scenarios. This study focuses on two objectives:

- minimizing the detection time of contaminations in MWNs
- minimizing the non-detection rate of contaminations in MWNs

Let $\mathcal{C}$ denote a finite set of contamination scenarios for a given MWN G. Further let $d(C, S)$ denote the time between contamination and first detection given a contamination scenario C and a sensor placement S. A contamination scenario C is detectable if $d(C, S) < +\infty$. The set of detectable scenarios using a sensor placement S over a set of contamination scenarios $\mathcal{C}$ is thus $D_S^{\mathcal{C}} = \{C \in \mathcal{C} |\ d(C, S) < +\infty\}$. Let $Z_1(S, C)$ denote the minimal contamination detection time of a sensor placement S on the detectable set of contaminations $D_S^{\mathcal{C}}$

$$Z_1(S, \mathcal{C}) = \frac{1}{|D_S^{\mathcal{C}}|} \sum_{C \in D_S^{\mathcal{C}}} d(C, S) \tag{1}$$

The non-detection rate of contaminations in a MWNs is defined accordingly, as

$$Z_2(S, \mathcal{C}) = 1 - \frac{|D_S^{\mathcal{C}}|}{|\mathcal{C}|} \tag{2}$$

This leads to the following multi-objective function to be minimized by placing $p \in \mathbb{N}$ sensors in a given MWN

$$Z(S) = \begin{pmatrix} Z_1(S, \mathcal{C}) \\ Z_2(S, \mathcal{C}) \end{pmatrix} \quad \longrightarrow \min_{\substack{S \subseteq V \\ |S| = p}} ! \tag{3}$$

3 Population Based ACO for the Sensor Placement Problem

Multi-criteria optimization problems such as the sensor placement problem in MWNs do not have a single best solution, due to the trade-off between the optimization goals. Therefore methods for solving such problems try to find a set of optimal non-dominated solutions (also called Pareto solutions). Optimal non-dominated solutions are solutions for whom a fitness improvement of one objective always leads to the decline of quality for at least one other objective. The set of all optimal non-dominated solutions is called the Pareto set.

In order to solve multi-criteria optimization problems with the ant colony optimization meta-heuristic, Guntsch et al. [13] proposed the PACO. As in the standard ACO a set of ants iteratively constructs solutions for the given problem taking into account local heuristics as well as pheromone information. However, rather than remembering only one best solution, all non-dominated solutions are stored in a set – the population. Pheromone is distributed according to this population and the population is updated if new found solutions dominate solutions already in the population.

The PACO has been adapted in this paper for solving the sensor placement problem in MWNs and is called here WSP-PACO. The algorithm uses m ants, which have to find sensor placements of p sensors in the network in order to minimize the objective functions outlined in section 2.2. An ant constructs a sensor placement for a given network iteratively using both local and pheromone heuristics. The local heuristic is dynamic, i.e., the location of the kth sensor depends on the locations of the $k-1$ sensors that the ant has already placed. After each ant has constructed a solution, the population and the pheromone values are updated.

More formally, given p sensors have to be placed in a water network $G = (V, E)$, $V = \{v_1, v_2, \ldots, v_n\}$ with under a given flow pattern F. The pheromones are placed on the vertices. Let τ_i denote the pheromone value of vertex v_i. Initially set to $\tau_i \leftarrow \tau_{init}$. The population which stores the found non-dominated solutions is denoted by Q and is initially empty (i.e., $Q = \emptyset$).

Solution construction. As pointed out above, an ant use the local and pheromone heuristics when deciding on a new partial solution. In WSP-PACO an ant iteratively creates a solution by placing a new sensor, taking into account the sensors it already placed in the network.

Local heuristic. Let η_{ik} denote the heuristic information of vertex v_i as the kth sensor placement

$$\eta_{ik} = \sum_{t=1}^{T} \left(\frac{dg_{in}^{t}(v_i)}{dg_{out}^{t}(v_i) + \epsilon} \right)^{\frac{p-k+1}{p+1}} \cdot \left(\min \left\{ \min_{s \in S_k} \mathrm{dis}_t(v_i, s), D_t \right\} \right)^{\frac{k}{p+1}} \tag{4}$$

where $dg_{in}^{t}(v_i)$ corresponds to the number of incoming edges of node v_i, $dg_{out}^{t}(v_i)$ denotes its outgoing edges, S_k denotes the set of nodes in the network already

equipped with a sensor, $\mathrm{dis}_t(v_i, s)$ is the time that water needs under the flow scenario at time t to flow from vertex v_i to the vertex where sensor s along a fastest directed path, $0 < \epsilon \leq 1$ is a small value to prevent division by zero, and D_t denotes the maximum minimal time that water needs to flow from one node to another node under the flow scenario at time t. Note that η_{ik} takes each flow pattern into account (this similar to Kessler et. al. [10]) and dependents on the sensors that have already been placed. When few sensors are present in the network a new sensor should be placed on nodes that can detect contamination from many sources — this is achieved by the first factor in the formula. When many sensors have already been placed in the network, new sensors should be more evenly distributed in the network — this is achieved by the second factor.

The two superscripts in formula 4

$$\mu(k) = \frac{p-k+1}{p+1} \quad , \quad \nu(k) = \frac{k}{p+1}. \tag{5}$$

guide the impact of the respective parts of the equation regarding the number of sensors that have already been placed k and cause the second factor to gain influence with increasing k.

Probability of placing a sensor. Given the local heuristic η_{ik} and the pheromone information τ_i, the probability that an ant places its kth sensor on node v_i is given by

$$p_{ik} = \frac{[\tau_i]^\alpha [\eta_{ik}]^\beta}{\sum_{i=1}^{n} [\tau_i]^\alpha [\eta_{ik}]^\beta}. \tag{6}$$

where α and β are parameters that determine the relative impact of the respective heuristics.

Population Update. Population Q is updated after each ant has constructed a new sensor placement for the network G. Let Q^{new} denote the new solutions that have been constructed by the ants. $Q^{all} = Q^{new} \bigcup Q$ thus denotes all possible members of Q in the next iteration. Further, let $\tilde{Q}$ denote all non-dominated solutions of Q^{all}. Solutions $S^i \in Q^{new}$ which are dominated by $> \kappa \times |\tilde{Q}|$ other solutions are removed from Q^{all} (if $\kappa = 0$ only non-dominated solutions will be kept in Q^{all}). Q is then set to Q^{all}. Having some dominated solutions in Q (i.e., $\kappa > 0$) can be beneficial for the optimization process as it creates more variation for the pheromone update. By choosing a small value κ, the dominated solutions in Q will typically be close to non-domination.

Pheromone Update. After the new population has been created the pheromone values are updated. Similar to Guntsch et al. [13] only $K >$ solutions of population Q are used for the update. As defined earlier, τ_{init} denotes the initial pheromone level placed on each node v_i. Let τ_{max} denote the maximal pheromone level that can be present on each node, $\tau_{init} < \tau_{max}$.

First a random solution $S_i \in Q$ is selected. Let P denote the set of S_i's closest neighbors according to some distance criteria with $P = Q$ if $k > \#Q$.

The pheromone values of all nodes are updated according to $\tau_i \leftarrow \tau_{init} + \Delta \cdot |\{S \in P | v_i \in S\}|$ where $\Delta = (\tau_{max} - \tau_{init})/K$ denotes the amount of pheromone that is added for each solution in the population.

For better understanding an outline of WSP-PACO is given in Algorithm 1.

Algorithm 1. WSP-PACO

```
while stop criterion not met do
    Constructing solutions
    for i = 1 : m do
        S^i = ∅
        for k = 1 : p do
            for j = 1 : n do
                calculate η_ik according to Eq. 4
                calculate p_jk according to Eq. 6
            end for
            choose random v_j ∈ V according to p_jk
            S^i = S^i ⋃ {v_j}
        end for
    end for
    Changing the population
    alter population Q
    update pheromone of solutions removed from the population
    choose P and update pheromones accordingly
end while
```

4 Experiments

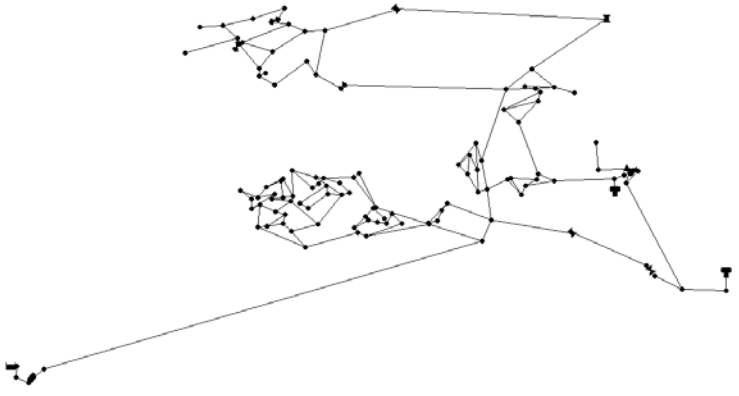
WSP-PACO was implemented and executed using MATLAB. Its performance was tested on two realistic water networks which have already been used in previous sensor placement studies [15,16] [1]. The solutions presented here constitute the non-dominated solutions found via WSP-PACO over 20 simulation runs with a single run lasting 10000 solution construction steps. Unless stated otherwise the following parameter settings were used: $m = 3$, $K = 5$, $\alpha = 1$, $\beta = 1$, $\tau_{init} = 1$, $\tau_{max} = 4$, $\kappa = 1$, and $\epsilon = 0.2$.

Test Networks

Network 1 was previously used in [15] and is depicted in Figure 2a. The network consists of $n = 129$ vertices and 177 edges. The corresponding flow scenario lasts for 96 hours (that is the time for one cycle) and consists of a sequence of 207 flow patterns. As in Ostfeld et al. [15] the algorithm's performance in this network is evaluated for two different numbers of sensors $p = \{5, 20\}$ under three sets of contamination events. The first two sets contain 500 random contamination events and the third set contains 37152 contamination events (taken from [15]).

[1] Both networks can be obtained at http://www.projects.ex.ac.uk/cws/

Network 2 was previously used in [16] and constitutes the water supply system of Richmond, Virginia, USA (see Figure 2b). The network contains $n = 872$ vertices and 957 edges. The flow scenario has a complete flow cycle that lasts for 24 hours and contains a sequence of 55 flow patterns. The number of sensors $p = 5$ was adopted from the previous study [16], and two sets of contamination events each one containing 2000 random contamination events each were used for performance evaluation.



(a) Network 1 previously used in [15]

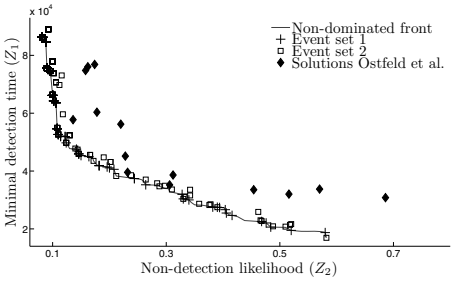
(b) Network 2 previously used in [16]

Fig. 2. Test-networks used to test the algorithms performance. Both networks were visualized using EpaNet [17].

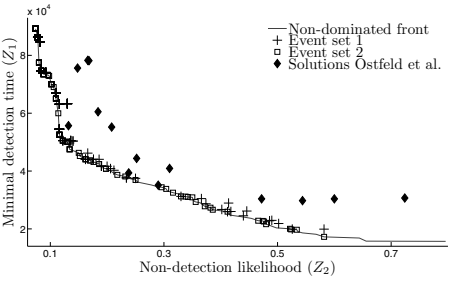
4.1 Network 1

Figure 3 depicts the fitness regarding objectives Z_1 and Z_2 of sensor placements found by WSP-PACO for Network 1. In order to get an estimate of the performance of our algorithm, the plots also contain the fitness of solutions previously presented in [15]. However, it should be noted that these solutions were evolved using a different contamination set and optimized for two more objectives, namely minimization of affected population and minimization of water consumption prior to detection (see [15] for more details).

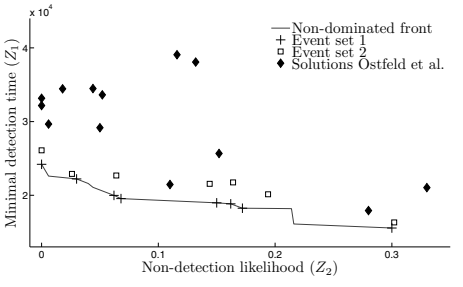
As can be seen in Figure 3, the increase in the number of sensors p placed in a network has a strong impact on the objectives as it minimizes the scale of the non-dominated front (compare Fig. 3a and 3c). Another observation that



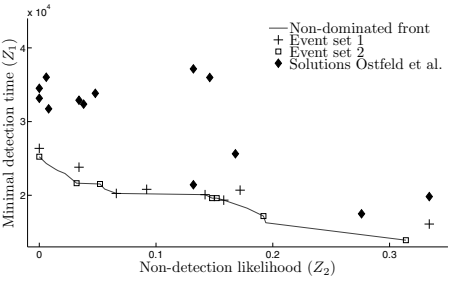
(a) Network 1, $p = 5$, contamination scenario set 1



(b) Network 1, $p = 5$, contamination event set 2

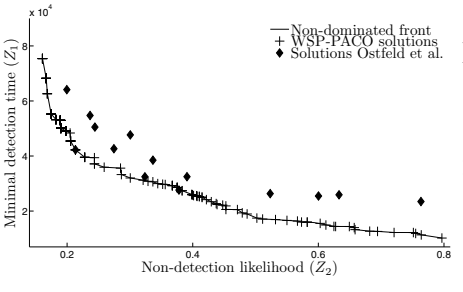


(c) Network 1, $p = 20$, contamination event set 1

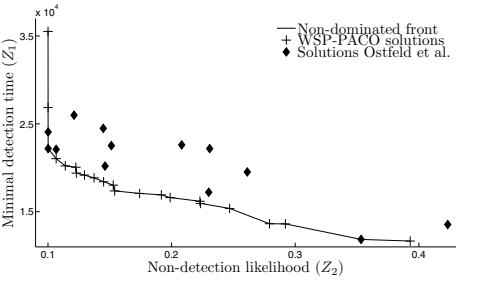


(d) Network 1, $p = 20$, contamination event set 2

Fig. 3. Network 1: Z_1 versus Z_2 for $p = \{5, 20\}$ for two contamination event sets. The non-dominated solutions found for each event set are also evaluated for the other event set. Previously found solutions [15] for this network are also evaluated for each event set.



(a) Network 1, $p = 5$, contamination event set from [15]

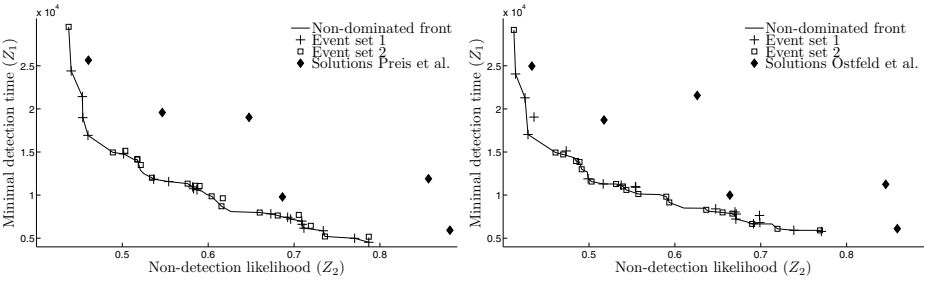


(b) Network 1, $p = 20$, contamination event set from [15]

Fig. 4. Network 2: Z_1 versus Z_2 for $p = \{5, 20\}$ using the contamination event set from Ostfeld et al. [15]

can be made is that solutions which are on the non-dominated front in one contamination event set are usually very close to the non-dominated front of solutions found for the other contamination event set (e.g., Fig. 3a and 3b).

The solutions found by WSP-PACO dominate solutions from [15] for both contamination event sets with random events used here. This is not surprising due to the reasons outlined above. In order to get a better performance comparison, WSP-PACO was used in the same contamination event set with specific events used in [15]. Figure 4 depicts the fitness regarding objectives Z_1 and Z_2 of sensor placements found by WSP-PACO for Network 1. Again the number of sensors placed in the water network has a strong effect on the detection rate and detection time. As can be seen, the contamination event set used does have an impact, as solutions from [15] are not as strongly dominated as was the case for the first two contamination event sets. However, the solutions found by WSP-PACO clearly dominate the other solutions.



(a) Network 2, $p = 5$, contamination scenario set 1

(b) Network 2, $p = 5$, contamination scenario set 2

Fig. 5. Network 1: Z_1 versus Z_2 for $p = 5$ for two contamination scenario sets. The non-dominated solutions found for each scenario set are also evaluated for the other scenario. Previously found solutions [16] for this network are also evaluated for each scenario set.

4.2 Network 2

Figure 5 depicts the fitness with respect to objectives Z_1 and Z_2 of sensor placements found by WSP-PACO for Network 2 for $p = 5$. Again, the non-dominating front obtained in one contamination event set shows very good performance when evaluated according to the other contamination event set. The solutions obtained using WSP-PACO clearly dominate the solutions which are given in Preis et al. [16] for this network. However, one must be careful when comparing the solutions as different contamination event sets were used. This can have an impact on the evolution of solutions, as we have seen before. Unfortunately we were not able to obtain the contamination scenario sets that were used by Preis et al. [16] to evolve their solutions [2]. Thus a more detailed comparison is not possible.

[2] Via email the authors told us that the specific event data are lost.

5 Conclusions

In this paper we proposed a population-based ant colony optimization approach called WSP-PACO to solve the sensor placement problem in water networks. Water networks are large scale networks that exhibit several time-dependent flow patterns and the identification of contamination in such systems is of high sanitary and economic importance. Sensor placement in water networks thus constitutes a challenging real-world optimization problem of high complexity. The approach presented here was designed to tackle two potential objectives of sensor placement in such networks, namely minimizing the time taken to detect contamination and minimizing the non-detection rate in a water network. WSP-PACO was tested on two realistic water network models under different sensor and contamination scenario settings. The algorithm showed good performance and is competitive regarding its optimization objectives in comparison to the two other studies compared here. Overall, this suggests that the ant colony optimization methodology is highly suitable for this problem domain.

Acknowledgments. This work was supported by the Human Frontier Science Program Research Grant "Optimization in natural systems: ants, bees and slime moulds". We are grateful to Cliodhna Quigley for assistance with the manuscript.

References

1. Al-Zahrani, M.A., Moied, K.: Optimizing water quality monitoring stations using genetic algorithms. The Arabian Journal for Science and Engineering 28(1B), 57–75 (2003)
2. Berry, J., Hart, W.E., Phillips, C.A., Uber, J.: A general integer-programming-based framework for sensor placement in municipal water networks. In: Proceedings of the World Water and Environmental Resources Conference (2004)
3. Berry, J., Hart, W.E., Phillips, C.A., Uber, J., Walski, T.M.: Water quality sensor placement in water networks with budget constraints. In: Proceedings of the World Water and Environmental Resources Conference (2005)
4. Berry, J., Hart, W.E., Phillips, C.A., Uber, J., Watson, J.-P.: Sensor placement in municipial water networks with temporal integer programming models. Journal of Water Resources Planning and Management 132(4), 218–224 (2006)
5. Dorigo, M., Di Caro, G.: New Ideas in Optimization. In: The Ant Colony Optimization Meta-Heuristic, pp. 11–32. McGraw-Hill, New York (1999)
6. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
7. Gueli, R.: Predator - prey model for discrete sensor placement. In: Proceeding In 8th Annual Water Distribution Systems Analysis Symposium (2006)
8. Hart, W.E., Berry, J., Boman, E., Phillips, C.A., Riesen, L.A., Watson, J.-P.: Limited-memory techniques for sensor placement in water distribution networks. In: Second International Conference on Learning and Intelligent Optimization, LION 2007 II, Trento, Italy, pp. 125–137 (2008)
9. Hart, W.E., Berry, J., Riesen, L.A., Murray, R., Phillips, C.A., Watson, J.-P.: Spot: A sensor placement optimzation toolkit for drinking water contamination warning system design. In: Proceedings of the World Water and Environmental Resources Conference (2007)

10. Kessler, A., Ostfeld, A., Sinai, G.: Detecting accidental contaminatins in municipal water networks. Journal of Water Resources Planning and Management 124(4), 192–198 (1998)
11. Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., Faloutsos, C.: Efficient sensor placement optimization for securing large water distribution networks. Journal of Water Resources Planning and Management 134(6), 516–526 (2008)
12. Kumar, A., Kansal, M.L., Arora, G., Ostfeld, A., Kessler, A.: Detecting accidental contaminations in municipial water networks. Journal of Water Resources Planning and Management 125(5), 308–310 (1999)
13. Middendorf, M., Guntsch, M.: Solving multi-criteria optimization problems with population-based ACO. In: Fonseca, C.M., Fleming, P.J., Zitzler, E., Deb, K., Thiele, L. (eds.) EMO 2003. LNCS, vol. 2632, pp. 464–478. Springer, Heidelberg (2003)
14. Ostfeld, A., Kessler, A., Goldberg, I.: A contaminant detection system for early warning in water distribution networks. Engineering Optimization 36(5), 525–538 (2004)
15. Ostfeld, A., Uber, J.G., Salomons, E., Berry, J.W., Hart, W.E., Phillips, C.A., Watson, J.-P., Dorini, G., Jonkergouw, P., Kapelan, Z., di Pierro, F., Khu, S.-T., Savic, D., Eliades, D., Polycarpou, M., Ghimire, S.R., Barkdoll, B.D., Gueli, R., Huang, J.J., McBean, E.A., James, W., Krause, A., Leskovec, J., Isovitsch, S., Xu, J., Guestrin, C., VanBriesen, J., Small, M., Fischbeck, P., Preis, A., Propato, M., Piller, O., Trachtman, G.B., Wu, Z.Y., Walski, T.: The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms. Journal of Water Resources Planning and Management 134, 556–569 (2008)
16. Preis, A., Ostfeld, A.: Multiobjective contaminant sensor network design for water distribution systems. Journal of Water Resources Planning and Management 134(4), 366–377 (2008)
17. Rossmann, L.A.: Epanet 2 users manual. US Environmental Protection Agency (2000)
18. van Bloemen Waanders, B. (ed.): Algorithm and Simulation Development in Support of Response Strategies for Contamination Events in Air and Water Systems (2006)
19. Watson, J.-P., Hart, W.E., Berry, J.: Scalable high-performance heuristics for sensor placement in water distribution networks. In: Proceedings of the World Water and Environmental Resources Conference (2005)

The Codebook Design of Image Vector Quantization Based on the Firefly Algorithm

Ming-Huwi Horng and Ting-Wei Jiang

Department of Computer Science and Information Engineering,
National Pingtung Institute of Commerce, PingTung, Taiwan
{horng,mh.horng}@npic.edu.tw

Abstract. The vector quantization (VQ) was a powerful technique in the applications of digital image compression. The traditionally widely used method such as the Linde-Buzo-Gray (LBG) algorithm always generated local optimal codebook. This paper proposed a new method based on the firefly algorithm to construct the codebook of vector quantization. The proposed method uses LBG method as the initial of firefly algorithm to develop the VQ algorithm. This method is called FF-LBG algorithm. The FF-LBG algorithm is compared with the other three methods that are LBG, PSO-LBG and HBMO-LBG algorithms. Experimental results showed that the computation of this proposed FF-LBG algorithm is faster than the PSO-LBG, and the HBMO-LBG algorithms. Furthermore, the reconstructured images get higher quality than those generated from the LBG and PSO-LBG algorithms, but there are not significantly different to the HBMO-LBG algorithm.

Keywords: Vector Quantization, LBG algorithm, Firefly algorithm, Particle warm optimization, Honey bee mating optimization.

1 Introduction

The codebook design of vector quantization (VQ) algorithms had been performed by many researchers; new algorithms continue to appear. A well-known method is the LBG algorithm [1]; however, the LBG algorithm is a local search procedure. It suffers from the serious drawback that its performance depends heavily on the initial starting conditions. Recently, the evolutionary optimization algorithms had been developed to design the codebook for improving the results of LGB algorithm. Chen et al [2] proposed an improvement based on the particle swarm optimization (PSO). The result of LBG algorithm is used to initialize global best particle by which it can speed the convergence of PSO. Feng et al [3] developed an evolutionary particle swarm optimization vector quantization learning scheme for vector quantization. Horng [4] and Jiang [10] applied the honey bee mating optimization to develop a new algorithm for vector quantization.

The firefly algorithm may also be considered as a typical swarm-based approach for optimization, in which the search algorithm is inspired by social behavior of fireflies and the phenomenon of bioluminescent communication. There are two important issues in the firefly algorithm that are the variation of light intensity and formulation of attractiveness. Yang [5] simplifies the attractiveness of a firefly is

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 438–447, 2010.

determined by its brightness which in turn is associated with the encoded objective function. The attractiveness is proportional to their brightness. Furthermore, every member of the firefly swarm is characterized by its bright that can be directly expressed as an inverse of a cost function for a minimization problem. Lukasik and Zak [6] applied the firefly algorithm for continuous constrained optimization. Yang [7] compared the firefly algorithm with the other meta-heuristic algorithms such as genetic and particle swarm optimization algorithms in the multimodal optimization. These two works had the same conclusions that the algorithm applied the proposed firefly algorithm is superior to the two existing meta-heuristic algorithms.

In this paper, we apply the firefly algorithm to the vector quantization. The method is called the FF-LBG algorithm. This paper compares the results of FF-LBG algorithm with the ones of other algorithms those are LBG, PSO-LBG and HBMO-LBG algorithms. The rest of this paper is organized as follows. Section 2 introduces the vector quantization and LBG algorithm. Section 3 presents the PSO-LBG algorithm for designing the codebook of the vector quantization. Section 4 introduces the HBMO-LBG algorithm. Section 5 presents the proposed method which searches for the optimal codebook using the FF-LBG algorithm. Performance evaluation is discussed in detail in Section 6. Conclusions are presented in Section 7.

2 LBG and PSO-LBG Algorithm

Vector quantization (VQ) is a lossy data compression technique in block coding. The generation of codebook is known as the most important process of VQ. Let the size of original image $Y = \{y_{ij}\}$ be $M \times M$ pixels that divided into several blocks with size of $n \times n$ pixels. In other words, there are $N_b = \left\lceil \frac{N}{n} \right\rceil \times \left\lceil \frac{N}{n} \right\rceil$ blocks that represented by a collection of input vectors $X = (x_i, i = 1,2,..., N_b)$. Let L= $n \times n$. The input vector x_i, $x_i \in \Re^L$ where $\Re^L$ is L-dimensional Euclidean space. A codebook C comprises N_c L-dimensional codewords, i.e., $C = \{c_1, c_2, ..., c_{N_c}\}$, $c_j \in \Re^L$, $\forall j = 1,2,.., N_c$. Each input vector is represented by a row vector $x_i = (x_{i1}, x_{i2}, ..., x_{iL})$ and each codeword of the codebook is denoted as $c_j = (c_{j1}, c_{j2}, ..., c_{jL})$. The VQ techniques assign each input vector to a related codeword, and the codeword will replace the associated input vectors finally to obtain the aim of compression. The optimization of C in terms of mean square error (MSE) can be formulated by minimizing the distortion f unction D. In general, the lower the value of D is, the better the quality of C.

$$D(C) = \frac{1}{N_b} \sum_{j=1}^{N_c} \sum_{i=1}^{N_b} \mu_{ij} \cdot \left\| x_i - c_j \right\|^2 \tag{1}$$

subject to the following constraints:

$$\sum_{j=1}^{N_c} \mu_{ij} = 1, \ \forall i \in \{1,2,....,N_b\} \tag{2}$$

$$\mu_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is in the } j^{th} \text{ cluster,} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$L_k \le c_{jk} \le U_k, k = 1,2,...,L \tag{4}$$

where L_k is the minimum of the k[th] component in the all training vectors, and U_k is the maximum of the k[th] component in all input vectors. The $\|x-c\|$ is the Euclidean distance between the vector x and codeword c.

An algorithm for a scalar quantizer was proposed by Lloyd [8]. This algorithm is known as LBG or generalized Lloyd algorithm (GLA). The LBG algorithm gives input vectors, x_i , $i = 1, 2,.., N_b$, distance function d, and an initial codewords $c_j(0)$, $j = 1,...,N_c$. The LBG iteratively applies the two conditions: (a) partition the input vectors into several groups using the minimum distance rule and (b) determine the centroids μ_{ij} of each partition. to produce optimal codebook.

The particle swarm optimization (PSO) is a new branch of evolutionary computation technique. In the multi-dimensional space, each particle represents a potential solution to a problem. There exists a fitness evaluation function that assigns a fitness values to this potential solution for designing the codebook *C* based on the Eq. (5).

$$Fitness(C) = \frac{1}{D(C)} = \frac{N_b}{\sum_{j=1}^{N_c}\sum_{i=1}^{N_b} \mu_{ij} \cdot \|x_i - c_j\|^2} \tag{5}$$

Two positions are recorded by every particle. One is named global best (*gbest*) position, which has the highest fitness value in the whole population. The other is called personal best (*pbest*) position, which has the highest fitness value of itself at present. The population of particles is flying in the search space and every particle changes his position according the *gbest* and the *pbest* with the Eqs. (6) and (7).

$$v_{ik}^{n+1} = v_{ik}^n + c_1 r_1^n (pbest_{ik}^n - x_{ik}^n) + c_2 r_2^n (gbest_k^n - x_{ik}^n) \tag{6}$$

$$x_{ik}^{n+1} = x_{ik}^n + v_{ik}^{n+1} \tag{7}$$

where k is the number of dimensions ($k=1,2,\ldots,L$) and i represents a particle of the population ($i=1,2,\ldots,s$). X means the position of particle in the search space; v is the velocity vector for the particle to change its position. Parameter c_1 and c_2 are the cognitive and social learning rates respectively. r_1 and r_2 are two random numbers that belongs to [0, 1]. The algorithm of PSO-LBG sets the result of LBG algorithm into an initial particle. Followings are the detail algorithm.

1. Run the LBG algorithm once.
2. Assign the result of LBG algorithm to one particle and initialize positions of rest particles and associated velocity of all particles randomly.

3. Calculate the fitness value of each particle according to based on the Eq. (5).
4. Compare each particle's fitness value with the previous particle's personal best value. If better, it updates *pbest* and takes record current position as the particle's personal best position.
5. Find the highest fitness value of the whole particles. If the value is better than *gbest*, replace *gbest* with this fitness value, and take record the global best position.
6. Change velocities and positions according to Eqs. (6) and (7) for each particle.
7. Repeat 3 to 7 until stop criteria are satisfied.

3 HBMO-LBG Vector Quantization Algorithm

A honeybee colony typically consists of a single egg-laying long-lived queen, anywhere from zero to several thousands drones and usually 10,000-60,000 workers [9]. A mating flight starts with a dance performed by the queen who then starts a mating flight during which the drones follow the queen and mate with her in the air. In order to develop the algorithm, the capability of workers is restrained in brood care and thus each worker may be regarded as a heuristic that acts to improve and/or take care of a set of broods. An annealing function is used to describe the probability of a drone (D) that successfully mates with the queen (Q) shown in Eq. (8).

$$P(Q,D) = \exp[-\Delta(f)/S(t)] \tag{8}$$

where $\Delta(f)$ is the absolute difference of the fitness of D and the fitness of Q, and the $S(t)$ is the speed of queen at time *t*. After each transition of mating, the queen's speed and energy are decayed according to the following equation:

$$S(t+1) = \alpha \times S(t) \tag{9}$$

where α is the decreasing factor ($\alpha \in [0,1]$). Workers adopt some heuristic mechanisms such as crossover or mutation to improve the brood's genotype. The fitness of the resulting genotype is determined by evaluating the value of the objective function of the brood genotype. In the HBMO-LBG algorithm, the solutions include the best solution; candidate solution and the trivial solution are represented in the form of codebook. The fitness function used also defined in Eq. (5). The details of HBMO-LBG algorithm can be found in [4].

4 FF-LBG Vector Quantization Algorithm

In the firefly algorithm, there are three idealized rules: (1) all fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex; (2) Attractiveness is proportional to their brightness, thus for any two flashing fireflies, the less brighter one will move towards the brighter one. If there is no brighter than a particular firefly, it will move randomly. As firefly attractiveness one should select any monotonically decreasing function of the distance $r_{i,j} = d(x_j, x_i)$ to the chosen jth firefly, e.g. the exponential function.

$$r_{i,j} = \left\| x_i - x_j \right\| \tag{10}$$

$$\beta \leftarrow \beta_0 e^{-\gamma r_{i,j}} \tag{11}$$

where the β_0 is the attractiveness at $r_{i,j}=0$ and γ is the light absorption coefficient at the source. The movement of a firefly *i* is attracted to another more attractive firefly *j* is determined by

$$x_{i,k} \leftarrow (1-\beta)x_{i,k} + \beta x_{j,k} + u_{i,k} \tag{12}$$

$$u_{i,k} = \alpha(rand1 - \frac{1}{2}) \tag{13}$$

If there is no brighter than a particular firefly x_i with maximum fitness, it will move randomly according to the following equation.

$$x_{i^{\max},k} \leftarrow x_{i^{\max},k} + u_{i^{\max},k}, \quad \text{for } k=1,2,\ldots,N_c. \tag{14}$$

$$u_{i^{\max},k} = \alpha(rand2 - \frac{1}{2}) \tag{15}$$

when $rand1 \approx U(0,1)$ $rand2 \approx U(0,1)$ are random numbers obtained from the uniform distribution; (3). The brightness of a firefly is affected or determined by the landscape of the fitness function $\phi(\bullet)$. For maximization problem, the brightness I of a firefly at a particular location x can be chosen as I(x) that is proportional to the value of the fitness function $\phi(x)$.

In the FF-LBG algorithm, the solutions (fireflies) are represented in the form of codebook shown as Figure 1. The fitness function used also defined in Eq. (9). The key point to generate a high quality solution Q is to find the perfect codebook which maximizes the fitness function for all input vectors. The details of FF-LBG algorithm show as follows:

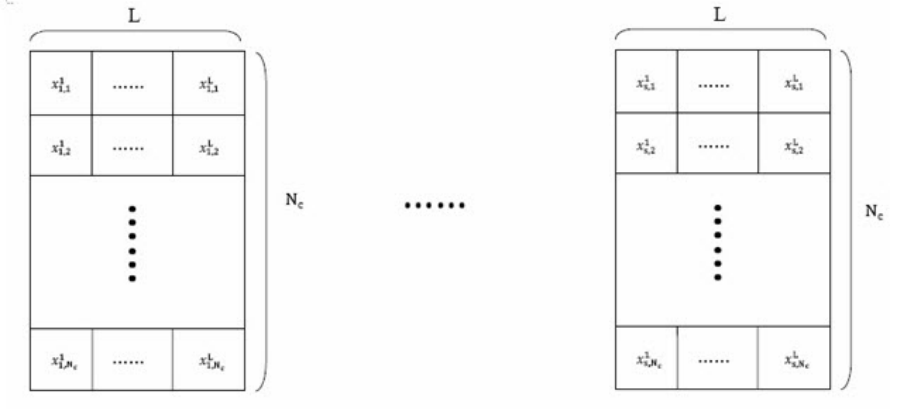


Fig. 1. The structure of solutions (codebook) of FF-LBG algorithm, the $x_{i,j}$ is the j-th codeword of the codebook x_i

Step 1. (Generate the initial solutions and given parameters)
In this step, the codebook of LBG algorithm is set to one of initial solution, and then the set of initial trivial solutions, x_i , ($i = 1,2,...,m-1$) are randomly generated. Each solution is the codebook with n_c codeword. Furthermore, the step gives the parameters of the α , β_0 , the maximum cycle number L and γ . Set *l*=0.

Step 2. (The best solution will randomly move to the different position)
Step 2 selects the best one from the all solutions and define as the $x_i^{\max}$, that is,

$$\begin{aligned} i^{\max} &= \arg\max_{i} Fitness(x_i); \\ x_i^{\max} &= \arg\max_{x_i} Fitness(x_i); \end{aligned} \tag{16}$$

Step 3. (The movement of a firefly x_j is attracted to another more attractive firefly x_i)
In step 3, each solution x_j computes its fitness value as the corresponding the brightness of firefly. For each solution x_j , this step selects other solution x_i with the more bright and then moves to it following the following equations.

$$r_{i,j} = \left\| x_i - x_j \right\| = \sqrt{\sum_{k=1}^{N_c} \sum_{l=1}^{L} (x_{i,k}^l - x_{j,k}^l)^2} \tag{17}$$

$$\beta = \beta_0 e^{-\gamma r_{i,j}} \tag{18}$$

$$x_{j,k}^l = (1-\beta) x_{i,k}^l + \beta x_{j,k}^l + u_{j,k}^l \text{ , k=1 ,2,…, } N_c \text{ ., } l=1, 2,....,L. \tag{19}$$

where the $u_{i,k}^l \sim U(0,1)$ is a randomly number.

Step 4. (The best solution randomly move its position)
The best solution $x_i^{\max}$ will randomly move its position based the following equation.

$$x_{i^{\max},k}^l \leftarrow x_{i^{\max},k}^l + u_{i^{\max},k}^l \text{ , k=1,2,…, } N_c \text{ ., l=1,….,L} \tag{20}$$

where the $u_{i,k}^l \sim U(0,1)$ is a randomly number.

Step 5. (Check the termination criterion)
If the cycle is equal to the maximum cycle number l then the algorithm is finished and output the best solution $x_i^{\max}$; otherwise l increases by one and go to Step 2.

5 Experimental Results and Discussion

The typical experiments for evaluating the methods used for codebook design are the grayscale image coding. Three 512×512 still images named “LENA”, “PEPPER”,

and "LAKE" with pixel amplitude resolution of 8 bits are shown in Fig. 2. In coding an image, the image is completely divided into immediate and non-overlapping blocks with 4×4 pixels. Each block is treated as a pattern input vector with 16 dimensions. Therefore, there exist a total of 16384 vectors to be encoded in an image. In experiments, four different codebook sizes that are 64, 128, 256 and 512 are implemented. We compare the FF-LBG method with four different algorithms, including the traditional LBG, the particle swarm optimization (PSO)-based LBG and the honey bee mating optimization (HBMO)-based LBG. The programs of the five algorithms are designed in language of Visual C++ 6.0 on a personal computer with 2.4GHz CPU, 1G RAM running window XP system. All experiments were conducted for $\beta_0 = 1$, $\alpha = 0.01$ and fixed $\gamma = 1.0$ of FF-LBG algorithm. The setup of parameters of PSO-LBG and HBMO-LBG are referred to the work of Jiang [10]. The size of initial solutions (fireflies) is assigned to be 50 and the maximum iteration number *l=200*. The bit rate is defined in Eq. (21).

$$bit\_rate = \frac{\log_2 N_c}{K} \tag{21}$$

(a). LENA (b). PEPPER (c). LAKE

Fig. 2. The test images: (a) LENA, (b) PEPPER, and (c) LAKE

The N_C represents the size of designed codebook and the K is the pixel number of block. Furthermore, the quality of encoded image was evaluated using the peak signal-to-noise ratio (PSNR). The PSNR is defined as

$$PSNR = 10\times\log_{10}(\frac{A^2}{MSE}) \text{ (dB)} \tag{22}$$

That A is the maximum of gray level and MSE is the mean square error between the original image and the decompressed image.

$$MSE = \frac{1}{M\times M}\sum_{i=1}^{M}\sum_{j=1}^{M}(y_{ij} - \overline{y_{ij}})^2 \tag{23}$$

where $M\times M$ is the image size, y_{ij} and $\overline{y_{ij}}$ denote the pixel value at the location *(i, j)* of original and reconstructed images, respectively.

Table 1. The PSNR values of the test images by using the five algorithms under different bit rates

Image (512×512)	Bit Rate	PSNR (dB)			
		LBG	PSO-LBG	HBMO-LBG	FF-LBG
LENA	0.375	25.657	25.617	26.830	26.725
PEPPER		25.891	27.101	27.206	27.206
LAKE		26.473	27.146	27.237	27.316
LENA	0.4375	25.740	27.058	27.562	27.562
PEPPER		25.976	27.791	27.814	27.653
LAKE		26.557	27.989	28.062	28.114
LENA	0.5	25.750	28.169	28.337	28.337
PEPPER		25.998	28.861	28.851	28.954
LAKE		26.593	28.615	28.745	28.825
LENA	0.5625	25.786	28.994	29.198	29.148
PEPPER		26.006	29.661	27.729	29.893
LAKE		26.599	29.434	29.522	29.547

Table 2. The computation times of the test images by using the five algorithms under different bit rates

Image (512×512)	Bit Rate	Computation times (sec)			
		LBG	PSO-LBG	HBMO-LBG	FF-LBG
LENA	0.375	15.43	83.45	98.65	76.82
PEPPER		16.45	87.89	94.25	73.55
LAKE		17.54	78.56	91.65	74.65
LENA	0.4375	63.23	321.42	352.25	289.34
PEPPER		65.45	335.57	356.29	297.54
LAKE		63.87	342.87	359.34	292.98
LENA	0.5	184.45	976.87	968.25	834.65
PEPPER		193.49	1013.72	982.63	816.52
LAKE		175.67	981.93	986.32	832.19
LENA	0.5625	425.87	3567.64	3525.65	3151.98
PEPPER		445.76	3498.19	3465.25	3168.34
LAKE		442.26	3467.65	3512.33	3205.71

Table 1 shows the PSNR values of test images by using the four different vector quantization algorithms. Obviously, the usages of the FF-LAB algorithm have higher PSNR value compared with original LBG and PSO-LBG algorithms. More precisely, the PSNR of LBG algorithm is the worst and the other three algorithms can significantly improve the results of LBG algorithm. Furthermore, the PSNR values of the FF-LBG and HBMO-LBG reveal that the two algorithms have no significant difference in the quality measure of compression. Table 2 shows the computation time of the four different vector quantization methods with different bit rates. Form this table, we find that the need of the computation time by using the original LBG algorithm is the least; however it has the smallest PSNR value. The computation time of using the FF-LBG is less than the ones of using the other four vector quantization. It is valuable to improve the efficiency in terms of the parallel fashion in the implementation of FF-LBG algorithm in further study.

6 Conclusion

This paper gives a detailed description of how the firefly algorithm is used to implement the vector quantization and enhance the performance of LBG method. All of our experimental results showed that the FF-LBG algorithm can increase the quality of reconstructive images with respect to other three methods such as the traditional LBG, the PSO-LBG and HBMO-LBG algorithm. The proposed FF-LBG algorithm can provide a better codebook with smallest distortion and the least computation time.

Acknowledgments. The author would like to thank the National Science council, ROC, under Grant No. NSC 99-2221-E-251-007 for support of this work.

References

1. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer design. IEEE Transaction on Communications 28, 84–95 (1980)
2. Chen, Q., Yang, J.G., Gou, J.: Image Compression Method using Improved PSO Vector Quantization. In: Wang, L., Chen, K., S. Ong, Y. (eds.) ICNC 2005. LNCS, vol. 3612, pp. 490–495. Springer, Heidelberg (2005)
3. Feng, H.H., Chen, C.Y., Ye, F.: Evolutionary Fuzzy Particle Swarm Optimization Vector Quantization Learning Scheme in Image Compression. Expert Systems with Applications 32, 213–222 (2007)
4. Horng, M.H.: Honey Bee Mating Optimization Vector Quantization Scheme in Image Compression. In: Deng, H., Wang, L., Wang, F.L., Lei, J. (eds.) AICI 2009. LNCS, vol. 5855, pp. 185–194. Springer, Heidelberg (2009)
5. Yang, X.S.: Nature-inspired metaheuristic algorithms. Luniver Press (2008)
6. Lukasik, S., Zak, S.: Firefly algorithm for continuous constrained optimization tasks. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 5–7. Springer, Heidelberg (2009)

7. Yang, X.S.: Firefly algorithms for multimodal optimization. In: Watanabe, O., Zeugmann, T. (eds.) SAGA 2009. LNCS, vol. 5792, pp. 169–178. Springer, Heidelberg (2009)
8. Lloyd, S.P.: Least Square Quantization in PCM's. Bell Telephone Laboratories Paper. Murray Hill, NJ (1957)
9. Abbasss, H.B.: Marriage in Honey-bee Optimization (HBO); a Haplo, etrosis Polygynous Swarming Approach. In: CEC 2001, pp. 207–214 (2001)
10. Jiang, T.W.: The application of image thresholding and vector quantization using honey bee mating optimization. Master thesis of National PingTung Institute of Commerce (2009)

Confronting Two-Pair Primer Design Using Particle Swarm Optimization

Cheng-Hong Yang[1,2], Yu-Huei Cheng[1], and Li-Yeh Chuang[3]

[1] Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
[2] Department of Network Systems, Toko University, Chiayi, Taiwan
[3] Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan
chyang@cc.kuas.edu.tw, yuhuei.cheng@gmail.com, chuang@isu.edu.tw

Abstract. Polymerase chain reaction with confronting two-pair primers (PCR-CTPP) is a novel PCR technique had been applied to many SNPs (Single Nucleotide Polymorphisms) genotyping experiments successfully in the recent years. The advantages of simplicity make it is a time- and cost-effective SNP genotyping method. However, computation of feasible CTPP primers is still challenging. In this study, a particle swarm optimization (PSO)-based method is proposed to design a feasible CTPP primer set. Overall, two hundred and eighty-eight SNPs in SLC6A4 gene were tested *in silicon* by the proposed method. The result indicates that the proposed method provide feasible CTPP primers effectively than the genetic algorithm (GA) in the literature. It can assist the biologists and researchers to obtain a feasible CTTP primer set.

Keywords: PCR-CTPP, primer design, PSO, SNP.

1 Introduction

SNPs (Single Nucleotide Polymorphisms) are important genetic variations used in association studies of diseases and cancers. Many high-throughput platforms of SNP genotyping such as real-time PCR [1] and SNP array [2] have been introduced, but PCR-restriction fragment length polymorphism (RFLP) genotyping [3-5] is still used to validate SNPs or novel mutations by most laboratories due to its inexpensive for the small-scale genotyping. However, the major shortcoming of PCR-RFLP is usually long digestion time in 2-3 hours for restriction enzymes [6, 7].

Recently, PCR with confronting two-pair primers (PCR-CTPP) was developed a restriction enzyme-free SNP genotyping technique [8, 9]. Many SNPs has been genotyped successfully using this technique [10, 11]. PCR-CTPP considerably lowers needs to consume restriction enzymes. However, the effective computation methods are still challenging to develop.

In the past, we introduced a genetic algorithm to design CTPP primer sets [12]. However, the computational result is not good (i.e., especially the most factor T_m difference) for most cases; thus the particle swarm optimization (PSO) [13] is proposed to apply to the problem.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 448–456, 2010.

2 Method

PSO is a population-based stochastic optimization technique developed by Kennedy and Eberhart in 1995 [13]. It simulates the social behavior of organisms, such as birds in a flock or fish in a school, and describes an automatically evolving system. In PSO, each single candidate solution can be considered "an individual bird of the flock", that is, a particle in the search space. Each particle makes use of its own memory as well as knowledge gained by the swarm as a whole to find the best (optimal) solution. All of the particles have fitness values, which are evaluated by an optimized fitness function. They also have velocities, which direct the movement of the particles. During movement, each particle adjusts its position according to its own experience, and according to the experience of a neighbouring particle, thus making use of the best position encountered by itself and its neighbour. PSO has been successfully applied in many areas, e.g., function optimization, artificial neural network training, fuzzy system control, and other application problems. A comprehensive survey of PSO algorithms and their applications can be found in Kennedy *et al.* [14].

2.1 Problem Formulation

The CTPP primer design problem can be described as follows. Let T_D be the DNA template sequence, which is composed of nucleotide codes with an identified SNP. T_D is defined by:

$$T_D = \{B_i \mid i \text{ is the index of DNA sequence}, \\ 1 \le i \le \iota, \exists! B_i \in \text{IUPAC code of SNP}\} \tag{1}$$

where B_i is the regular nucleotide code ('A', 'T', 'C', or 'G') mixed with a single IUPAC code of SNP ('M', 'R', 'W', 'S', 'Y', 'K', 'V', 'H', 'D', 'B' or 'N') ($\exists!$ is the existence and uniqueness). For the target SNP, we focused only on true SNPs described in dbSNP [15] of NCBI, i.e., deletion/insertion polymorphism (DIP) and multi-nucleotide polymorphism (MNP) are not included.

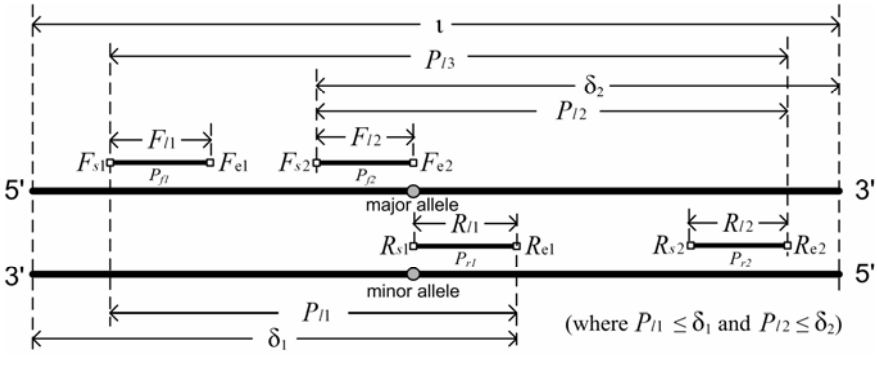


Fig. 1. Parameters of the DNA template and the CTPP primer set. Symbols indicate: *F*: Forward primer; *R*: Reverse primer; *s*: Start nucleotide position; *e*: End nucleotide position; *P*: Length of PCR product using a primer set (F/R); *l*: Length of primer or product;ι: Length of DNA template; δ_1: Length from the R_{s1} end to downstream of DNA template; δ_2: Length from F_{s2} to the downstream end of DNA template.

The CTPP primer design requires two pairs of short sequences which are constraining T_D based on a defined SNP site as illustrated (Fig. 1). The forward primer 1 (P_{f1}) is a short sense sequence in the upstream (5' end) of a defined SNP site for some distances, the reverse primer 1 (P_{r1}) is a short antisense sequence which contains a nucleotide (the minor allele of the defined SNP site) located at its 3' end, the forward primer 2 (P_{f2}) is a short sense sequence which contains a nucleotide (the major allele of the defined SNP site) located at its 3' end, and the reverse primer 2 (P_{r2}) is the antisense sequence in the upstream of a defined SNP site for some distances. These four primers are defined as follows:

$$P_{f1} = \{B_i \mid i \text{ is the index of } T_D,\ F_{s1} \le i \le F_{e1}\} \tag{2}$$

$$P_{r1} = \{\overline{B_i} \mid i \text{ is the index of } T_D, R_{s1} \le i \le R_{e1}\} \tag{3}$$

$$P_{f2} = \{B_i \mid i \text{ is the index of } T_D, F_{s2} \le i \le F_{e2}\} \tag{4}$$

$$P_{r2} = \{\overline{B_i} \mid i \text{ is the index of } T_D, R_{s2} \le i \le R_{e2}\} \tag{5}$$

where both P_{f1}/P_{r1} and P_{f2}/P_{r2} are two sets of primer pairs. F_{s1} *vs.* F_{e1} and R_{s1} *vs.* R_{e1} indicate the start index *vs.* the end index of P_{f1} and P_{r1} in T_D, respectively. F_{s2} *vs.* F_{e2} and R_{s2} *vs.* R_{e2} indicate the start index *vs.* the end index of P_{f2} and P_{r2} in T_D, respectively. $\overline{B_i}$ is the complementary nucleotide of B_i, which is described in formula (1). For example, if B_i = 'A', then $\overline{B_i}$ = 'T'; if B_i = 'C', then $\overline{B_i}$ = 'G', and *vice versa*.

The SNP site is defined at the 3' end positions of P_{f2} and P_{r1}, which are indicated by the symbols F_{e2} and R_{s1}, respectively. As described in Fig. 1, a vector (v) with F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2} and R_{l2} is essential to design the CTPP primer sets. This vector is defined as follows:

$$P_v = (F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2}, R_{l2}) \tag{6}$$

F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2} and R_{l2} represent the number of nucleotides of the forward primer 1, product length between P_{f1} and P_{r1}, reverse primer 1, forward primer 2, product length between P_{f2} and P_{r2} and reverse primer 2, respectively. Consequently, the forward and the reverse primers can be acquired from P_v which is the prototype of a particle in PSO and is used to perform evolutionary computations as described in the following sections.

2.2 CTPP Design Method

The flowchart of the proposed method is shown as Fig. 2. The proposed method consists of six processes: (1) particle swarm initialization, (2) fitness evaluation, (3) *pbest* and *gbest* finding (4) particle updating, and (5) judgment on termination conditions, are described below.

(1) Particle swarm initialization. To start the algorithm, particles $P_v = (F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2}, R_{l2})$ of particular number are randomly generated for an initial population without duplicates. F_{l1}, R_{l1}, F_{l2} and R_{l2} are randomly generated between the minimum and the maximum length of the primer length constraint. The minimum and maximum

primer length constraints are set to between 16 bp and 28 bp, respectively. The PCR product lengths, P_{l1} and P_{l2} are randomly generated between 100 bp and δ_1, and between 100 bp and δ_2, respectively. (δ_1 and δ_2 are maximum tolerant PCR product length of P_{l1} and P_{l2} shown in Fig.1)

(2) *Fitness evaluation.* The fitness value in the fitness function is used to individually ascertain that a particle (i.e., solution) is either good or bad. We use formula (7) [12] to evaluate the fitness values of all particles in the population for related operations later.

$$\begin{aligned} Fitness(P_v) = {} & 3 * (Len_{diff}(P_v) + GC_{proportion}(P_v) + GC_{clamp}(P_v)) \\ & + 10 * (dimer(P_v) + hairpin(P_v) + specificity(P_v)) \\ & + 50 * (\mathrm{Tm}(P_v) + \mathrm{Tm}_{diff}(P_v)) + 100 * Avg\_\mathrm{Tm}_{diff}(P_v) \\ & + 60 * PCR\mathrm{len}_{ratio}(P_v) \end{aligned} \tag{7}$$

The weights (3, 10, 50, 60 and 100) of the fitness function are applied to estimate the importance of the primer constraints. These weights are set according to the experiential conditions for PCR-CTPP. They also accept adjustment based on the experimental requirements.

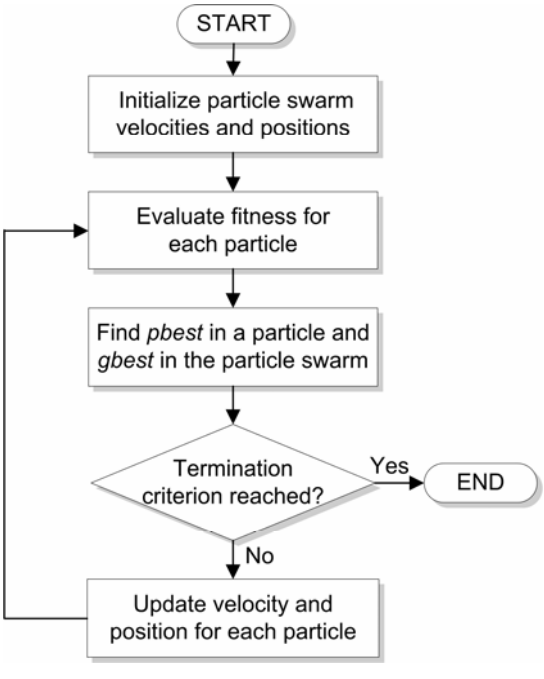


Fig. 2. Flowchart of the PSO-based CTPP primer design. At first, the velocities and positions of a specific number of particles are generated randomly. And then all fitness values of all particles are calculated by the fitness function. A judgment on termination conditions is carried out, and if the termination conditions are reached then the algorithm will be finished, else the algorithm proceeds with the following processes. Find out the *pbest* from each particle and find out *gbest* from all particles and then update velocities and positions for all particles according to the updating formulas. Repeat related steps shown as the figure until the best solution is found or the preset generation number is reached.

Primer length
A feasible primer length for a PCR experiment is set between 16 bp and 28 bp. Since the random values of F_{l1}, R_{l1}, F_{l2} and R_{l2} have been limited by the constraint condition, the primer length estimation does not be considered to join to the fitness function. A length difference (Len_{diff}) less than or equal to 3 bp between the F_{l1}/R_{l1}, F_{l2}/R_{l2}, and F_{l1}/R_{l2} primer sets is considered optimal. The $Len_{diff}(P_v)$ function is used to judge the constraint.

GC content
In general primer design, the typical GC proportion constraint is set between 40% and 60%. However, the designed CTPP primers contain the target SNP limiting the range of the GC proportion. To relax this constraint, the constraint of GC proportion in a primer is adjusted to between 20% and 80%. The $GC_{proportion}(P_v)$ function is proposed to lead the GC proportion of CTPP primers corresponding this constraint.

GC Clamp
To meet the presence of 'G' or 'C' at the 3' terminal of a primer to ensure a tight localized hybridization bond, the $GC_{clamp}(P_v)$ function is proposed to meet the criterion.

Melting temperature
The melting temperature (T_m) for each CTPP primer must be considered carefully for PCR experiment. The T_m calculation formula for a primer is described as follows:

$$Tm_{BM}(P) = 81.5 + 16.6 * (\log_{10}[Na^+]) + 0.41 * (GC\%) - 675 / |P| \quad (8)$$

where P represents a primer and $|P|$ represents the length of primer P; Na^+ is the molar salt concentration. The suffix BM represents the formula which was proposed by Bolton and McCarthy [16].

The $Tm(P_v)$ function is proposed to confined a CTPP primer set ranging from 45°C to 62°C. Similar T_m between a primer pair is important to experiment in a tube. The $Tm_{diff}(P_v)$ function is proposed to guide the difference of the melting temperatures to less than or equal to 1°C. In order to balance the T_m values among a CTPP primers, the $Avg\_Tm_{diff}(P_v)$ function is proposed to calculate the average T_m difference.

Dimer and hairpin
Primer dimers (annealing of two primers), such as cross-dimers (a forward primer and a reverse primer) and self-dimers (two forward primers or two reverse primers) must also be avoided. To check for the occurrence of primer dimers, the function $dimer(P_v)$ is proposed. In addition, the hairpin check is also implemented to avoid annealing to itself. To check for the presence of a hairpin structure in CTPP primers, the $hairpin(P_v)$ function is proposed.

Specificity
Subsequently, the function $specificity(P_v)$ is proposed to check for each CTPP primer if reappearance in the template DNA sequence to ensure its specificity. The PCR experiment may fail when a designed primer is not sequence-specific (i.e., it appears more than once in the DNA template).

PCR product length
Finally, the $PCR\text{len}_{ratio}(P_v)$ function is proposed to calculate the appropriate PCR product length. Three ratios, i.e., ratio1, ratio2 and ratio3, are introduced to the function $PCR\text{len}_{ratio}(P_v)$ representing P_{l1}, P_{l2} and P_{l3}, respectively. The minimum PCR product length needs to be greater than 100 bp.

(3) pbest and gbest finding. One of the characteristics of PSO is that each particle has a memory of its own best experience. Each particle finds its personal best position and velocity (called *pbest*) and the global best position and velocity (called *gbest*) when moving. If the fitness of a particle P_v is better than the fitness of *pbest* in the previous generation, *pbest* will be updated to P_v in the current generation. If the fitness of a particle P_v is better than *gbest* in the previous generation and is the best one in the current generation, *gbest* will be updated to P_v. Each particle adjusts its direction based on *pbest* and *gbest* in the next generation.

(4) Particle updating. In each generation, the particles will change their position and velocity. Equations (9) and (10) give the updating formulas for each particle.

$$v_i^{next} = w \times v_i^{current} + c_1 \times r_1 \times (s_i^p - s_i^{current}) + c_2 \times r_2 \times (s^g - s_i^{current}) \tag{9}$$

$$s_i^{next} = s_i^{current} + v_i^{next} \tag{10}$$

In equations (9) and (10), v_i^{next} is the updated velocity of the *i*th particle; $v_i^{current}$ is the current velocity of the *i*th particle; c_1 and c_2 are the acceleration constants; w is the inertia weight; r_1 and r_2 are a number which is randomly generated within 0~1; s_i^p is the personal best position of the *i*th particle; s^g is the global best position of the particles; $s_i^{current}$ is the current position of the *i*th particle; s_i^{next} is the updated position of the *i*th particle. In order to prevent a particle from overshooting the limits of F_s, F_l, P_l and R_l during the update process, we randomly reset the particle according to the primer constraints.

(5) Judgment on termination conditions. The algorithm is terminated when *gbest* has achieved the best position, i.e., its fitness value is 0, or when a maximum number of generations have been reached.

3 Results and Discussion

3.1 Template Sequence

A point mutation in the SLC6A4 gene was recently identified and shown to be associated with psychosis [17], and bipolar [18] patients. Overall, two hundred and eighty-eight SNPs which exclude the deletion/insertion polymorphism (DIP) and multi-nucleotide polymorphism (MNP) in SLC6A4 gene were used to estimate the efficiency of the proposed method. All SNPs were retrieved with 500 bp flanking length (at both sides of SNP) from SNP-Flankplus (http://bio.kuas.edu.tw/snp-flankplus/) [19] as template sequences.

3.2 Parameter Settings

Four main parameters are set for the proposed method, i.e., the number of iterations (generations), the number of particles, the inertia weight w, and the acceleration constants c_1 and c_2. Their values were set to 50, 10, 0.8, 2 and 2, respectively. For GA-based method, the number of iterations (generations), the population size, the probability of crossover and the probability of mutation were respective 1000, 50, 0.6 and 0.001; the values are based on DeJong and Spears' parameter settings [20].

3.3 Results for the PSO-Based and GA-Based CTPP Primer Design Methods

The statistics of the entire CTPP primers based on the common constraints are shown in Table 1. For the 288 SNPs, the primer lengths are all between 16 bp and 28 bp. For PSO, 82.87% designed primers satisfy the length difference criterion. Most of the primer length differences were between 0 and 5 bp (data not shown). For GC%, 96.61% primers satisfy the criterion; only 15 primers were less than 20%, 28 primers were more than 80% (data not shown). There are 57.73% primers satisfy the GC clamp criterion. Most of the designed primers also had the satisfied T_m (95.40%); more than half of the primer pairs are satisfied with the T_m difference criteria (58.80%). The criterion for product length was satisfied in 57.06% of the designed primer pairs. For the criteria for primer dimer, hairpin and specificity, only few primers were problematic (3.85%, 13.37% and 1.91%, respectively).

For GA, the parameter settings are based on DeJong and Spears, 75.12% designed primers satisfy the length difference criterion. Most of the primer length differences were between 0 and 5 bp (data not shown). For GC%, 96.09% primers satisfy the criterion; only 30 primers were less than 20%, 25 primers were more than 80% (data not shown). There are 55.99% primers satisfy the GC clamp criterion. Most of the designed primers also had the satisfied T_m (86.63%); however, only a few the primer pairs are satisfied with the T_m difference criteria (23.61%). The criterion for product length was satisfied in 71.18% of the designed primer pairs. For the criteria for primer dimer, hairpin and specificity, only few primers were problematic (4.44%, 14.06% and 3.04%, respectively).

Table 1. The statistics of the designed CTPP primers showing the accuracy (%) for primers satisfied the common constraints for SNPs of the SLC6A4 gene based on GA and PSO method

Method	primer length difference	GC%	GC clamp	T_m	T_m difference	product length	dimer	hairpin	specificity
PSO	82.87	96.61	57.73	95.40	58.80	57.06	96.15	86.63	98.09
GA	75.12	96.09	55.99	86.63	23.61	71.18	95.56	85.94	96.96

3.4 Compare the Results of PSO-Based with GA-Based CTPP Methods

From Table 1, almost all satisfied accuracy of the primer constraints using PSO-based method is better than GA-based method, only the product length criterion is not. The satisfied primer length difference of PSO-based method is higher 7.75% than GA-based method. The satisfied GC% of PSO-based method is lightly lower 0.52% than

GA-based method. The satisfied GC clamp of PSO-based method is higher 1.74% than GA-based method. The satisfied T_m of PSO-based method is lightly higher 8.77%, and the satisfied T_m difference of PSO-based method is greatly higher 35.19% than GA-based method. The satisfied dimer, hairpin and specificity of PSO-based method are lightly higher 0.59%, 0.69%, and 1.13% than GA-based method. However, the satisfied product length of PSO-based method is lower 14.12% than GA-based method. Those shows PSO-based method is superior to GA-based method. In addition, the most factor (i.e., T_m difference) is 35.19% improved.

4 Conclusions

In PCR-CTPP, the melting temperature is the most factor to affect the successful rate of genotyping experiment. The PSO-based CTPP primer design method has provided better melting temperature and common primer constraints estimation. It can assist the biologists and researchers to obtain a more feasible CTTP primer set than GA-based method. The experimental flexibility of the PSO-based designed PCR-CTPP primers for 288 polymorphisms has been confirmed by *in silicon* simulations. Due to the lower costs and shorter genotyping times, PCR-CTPP may replace PCR-RFLP in the future [21]. Recently, we have been enthusiastic at the development of PCR-CTPP primer design methods to facilitate PCR-CTPP for validating SNPs or novel mutations. In conclusion, the proposed PSO-based method is a useful tool to design feasible CTPP primers since it conforms to the most of the PCR-CTPP constraints. The *in silicon* simulation results indicate that PSOs applied to the design of CTPP primer sets outperform GAs.

Acknowledgments. This work is partly supported by the National Science Council in Taiwan under grant NSC96-2221-E-214-050-MY3, NSC98-2221-E-151-040-, NSC 98-2622-E-151-001-CC2 and 98-2622-E-151-024-CC3.

References

1. Hui, L., DelMonte, T., Ranade, K.: Genotyping using the TaqMan assay. Curr. Protoc. Hum. Genet. ch. 2, Unit 2 10 (2008)
2. Jasmine, F., Ahsan, H., Andrulis, I.L., John, E.M., Chang-Claude, J., Kibriya, M.G.: Whole-genome amplification enables accurate genotyping for microarray-based high-density single nucleotide polymorphism array. Cancer Epidemiol. Biomarkers Prev. 17, 3499–3508 (2008)
3. Chang, H.W., Yang, C.H., Chang, P.L., Cheng, Y.H., Chuang, L.Y.: SNP-RFLPing: restriction enzyme mining for SNPs in genomes. BMC Genomics 7, 30 (2006)
4. Lin, G.T., Tseng, H.F., Yang, C.H., Hou, M.F., Chuang, L.Y., Tai, H.T., Tai, M.H., Cheng, Y.H., Wen, C.H., Liu, C.S., Huang, C.J., Wang, C.L., Chang, H.W.: Combinational polymorphisms of seven CXCL12-related genes are protective against breast cancer in Taiwan. OMICS 13, 165–172 (2009)
5. Chang, H.W., Cheng, Y.H., Chuang, L.Y., Yang, C.H.: SNP-RFLPing 2: an updated and integrated PCR-RFLP tool for SNP genotyping. BMC Bioinformatics 11, 173 (2010)

6. Chuang, L.Y., Yang, C.H., Tsui, K.H., Cheng, Y.H., Chang, P.L., Wen, C.H., Chang, H.W.: Restriction enzyme mining for SNPs in genomes. Anticancer Res. 28, 2001–2007 (2008)
7. NCBI: Restriction Fragment Length Polymorphism (RFLP), http://www.ncbi.nlm.nih.gov/genome/probe/doc/TechRFLP.shtml (accessed September 2009)
8. Hamajima, N., Saito, T., Matsuo, K., Kozaki, K., Takahashi, T., Tajima, K.: Polymerase chain reaction with confronting two-pair primers for polymorphism genotyping. Jpn. J. Cancer Res. 91, 865–868 (2000)
9. Tamakoshi, A., Hamajima, N., Kawase, H., Wakai, K., Katsuda, N., Saito, T., Ito, H., Hirose, K., Takezaki, T., Tajima, K.: Duplex polymerase chain reaction with confronting two-pair primers (PCR-CTPP) for genotyping alcohol dehydrogenase beta subunit (ADH2) and aldehyde dehydrogenase 2 (ALDH2). Alcohol 38, 407–410 (2003)
10. Katsuda, N., Hamajima, N., Tamakoshi, A., Wakai, K., Matsuo, K., Saito, T., Tajima, K., Tominaga, S.: Helicobacter pylori seropositivity and the myeloperoxidase G-463A polymorphism in combination with interleukin-1B C-31T in Japanese health checkup examinees. Jpn. J. Clin. Oncol. 33, 192–197 (2003)
11. Togawa, S., Joh, T., Itoh, M., Katsuda, N., Ito, H., Matsuo, K., Tajima, K., Hamajima, N.: Interleukin-2 gene polymorphisms associated with increased risk of gastric atrophy from Helicobacter pylori infection. Helicobacter 10, 172–178 (2005)
12. Yang, C.H., Cheng, Y.H., Chuang, L.Y., Chang, H.W.: Genetic Algorithm for the Design of Confronting Two-Pair Primers. In: Ninth IEEE international Conference on BioInformatics and BioEngineering (BIBE), pp. 242–247 (2009)
13. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
14. Kennedy, J.F., Eberhart, R.C., Shi, Y.: Swarm intelligence. Springer, US (2001)
15. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbSNP: the NCBI database of genetic variation. Nucleic. Acids Res. 29, 308–311 (2001)
16. Sambrook, J., Fritsch, E.F., Maniatis, T.: Molecular cloning. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY (1989)
17. Goldberg, T.E., Kotov, R., Lee, A.T., Gregersen, P.K., Lencz, T., Bromet, E., Malhotra, A.K.: The serotonin transporter gene and disease modification in psychosis: Evidence for systematic differences in allelic directionality at the 5-HTTLPR locus. Schizophr. Res. 111, 103–108 (2009)
18. Mandelli, L., Mazza, M., Martinotti, G., Di Nicola, M., Daniela, T., Colombo, E., Missaglia, S., De Ronchi, D., Colombo, R., Janiri, L., Serretti, A.: Harm avoidance moderates the influence of serotonin transporter gene variants on treatment outcome in bipolar patients. J. Affect Disord. 119, 205–209 (2009)
19. Yang, C.H., Cheng, Y.H., Chuang, L.Y., Chang, H.W.: SNP-Flankplus: SNP ID-centric retrieval for SNP flanking sequences. Bioinformation 3, 147–149 (2008)
20. De Jong, K.A., Spears, W.M.: An analysis of the interacting roles of population size and crossover in genetic algorithms, vol. 1, pp. 38–47. Springer, Heidelberg (1990)
21. Hamajima, N., Saito, T., Matsuo, K., Tajima, K.: Competitive amplification and unspecific amplification in polymerase chain reaction with confronting two-pair primers. J. Mol. Diagn. 4, 103–107 (2002)

Strategic Health Information Management and Forecast: The Birdwatching Approach

Arash Shaban-Nejad and Volker Haarslev

Department of Computer Science and Software Engineering, Concordia University, H3G1M8 Montreal, Quebec, Canada
{arash_sh,haarslev}@cs.concordia.ca

Abstract. To facilitate communication and the exchange of information between patients, nurses, lab technicians, health insurers, physicians, policy makers, and existing knowledge-based systems, a set of shared standard terminologies and controlled vocabularies are necessary. In modern health information management systems, these vocabularies are defined within formal representations called ontologies, where terminologies are only meaningful once linked to a descriptive dataset. When the datasets and their conveyed knowledge are changed, the ontological structure is altered accordingly. Despite the importance of this topic, the problem of managing evolving ontological structures is inadequately addressed by available tools and algorithms, partly because handling ontological change is not a purely computational affair. In this paper, we propose a framework inspired by a social activity, birdwatching. Using this model, the evolving ontological structures can be monitored and analyzed based on their state at a given time. Moreover, patterns of changes can be derived and used to predict and approximate a system's behavior based on potential future changes.

Keywords: Change management, Biomedical ontologies, Multi-agent system, Health information management.

1 Introduction

> *"When you know what the habitat and the habits of birds are watching them is so much more interesting."*
>
> *The Beginners Guide to Bird Watching*[1]

Strategic information systems (SIS) are widely used in the healthcare industry to support real-time decision making and consistent maintenance of various changes in strategic vision. Many strategic information systems have employed various controlled vocabularies, ontologies, and knowledge bases as their conceptual backbone to standardize and facilitate human-human, human-agent, and agent-agent interactions and communications (Figure 1).

[1] http://birdwatchingforbeginners.info/

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part III, LNAI 6423, pp. 457–468, 2010.

Using biological classification and clinical vocabularies/lexicons has a long history in medicine and life science dating back to Aristotle's *scala naturae* [1] (scale of nature), which was a very simple method of dividing organisms into groups, ranging from the simple species to more complex ones, based on their appearance. In the 17th century, Carl Linnaeus, who is often referred as the father of modern taxonomy, developed his classification system for the naming and classification of all organisms. Linnaeus represented his classification method based on binomial nomenclature (e.g., humans are identified by the binomial *Homo sapiens*). Later, as the understanding of the relationships between organisms changed, taxonomists converted the five ranks into the seven-rank hierarchy by adding the two ranks of "Phylum" (between Kingdom and Class) and "Family" (between Order and Genus).

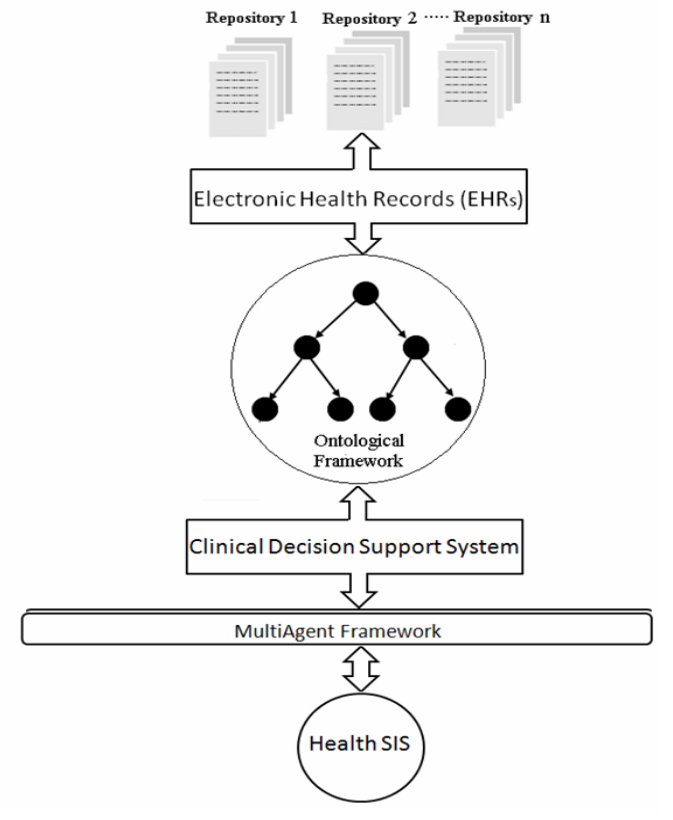


Fig. 1. An abstract representation of the interactions in a typical knowledge-based Health Strategic Information System (SIS)

Change in the taxonomic ranks is still an ongoing process. Due to advances in knowledge and the influence of evolutionary techniques as the mechanism of biological diversity and species formation, taxonomists needed a new classification scheme to reflect the phylogeny of organisms. Also, recruitment of new criteria, besides structural similarities, such as genetic codes and molecular features, and advances in tools

and techniques resulted in the discovery of various organisms, altering the older structures and forming new kingdoms with new branches and terminologies [2]. The biomedical classifications and terminologies have been organized in several models [3] as Controlled Vocabularies, Thesauri, Taxonomies, and Ontologies.

A relatively new trend is emerging to use ontology, as defined by Gruber [4] ("specification of conceptualization"), to provide an underlying discipline of sharing knowledge and modeling biomedical applications by defining concepts, properties and axioms. Modifying and adjusting ontologies in response to changing data or requirements are significant barriers to the implementation of efficient biomedical ontologies in real clinical environments. Depending on the size and complexity of the ontological structures, their maintenance can be very expensive and time consuming. In this paper, we introduce the sociotechnical aspects of our agent-based framework, which aims to assist and guide ontology engineers through the change management process in evolving biomedical ontologies [5] with minimal human intervention.

2 The Birdwatching: A Nature Inspired Approach

Since the existing biomedical knowledge bases are being used in various organizational and geographical levels (i.e. institutional, local, regional, national and international), any change management framework should be able to address this decentralization and distribution nature. One of the critical tasks in any change management framework is *traceability,* which provides transparent access to different versions of an evolving system. It also aids in understanding the impact of a change, recognizing a change and alerting upon occurrence, improving the visibility, reliability, auditability, and verifiability of the system, propagating a change [6], and reproducing results for (or undoing effects of) a particular type of change. Advances in impact analysis gained by traceability facilitate predictability in the post-change analysis stage in an ontology maintenance framework.

To explain our method for change management more intuitively we use a conceptual metaphor based on Birdwatching activity. Birdwatching as a recreational and social activity is the process of observation and study of birds through a particular time frame using different auditory devices. Figure 2 shows a sequence[2] of typical activities recommended for Birdwatching.

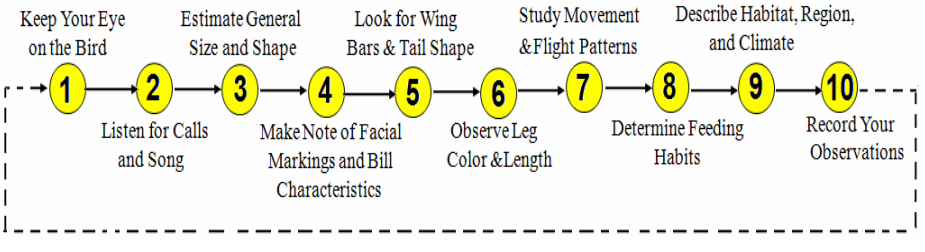


Fig. 2. A series of activities in Birdwatching

[2] Bird Watching Tips for Beginners:
http://animals.about.com/od/birding/tp/birdidtips.htm

Looking at the list of activities presented in Figure 2 one can discover that the central idea of Birdwatching, which is tracking the position of the birds at different time points and predicting their path by deriving a flight pattern based on recorded observed information, is quite close in spirit to monitoring any dynamic spatial-temporal system. Inspired by this metaphor we have designed a multi-agents framework called RLR [7], which aims for Representation, Legitimation and Reproduction of changes in ontological structures. Using intelligent agents reduces several issues related to human intervention in dynamic e-health systems [8].

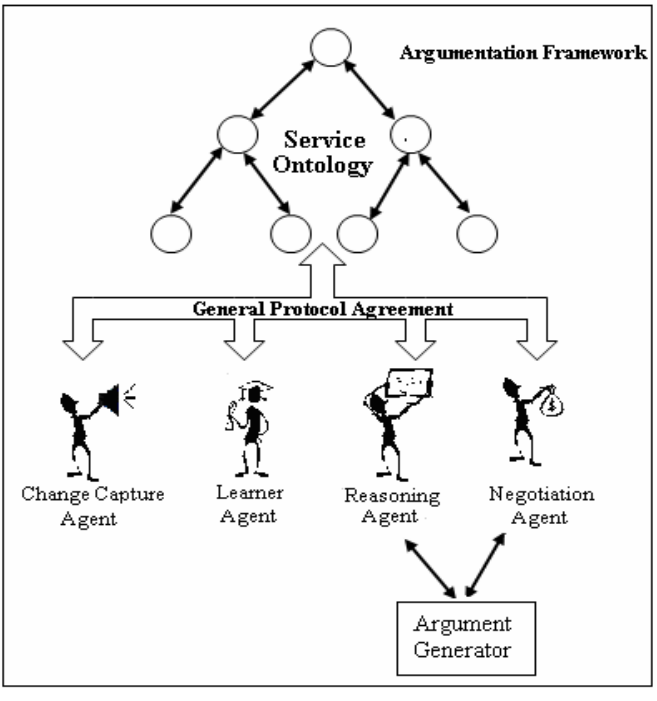


Fig. 3. The RLR framework with a service ontology providing consensus between agents. To reach an agreement among the agents and provide a common understanding, the service ontology is needed, so that updating this ontology generates a new understanding for the software agents, which can then update and adjust their beliefs based on new knowledge.

As part of RLR, we have defined a set of change capture agents, learning agents, negotiation agents, and reasoning agents within an argumentation-based framework (Figure 3) that enables agents with conflicting interests to cooperate. To reach an agreement among the agents and provide a common understanding between them, a service ontology, as shown in Figure 3, is needed so that updating this ontology generates a new understanding for the software agents, which can then update and adjust their beliefs based on new knowledge. Employing service ontologies to automatically provide a service profile to describe the supported services and the related communicative transactions and invoke the services for service-seeking agents is currently being considered as a solution to overcome some of the issues related to overreliance

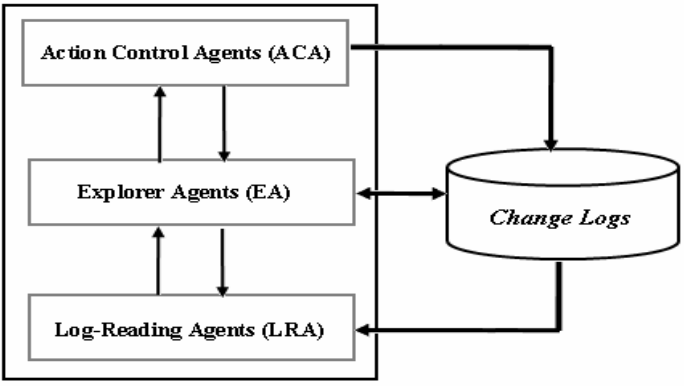


Fig. 4. The cooperation between the change capture agents

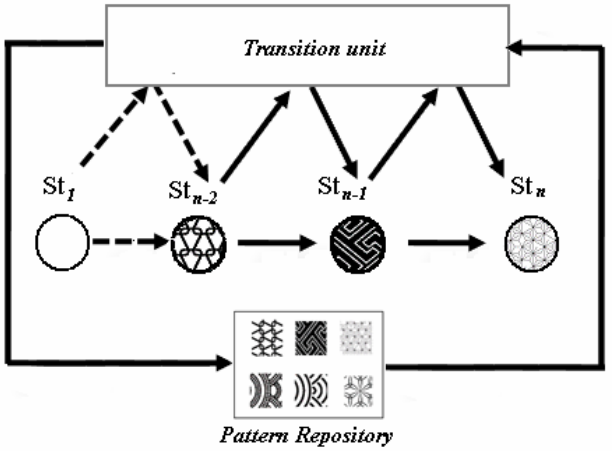


Fig. 5. A generic transition system in a multi-agent system. A system changes its state from St_1 to St_2 via a transition unit and a rigorous argumentation process between the agents to choose proper patterns from the change pattern repository for implementing a certain type of change.

on human intervention. However, these ontologies will not remain static and unchanged throughout their life cycle, and managing their dynamic structure would be part of the whole problem itself.

In the RLR framework the change capture Agents (Figure 4) are responsible for Tasks 1 and 2, represented in Figure 2. The changes logs store the information about the changes (Task 4). The learning agents start with limited knowledge (Task 5 and 6) and improve themselves by gaining inferred knowledge (Tasks 8 and 9) based on the semantics provided by the ontological backbone. Moreover, the learning agents along with negotiation agents (which manage the negotiation process to find a proper way to implement a change) and reasoning agents (which check for inconsistencies and perform final validations) can derive a pattern of change using the information stored in the change logs and the background and derived knowledge (Task 7). Using this

pattern one can achieve a practical estimate for expected changes (Task 3). Finally the result of the observation will be stored to be used for future inferencings (Task 10), and to choose an appropriate pattern (Task 7) in the reproduction phase.

The final outcome, which has been generated through a rigorous argumentation process over generally accepted arguments, has an implicit link to the archived historical processes that can be reused to choose a proper pattern (Task 7) in the reproduction phase (Figure 5).

3 Formalization of the Framework

Looking at the different tasks in Birdwatching one can discover that the central idea of Birdwatching is quite close in spirit to monitoring any dynamic spatial-temporal system. The Galileo's dialogue [24] for explaining motion for the first time stated that for capturing and tracking a moving object one needs to record the position of that object in each instance of time.

3.1 Categorical Representation

Here we use category theory [9], as an algebraic notation independent of any implementation language, to study the ontological dynamism by mapping from a category of times to a category of states or back to our Birdwatching metaphor, the bird's flight (motion) can be represented by mapping from a category of times to a category of spaces (Figure 6). The role of time is not usually taken into account in current ontology evolution studies. Considering time in ontologies can increase the complexity and needs a very expressive ontology language to represent it. In our approach we represent conceptualization of things indexed by times and we use categorical constructors for capturing the states of ontologies at different time points.

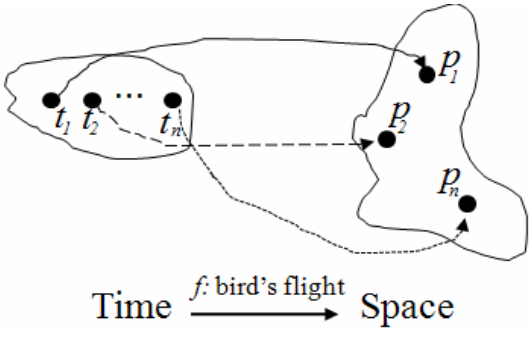


Fig. 6. A map from category of time points to category of positions in space for describing a bird's flight categorical perspective [25]

Similarly, the behavior of an individual ontological element (state) can be monitored by function g, which maps the time points to the set of positions for the element in the ontology.

$$\text{Time} \xrightarrow{g:element's\_behavior} \text{Ontology}$$

Moreover an ontology has different states and behaves in a distributed semantic web environment.

$$\text{State} \xleftarrow{i:has\_state} \text{Ontology} \xrightarrow{h:behavies} \text{Semantic Web}$$

Composing these diagrams one can see that a behavior of an individual ontological element should be studied in close relations with time, the state and the behavior of the whole ontological structure in a semantic web environment (Figure 7).

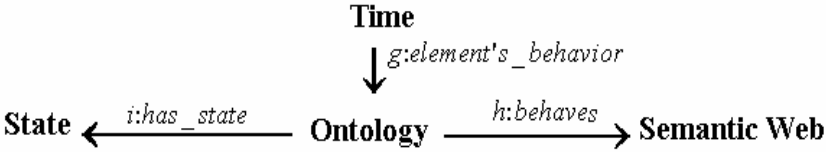


Fig. 7. A temporal diagram for studying the behavior of ontologies

Using category theory enables us to formally represent and track the evolving ontological structure and the argumentation network. It also provides a formal basis to be used by the RLR system for recommendations and conflict resolution. We have also employed the category theoretical distributed graph transformation techniques [10] to analyze the model transition and transformation using certain conditions, which are specified via transformation rules. As an example, in Figure 8, consider two taxonomies related to ontologies O_1 (source ontology) and O_2 (target ontology), where each node represents a concept, which is identified with a label along with a set of corresponding attributes. After discovering similarities and differences between these two taxonomies, we need to find a proper transformation that has been transformed O_1 to O_2. To start this procedure, the two taxonomies need to be aligned and brought into a mutual agreement, based on the matching concepts (the ones that affected less in the transformation) within the ontologies. The matching will be computed based on the degree of similarities between two concepts. From the categorical point of view, the problem of comparing two hierarchical structures can be studied by exploring isomorphisms in their structures.

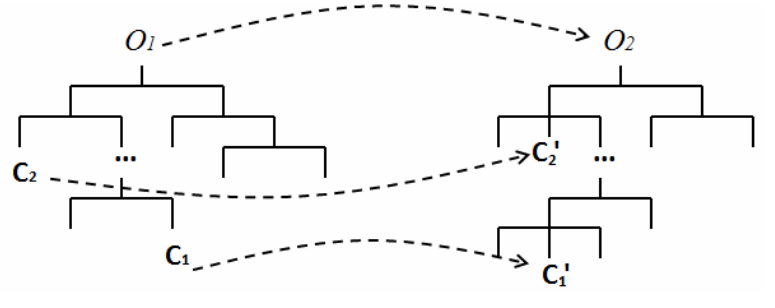


Fig. 8. The alignments between some concepts in two ontologies O_1 and O_2

The use of graph transformations helps us discover the set of operations that transforms the hierarchy indicating the old version of an ontology into the hierarchy indicating the new one.

3.2 The RLR Dialectic Change Management

In debates on distinguishing between "Dependent" and "Independent" entities in the real world, the two concepts of *Ontological Philosophy* and *Dialectic Change* attracted our attention. The concept of Ontological Philosophy [11] focuses on the wholeness and unity of the world and considers change as an aspect of substances in the real world. From the other side, the concept of Dialectical Change [12] tries to represent a change as new forms built upon the old and by combining the new and the old without total replacement, implying both newness and continuity. In this theory, any change needs a cause and can be placed through a process. Holsti [12] used the Marxist idiom, the synthesis, as a metaphor for this processes.

Using the concept of "*Dialectic Change*" as a metaphor, we can introduce our formal agent-based argumentative framework, where "synthesis" takes place, for studying ontology evolution and shifting as model transformation. This transformation results from quantitative changes accumulated over a period of time and generates a new form out of old patterns ("coexistence of both old and new") [12]. In fact, most of the changes that occur in an ontological structure, which lead to a new state, emerge from the preceding states[3]. In other words, the change lies within the system [13]. Therefore, "learning" about different actions in different states of a system seems to be a key factor for starting a successful change management mechanism.

3.3 Models of Learning

By determining the tradeoffs between losses and benefits that can result from agents' actions, we will be able to have a mathematical model to foresee the agents' (software or human) behavior. A state of "Nash equilibrium" [14] is one of the popular approaches in evolutionary game theory for modeling the most beneficial (or least harmful) set of actions for a set of intelligent agents. For the sake of prediction, Nash equilibrium can be understood as "a potential stable point of a dynamic adjustment process in which individuals adjust their behavior to that of the other players in the game, searching for strategy choices that will give them better results" [15]. In our framework the intelligent agents decide on the proper actions and able to change and improve their decisions based on what they learn. Here, following the approach given by [23], for each learning agent, we define an internal state; a function that shows how an agent decides and chooses actions based on its internal state (decision-making); the functions showing the payoff dominance (loss/benefit); and a state update function, specifying how an agent updates its state based on the payoff received from previous iterations. The state of each agent depends on the probability distribution over all the possible situations [23], and the one with the highest probability can specify the final decision. Another technique for automating the learning process is through inductive bias. The inductive bias of learning [16] in neural net-

[3] A "state" in this manuscript is being used to express a situation describing a part of the real (dynamic) world in a specific instance of time.

works is a set of assumptions, given as input, that the learner uses to predict and approximate the target outputs (even for unseen situations) through a series of training instances and their generalization.

3.4 Anomaly Pattern Analysis

Intelligent agents in RLR also detect and generate patterns of anomalies, either syntactic or semantic, by assessing and analyzing consistent common errors that occur through different revisions. After the anomalies have been flagged by change capture agents, the learner agent can then be taught the proper route for performing the revisions through a set of pattern mining algorithms (see [17] as an example of techniques for mining dynamic patterns). This task is crucial in a wide variety of applications, such as biosurveillance for disease outbreak detection [18] and cancer diagnosis. The learner agents not only enable the RLR framework to manage potential, expected, and prescheduled changes, but also prepare it for dealing with random and unexpected alterations. However, human supervision and participation will be anticipated for the former case.

3.5 The Change Analysis Model in RLR

Our change analysis model is composed of a set of states that are linked to their predecessors and successors through some defined relationships. This allows us to check backward and forward compatibilities for one specific ontological structure from a given state. This is determined by defining various conditions and constraints for an event. The conditions can later be used to restore the previous state based on the insights gained for each event. Somehow it means a revision or review of the past, or an attempt to define an alternate (parallel) past [19]. Since ontological assertions are based on open world assumptions, neither past nor future knowledge about the world is complete. One can always ask questions (e.g., "Could a specific mutation, under certain circumstances, lead to the species X or Y?") and draw a different path from the previous states to the subsequent states. This iterative process of switching between the future, current, and revised past states has been regarded in [19] as the process of "rolling back to some previous state and then reasoning forward" in the form of queries such as, "Is there some future time in which p is true?" [19].

To deal with forward and backward compatibility, in our research we have employed graph transformation techniques, which enable us to analyze different states of the graphs based on the given initial states and the transformation rules. Indeed, graph transformation offers many benefits, but lacks sufficient expressivity and semantics to deal with all aspects of ontology change management. Our approach for this issue can be improved by recruiting a formal mathematical representation such as category theory. The enhancement can be done in two aspects: 1) the transformation rules can impose restrictions on ontology transformation in the way that, for example, some alteration can be prohibited, or some changes, which have less impact on ontological elements, can be excluded in the related change analysis (e.g., the transition of a fungus from one genus to another does not affect its physical appearance); 2) the changes in states can be scheduled to occur simultaneously, sequentially, or in parallel.

3.6 Identity Preservation in RLR

The identity of a concept can be determined by those properties and facts that remain stable through time, even during multiple ontological changes. If ontologies are able to maintain their conceptual stability, they can better preserve their intended truth. To this end, the RLR framework employs a defensive mechanism to prevent harmful changes and reduce the risk of potentially dangerous actions by incrementally adapting to the changes at different levels. If a destructive change is about to happen in the ontology (e.g., deleting a concept, such as "fungi", when other dependent concepts, such as "fungal infection", exist), a warning signal will be sent to the agents based on the knowledge within the ontology (e.g., "fungi are the cause of fungal infections") to infer the potential threat and prepare them to plan for a proper action. This mechanism works much like the self-awareness system inside rational animals, which helps them avoid possible dangers without actually experiencing their life threatening influences. For example, as pointed out in [20], a person who is confronted with fire does not have to experience the burning sensation and can run away as a counteraction, since the person has been taught that smoke indicates fire and that fire can kill humans.

4 Discussion and Future Works

To avoid the fatal errors caused by uncontrolled changes in biomedical knowledge-based systems, a consistent change management process with minimum human intervention is vital. In this paper, we have described a method based on a metaphor taken from a recreational activity, birdwatching, to highlight the temporal aspects of ontologies by representing the conceptualization of things indexed by times, which enables one to control forward and backward compatibilities for taxonomic revisions. In fact, our introduced approach, based on the insights from category theory, can be employed to develop algorithms and tools to assist ontology change management. In our recent experiments, we have applied the introduced agent-based method, formalized with category theory, in several biomedical applications, including the management of requirement volatility in e-health systems [21] and analyzing the evolutionary relationships between fungal species [22]. Currently, we are working to improve our rule-based graph transformation method and extend it to cover hierarchical distributed graphs, which support nested hierarchies in different levels of abstraction.

References

1. Verhagen, F.C.: Worldviews and Metaphors in the human-nature relationships: An Ecolinguistic Exploration Through the Ages. Language & Ecology 2(3) (2008)
2. eHealth: standardized terminology. Executive Board 118th Session, EB118, Provisional agenda 8.4. World Health Organization (May 25, 2006), http://apps.who.int/gb/ebwha/pdf_files/EB118/B118_8-en.pdf

3. Hedden, H.: Controlled Vocabularies, Thesauri, and Taxonomies. The Indexer 26(1), 33–34 (2008)
4. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition 5(2), 199–220 (1993)
5. Shaban-Nejad, A., Haarslev, V.: Bio-medical Ontologies Maintenance and Change Management. In: Sidhu, A.S., Dillon, T.S. (eds.) Biomedical Data and Applications. Studies in Computational Intelligence, vol. 224, pp. 143–168. Springer, Heidelberg (2009)
6. Smith, M.J., Dewar, R.G., Kowalczykiewicz, K., Weiss, D.: Towards Automated Change Propagation; the value of traceability. Technical Report, Heriot Watt University (2003)
7. Shaban-Nejad, A., Haarslev, V.: Incremental biomedical ontology change management through learning agents. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS (LNAI), vol. 4953, pp. 526–535. Springer, Heidelberg (2008)
8. Shaban-Nejad, A., Haarslev, V.: Human Factors in Dynamic E-Health Systems and Digital Libraries. In: Pease, W., Cooper, M., Gururajan, R. (eds.) Biomedical Knowledge Management: Infrastructures and Processes for E-Health Systems. Information Science Reference – ISR, pp. 192–203. IGI Global (2010)
9. Asperti, A., Longo, G.: Categories, types, and structures: an introduction to category theory for the working computer scientist. MIT Press, Cambridge (1991)
10. Ehrig, H., Ehrig, K., Prange, U., Taentzer, G.: Fundamentals of Algebraic Graph Transformation. In: Monographs in Theoretical Computer Science. An EATCS Series. Springer, New York (2006)
11. Scribner, P.: Introduction to Ontological Philosophy (1999), http://www.twow.net/MclOtaI.htm
12. Holsti, K.J.: The Problem of Change in International Relations Theory. Paper No. 26 from CIR Working Paper Series (1998)
13. Gilbert, M.C.: The Dialectics of Knowledge Management (2006), http://news.gilbert.org/DialecticsKM
14. Osborne, M.J.: Nash Equilibrium: Theory. In: An Introduction to Game Theory. Oxford University Press, USA (2003)
15. Holt, C.A., Roth, A.E.: The Nash equilibrium: A perspective. PNAS 101(12), 3999–4002 (2004)
16. Mitchell, T.M.: The need for biases in learning generalizations. In: Shavlik, J.W., Dietterich, T.G. (eds.) Readings in Machine Learning, pp. 184–191. Morgan Kaufmann, San Francisco (1990)
17. Chung, S., McLeod, D.: Dynamic Pattern Mining: An Incremental Data Clustering Approach. J. Data Semantics 2, 85–112 (2005)
18. Wong, W.K., Moore, A.W., Cooper, C.F., Wagner, M.: Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. In: 20th International Conference on Machine Learning (ICML 2003), pp. 808–815. AAAI Press, Menlo Park (2003)
19. Mays, E.: A Modal Temporal Logic for Reasoning about Change. In: 21st Annual Meeting of the Association for Computational Linguistics (ACL), Cambridge, MA, US, pp. 38–43 (1983)
20. Heylighen, F.: Representation and Change. A Metarepresentational Framework for the Foundations of Physical and Cognitive Science. Communication & Cognition, Gent. 200 (1990)
21. Shaban-Nejad, A., Ormandjieva, O., Kassab, M., Haarslev, V.: Managing Requirements Volatility in an Ontology-Driven Clinical LIMS Using Category Theory. International Journal of Telemedicine and Applications, Article ID 917826, 14 (2009) doi: 10.1155/2009/917826

22. Shaban-Nejad, A., Haarslev, V.: Ontology-inferred phylogeny reconstruction for analyzing the evolutionary relationships between species: Ontological inference versus cladistics. In: 8th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2008), Athens, Greece, pp. 1–7. IEEE Press, Los Alamitos (2008)
23. Wang, J.: Computational Approaches to Linguistic consensus. Dissertation at University of Illinois at Urbana-Champaign (2006)
24. Galileo, G.: Dialogue Concerning the Two Chief Systems of the World - Ptolemaic and Copernican (1632), http://www.gap-system.org/~history/Extras/Galileo_Dialogue.html
25. Lawvere, F.W., Schanuel, S.H.: Conceptual Mathematics: A First Introduction to Categories, 2nd edn. Cambridge University Press, Cambridge (2009) (The 1st ed. published on 1997)

Author Index