

David C. Wyld
Jan Zizka (Eds)

Computer Science & Information Technology

International Conference on Foundations of Computer Science &
Technology (CST 2014)
Zurich, Switzerland, January 02 ~ 04 - 2014



AIRCC

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN : 2231 - 5403
ISBN : 978-1-921987-24-3
DOI : 10.5121/csit.2014.4101 - 10.5121/csit.2014.4136

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The International Conference on Foundations of Computer Science & Technology (CST 2014) was held in Zurich, Switzerland, during January 02~04, 2014. The Third International Conference on Information Technology Convergence and Services (ITCS 2014), Third International Conference on Software Engineering and Applications (JSE 2014) , The Third International Conference on Signal and Image Processing (SIP 2014), International Conference on Artificial Intelligence & Applications (ARIA 2014), Fifth International conference on Database Management Systems (DMS 2014) were collocated with the CST-2014. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CST 2014, ITCS 2014, JSE 2014, SIP 2014, ARIA 2014, DMS 2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CST 2014, ITCS 2014, JSE 2014, SIP 2014, ARIA 2014, DMS 2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CST 2014, ITCS 2014, JSE 2014, SIP 2014, ARIA 2014, DMS 2014.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld
Jan Zizka

Organization

General Chairs

David C. Wyld
Natarajan Meghanathan

Southeastern Louisiana University, USA
Jackson State University, USA

Steering Committee

Abdul Kadir Ozcan
Brajesh Kumar Kaushik
Dhinaharan Nagamalai
Eric Renault
John Karamitsos
Khoa N. Le

The American University, Cyprus
Indian Institute of Technology - Roorkee, India
Wireilla Net Solutions PTY Ltd, Australia
Institut Telecom–Telecom SudParis, France
University of the Aegean, Samos, Greece
University of Western Sydney, Australia

Program Committee Members

A Vadivel
A.G.Ananth
A.Kannan
Abdellatif BERKAT
Achhman Das Dhomeja
Ajay K Sharma
Alejandro Regalado Mendez
Alvin Lim
Amandeep Singh Thethi
Asghar gholamian
Ashok kumar Sharma
Ayad salhieh
Azween Bin Abdullah
Balaji Raj N
Binod Kumar Pattanayak
Buket Barkana
Carlos E. Otero
Ch.V.Rama Rao
Choudhari
D.Minnie
Deepak Laxmi Narasimha
Denivaldo LOPES
Dinesh Chandrajain
Ferdin Joe J
G.M. Nasira

National Institute of Technology Trichy, India
R.V. College of Engineering-Bangalore, India
K.L.N. College of Engineering, India
Abou-Bekr Belkadd University (Tlemcen), Algeria
University of Sindh, Pakistan
Dr B R Ambedkar NIT, India
Universidad del Mar. Mexico
Auburn University, USA
Guru Nanak Dev University Amritsar, India
Babol University of Technology, Iran
YMCA Institute of Engineering, India
Australian College at Kuwait, Kuwait
Universiti Teknologi Petronas, Malaysia
JJ College of Engineering and Technology, India
Siksha O Anusandhan University, India
University of Bridgeport, USA
The University of Virginia's College at Wise, USA
Gudlavalleru Engineering College, India
Bhagwati Chaturvedi College of Engineering, India
Madras Christian College, India
University of Malaya, Malaysia
Federal University of Maranhao - UFMA, Brazil
University of RGPV, India
Prathyusha Institute of Tech. & Management, India
Sasurie College of Engineering, India

Hao Shi	Victoria University, Australia
Hao-En Chueh	Yuanpei University, Taiwan, R.O.C.
Henrique J. A. Holanda	UERN - Universidade do Estado do Rio Grande do Norte
Indrajit Bhattacharya	Kalyani Govt. Engg. College, India
Jalel Akaichi	University of Tunis, Tunisia
Jestin Joy	Federal Institute of Science and Technology, India
Jyoti Singhai	Electronics and Communication Deptt-MANIT, India
Jyotirmay Gadewadikar	Alcorn State University, USA
K. Chitra	Govt Arts College for Women, India
kalikiri nagi reddy	NBKR Institute of Science & Technology, India
Khoa N. Le	University of Western Sydney, Australia
Krishna Prasad E S N Ponnekanti (KP)	Aditya Engineering College-Kakinada, India
Krishnaveni	Avinashilingam University for Women, India
L.Jaba Sheela	Anna University, India
lakshmi Rajamani	Osmania University, India
Lylia Abrouk	University of Burgundy, France
M. Dinakaran	VIT University – Vellore, India
M. P. Singh	National Institute of Technology Patna, India
M.Hemalatha	Karpagam University, India
M.P Singh	National Institute of Technology, India
M.Pravin Kumar	K.S.R College of Engineering, India
Madhan KS	Infosys Technologies Limited, India.
Michel Owayjan	AUST, Lebanon
Mohammed Ali Hussain	Sri Sai Madhavi Institute of Science & Tech., India
Mohd. Ehmer Khan	Al Musanna College of Technology, Sultanate of Oman
Monika Verma	Punjab Technical University, India
Narottam C. Kaushal	NIT Hamirpur, India
Nitiket N Mhala	B.D.College of Engineering - Sewagram, India
Nour Eldin Elmadany	Arab Academy for Science and Technology, Egypt
P.Ashok Babu	D.M.S.S.V.H. College of Engineering, India
P.Shanmugavadivu	Gandhigram Rural Institute - Deemed University, India
P.Thiyagarajan	Pondicherry University, India
Patrick Seeling	University of Wisconsin, USA
Pravin P. Karde	HVPM's College of Engg. & Tech. - Amravati, India
Premanand K.Kadbe	Vidya Pratishthan's College of Engineering, India
R. Murali	Dr. Ambedkar Institute of Technology, Bangalore
R.Baskaran	Anna University - Chennai, India
Rahul Vishwakarma	Tata Consultancy Services,ACM, India
Raman Maini	Punjabi University, India
Richard Millham	University of Bahamas, Bahamas
Roberts Masillamani	Hindustan University, India
S.Sapna	K.S.R College of Engineering, India
S.Senthilkumar	NIT - Tiruchirappalli, India
Salman Abdul Moiz	Centre for Development of Advanced Computing, India
Sandhya Tarar	Gautam Buddha University, India
Sanjay K, Dwivedi	Ambedkar Central University Lucknow, India
Sanjay Singh	Manipal University, India

Sanjoy Das
Sherif S. Rashad
Shin-ichi Kuribayashi
Shrirang.Ambaji.Kulkarni
Sundarapandian V
T Venkat Narayana Rao
Tien D. Nguyen
Tuli Bakshi
Utpal Biswas
V.Radha
Vijayanandh. R
Wichian Sittipraporn
wided oueslati
Zuhal Tanrikulu

Jawaharlal Nehru University, India
Morehead State University, USA
Seikei University, Japan
National Institute of Engineering, India
Vel Tech Dr. RR & Dr. SR Technical University, India
Hyderabad ITM , India
Coventry University, UK
Calcutta Institute of Technology(WBUT), India
University of Kalyani, India
Avinashilingam University, India
Bharathiar Univ, India
Mahasarakham University, Thailand
l'institut superieur de gestion de tunis, Tunisia
Bogazici University, Turkey

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)



Software Engineering & Security Community (SESC)



Digital Signal & Image Processing Community (DSIPC)



Organized By



ACADEMY & INDUSTRY RESEARCH COLLABORATION CENTER (AIRCC)

TABLE OF CONTENTS

Foundations of Computer Science & Technology

On Selection of Periodic Kernels Parameters in Time Series Prediction.....	1
<i>Marcin Michalak</i>	

Research on the Mobile Robots Intelligent Path Planning Based on ANT Colony Algorithm Application in Manufacturing Logistics.....	11
<i>GUO Yue, SHEN Xuelian and ZHU Zhanfeng</i>	

Dictionary-Based Concept Mining : An Application for Turkish.....	27
<i>Cem Rifki Aydin, Ali Erkan, Tunga Güngör and Hidayet Takçi</i>	

Finding Important Nodes in Social Networks Based on Modified Pagerank.....	39
<i>Li-qing Qiu, Yong-quan Liang and Jing-Chen</i>	

A SAT Encoding for Solving Games with Energy Objectives.....	45
<i>Raffaella Gentilini</i>	

A Content Based Watermarking Scheme Using Radial Symmetry Transform and Singular Value Decomposition.....	53
<i>Lakehal Elkhamssa and Benmohammed Mohamed</i>	

Analysis of Intraday Trading of Index Option in Korean Option Market.....	65
<i>Young-Hoon Ko</i>	

Quality of Service Management in Distributed Feedback Control Scheduling Architecture Using Different Replication Policies.....	75
<i>Malek Ben Salem, Emna Bouazizi, Rafik Bouaziz and Claude Duvalet</i>	

Information Technology Convergence and Services

Emotion Teaching Interface for Finger Braille Emotion Teaching System.....	89
<i>Yasuhiro Matsuda and Tsuneshi Isomura</i>	

Distributed System for 3D Remote Monitoring Using KINECT Depth Cameras....	101
<i>M. Martínez-Zarzuela, M. Pedraza-Hueso, F.J. Díaz-Pernas, D. González-Ortega and M. Antón-Rodríguez</i>	

Color Satellite Image Compression Using the Evidence Theory and Huffman Coding.....	113
<i>Khaled SAHNOUN and Noureddine BENABADJI</i>	

On-Board Satellite Image Compression Using the Fourier Transform and Huffman Coding.....	119
<i>Khaled SAHNOUN and Noureddine BENABADJI</i>	

Amino Acid Interaction Network Prediction Using Multi-Objective Optimization.....	127
<i>Md. Shiplu Hawlader and Saifuddin Md. Tareeq</i>	

Software Engineering and Applications

Flight Trajectory Recreation and Playback System of Aerial Mission Based on OssimPlanet.....	141
<i>Wu Wu, Jiulin Hu, Xiaofang Huang, Huijie Chen and Bo Sun</i>	

Audit Maturity Model.....	155
<i>Bhattacharya Uttam, Rahut Amit Kumar and De Sujoy</i>	

A Structural Approach to Improve Software Design Reusability.....	163
<i>Tawfig M. Abdelaziz, Yasmeen.N.Zada and Mohamed A. Hagal</i>	

Quality-Aware Approach for Engineering Self-Adaptive Software Systems.....	173
<i>Mohammed Abufouda</i>	

Evaluation of the Software Architecture Styles from Maintainability Viewpoint.....	183
<i>Gholamreza ShahMohammadi</i>	

A Similarity Measure for Categorizing the Developers Profile in a Software Process.....	199
<i>Hamid Khemissa, Mohamed Ahmed-nacer and Abdelkader Belkhir</i>	

Signal and Image Processing

Image Acquisition in an Underwater Vision System with NIR and VIS Illumination.....	215
<i>Wojciech Bieganski and Andrzej Kasinski</i>	

Variation-Free Watermarking Technique Based on Scale Relationship.....	225
<i>Jung-San Lee, Hsiao-Shan Wong and Yi-Hua Wang</i>	
Region Classification Based Image Denoising Using Shearlet and Wavelet Transforms.....	241
<i>Preety D. Swami, Alok Jain and Dharendra K. Swami</i>	
Synthetical Enlargement of MFCC Based Training Sets for Emotion Recognition.....	249
<i>Inma Mohino-Herranz, Roberto Gil-Pita, Sagrario Alonso-Diaz and Manuel Rosa-Zurera</i>	
A Modified Histogram Based Fast Enhancement Algorithm.....	261
<i>Amany A. Kandeel, Alaa M. Abbas, Mohiy M. Hadhoud and Zeiad El-Saghir</i>	
Fingerprints Image Compression by Wave Atoms.....	271
<i>Mustapha Delassi and Amina Serir</i>	
 Artificial Intelligence & Applications	
Visual Tracking Using Particle Swarm Optimization.....	279
<i>J.R.Siddiqui and S.Khatibi</i>	
Fuzzy Inference System for VOLT/VAR Control in Distribution Substations in Isolated Power Systems.....	293
<i>Vega-Fuentes E, León-del Rosario S, Cerezo-Sánchez J M and Vega-Martínez A</i>	
Towards Universal Rating of Online Multimedia Content.....	305
<i>Lawrence Nderu, Nicolas Jouandeau and Herman Akdag</i>	
Query Proof Structure Caching for Incremental Evaluation of Tabled Prolog Programs.....	311
<i>Taher Ali, Ziad Najem and Mohd Sapiyan</i>	
Design, Implement and Simulate an Agent Motion Planning Algorithm in 2D and 3D Environments.....	323
<i>Haissam El-Aawar and Hussein Bakri</i>	
A ROS Implementation of the Mono-Slam Algorithm.....	339
<i>Ludovico Russo, Stefano Rosa, Basilio Bona and Matteo Matteucci</i>	
Pre-Ranking Documents Valorization in the Information Retrieval Process.....	353
<i>Chkiwa Mounira, Jedidi Anis and Faiez Gargouri</i>	

Auto Landing Process for Autonomous Flying Robot by Using Image Processing Based on Edge Detection..... 361

Bahram Lavi Sefidgari and Sahand Pourhassan Shamchi

Intelligent Multi-Agent Fuzzy Control System Under Uncertainty..... 369

Ben Khayut, Lina Fabri and Maya Abukhana

Database Management Systems

Intelligent and Pervasive Archiving Framework to Enhance the Usability of the Zero-Client-Based Cloud Storage System..... 381

Keedong Yoo

Efficiently Processing of Top-K Typicality Query for Structured Data..... 391

Jaehui Park and Sang-goo Lee

ON SELECTION OF PERIODIC KERNELS PARAMETERS IN TIME SERIES PREDICTION

Marcin Michalak

Institute of Informatics, Silesian University of Technology,
ul. Akademicka 16, 44-100 Gliwice, Poland
Marcin.Michalak@polsl.pl

ABSTRACT

In the paper the analysis of the periodic kernels parameters is described. Periodic kernels can be used for the prediction task, performed as the typical regression problem. On the basis of the Periodic Kernel Estimator (PerKE) the prediction of real time series is performed. As periodic kernels require the setting of their parameters it is necessary to analyse their influence on the prediction quality. This paper describes an easy methodology of finding values of parameters of periodic kernels. It is based on grid search. Two different error measures are taken into consideration as the prediction qualities but lead to comparable results. The methodology was tested on benchmark and real datasets and proved to give satisfactory results.

KEYWORDS

Kernel regression, time series prediction, nonparametric regression

1. INTRODUCTION

Estimation of a regression function is a way of describing a character of a phenomenon on the basis of the values of known variables that influence on the phenomenon. There are three main branches of the regression methods: parametric, nonparametric, and semiparametric. In the parametric regression the form of the dependence is assumed (the function with the finite number of parameters) and the regression task simplifies to the estimation of the model (function) parameters. The linear or polynomial regression are the most popular examples. In the nonparametric regression any analytical form of the regression function can be assumed and it is built straight from the data like in Support Vector Machines (*SVM*), kernel estimators, or neural networks. The third group is the combination of the two previously described. The regression task in this case is performed in two steps: firstly the parametric regression is applied followed by the nonparametric.

Time series are a specific kind of data: the observed phenomenon depends of some set of variables but also on the laps of time. The most popular and well known methods of time series analysis and prediction are presented in [1] which first edition was in 60's of the 20th century.

In this paper the semiparametric model of regression is applied for the purpose of time series prediction. In the previous works kernel estimators and *SVM* were used for this task [2][3] but

these methods required mapping of the time series into a new space. Another approach was presented in [4] where the Periodic Kernel Estimator (*PerKE*) was defined. It is also the semiparametric algorithm. In the first step the regression model is built (linear or exponential) and for the rests the nonparametric model is applied. The final prediction is the compound of two models. The nonparametric step is the kernel regression with the specific kind of kernel function called periodic kernel function. In the mentioned paper two kernels were defined.

Because each periodic kernel requires some parameters in this paper the analysis of the influence of kernel parameters on prediction error becomes the point of interest. The paper is organized as follows: it starts from a short description of prediction and regression methods, then the *PerKE* algorithm is presented. Afterwards, results of the experiments performed on time series are given. The paper ends with conclusions and the description of further works.

2. PREDICTION AND REGRESSION MODELS

2.1. ARIMA (SARIMA) Models

SARIMA (*Seasonal ARIMA*) model generalizes the Box and Jenkins *ARIMA* model (*AutoRegressive Integrated Moving Average*) [1] as the connection of three simple models: autoregression (*AR*), moving average (*MA*) and integration (*I*).

If B is defined as the lag operator for the time series x ($Bx_t = x_{t-1}$) then the autoregressive model of the order p (a_t is the white noise and will be used also in other models) is given by the formula:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + a_t$$

and may be defined as:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p) x_t = a_t$$

In the *MA* models the value of time series depends on random component a_t and its q delays as follows:

$$x_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

or as:

$$a_t (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) = x_t$$

For the non-stationary time series the d operation of its differentiation is performed, described as the component $(I - B)^d$ in the final equation. The full *ARIMA*(p, d, q) model takes the form:

$$(1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p) (1 - B)^d x_t = a_t (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)$$

The *SARIMA* model is dedicated for time series that have strong periodic fluctuations. If s is the seasonal delay the model is described as *SARIMA*(p, d, q)(P, D, Q) ^{s} where P is the order of seasonal autoregression ($1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_p B^{Ps}$), Q is the order of seasonal moving average ($1 - \Theta_1 B^s - \Theta_2 B^{2s} - \dots - \Theta_q B^{qs}$) and D is the order of seasonal integration ($1 - B^s - B^{2s} - \dots - B^{Ds}$).

2.2. Decomposition Method

This method tries to separate several components of the time series, each of them describing the series in the different way. Most important components are as follows:

- trend component (T): the long-time characteristic of the time series,
- seasonal component (S): the periodic changes of the time series,
- cyclical component (C): repeated but non-periodic changes of the time series,
- irregular (random) component (e).

Components are usually aggregated. It may be an additive aggregation when the final predicted value is a sum of all time series components or multiplicative aggregation when the final value is calculated as a multiplication of all time series components. First is called additive and the final predicted value is the sum of component time series values and the second is called multiplicative (aggregation is the multiplication of time series values).

2.3. Periodic Kernels

Periodic kernels belong to a wide group of kernel functions that are applied for the task of estimation of the regression function. They fulfil the typical conditions for the kernel function and some of the specific ones. As the most important typical features of the kernel function the following should be mentioned [5]:

- $\int_{\mathbb{R}} K(u) du = 1$
- $\forall u \in \mathbb{R} K(u) = K(-u)$
- $\int_{\mathbb{R}} uK(u) du = 0$
- $\forall u \in \mathbb{R} K(0) \geq K(u)$
- $\int_{\mathbb{R}} u^2 K(u) du < \infty$

Furthermore, if we assume that the period of the analysed time series is T then there are the following specific conditions for the periodic kernel function:

- for each $k \in \mathbb{Z}$ the value $K(kT)$ is the strong local maximum,
- for each $x \in \mathbb{R} \setminus \{0\}$ $K(0) > K(x)$,
- for each $n_1, n_2 \in \mathbb{N}$ that $n_1 < n_2$ $K(n_1) > K(n_2)$.

In the paper [4] two periodic kernels were defined, named First Periodic Kernel (FPK) and Second Periodic Kernel (SPK). The formula of FPK is the multiplication of the exponential function and the cosine:

$$FPK(x) = \frac{1}{C} e^{-a|x|} (1 + \cos bx)$$

The constant C assures that K is integrable to one. This value depends on the values a and b as follows:

$$C = 2 \int_0^{\infty} e^{-ax} (1 + \cos bx) dx = \frac{4a^2 + 2b^2}{a(a^2 + b^2)}$$

In order to define the FPK it is to substitute a and b with the period T and parameter θ that is a function attenuation (the ratio of the two consecutive local maxima):

$$b = \frac{2\pi}{T}$$

$$\theta = \frac{K(t+T)}{K(t)} \Rightarrow -aT = \ln \theta \Rightarrow a = -\frac{\ln \theta}{T}$$

Based on this substitution the following formula is obtained:

$$K(x) = \frac{1}{C} e^{\frac{\ln \theta}{T}|x|} \left(1 + \cos \frac{2x\pi}{T} \right)$$

$$C = \frac{4T \ln^2 \theta + 4T\pi^2}{-\ln^3 \theta - 4\pi^2 \ln^2 \theta}$$

On the Figure 1. the sample *FPK* is presented.

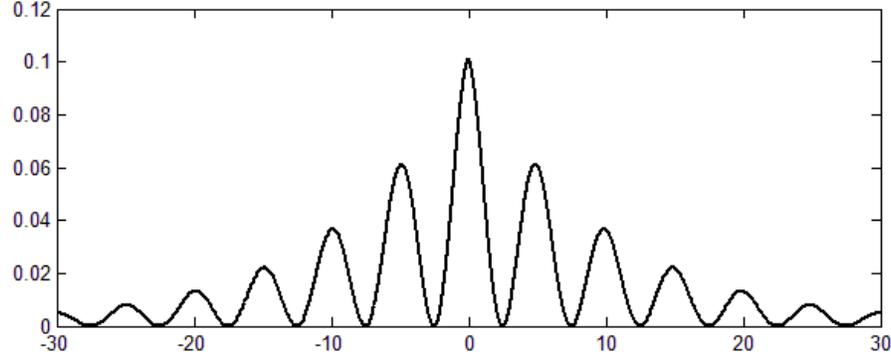


Figure 1. First Periodic Kernel generated with $T=5$, $\theta = 0.6$

The second kernel (*SPK*) has a following formula:

$$SPK(x) = \frac{1}{C} e^{-a|x|} \cos^n bx$$

where

$$C = 2 \int_0^{\infty} e^{-ax} \cos^n bx \, dx = 2[I_n]_0^{\infty}$$

and I_n is an integral:

$$I_n = \int e^{-ax} \cos^n bx \, dx$$

The final formula for the constant C calculated recurrently is following:

$$C = \left(-\frac{1}{a} - \sum_{i=1}^n \frac{a}{(a^2 + 4i^2) \prod_{k=0}^i \mu_k} \right) \prod_{i=1}^n \mu_i$$

with

$$\mu_0 = 1, \quad \mu_i = \frac{2i(2i-1)}{a^2 + 4i^2}$$

It is possible to calculate the value of the C in the analytical way when the software allows symbolic calculation. Experiments presented in this paper were performed in Matlab and the C was calculated in the symbolic way.

This kernel also may be defined with the period T and the attenuation θ :

$$K(x) = \frac{1}{C} e^{-\frac{\ln \theta}{2}|x|} \cos^n \frac{\pi x}{T}, \quad b(T) = \frac{\pi}{T}, \quad a(\theta) = -\frac{\ln \theta}{T}$$

The role of n parameter is to describe the „sharpness“ of the function in the local maxima. On the Figure 2. the sample *SPK* is given.

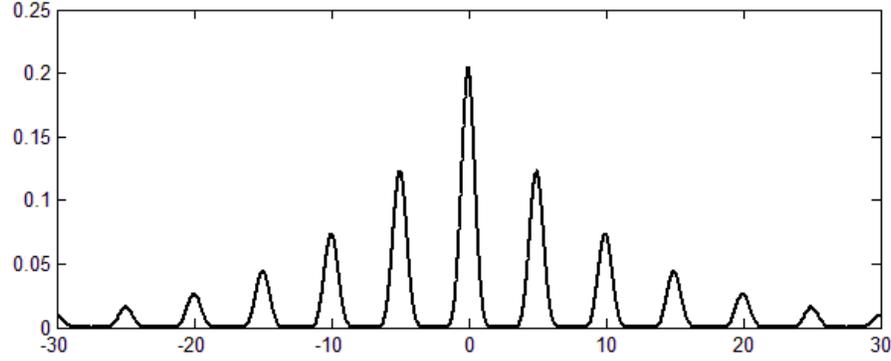


Figure 2. Second Periodic Kernel generated with $T=5$, $\theta = 0.6$ and $n = 10$

3. PERKE ALGORITHM

Periodic Kernel Estimator is a member of the group of semiparametric (two step) methods [6][7]. Methods from this group consist of the initial parametric step and the final nonparametric one. After the parametric step the residuals are calculated and the nonparametric part of the model tries to explain the only variation of residuals. The final model can consist of addition or multiplication of the basic results. In this aspect the semiparametric method is similar to the decomposition method.

The *PerKE* models the residual part of the time series with the following formula:

$$x(t) = \frac{\sum_{i=1}^k x_{t-i} K(t-i)}{\sum_{i=1}^k K(t-i)}$$

where k is the number of previous observation in the train sample of the time series.

It may be noticed that this equation is derived from the Nadaraya-Watson kernel estimator [8][9] but the smoothing parameter h was removed. This may cause the situation of oversmoothing the data. It is observed in two variants: the predicted values are overestimated (bigger than real values) or underestimated (smaller than real values). In order to avoid this situation the parameter called underestimation α is introduced. It is the fraction of the predicted and original value:

$$\alpha_i = \frac{\tilde{x}_i}{x_i}$$

The underestimation is trained in the following way: if p is an interesting prediction horizon the last p observations from the train set are considered as the test set and predict them on the basis of the rest of the train set. Then the vector of underestimations is defined as the vector of fractions of

predicted and real values. In the final prediction the values coming from the nonparametric step are divided by the corresponding α .

4. SELECTION OF KERNEL PARAMETERS

4.1. Discretisation of Periodic Kernels

In the experiments a simplified –a discretized – form of periodic kernels was used. Let assume that only the values of the kernel for the period multiple are interesting: $K(x)$ where $x = kT, k \in Z$. Then the formula for *FPK* simplifies to the following one:

$$K(kT) = \frac{2}{C} e^{|k| \ln \theta}$$

Discretisation of the *SPK* leads to the same formula. The only difference between two discretized kernels is the value of the C constant which can be tabularised before the experiments. It speeds up calculation because each constant C (for each demanded form of periodic kernel) was calculated once and was read in a constant time.

On the basis of the discretized form of periodic kernels and the kernel regression formula of residual part of the series, it might be claimed, that both types of periodic kernels give the same results.

4.2. The Error Evaluation

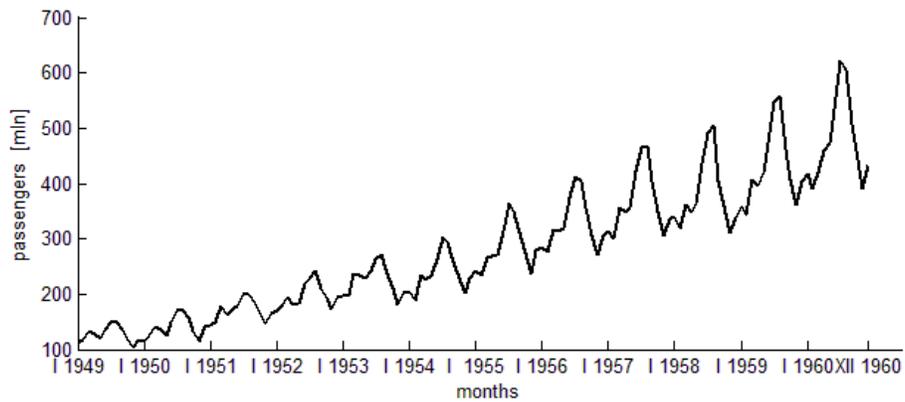
The error of prediction was measured with two different quality functions:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \tilde{y}_i|}{|y_i|} \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2}$$

Each of them describes a different kind of an error. The first one points the averaged absolute error and is more resistant when the test samples have values from very wide range. The second one measures the error in the unit of the analysed data so it can be more interpretable in some cases.

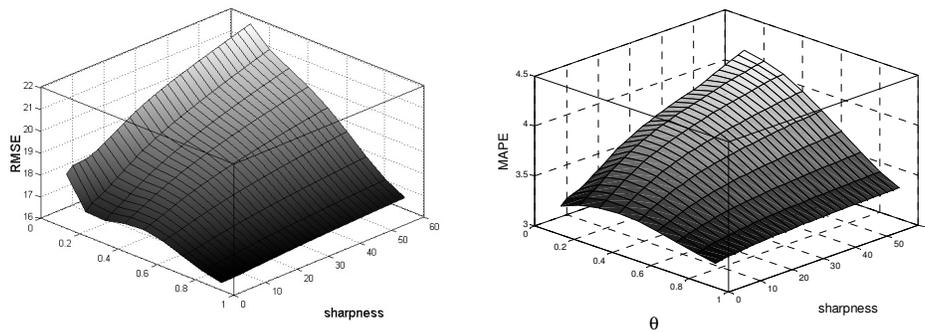
4.3. Setting the Parameters for *SPK*

Let's consider the very popular time series describing the number of passengers in America (*G* series from Box and Jenkins [1]). It contains 144 monthly values of number of passengers (in millions) between 01.1949 and 12.1960. Its natural period is 12. This time series is presented on the Figure 3.

Figure 3. *G* time series

For the purpose of the analysis of an influence of the *SPK* parameters on the prediction accuracy the following optimization step was performed. Instead of calculation of the *C* value for each prediction task, the array of *C* values for the predefined periodic kernel parameters was created. The attenuation was changing from $\theta = 0.1$ to $\theta = 0.9$ with the step 0.1. The sharpness was changing from $n = 2$ to $n = 60$ with the step 2.

The error of the prediction depending on the kernel parameters is shown on the Figure 4.

Figure 4. *G* time series prediction error (*MAPE* on the left and *RMSE* on the right) as the function of θ and sharpness

In general, it may be seen that the error of the prediction decreases when the θ increases. Additionally, it is observed that the influence of the sharpness is opposite. In other words the decrease of the sharpness implies the decrease of the error.

Because the period of this series is 12 (the number of months) periodic kernel parameters were established on the basis of prediction on 144 – 12 *G* series values (all data without the last 12 values). Both error measures were considered. The smaller time series were called train series.

Table 1 compares the errors on the train series and on the whole series. The best results (typed with bold font) for the train series were for $\theta = 0.9$ and sharpness = 2. Performing the grid experiment for the whole series the best results were for 0.9 and 2 (with *MAPE*) and for 0.9 and 4 (with *RMSE*) respectively. It can be seen, that on the basis of the *MAPE* results for train data the

best values of parameters (with the assumed grid steps) were found and with the *RMSE* results – almost the best.

Table 1. Comparison of best results and kernel parameters for train and whole time series.

Train series				Whole series			
θ	sharpness	<i>MAPE</i>	<i>RMSE</i>	θ	Sharpness	<i>MAPE</i>	<i>RMSE</i>
0.9	2	3.6877	17.8085	0.9	2	3.1989	16.1038
0.9	4	3.6938	17.8752	0.9	4	3.2084	16.0972

5. REAL DATA APPLICATION

Selection of periodic kernel parameters was applied for the real time series, describing the monthly production of heat in one of the heating plant in Poland. This series (denoted as *E*) contained 97 values. The series is presented on the Figure 5.

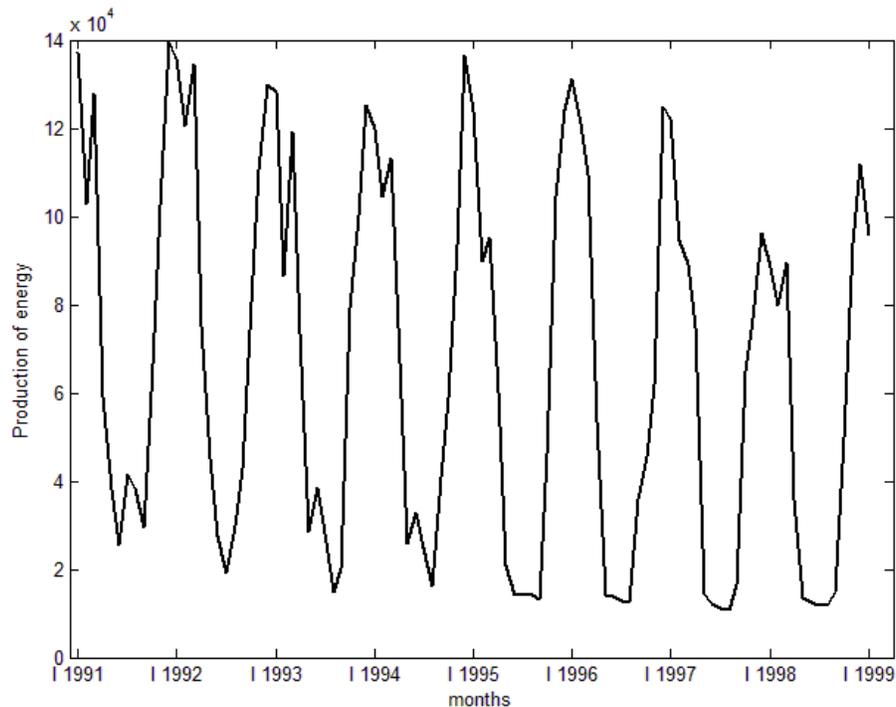


Figure 5. *E* time series prediction – monthly production of heat.

PerKE algorithm was performed in three ways: periodic kernels with arbitrarily set kernel parameters (two types of periodic kernels) and the *SPK* with the presented methodology of parameters setting. Additionally, two popular time series prediction methods were used as the reference points for kernel prediction results: SARIMA and decomposition method.

The results of all experiments are shown in the Table 2. (*G* series) and Table 3. (*E* series). In the first case periodic kernel parameters did not depend on the chosen measure. The final prediction quality is still better than the quality of other popular prediction methods.

Table 2. Comparison of the G time series prediction results.

method	MAPE	RMSE	annotations
SARIMA	4.80%	26.95	$(1,0,0)(2,0,0)^{12}$
decomp.	4.51%	26.60	exponential+multiplicative
FPK	3.20%	16.10	
SPK	3.72%	21.00	$T=12, \theta=0.4, n=60$
$SPK(MAPE/RMSE)$	3.20%	16.10	$T=12, \theta=0.9, n=2$

Table 3. Comparison of the E time series prediction results.

method	MAPE	RMSE	annotations
SARIMA	20.95%	10 115.91	$(1,0,0)(2,0,0)^{12}$
decomp.	22.10%	9 010.87	linear+additive
FPK	69.13%	19 855.28	
SPK	20.08%	8 638.12	$T=12, \theta=0.9, n=80$
$SPK(MAPE)$	19.13%	14 735.66	$T=12, \theta=0.9, n=2$
$SPK(RMSE)$	18.26%	15 861.22	$T=12, \theta=0.1, n=2$

In the second case (E series) the selected set of periodic kernel parameters depended on the quality measure. But for each of them the decrease of relative error is observed.

6. CONCLUSIONS AND FURTHER WORKS

In the paper the analysis of the periodic kernel parameters influence on the prediction error was analysed. Two types of periodic kernels were taken into consideration and the error of the prediction was measured with two different methods. On the basis of the analysis of the G time series and the E time series it may be said that the methodology of finding the periodic kernel parameters gives satisfying results.

Further works will focus on the application of *PerKE* and periodic kernels to time series with the different time interval between observations. It is expected that more differences between the two kernels will occur. It is also possible that the sharpness will have the bigger influence on the prediction error.

ACKNOWLEDGEMENTS

This work was supported by the European Union from the European Social Fund (grant agreement number: UDA-POKL.04.01.01-106/09).

REFERENCES

- [1] Box, George & Jenkins, Gwilym (1970) Time series analysis. Holden-Day, San Francisco.
- [2] Michalak, Marcin (2011) "Adaptive kernel approach to the time series prediction", Pattern Analysis and Application, Vol. 14, pp. 283-293.
- [3] Michalak, Marcin (2009) "Time series prediction using new adaptive kernel estimators", Advances in Intelligent and Soft Computing, Vol. 57, pp. 229-236.
- [4] Michalak, Marcin (2011) "Time series prediction with periodic kernels", Advances in Intelligent and Soft Computing, Vol. 95, pp. 137-146.
- [5] Scott, David (1992) Multivariate Density Estimation. Theory, Practice and Visualization, Wiley & Sons.
- [6] Abramson, Ian (1982) "Arbitrariness of the pilot estimator in adaptive kernel methods", Journal of Multivariate Analysis, Vol. 12, pp. 562-567.

- [7] Hjort, Nils & Glad, Ingrid (1995)“Nonparametric density estimation with a parametric start”, *Annals of Statistics*, Vol. 23, pp. 882-904.
- [8] Nadaraya, Elizbar (1964)“On estimating regression”,*Theory of Probability and Its Applications*,Vol. 9, pp.141-142.
- [9] Watson, Geoffrey (1964)“Smooth regression analysis”,*Sankhya - The Indian Journal of Statistics*,Vol. 26, pp. 359-372.

AUTHOR

Marcin Michalak was born in Poland in 1981. He received his M.Sc. Eng. in computer science from the Silesian University of Technology in 2005 and Ph.D. degree in 2009 from the same university. His scientific interests is in machine learning, data mining, rough sets and biclustering. He is an author and coauthor of over 40 scientific papers.



RESEARCH ON THE MOBILE ROBOTS INTELLIGENT PATH PLANNING BASED ON ANT COLONY ALGORITHM APPLICATION IN MANUFACTURING LOGISTICS

GUO Yue , SHEN Xuelian and ZHU Zhanfeng

Management Engineering Institute, Ningbo University of Technology, No.201
Fenghua Road, Ningbo. 315211 Zhejiang, China
guoyue@nbut.edu.cn shenxuelian@nbut.edu.cn
zhuzhanfeng@nbut.edu.cn

ABSTRACT

With the development of robotics and artificial intelligence field unceasingly thorough, path planning as an important field of robot calculation has been widespread concern. This paper analyzes the current development of robot and path planning algorithm and focuses on the advantages and disadvantages of the traditional intelligent path planning as well as the path planning. The problem of mobile robot path planning is studied by using ant colony algorithm, and it also provides some solving methods.

KEYWORDS

Manufacturing Logistics; Mobile robots; Path planning; Ant colony algorithm

1. INTRODUCTION

The research of mobile robot started from the late 1960s. The Stanford Institute successfully developed the autonomous mobile robot—Shakey robot in 1966, The robot has independent reasoning, planning, control and other functions in complex conditions with the application of artificial intelligence. At the end of the 1970's, the application of computer and sensor technology researches on mobile robot reach to a new high tide as a result of the development. The mid 1980's, a large number of world famous company started to develop mobile robot platform. The mobile robot is mainly used as the mobile robot experiment platform in university laboratories and research institutions, and promoting the multi-directional learning of the mobile robot. Since the 1990s, the symbol of environment information sensor and information processing technology development of high level, high adaptability of mobile robot control technology, programming technology under the real environment has emerged, and the higher level research of mobile robotics able to be conducted. In recent years, mobile robots are widely used in space exploration, ocean development, atomic energy, factory automation, construction, mining, agriculture, military, and service, etc. Research on mobile robot has become a hot research issue and the concern of the international robot.

Intelligent mobile robot is a set of integrated system of multiple functions which consists of environment perception, dynamic decision-making and planning, behavior controlling and executing. In recent years, mobile robot has wide application prospect in space exploration, ocean development, atomic energy, factory automation, construction, mining, agriculture, military, and service, etc. China started the research on the intelligent robots later than some developed countries, and there still existed a big gap within China and developed countries. In recent years, the research theory and method for robot have reached the international advanced level by the China Robotics Lab, and achieved a number of important scientific research achievements in robotics frontier exploration and demonstration application etc. Because the state and society paid much attention on the robot field, which including: all-weather 120Kg suspended wing UAV system, polar research snow mobile robot, Ling Lizard-anti-terrorism and anti-riot robot, robot nano operating system, etc.

Mobile robot's path programming technology is one of the core technology in the field of robot research, which study of the algorithms is advantageous to the improvement of robot planning to meet the needs of practical applications. The path programming is that, in the obstacle environment, according to a certain evaluation standard, finding collision free path from the initial state to the target state. The main issues include finding the optimal or approximate optimal one from the initial state to the target state collision through the free path and an algorithm to built reasonable model by using of the mobile robot environmental information. In the model which being able to cope with uncertain factors and path tracking errors in the environment, making the influence of external objects to the robot reduced to a minimum: how to use all the information known to guide the robot motion, resulting in a better decision [1] .

2. LITERATURE REVIEW

The traditional method of path programming is carried out simulation test based on graph. The general approach is based on the global path planning. At present domestic and foreign common methods include grid method, topology, visibility graph, Voronoi graph, method, the artificial potential field method, A* algorithm etc.

Grid method proposed by Howden in 1968 [2] , decomposing the robot planning space into a number of information network unit working space is divided into unit after the use of heuristic algorithm to search the safe path in the unit. [3] The search process always uses work space with four quadtree or octree. Consistency and standards grid makes simple adjacency relation in raster space. After giving each grid traffic factor, path planning problem turns into a problem of searching optimal path with two grid nodes in the grid network. Topological method is mainly divided space with topological feature subspace, and then look for topological path is the starting point to the target point based on the topology of the network, and finally find the path geometry by the topological path.[4] The basic idea of the method is to find paths in high dimensional space transformed into the problem of determining the connectivity of the problem of Low Dimensional Topology space.[5] The visibility graph method is a kind of configuration space method,[6] it mainly regards the robot as a particle processing, expanding the boundary of the corresponding outward obstacles in the work environment, and the boundary is formed with vertices of polygons, determining its vertices, including the robot starting point and the target point. These points connect, but each vertex cannot connect across each other, forming a visibility graph.

The Voronoi plot method is first discovered by the Russian mathematician Voronoi which can be applied to static random environment, which is to say in the process of robot running, the environment is static.[7] All the obstacles are motionless, but environment is uncertain before the robot starts its path planning, the size and the location parameter of the obstacles in the environment is changeable. This method using the path which may be far away from obstacles to

show walls arc, that results the path will increase from the initial node to the target node. Artificial potential field method is a method of local path planning, proposed by Khatib etc.[8] The basic idea of the method is regarding the motion of the robot in the environment as a virtual artificial force field motion.[9] Obstacles generate repulsive force on the robot, and attraction on the target. The joint force of attraction and repulsive force controls robot motion direction, which will determine the position of the robot[10].

In recent years, with the rapid development of in-depth research and the modern computing technology in mobile robot path programming, the traditional path programming is hard to meet its requirements and failure in meeting the need of actual environment changes. Therefore intelligent path algorithms have been studied and used in robot path programming widely. The artificial intelligent path programming algorithm improves the accuracy of robot obstacle avoidance path programming greatly and accelerates the programming speed, all these are to be met the needs of practical application. Intelligent path programming algorithm includes genetic algorithm[11][12], particle swarm algorithm[13], fuzzy logic[14][15], neural network[16][17], artificial immune algorithm[18] and hybrid algorithm[19][20][21]. The above algorithms have been made certain achievements for the robot protecting the obstacle in known or unknown circumstances.

Italian scholar Dorigo and Colorni proposed a heuristic optimization algorithm in 1991, which is biologically inspired.[22] It simulated and reference the behavior of ants in the real world to solve combinatorial optimization problems under distributed environment.[23] It also solves the problems of large cost when robot in complex environment contains a large number of irregular obstacles in the path programming[24].

Ant colony algorithm is produced to simulate the process of ants foraging. Ants release specials in the search of path when confronted with a no through road, they will randomly select one while releasing hormone information of path length. When the ants again encountered this intersection, optimal path on the pheromone concentration increase, while the other pheromone concentration is cutting with the passage of time.[25] At the same time, the ants can adapt to changes in the environment when obstacles emerge, they will find an optimal path to go. Ant colony algorithm has the features of group cooperation, positive feedback and distributed computing. Group cooperation is a cooperation for better optimization task. Although each artificial ant can build a solution, but the solution with high quality is always produced by ant colony cooperation. The feedback mode of the algorithm is used in the optimum solution which leaves more pheromone on a path, and more pheromone in turn attracts more ants. The positive feedback process guides the system towards the optimal solution of the evolution direction. Distribute Computing of Ant colony algorithm can calculate each artificial ant at the multiple points in the problem space, at the same time, it began to separate the structural problems of solutions. The result will not be affected only because of one artificial ant cannot successfully obtained the solution Distributed Computing makes the algorithm easy to be implemented.[26] These characteristics make the ant colony algorithm suitable for solving complex combinatorial optimization problems.

The path programming is a kind of combinatorial optimization problem, so the ant colony algorithm is suitable to solve the path planning problems.

3. QUESTIONS

3.1 Problems in Design

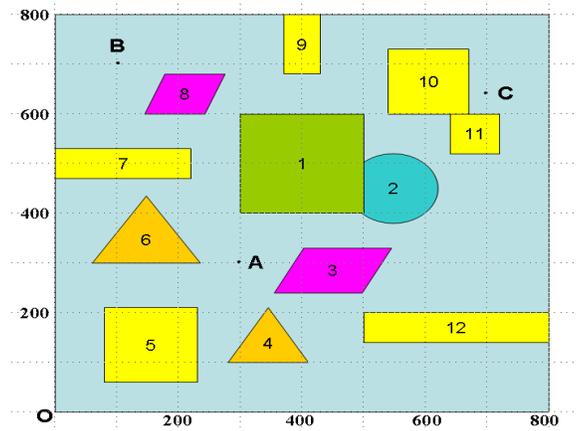
In a 800×800 planar scene graph, there is a robot at the origin of $O(0, 0)$, which can only activities in the planar scene range. The 12 different shapes of the regions are the obstacles that the robot cannot collide. The description of the mathematical are as shown in the following table:

Table 1. Planar scene graph description

Region No.	Obstacles' name	vertex coordinates at left corner	Other description of the characteristics
1	square	(300, 400)	length of 200
2	circular		center coordinates (550, 450) a radius of 70
3	parallelogram	(360, 240)	Base length140, the coordinates of the vertices on the left top(400, 330)
4	triangles	(280, 100)	top vertex coordinates (345, 210), the lower right vertex coordinates (410, 100)
5	square	(80, 60)	length 150
6	triangles	(60, 300)	top vertex coordinates (150, 435), the lower right vertex coordinates(235, 300)
7	rectangular	(0, 470)	length 220, width 60
8	parallelogram	(150, 600)	Base length 90, the coordinates of the vertices on the left top(180, 680)
9	rectangular	(370, 680)	length60, width120
10	square	(540, 600)	length130
11	square	(640, 520)	length80
12	rectangular	(500, 140)	length300, width60

3.2 Maintaining the Integrity of the Specifications

Specify a point outside the obstacles as the target for the robot to reach (target point and the distance to the obstacle are at least more than 10 units). Set the rule that walking path of the robot should be by lines and arcs. The arc parts are robot's turning path. A robot cannot turn by line. The turning path consists with straight path tangent to a circle, and can also be composed of two or more circular arc path, but the minimum radius of each circular arc path is 10 units in order not to collide with obstacles. It also requires the distance between the robot walking route and obstacle is no more than 10 units, or a collision will occur and the robot can not complete the straight walking. 4 point O in the scene graph(0, 0), A(300, 300), B(100, 700), C(700, 640).

Fig1. 800×800 planar scene graph

The maximum speed of the robot walking straight into 5 unit / sec. and maximum turning speed

is $v = v(\rho) = \frac{v_0}{1 + e^{10 - 0.1\rho^2}}$ where P is radius of turn . If the speed is higher than that, the robot will rollover, and is unable to walk.

Question: Start from $O(0, 0)$, which is the shortest path of $O \rightarrow A$, $O \rightarrow B$?

4. METHODOLOGY

In order to find the shortest path to the target from $(0,0)$ with certain rules walking around obstacles ,we can draw the envelope of robot walking hazardous area, just around the corner with a radius of 10 units a quarter arc, by the method of the rope then to find the shortest possible path (for example, seek the shortest path between O and A , can be connected to the section of rope between O and A , to the arc of the corner support taut, then the length of this wire is O to a shortest possible path (A),and then list shortest path possible paths to each target point with Brute-force method.

Designated $O(0,0)$ after the middle of a number of points around obstacles to reach the target point according to certain rules in the back of O , which allows us to consider not just obstacles inflection point, should be considered after the target point in the path at the turn of the problem.Simple line circle structure can not solve this problem,so we have adopted the form of a minimum turning radius at the inflection point and the target point on the way.We can also be appropriate to transform the inflection point of the turning radius, so that the robot can along straight line through the target point of the way, and then create optimization model to optimize these two programs, and ultimately obtained the shortest path.

The model assumes and symbol description are analyzed by the following assumptions:

- (1) Assuming the width of the robot itself is negligible. Thus, the movement of the robot can be regarded as a point moves.
- (2) Assuming the robot walk straight and turn at maximum speed.
- (3) Assuming that the obstacle is always subject to 12 different shapes of the area and the nature of the location, size, etc. has been the same.

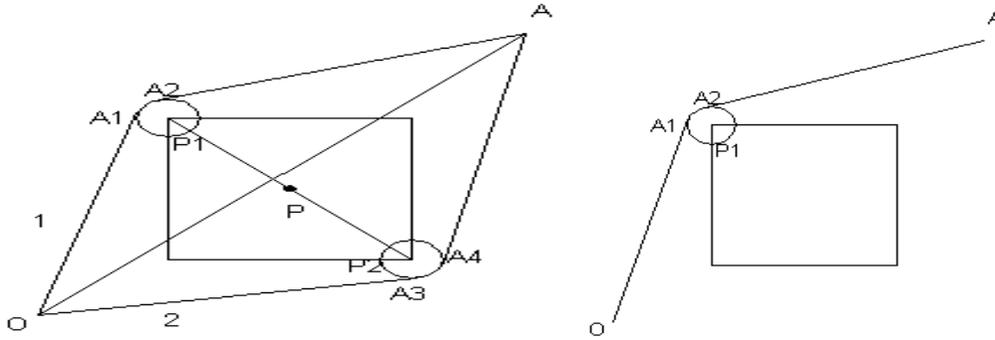
Table2 Symbol & Description

Symbol	Description
v_0	Maximum speed straight line when walking
ρ	Turning radius
S_i	The i-th sub-length of the shortest path from $O \rightarrow A$
L_i	The i-th sub-length of the shortest path from $O \rightarrow B$
L'_i	The i-th sub-length of the shortest path from $O \rightarrow C$
t_{\min}	The shortest time from $O \rightarrow A$
$\overline{B_i B_j}$	Arc length from $B_i \rightarrow B_j$
$ B_i B_j $	Length of the line from $B_i \rightarrow B_j$

5. FINDINGS AND INTERPRETATIONS

5.1 The Shortest Path From $O \rightarrow A$ Optimization Model

Two points within the plane of the shortest path based on the length of the line segment as the endpoint, but the connection of these two segments with obstacles intersect, so try to attempt to bypass the obstacle and its hazardous areas other path. Obstacle is a square, the center of this square is located in the lower part of the connection, so the robot to bypass the obstacle from the top of the obstacle path is the shortest path.

Figure 2. $O \rightarrow A$ path

Shown in Figure 2, the shortest path from $O \rightarrow A$ is constituted by straight line OA_1 and A_2A and a tangent arc $\overline{A_1A_2}$ wherein the cut point. Which A_1A_2 as the cutoff point. Arc $\overline{A_1A_2}$ thought the center of the circle $P_1(80,210)$, the radius is 10. Set cut-point coordinates $A_1(x_1, y_1)$ $A_2(x_2, y_2)$, these three sections of the path length can be calculated:

$$A_2A = s_3 = \sqrt{(x_2 - 300)^2 + (y_2 - 300)^2} \quad OA_1 = s_1 \quad (1)$$

$$OA_1 = s_1 = \sqrt{(x_1 - 0)^2 + (y_1 - 0)^2} \quad (2)$$

$$\angle A_1 A_2 = s_2 = 2 \arcsin \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{20} \cdot r \quad (3)$$

Here we have x_1, y_1, x_2, y_2 for the decision variables, the total length of the shortest path as the

$$s = \sum_{s_1}^3 s_i$$

objective function: Min

Constraints: radius $r = 10$, $OA_1^2 + A_1P^2 = OP^2$, $A_2A_1^2 + PA_1^2 = A_2P^2$, points A_1, A_2 on the arc. In summary, the structure optimization model as follows:

$$\text{Min } s = \sum_{s_1}^3 s_i = \sqrt{x_1^2 + y_1^2} + \sqrt{(x_1 - 300)^2 + (y_2 - 300)^2} + 20t \quad (4)$$

$$s.t. \begin{cases} x_1 \leq 80 \\ x_1 \geq 70 \\ y_1 \leq 220 \\ y_1 - \sqrt{100 - (x_1 - 80)^2} = 210 \\ x_1^2 + y_1^2 + 100 = 50500 \\ (x_1 - 300)^2 + (y_2 - 300)^2 + 100 = 56500 \\ \sin t = \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{20} \end{cases} \quad (5)$$

Solving the above model (see Appendix 1 lingo procedures), the results are as follows:

1) The OA length of the shortest path: $s_{\min} = 471.0372$

$$A_1 = (70.50596, 213.1406)$$

2) Two arc tangent point coordinates: $A_2 = (76.60640, 219.4066)$

Robot shortest path from point O to point A is reached can be expressed in the table3:

Table3. the shortest path from $O \rightarrow A$

No	Start	End	Types of segments	Length
1	(0, 0)	(70.50596,213.1406)	Straight line	224.4994
2	(70.50596,213.1406)	(70.6064,219.4066)	(80,210) as the center of the arc	9.1105
3	(70.6064,219.4066)	(300, 300)	Straight line	237.4273
Total length				471.0372

5.2 The Shortest Path from $O \rightarrow B$ Model

In this section, we will simplify the roadmap as an empowered network diagram, and use the ant colony algorithm to find the approximate route of the shortest path. The ant colony algorithm is a bionic algorithm derived from the nature of ants routing mode simulation. Ants in the process of movement will leave a substance called Pheromone on its path through the information transferred. Ants can perceive this substance in the course of the campaign as their movement direction. Therefore, a large number of ants showed an information feedback phenomenon: more ants walking on a path, choose the after the greater the probability.

We put on a plane at some point number, and the relationship of the distance between them simplified network chart. If the node can directly reach in a straight line rather than an obstacle, the weights of the edges between them weight of the straight-line distance, otherwise there is no edge between them. As follows:

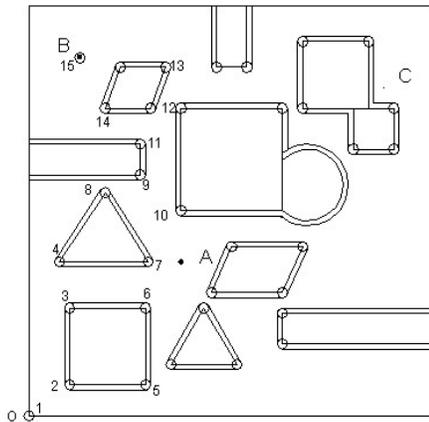


Figure4. Nodes Numbers Figure

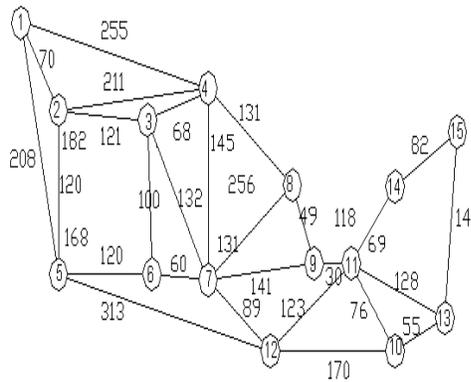


Figure 5. Empowering network chart

Using ant colony algorithm to select the shortest route from $O \rightarrow B$ from the simplified network chart.

1) Ant colony algorithm model

Value the point $1 \rightarrow 15$, 0-1 whether if it on the path, form 15 bit sequence 0,1, thereby calculating the distance of this path. The distance as a mapping of the pheromone variable, due to the requirements of the most short-circuit, so you can use the countdown or relative distance as the pheromone concentration. Then you get each ant transition probability. If transition probability is greater than the global transfer factor, then the global transfer; otherwise transfer must have step. So that you can step to the global optimal solution close.

2) Perform steps

- The first step to initialize N ants. In fact N road, and calculate the current position of ants.
- The second step initialization of operating parameters, start the iteration.
- The third step in the iterative complement the range of calculated transition probabilities, less than the global transition probability for small-scale search, or a wide range of search.
- The fourth step is to update the pheromone, records state, ready for the next iteration.
- The fifth step is to enter the third step
- The Sixth step output and programming.

3) Analysis of results

The initial state of 50 ants are disorderly distribution, optimized the final position to the polarization, so that we get the optimal solution.

Figure6 are average and optimal curve, from which you can know that the algorithm converges very fast, the effect is better. The shortest path is $1 \rightarrow 4 \rightarrow 8 \rightarrow 9 \rightarrow 11 \rightarrow 14 \rightarrow 15$. Chromosome: 100100011010011, running time is 0.3910.

Therefore, the shortest path from $O \rightarrow B$ as shown below:

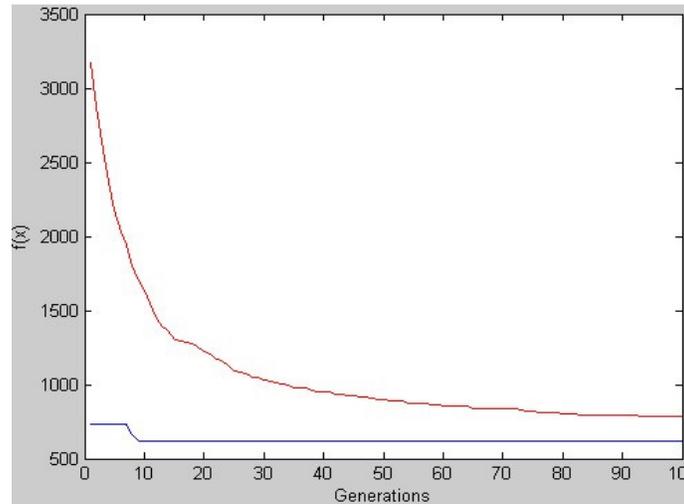
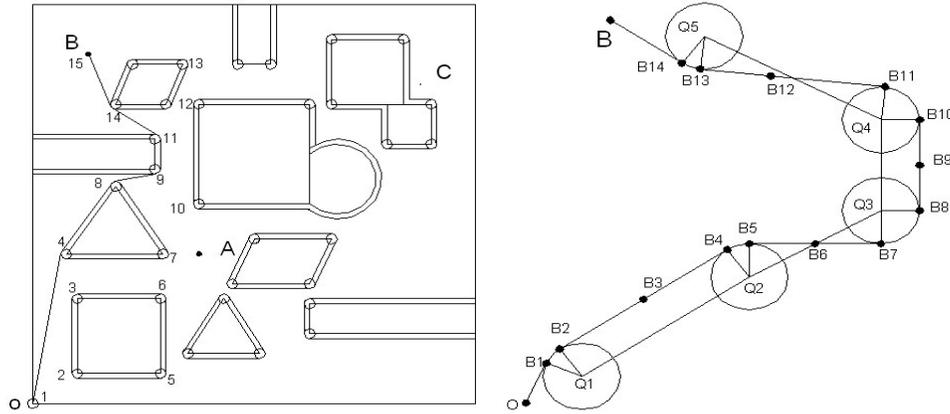


Figure 6. Pheromone concentration average value and the optimum value

5.3 The Segmented Path Length from $O \rightarrow B$ Optimization Model for Solving

In the previous section, we have determined to run the route of the shortest node $1 \rightarrow 4 \rightarrow 8 \rightarrow 9 \rightarrow 11 \rightarrow 14 \rightarrow 15$ in Figure 3, but this simplified diagram from only consider a straight line, without regard to the actual deployment of the arc length. Therefore, we put this route segment, making each piece only route to bypass an obstacle.

Based on the above analysis, from this route $O \rightarrow B$ is divided into five sections $L_1(O \rightarrow B_3)$, $L_2(B_3 \rightarrow B_6)$, $L_3(B_6 \rightarrow B_9)$, $L_4(B_9 \rightarrow B_{12})$, $L_5(B_{12} \rightarrow B)$ calculate their length, then the sum thus obtained the shortest path from $O \rightarrow B$.

Figure7. Seeking from $O \rightarrow B_3$ shortest path

1) Seeking from $O \rightarrow B_3$ shortest path

seek the shortest path $O \rightarrow A$ structure optimization model is similar to when seeking the shortest path $O \rightarrow B_3$, we will coordinate O , B_3 , Q_1 as the route start point (a, b) , end point (c, d) and the arc center (m, n) coordinates variable values.

Namely: $a = 0, b = 0; c = 100, d = 378; m = 60, n = 300$.

The coordinates of the cut-off point $B_1(x_1, y_1)$, $B_2(x_2, y_2)$ as a decision variable, $O \rightarrow B_3$ as the objective function of the length of the shortest structure optimization model as follows:

$$\text{Min } s = \sqrt{(x_1 - m)^2 + (y_1 - n)^2} + \sqrt{(x_1 - c)^2 + (y_2 - d)^2} + 20t \quad (6)$$

$$s.t. \left\{ \begin{array}{l} x_1 \leq m + 10 \\ x_1 \geq m \\ y_1 \leq n + 10 \\ y_1 - \sqrt{100 - (x_1 - m)^2} = n \\ (x_1 - a)^2 + (y_1 - b)^2 + 100 = m^2 + n^2 \\ (x_1 - c)^2 + (y_2 - d)^2 + 100 = (m - c)^2 + (n - d)^2 \\ \sin t = \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{20} \end{array} \right. \quad (7)$$

Solving the above model, the results are as follows:

The length OB_3 of the shortest path: $s_{\min} = 397.0986$

Two arc tangent point coordinates:

$$B_1 = (50.1353, 301.6396)$$

$$B_2 = (51.6795, 305.547)$$

2) Seeking $B_3 \rightarrow B_6$ shortest path model

The optimization model with the same route as the previous paragraph, the coordinates of B_3 , B_6 , Q_2 as the beginning (a,b) of the route, the end (c,d) and the arc center (m,n) coordinates of the variable value. Namely:

$$a = 100, b = 378; c = 185, d = 452.5; m = 150, n = 435.$$

Using the lingo program solving the optimization model, the length of the shortest path. Similarly, we can calculate the shortest path of the other sub-routes. In summary, we have come to the shortest path.

Table4. The result of the length of the shortest path

Segmented	Start	End	Types of segments	Length
1	(0,0)	(50.1353,301.6396)	Straight line	305.7777
2	(50.1353,301.6396)	(51.6795,305.547)	(60, 300) as the center of the arc	5.88
3	(51.6795,305.547)	(141.6795,440.547)	Straight line	162.2498
4	(141.6795,440.547)	(147.9621,444.79.0)	(150, 435) as the center of the arc	7.7756
5	(147.9621,444.79.0)	(222.0379,460.2099)	Straight line	75.6637
6	(222.0379,460.2099)	(230,470)	(220, 470) as the center of the arc	13.6557
7	(230,470)	(230,530)	Straight line	60
8	(230,530)	(225.5026,538.3538)	(220, 530) as the center of the arc	9.8883
9	(225.5026,538.3538)	(144.5033,591.6462)	Straight line	96.9536
10	(144.5033,591.6462)	(140.6892,596.3523)	(150, 600) as the center of the arc	6.1545
11	(140.6892,596.3523)	(100,700)	Straight line	110.377
Total length				854.3759

6. CONCLUSIONS

With the continual development of robotic research in the field of artificial intelligence, the use of ant colony algorithm effectively solves the problem of robot path planning in the practical work of calculation. Our studies show that in a certain range, the optimized model for ant colony algorithm can be used to calculate and design the shortest path when a robot moves from a

starting point beyond some obstacles and reaches the specified target points opposite the obstacles without any collision. Nevertheless, further study is necessary in that some limitations still exist in mobile robot path planning via ant colony algorithm, e.g. the model for the shortest path planning remains to be optimized, and whether there are other algorithm solutions to mobile robot path planning etc.

ACKNOWLEDGMENT

We would like to thank the referees very much for their valuable comments and suggestions, there also got some help from Miss Lu Yan and Ms. Xu Si.

REFERENCES

- [1] TAN Min, WANG Shuo.(2013) Research Progress on Robotics. ACTA AUTOMATICA SINICA, 39(7):pp.963-972.
- [2] Yahja A., Singh S., Stentz A.(2000) An Efficient Online Path Planner For Outdoor Mobile Robots. Robotics And Autonomous Systems, 32(2):pp.129-143.
- [3] LU Qing.(2007) Research of Path Planning for Car-Like Robot Based on Grid Method. Computer and Information Technology, 15(6):pp.24-27.
- [4] Oommen B., Iyengar S., Rao N.,Kashyap R.(1987) Robot Navigation In Unknown Terrains Using Learned Visibility Graphs. IEEE Journal of Robotics and Automation, 3(6):pp.672-681.
- [5] LI Shan-shou, FANG Qian-sheng, XIAO Ben-xian,QI Dong-liu.(2008) Environment Modeling in Global Path Planning Based on Modified Visibility Graph. Journal of East China Jiaotong University, 25(6):pp.73-77.
- [6] CAI Xiao-hui.(2007)The Path Planning of Mobile Robots Based on Intelligent Algorithms. Zhejiang University Master Degree Thesis, 5:pp.6-12.
- [7] Takahashi O. and R. Scilling.(1989) Motion Planning in a Plane using Generalised Voronoi Diagrams. IEEE Transactions on Robotics and Automation, 5,(2):pp.169-174.
- [8] JIN Lei-ze, DU Zhen-jun, JIA Kai.(2007) Simulation study on mobile robot path planning based on potential field. Computer Engineering and Applications, 43(24):pp.226- 229.
- [9] Khatib O.(1985) Real Time Obstacle Avoidance For Manipulators And Mobile Robots[J].The International Journal of Robotics Research, 5(2):pp.500-505.
- [10] Koren Y, And Borenstein J.(1991) Potential field methods and their inherent limitations for mobile robot navigation[C]. Proceedings of the IEEE Conference on Robotics and Automation, Sacramento, California, , April 7-12: pp.1398- 1404.
- [11] AI-Taharwa I.,Sheta A.,AI-Weshan M.(2008) A Mobile Robot Path Planning Using Genetic Algorithm In Static Environment. Journal of Computer Sciences, 4(4):pp.341-344.
- [12] YANG Lin-quan, LUO Zhong-wen, TANG Zhong-hua, LV Wei-xian. (2008) Path Planning Algorithm For Mobile Robot Obstacle Avoidance Adopting Bezier Curve Base on Genetic Algorithm. 2008 Chinese Control and Decision Conference:pp.3286-3289.
- [13] Gondy Leroy, Ann M. Lally and Hsin Chun Chen.(2003) The use of dynamic contexts to improve casual Internet searching. ACM Transactions on Information System, 21(3):pp.229-253.
- [14] St.Preitl, R.E.Precup, J.Fodor, B.Bede.(2006) Feedback Tuning In Fuzzy Control System.Theory and Applications, 3(3):pp.81-96.
- [15] P.Vadakkepat, O.C.Miin, X.Peng, T.H.Lee.(2004) Fuzzy Behavior-based Control of Mobile Robots[J]. IEEE Transactions on Fuzzy Systems, 12(4):pp.559-564.
- [16] CHEN Huahua, DU Xin, GU Weikang.(2004) Neural Network and Avoidance Genetic Algorithm Based Dynamic Obstacle and Path Planning for A Robot. Journal of Translucation Technology, (4):pp.551-555.
- [17] NI Bin, CHEN Xiong, LU Gongyu.(2006) A Neural Networks Algorithm for Robot Path Planning in Unknown Environment. Computer Engineering and Applications, (11):pp.73-76+109.
- [18] XU Xin-ying, XIE Jun, XIE Ke-ming.(2008) Path Planning of Mobile Robot Based on Artificial Immune Potential Field Algorithm.Journal of Beijing University of Technology, 34(10):pp.1116-1120.

- [19] CHEN Xi, TAN Guan-zheng, JIANG Bin.(2008) Real-time optimal path planning for mobile robots based on immune genetic algorithm. Journal of Central South University (Science and Technology), 39(3):pp.577-583
- [20] DING Wei.(2007) Path Planning Based On Immune Evolution And Chaotic Mutation For Mobile Robot. Master Degree in Engineering Dissertation, Harbin University of Science and Technology, March:pp.19-24.
- [21] CUI Shi-gang, GONG Jin-feng, PENG Shang-xian, WANG Jun-song.(2004) Hybrid intelligent algorithm based 3-D of robot path planning. Manufacturing Automation, (2):pp.49-51.
- [22] GUO Yu, LI Shi-yong.(2009) Path Planning for Robot Based on Improved Ant Colony Algorithm. Computer Measurement and Control, 17(1):pp.187-190.
- [23] ZHU Qing-bao.(2005) Ant Colony Optimization Parallel Algorithm And Based On Coarse-grain Model.Computer Engineering, 31(1):pp.157-159.
- [24] FAN Lu-qiao YAO Xi-fan BIAN Qing-qing JIANG Liang-zhong.(2008) Ant Colony Algorithm and the Application on Path Planning For Mobile Robot. Robotics Technology, 23:pp.257-259+261.
- [25] XIE Min,GAO Li-xin.(2008) Ant algorithm applied in optimal path planning.Computer Engineering and Applications. Computer Engineering and Applications, 44(8):pp.245-248.
- [26] ZHU Qing-bao.(2005) Ants Predictive Algorithm For Path Planning Of Robot In A Complex Dynamic Environment. Chinese Journal of Computers, 28(11):pp.1898-1906.

Appendix: Matlab programming of Ant Colony Algorithm

```

function shortroad_ant_main
% Ant main program
clear all;close all;clc;%clear all
tic;%time start
Ant=50;Ger=100;%
    Running parameter
    initialization
power=[0      70      1000    276      1000    1000    1000    1000    313
      208      1000    1000    1000      1000    1000    89      1000    1000
      1000    1000    1000    1000      170     123     0      1000
      1000    1000    1000    1000      1000    1000
70      0      141      211      120      182      1000    1000    1000    1000    1000
      1000    1000    1000    1000      1000    1000    1000    1000    1000
      1000    1000    1000    1000      55     128     1000    0
      1000
1000    141      0      68      168      1000    1000    1000    1000    1000
      100      132     1000     500      1000    1000    555     118
      1000    1000    1000    1000      1000    1000    69      1000    1000
      1000
276     211      68      0      1000     1000    1000    1000    1000    1000
      1000    145     131     1000      1000    1000    1000    1000    1000
      1000    1000    1000    1000      1000    1000    1000    141
      1000
208     120      168      1000     0      1000    1000    1000    1000
      120     1000    1000    1000      1000    1000    313     1000
      1000    1000
1000    182      100      1000     120      1000    1000    1000    1000
      0      60      1000    1000      1000    1000
      1000    1000
1000    1000     132     145     1000
      60      0      131     141
      1000    1000     89     1000
      1000    1000
1000    1000     500     131     1000
      1000    131     0      49
      1000    1000    1000    1000
      555     1000
1000    1000     1000    1000     1000
      1000    141     49     0
[PM PN]=size(power);
% Initialization Ant place
v=init_population(Ant,PN);
v(:,1)=1;v(:,PN)=1;% The beginning
and end points in the path
% The distance when the information
factors concentration
fit=short_road_fun(v,power);
% Distance as small as
possible, so and information
factors concentration
corresponding
T0 = max(fit)-fit;
% Draw the picture

```



```

for j=1:n-1
    fit(i)=fit(i)+power(I(j),I(j+1));%
        Find distance of the path
end
end
end
END
%%
%Function init_population
function v=init_population(n1,s1)
v=round(rand(n1,s1));% Initializes all
    the Ants

```

AUTHORS

GUO Yue

PhD in Management Science and Engineering, MAIB, BBA Professor, Dean of Department of Enterprising Management ,Ningbo University of Technology In 2010 Dr. Guo award Chinese Provincial New Century Hundred-Thousand-Ten thousand Engineering Talent, in 2011 he also got Chinese Provincial Backbone Middle-Young Aged Teachers. Dr. Guo has been published over 50 papers and 8 books in his academic field. In management science and engineering field he also do some consulting services to solve the actual problems for local companies and enterprises.



SHEN Xuelian

PhD graduated from Southwest University of Finance and Economics, as a lecturer teaching in economics field courses at Ningbo University of Technology. He got his master degree from Swiss University.



ZHU Zhanfeng

Postdoctoral fellow, PhD in Management Science and Engineering, MBA, BA Professor, Headof Department of Economics & Management, Ningbo University of Technology He was visiting scholar of University of magdeburg in Germany; and got advanced logistics management division, registered senior consultant and Doctoral tutor. Also presently for ningbo college of engineering management, and director of the institute of the ningbo college of engineering development of small and medium-sized enterprises. At the same time, he was also the ministry of education of institutions of higher learning MBA class teaching steering committee of logistics and electronic commerce, deputy director of the national demonstration on repository construction project logistics management specialty teaching work committee.



INTENTIONAL BLANK

DICTIONARY-BASED CONCEPT MINING: AN APPLICATION FOR TURKISH

Cem Rıfki Aydın¹, Ali Erkan¹, Tunga Güngör¹ and Hidayet Takçı²

¹Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
cemrifkiaydin@gmail.com, alierkan@gmail.com, gungort@boun.edu.tr

²Department of Computer Engineering, Cumhuriyet University, Sivas, Turkey
htakci@gmail.com

ABSTRACT

In this study, a dictionary-based method is used to extract expressive concepts from documents. So far, there have been many studies concerning concept mining in English, but this area of study for Turkish, an agglutinative language, is still immature. We used dictionary instead of WordNet, a lexical database grouping words into synsets that is widely used for concept extraction. The dictionaries are rarely used in the domain of concept mining, but taking into account that dictionary entries have synonyms, hypernyms, hyponyms and other relationships in their meaning texts, the success rate has been high for determining concepts. This concept extraction method is implemented on documents, that are collected from different corpora.

KEYWORDS

Concept mining, Morphological analysis, Morphological disambiguation

1. INTRODUCTION

Concept can be thought of as a general idea or something specific conceived through the mind. A human can easily have a general opinion concerning something which he/she is reading, watching or to which she is listening. But as for computers, since they lack the functionality of human brain that can easily classify many patterns and have conceptual ideas on an object, they have to process many statistical or machine learning methods.

Concept mining is a field of study where text, visual or audio materials are processed and general concepts are extracted. Amongst its applications, the most commonly used approach for extracting concepts is concerned with text materials. In this paper we are interested in extracting concepts from textual documents. Concept mining is a hard and very beneficial operation. In many areas it can be used in an automatic, efficient and computerized manner. For example it can ease categorizing the judiciary classes, such as adult court, appellate court and many others [1]. Also in medical field, it can help both the patients and doctors in many different ways [2, 3]. Similarly, banks and other financial institutions may use concept mining to track the profiles of some creditworthy customers.

During last decades, through the enhancement in natural language processing, many artificial intelligence methods have been developed and used for extracting concepts, such as Latent Dirichlet Allocation (LDA) that aims to determine topics of a textual material. Besides those

methods, the most widely used lexical database for this field of study is WordNet. In WordNet the relationships between the words are stored in groups called synsets. These sets include hypernymy, hyponymy, synonymy relationships which are the most important and commonly used ones amongst many others. Hypernymy of a word can give us a general meaning of that word, hereby this relational property is the most widely used one for concept mining. This database is widely used for English concept extraction, but as for languages that are not as widely spoken as English, WordNet is not properly built and many word relationships (such as hypernymy) lack. That is why in these languages, the use of this lexical database would not yield as successful and meaningful results as it does in English.

In this paper, we used a novel approach for extracting concepts in Turkish through the use of a dictionary. Dictionaries have word entries, meaning texts, lexical class (noun, adjective, etc.) and some other properties. The meaning text of a word does, in fact, include its many relationally-relevant words, such as hypernyms or synonyms. So it is useful to benefit from the meaning texts in this dictionary to extract concepts. But these dictionary data are in an unstructured form, so they are pre-processed in this work, then concept mining method is run. The dictionary that is used is TDK (Türk Dil Kurumu - Turkish Language Association) Dictionary, which is the official Turkish dictionary that is the most respectful, accurate and most widely used one. We used the *bag-of-words* model for this study and we developed an algorithm which takes the frequencies of words in the document into account, since concepts do generally have to do with the most frequent words and the words related to those words present in the textual material. The success rate we observed is high compared with the other works done in concept mining for Turkish.

The remainder of the paper is as follows. In Section 2 related work concerning this field of study is examined. Section 3 elaborately examines the method developed in this work. Section 4 gives the evaluation results and the success rates achieved with the proposed method. Finally, the conclusion and future work are given in Section 5.

2. RELATED WORK

Numerous comprehensive studies are done in widely used languages in the domain of concept mining. In studies carried out in this domain, generally machine learning algorithms are used. Also WordNet, a lexical database, is commonly used. As for Turkish, the concept mining domain is still immature and also in this language some machine learning methods are used to extract concepts.

In one study, it is proposed that using WordNet's synset relations and then implementing clustering process may help obtaining meaningful clusters to be used for concept extraction [4]. The synset relationships are used as features for clustering documents and then these are used for succeeding clustering. But in this study word disambiguation has not been performed and it is observed that using synset words has decreased the clustering success.

There have been also developed some toolkits for concept mining, one of which is ConceptNet [5]. In accordance with this toolkit, similar to the WordNet synset relationships, spatial, physical, social, temporal and psychological aspects of everyday life are taken into account. In this toolkit, concepts are extracted by processing these relationships and a data structure is built. For instance a hierarchical relationship is named "IsA" and in a document the counterpart relational word of the word "apple" may be "fruit" in accordance with this relationship. If there are many common words that are extracted through relationships in a document, by taking also their frequencies into accounts, successful results may be achieved. ConceptNet is much richer than WordNet in terms of its relational structure.

As for Turkish, there have been a few studies concerning this study. It was proposed that meaningful concepts could be extracted without the use of a dictionary and with the clustering method processed on documents in corpora [6]. Here clusters would have initially been assigned concepts and the documents to which clusters are assigned, would have those concepts.

In an another study, a method is developed for extracting concepts from web pages [7]. The frequencies of words in the document are taken into account and words are assigned different scores according to their html tags. Words between specific tags such as "" and "<head>" are assigned higher scores. Then the words exceeding specified thresholds are determined as concepts. Success rates are reported to be high.

A study is done concerning extracting concepts from constructing digital library, documents in this library are categorized through clustering in accordance with those concepts [8]. An equation is created that multiplies term frequency (tf), inverse document frequency (idf), diameter, length, position of the first occurrence, and distribution deviation values of the keywords. Whichever words give the highest scores in accordance with this equation, they are selected as probable concepts. A higher success rate is achieved as compared with methods that take only tf and idf into account used for extracting concepts.

There have been used mostly lexical, relational databases and clustering methods for Concept Mining, besides those also Latent Dirichlet Allocation [9] and some other artificial intelligence methods have been used for extracting topics and concepts. Our method uses a statistical method that makes use of a dictionary, which has not been used in Turkish so far.

3. METHODOLOGY

In the concept mining domain, several dictionaries and lexical databases such as WordNet are used. Structural and semantic relationships between words can give us general idea (concept) about the words. In WordNet, especially hypernymy relation [10] is preferred in concept extraction, since it is the most relevant relation to a generalization of a word. There are also other relationships in WordNet such as meronymy and synonymy, those relationships give us other semantic relevances that may be useful in determining the concept of a word. For example, the synonym of the word "attorney" is "lawyer" and if a document from which concepts are to be extracted has many occurrences of the word "attorney", one probable concept may be "lawyer".

So far, the use of WordNet in Concept Mining has been extremely dominant. Taking it into account, in this work we used a novel approach, that is we used dictionary to help extract concepts from documents.

3.1 Dictionary Structure

The dictionary we used in this work includes the words and the meaning text of those words, where the words in the meaning carry specific relationships with each other. Figure 1 shows the XML structure of a dictionary item.

```

<entry>
  <name> jaguar </name>
  <affix>undefined</affix>
  <lex_class>isim, zooloji </lex_class>
  <stress>undefined</stress>
  <pronunciation> Fransızca jaguar </pronunciation>
  <origin> Fransızca</origin>
  - <meaning>
    <meaning_class>undefined</meaning_class>
    <meaning_text> Kedigillerden, Orta ve Güney
    Amerika'da yaşayan, postu iri benekli memeli
    türü (Felis onca).</meaning_text>
    - <quotation>
      <author>undefined</author>
      <quotation_text>undefined</quotation_text>
    </quotation>
  </meaning>
  <atasozu_deyim_bilesik>undefined</atasozu_deyim_bilesik>
  <birlesik_sozler>undefined</birlesik_sozler>
</entry>

```

Figure 1. Structure of a dictionary entry, "jaguar", in XML format

In Figure 1, the tag "<atasozu_deyim_bilesik>" stands for "proverb, idiom, compound" and the tag "<birlesik_sozler>" stands for "compound phrases" in Turkish. In this work we took only "name", "lex_class" and "meaning" tag elements into account. We processed only the words, "lex_class" of which are nouns, also analogies are made based on the meaning text nouns. There may be several meanings of a word, we took all of them into account and selected only one of them after disambiguation. The main relationships that would be encountered in dictionary item meaning texts can be summarized as follows:

- Synonymy: It is a relation that two words have equivalent meanings. It is a symmetrical relationship. (For example the words "intelligent" and "smart" have this relationship.)
- Meronymy: It is a relation that one of the words is a constituent of the other word. It is not a symmetrical relationship. (For example the words "hand" and "finger" have this relationship.)
- Location: It is a relation that shows the location of a word with respect to the other word. (For example the words "kitchen" and "house" have this relationship.)
- Usability: It is a relation that one word is used for an aim. (For example "toothbrush" is used for "brushing teeth".)
- Effect: It is a relation that one action leads to a result. (For example taking medication leads to a healthy state.)
- Hypernymy: It is a relation that one word is a general concept of an another word. (For example the word "animal" is the hypernym of the word "cat".)
- Hyponymy: It is a relation that one word is a more specific concept of an another word. (For example the word "school" is hyponym of the word "building".)
- Subevent: It is a relation that one action has a sub-action. (For example waking up in the morning would make one yawn.)
- Prerequisite relation: It is a relation that one action is a prerequisite condition for another one. (For example waking up in the morning is a prerequisite condition for hitting the road for job.)
- Antonymy: It is a relation that one word is the opposite concept of an another word. (For example the word "happy" is the antonym of the word "sad".)

This dictionary model may be used in many forms and has advantages as compared with the use of specific WordNet synset relations. For example, if we want to relate words to one another and

try making analogies between those words, the use of dictionary would be very helpful. Through implementation of clustering, many meaningful clusters can be built, because the analogies between the properties of words (such as the words 'finger' and 'hand' have meronymy relation and those would be in same cluster) may connect them in a semantic relationship. But instead of clustering process, we followed a simple statistical method, which would yield successful results.

3.2 Dictionary Preprocessing

In WordNet, relations are held in specific synsets, that is they have a structure based on word to word or word to noun-phrase relationships. So no parsing or disambiguation is needed for WordNet since words are in their root forms. However in basic dictionaries, there are no specific data structures that separately hold different synset relations. For example the meaning text of the word "cat" is as follows:

"A small carnivorous mammal (Felis catus or F. domesticus) domesticated since early times as a catcher of rats and mice and as a pet and existing in several distinctive breeds and varieties."

In the above example, the noun "rats" must be handled as the word "rat", whereas "mice" should be normalized as "mouse". The words should go through tokenization, stemming and normalization processes and then be taken into account. English is a language that has no much complexity concerning stemming, but as for Turkish, it is an agglutinative language and the parsing and disambiguation processes would quite matter. We used the parser and disambiguator tools, that are BoMorP and BoDis, developed for Turkish by Hasim Sak, Boğaziçi University [11, 12]. BoMorP implements the parsing operation and analyses the inflectional and derivational morphemes of a word and proposes part-of-speech (POS) tags. BoDis disambiguates amongst those POS tags and returns the one having the highest score according to an averaged perceptron-based algorithm.

A concept may be concerning a general idea of an abstract or concrete object, so the most probable concepts are generally nouns. So in this study we assumed that nouns represent concepts more than do any other word categories, so we took only nouns as concepts in the documents into account.

3.3 Document Analysis using Dictionary

A dictionary item may have many meanings and we have to determine which meaning is used in the context of a document. For example the word "cat" has many meanings and if the document contains this word we have to extract which meaning of this word is used in the document we are handling. For this, we looked up in the dictionary meaning text nouns as well as the nouns in the context of the word in the document and took the one yielding the highest similarity measure that has most common words. This formula is as follows:

$$\operatorname{argmax}_m \operatorname{Similarity}(m, c_w) = \operatorname{CommonCount}(m, c_w)$$

In this formula, m denotes a specific meaning of a word in a dictionary, whereas C_w denotes the context of a word in a document. In this work, we used a context size of 30 words, such that 15 words that are on the left and 15 words that are on the right of the candidate word are taken into account. If amongst those words, many occurrences are encountered also in a specific meaning of the word to be disambiguated, that meaning would be selected as the true meaning. We also normalized the score (CommonCount) by dividing it by the number of nouns in the dictionary meaning text to get more accurate results.

3.4 Mapping Concepts

3.4.1 Simple Frequency Algorithm

A concept that can be extracted from a document has generally to do with frequency of the specific word encountered in that document. For example, if we encounter a document that abounds with the word "football", we may be inclined to think that the concept of this document would be concerning "sport". Taking frequency into account, we developed a statistical method, that extracts concepts favouring the words that are more frequent.

We first take all the nouns in the document(s) and label them as pre-concepts. Here we eliminate other types of words, such as adjectives and verbs. Then we start building a matrix. This matrix has rows representing the nouns encountered in the document and columns representing the nouns encountered in the meaning text sentences of those row words. But we also have to take into account that the rows on the words are also added as column items. For example the word 'football' may be very frequent in the document, so this word should be regarded as a probable concept as well.

The cells in the matrix are filled as follows: After we built the matrix, we fill the cells by 1 or 0 depending on whether the column word appears in the row word's meaning text or not. Then we implement frequency operation: We multiply all the cell values in the matrix by the corresponding row word frequency. For instance, if the word "football" is encountered 10 times in a document and its meaning text nouns in the dictionary are "sport", and "team", then those columns' ("sport", "team" and "football") values in the correspondent row "football" would be updated as 10, whereas the other columns would be updated as 0.

Then we summed up the column values in the matrix and took the column word that yields the highest summation as the probable concept. This is meaningful since the concept may or may not be present in the document, so the use of dictionary would be beneficial. An example showing the mapping of terms into concepts is shown in Figure 2.

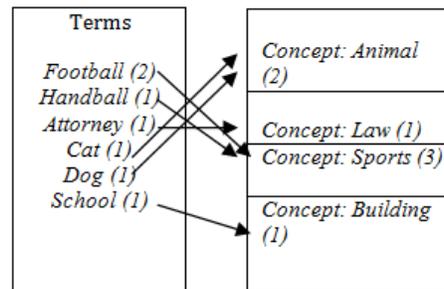


Figure 2. An example showing the mapping of words in a document into probable concepts.

In the above example, the column to the left is a representative of words encountered in the document, whereas the column to the right includes the representative nouns encountered in their meaning texts. Due to that the words "football" and "handball" are frequent in the above example and the word "sports" is present in their meaning texts, its score would be three and this word would be assigned as the probable concept of the document or corpus.

As mentioned above, we benefit from dictionary to extract concepts from documents, but instead of using just meaning text nouns for this concept extraction process, we also built a hierarchical data structure that contains 2, 3 and 4 levels. In accordance with this structure, the main word is atop the hierarchy, then the meaning text nouns of this word is in the lower level, whereas the

respective meaning text nouns of these meaning text nouns are in the lower levels. An example of this data structure with 3-levels is depicted in Figure 3.

This hierarchical structure may have some specific features, for example each word in different levels may be assigned a different coefficient and we may take this coefficient factor into account when building up the matrix. If we construct 3-level hierarchies built through the dictionary, we may assign high values for the top levels and low values for lower levels. This is the case because the semantic relationship between the main word and the lower level nouns weakens while going down through the hierarchy structure. We multiplied the top-level words in the matrix by 1, the second-level words by 0.5 and the lowest-level words by 0.25. We used this geometric approach since the meaning text nouns' frequencies increase geometrically from one level to the below one. But we noticed that 3-level structure gives a very low success rate, so we preferred 2-level structure with no coefficients yielding high accuracy values. Also 4-level structure gave even worse results than did 3-level one, so using a 2-level structure was the best choice.

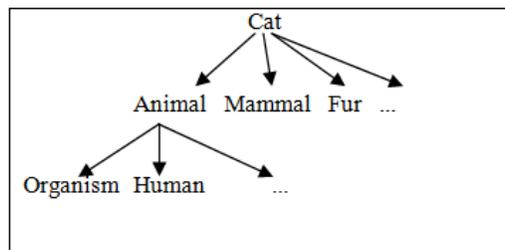


Figure 3. A hierarchical data structure with 3-levels of the word 'cat' in the dictionary.

We filled the matrix cells, as mentioned above, without taking frequency into account and saw that the results yielded were much less successful. That shows the importance of taking frequency into account.

We also have to take into account that some words are quite common in the dictionary, such as "situation", "thing", "person" and so on. Here we determined the top 1% most frequent words in the meaning texts in the dictionary as stop-words and eliminated them. Generally tf-idf is used for elimination of words, but since we make use of the dictionary as a base, top words elimination is sufficient.

3.4.2. Context and Frequency Algorithm

Although we noticed the algorithm we developed stated above gave meaningful concepts, drawbacks can be clearly seen. For example, let's assume there is a document containing the noun "football" and there is no other noun and its meaning text in the dictionary is as follows:

"A game played by two teams of 11 players each on a rectangular, 100-yard-long field with goal lines and goal posts at either end, the object being to gain possession of the ball and advance it in running or passing plays across the opponent's goal line or kick it through the air between the opponent's goal posts."

According to the algorithm stated above (Section 3.4.1), we take the nouns in this meaning text into account and build up a matrix containing those nouns, including the word 'football'. Since the word "football" is seen 3 times, the column labeled "goal" has a value of 3 as well and at the end the probable concept may be the word "goal", as well as the other concepts may be "game", "team", "line" and other nouns in the meaning text. (This is the case since the matrix would be of size $1 \times \text{CountNoun}(\text{MeaningTextOf}(\text{Football}))$), indicating that there is only one noun, that is

"football", in the document.) Having a concept of "goal" through this document would be a bit nonsense, hence we modified the algorithm in the following manner:

All the dictionary meaning text nouns would not be useful in determining the general idea of the main word, so some of those nouns have to be eliminated. In order to determine which meaning text noun is relevant in the use of the main word, we used a corpus-based context analysis. We had a few corpora and for each corpus, we used a 30-word window size context analysis, that is we looked up 15 words on the left of the test words and 15 words on the right of the test words. Hereby we eliminated the context words which are not nouns, because we think of concepts as only nouns. Then we assumed that if a context word is also present in the meaning text of the main word in dictionary, we take this context word into account. After scanning the whole corpus, whichever context word is seen most, given that context word is also seen in the meaning text of the main word, we add this word as a column word in the matrix corresponding to the row word. Then, similar to what we have done in (Section 3.4.1), we multiply the row elements values by the frequency of the row representative word and sum up the columns values. Whichever column value has the maximum value, we define that column representative word as the probable concept. In this case, we take mostly two words for each word in the document: The word itself and the word in the meaning text of this word that is most widely seen in the contexts in the corpus. We again, of course, firstly eliminated the stop words present in the TDK Dictionary.

This approach makes sense, since all meaning text nouns would not be useful in determining the general idea, that is concept, of a word. Also the corpus-based approach shows that the most relevant word in the meaning text of a test word is extracted through the context analysis. Selecting at most two words, that are the word itself and the most frequent word in the contexts of the word that is also present in the meaning text of the row word in the matrix rather than taking into account all nouns in the meaning text of a word increased the success rate for three of the corpora.

4. EVALUATION AND RESULTS

We tested our algorithms on four corpora in Turkish, which are as follows: Gazi University, Sport News, Forensic News and Forensic Court of Appeals Decisions. The corpora we handled are in different fields of topics as follows:

- **Gazi University Corpus:** This corpus includes 60 text files. There is no specific topic in this corpus, instead the files collected represent different fields of topics, such as tutorials on specific architectural or engineering subjects and many others.
- **Sport News Corpus:** This corpus includes 100 text files. Each file has a topic concerning sport news, especially Turkish football.
- **Forensic News Corpus:** This corpus includes 100 text files. These files are collected from news that are concerning forensic domain.
- **Forensic Court of Appeals Decisions Corpus:** This corpus has 108 text files. These files are concerning the forensic decisions and this corpus is similar to the corpus Forensic News. The difference is that this one is not collected from news documents database.

Initially we parsed and disambiguated the files in those corpora through the tools BoMorp and BoDis to extract the nouns needed for concept mining. Then we discarded the stop-words that abound in the dictionary.

4.1 Evaluation Metric

As an evaluation metric, we made use of the accuracy metric, which can be defined as follows:

$$accuracy = \frac{true\ positive + true\ negative}{true\ positive + false\ positive + true\ negative + false\ negative}$$

In order to evaluate the concepts extracted through the algorithm we developed, we also manually determined the concepts in those corpora. Then we compared the concepts with one another. We handled many ways to compare the concepts suggested by the algorithm with the ones extracted manually: For documents we made comparisons in 3, 5, 7, 10 and 15 window sizes. These windows include the top concepts, for example the one with size 5, includes the top 5 concepts. A comparison example table is shown below:

Table 1. An example showing the top 3 concepts in 2 documents.

Documents	Algorithm	Manual
Document 1	Sport, Game, Match	Sport, Match, Politics
Document 2	Court, Attorney, Judge	Attorney, Accused, Match

Table 1 shows the top three concepts for two documents, extracted both manually and algorithmically. In the first document, it can be seen that the success rate is $2 / (2 + 1) = 0.66$, since there are two words in common, which are "sport" and "match" that are found both in concept clusters extracted manually and algorithmically. However the word "game" is not in the top three concept cluster yielded manually, so it decreases the success rate. In Document 2, the success rate is 0.33, since only the word "attorney" is common amongst the three top concepts.

We also made comparison in a manner which compares the top 3, 5, 7, 10 and 15 concepts found through the algorithm with all the concepts found manually and evaluated the success ratio. It generally gave the highest success rates.

4.2 Results

Table 2 summarizes the accuracy results for different corpora. Window sizes 3, 5, 7 and 10 are tested for comparison. For algorithm 1 (Simple Frequency Algorithm) the most successful results are achieved through method, which uses 2-level structure taking frequency into account. All values shown in Table 2 for algorithm 1 are that achieved by 2-level structure taking frequency factor into account. For this algorithm, the sub-methods taking 3-level structure into account or eliminating frequency factor gave unsuccessful results. It is seen, on average, comparison with window size of three gives the highest success rate. The second algorithm gives, on average, better evaluation results. It can clearly be seen that results for different corpora vary a lot which shows that concept mining would be biased for documents concerning specific topics. The first algorithm gives, on average, a success rate of 63.62%, whereas the second one gives a success rate of 75.51%.

Table 2. Performance results for different corpora in terms of accuracy percentage.

Corpora	Comparison Window Size	Simple Frequency Algorithm	Frequency and Context Algorithm
Gazi	k = 3	58.40	69.30
	k = 5	55.70	67.00
	k = 7	54.70	64.30
	k = 10	53.90	62.60
Sport News	k = 3	57.40	55.40
	k = 5	56.90	54.20
	k = 7	56.00	53.40
	k = 10	55.30	52.68
Forensic News	k = 3	58.74	77.53
	k = 5	56.81	71.72
	k = 7	55.52	68.53
	k = 10	55.18	67.79
Forensic Decisions	k = 3	76.81	95.74
	k = 5	67.45	91.54
	k = 7	63.21	84.86
	k = 10	59.44	78.95

Table 3 shows, as an example, the results for Forensic Decisions corpus comparing all methods. First Algorithm (2-levels, 1-0) is the method that fills the matrix with values 1 or 0 using 2-level structure. If a column word is present the cell value in matrix is 1, if it is not present, cell value is 0. Frequency factor is overlooked in this sub-method. First Algorithm (2-levels, frequency) takes frequency into account. First Algorithm (3-levels, coefficients) takes frequency into account using 3-level structure. This algorithm assigns different scores for different hierarchical levels of a word. The second algorithm is the algorithm explained in Section 3.4.2.

5. CONCLUSION AND FUTURE WORK

In this work, we developed two novel algorithms which make use of the dictionary. So far, generally WordNet has been preferred for its synset relations, especially hypernymy, in concept mining, because the general idea of a word is generally related to its hypernym word. But only taking a solitary synset relation into account may be insufficient, also other relationships may give us an idea concerning a word, that is why we made use of a general language dictionary.

Table 3. Performance results of four methods for Forensic Decisions corpus with different comparison windows sizes.

Algorithms	Comparison Window Size			
	k = 3	k = 5	k = 7	k = 10
First Algorithm (2-levels, 1-0)	76.81	67.45	63.21	59.44
First Algorithm (2-levels, frequency)	68.84	65.43	66.72	63.92
First Algorithm (3-levels, coefficients)	64.81	60.73	57.87	55.72
Second Algorithm	95.74	84.77	78.58	73.11

Besides the use of a dictionary, we took the frequencies into account, because the more frequent a word in the document, the higher contribution of this word to the general idea and concept of documents amongst many. We also developed context-based algorithm which eliminates some of the meaning text nouns in dictionary from probable candidate concepts set, because some of those words would not be convenient in determining the general idea, that is concept, of a word. This approach increased our success rate.

As a future work, we would improve the second algorithm, where we can take not only the word and the noun that is most frequently found in its contexts in all corpus and present in its meaning text in dictionary into account, but also the other nouns present in both meaning text and contexts. We can give them coefficients in accordance with their frequencies instead of eliminating them. Apart from the direction of this paper's algorithm, for concept mining we may also implement clustering through all dictionary words, in accordance with the analogies between their meaning texts. This would approach the whole dictionary word-set as a training data.

ACKNOWLEDGEMENT

This work was supported by the Boğaziçi University Research Fund under the grant number 5187, and the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant number 110E162. Cem Rıfık Aydın is supported by TÜBİTAK BİDEB 2210. We thank Hasim Sak for providing us the tools for pre-processing and morphological analyses.

REFERENCES

- [1] Marie-Francine, Moens & Roxana, Angheluta, (2003) "Concept Extraction from Legal Cases: The Use of a Statistic of Coincidence", International Conference on Artificial Intelligence and Law, ICAIL, ACM
- [2] Vance, Faber & Judith G., Hochberg & Patrick M., Kelly, Timothy R. Thomas, James M. White, (1994) "Concept Extraction – a data-mining technique", Los Alamos Science
- [3] Nuala A., Bennett & Qin, He & Conrad T. K., Chang & Bruce R., Schatz, (1999) "Concept Extraction in the Interspace Prototype", Technical Report, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL
- [4] David M., Pennock & Kushal, Dave & Steve, Lawrence, (2003) "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Twelfth International World Wide Web Conference (WWW'2003), ACM
- [5] H., Liu & P., Singh, (2004) "ConceptNet - a practical commonsense reasoning toolkit", BT Technology Journal, Vol 22, No 4
- [6] Meryem, Uzun, (2011) "Developing a concept extraction system for Turkish", International Conference on Artificial Intelligence, ICAI'11
- [7] Paul M., Ramirez & Chris A., Mattmann, (2004) "ACE: Improving Search Engines via Automatic Concept Extraction", Information Reuse and Integration
- [8] Zhang, Chengzhi & Wu, Dan, (2008) "Concept Extraction and Clustering for Topic Digital Library Construction", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
- [9] Loulwah, AlSumait & Daniel, Barbar'a & Carlotta, Domeniconi, (2008) "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking", ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining
- [10] Zakaria, Elberrichi & Abdelattif, Rahmoun & Mohamed Amine, Bentaalah, (2008) "Using WordNet for Text Categorization", The International Arab Journal of Information Technology, Vol. 5, No. 1
- [11] Haşim, Sak & Tunga, Güngör & Murat, Saraçlar, (2008) "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus", GoTAL 2008, vol. LNCS 5221, pp. 417-427, Springer.
- [12] Haşim, Sak & Tunga, Güngör & Murat, Saraçlar, (2007) "Morphological disambiguation of Turkish text with perceptron algorithm", CICLing 2007, vol. LNCS 4394, pp. 107-118.

AUTHORS

Cem Rıfık Aydın received his B.Sc. degree from Bahçeşehir University, Istanbul, Turkey, in Department of Computer Engineering. He is currently an M.Sc. student in Department of Computer Engineering at Boğaziçi University, and awarded with TÜBİTAK BİDEB scholarship throughout his M.Sc. studies. His areas of interest include natural language processing, artificial intelligence applications, pattern recognition, game programming, information retrieval and genetic algorithms.



Ali Erkan received his B.Sc. and M.Sc. degrees from Department of Industrial Engineering, Bilkent University, Ankara, Turkey, and he received M.Sc. degree in Software Engineering from Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey, in 2010. He is currently studying for Ph.D. degree at Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey. His research interests include natural language processing, machine learning, pattern recognition, bioinformatics and statistics.



Tunga Güngör received his Ph.D. degree from Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey, in 1995. He is an associate professor in the department of Computer Engineering, Boğaziçi University. His research interests include natural language processing, machine translation, machine learning, pattern recognition, and automated theorem proving. He published about 60 scientific articles, and participated in several research projects and conference organizations.



Hidayet Takçı is an academician at Cumhuriyet University, Sivas, Turkey. He studies on some fields such as data mining, text mining, machine learning and security. Hitherto he has lectured some courses such as data mining and applications, text mining, neural nets, etc. In addition, he has many papers and projects in the field of text classification and text mining, information retrieval and web security.



FINDING IMPORTANT NODES IN SOCIAL NETWORKS BASED ON MODIFIED PAGERANK

Li-qing Qiu¹, Yong-quan Liang², Jing-Chen³

^{1,2}College of Information Science and Technology, Shandong University
of Science and Technology, Qingdao, China

³Shandong labor vocational and technical college Jinan, China

liqingqiu2005@126.com¹, lyq@sdust.edu.cn²
wfchenj@126.com³

ABSTRACT

Important nodes are individuals who have huge influence on social network. Finding important nodes in social networks is of great significance for research on the structure of the social networks. Based on the core idea of Pagerank, a new ranking method is proposed by considering the link similarity between the nodes. The key concept of the method is the use of the link vector which records the contact times between nodes. Then the link similarity is computed based on the vectors through the similarity function. The proposed method incorporates the link similarity into original Pagerank. The experiment results show that the proposed method can get better performance.

KEYWORDS

social networks, Pagerank, link similarity

1. INTRODUCTION

In modern society, social networks play an important role in a quickly changing world, and more and more people prefer to obtain information from social networks. This explains the increasing interest in social networks analysis which examines topology of a network in order to find interesting structure within it. Recent works have pointed that some very active nodes have huge impacts on other nodes [1]. Therefore, the problem of how to appropriately find important nodes in social networks among the huge nodes becomes an important issue.

Pagerank[2,3] is a well known method for identifying authoritative pages in a hyperlink network of web pages. Pagerank relies on the democratic nature of the web by using its topology as an indicator of the value to be attached to any page. Pagerank is so useful that we can also apply it to social networks in that the mutual relationship in social networks can be structured as links to the microblogs, and nodes can also be regarded as websites in Pagerank algorithm [4]. In the paper, we present a new importance ranking method based on the modified Pagerank to find the important nodes in social networks. In other words, the proposed method is derives from

Pagerank by considering the link similarity which measures the similarity between the nodes. The paper is organized as follows. In section 2 we describe our generalization of new measurement, and the experimental results together with the experimental settings are given in section 3. At last we conclude the paper by summarizing our findings in section 4.

2. PROPOSED METHOD

In the section, we propose a new method to find important nodes based on modified Pagerank, by considering the link similarity. We introduce the modified Pagerank model in section 2.1, and then we analyze the importance of nodes using the model in section 2.2.

2.1 Modified Pagerank

The core idea of Pagerank is that of introducing a notion of page authority. In pagerank, the authority reminds the notion of citation in the scientific literature. In particular, the authority of a page p depends on the number of incoming hyperlinks and on the authority of the page q which cites p with a forward link. Moreover, selective citations from q to p are assumed to provide more contribution to the value of p than uniform citations. Therefore, the Pagerank value PR_p of p is computed by taking into account the set of pages $pa[p]$ pointing to p . The Pagerank value PR_p is defined as follows:

$$PR_p = (1-d) + d \sum_{q \in pa[p]} \frac{PR_q}{h_q} \quad (1)$$

Here $d \in (0,1)$ is a dumping factor which corresponds to the probability with which a surfer jumps to a page picked uniformly at random.

The quality of the links, as measured by Pagerank, is a good choice for ranking nodes but we think there are some other features that can incorporate the activity of the node. We propose to incorporate the features via the link similarity taking into account contact times of the node. The idea behind is that the node has higher similarity must be prized with a higher value. Our main idea consists in constructing the link vector that records the contact times of the nodes, defining a link similarity function to measure the similarity of the nodes according to the link vector, and then reconstructing Pagerank model by considering the link similarity. These ideas are a work in progress.

Definition 1. For node v_i , its link vector is defined as:

$$V_i = \{t_{i1}, t_{i1}, \dots, t_{i-1}, t_{i+1}, \dots, t_{in}\} \quad (2)$$

Where t_{im} ($0 \leq m \leq n$) is the contact times of v_i and v_m .

Definition 2. For node v_i and v_j , the link similarity is defined as:

$$Similarity(V_i, V_j) = \bar{V}_i \cdot \bar{V}_j = \frac{\sum_{m=1}^n V_{im} \cdot V_{jm}}{\sqrt{\sum_{m=1}^n V_{im}^2} \sqrt{\sum_{m=1}^n V_{jm}^2}} \quad (3)$$

Obviously, the link similarity measures are functions that take two link vectors as arguments and compute a real value in the interval [0..1], the value 1 means that the two nodes are closely related while the value 0 means the nodes are quite different.

Definition 3. For node v_i , its modified Pagerank value is defined :

$$PR_{v_i} = (1-d) + d \sum_{v_j \in pa[v_i]} \frac{PR_{v_j} \cdot Similarity(V_i, V_j)}{h_{v_j}} \quad (4)$$

Here $pa[v_i]$ is the set of nodes which point to v_i , and the dump factor d is set to 0.15 in our experiments.

In the Modified Pagerank, the link similarity of the nodes is taken into account, which is a positive indicator to modify the original Pagerank(see Equation 1).

2.2 Node Analysis based on modified Pagerank

The node analysis based on modified Pagerank is composed of 3 steps as follows:

- Step1: read network as graph;
- Step2: computer modified Pagerank value according to Equation 4;
- Step3: obtain node modified Pagerank value list.

From above steps, we can see that a bit change is made to the original Pagerank by adding the link similarity indicator. The modified Pagerank assumes that the initial Pagerank values of all nodes are the same. Firstly, we calculate the first iterative ranking of each node based on the initial Pagerank values, and then calculate the second rank according to the first iteration. The process continues until the termination condition is satisfied. At last the Pagerank estimator converges to its practical value, which has proven no matter what the initial value is. The whole processing is implemented without manual intervention.

3. EXPERIMENTS

In the following, we use several different networks to study the performance of our generalization of novel algorithm for detecting local community structure.

3.1 First Experiment

We start with the first synthetic dataset, which is shown as Figure 1, to illustrate the process of the proposed method in detail. The network contains 5 nodes, and the contact times are associated

with corresponding edges.

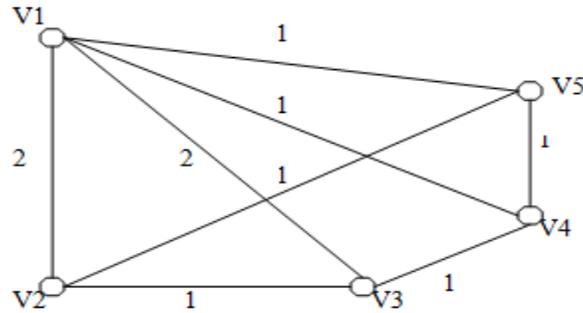


Figure1. A simple network of 5 nodes and 8 edges.

So the mutual relationship matrix can be produced as briefly shown in Table 1.

Table1. Mutual relationship matrix, A.

	V1	V2	V3	V4	V5
V1		2	2	1	1
V2	2		1	0	1
V3	2	1		1	0
V4	1	0	1		1
V5	1	1	0	1	

Parameters in the matrix a_{ij} is defined as the contact times between the nodes. Then the link vectors are: $V1 = \{2, 2, 1, 1\}$, $V2 = \{2, 1, 0, 1\}$, $V3 = \{2, 1, 1, 0\}$, $V4 = \{1, 0, 1, 1\}$, $V5 = \{1, 1, 0, 1\}$. Thereby, the link similarity can be computed as:

$$Similarity(V1, V2) = \frac{2*2 + 2*1 + 1*0 + 1*1}{\sqrt{2^2 + 2^2 + 1^2 + 1^2} * \sqrt{2^2 + 1^2 + 0^2 + 1^2}} = 0.904$$

The process continues until we get the total link similarity of the network. The link similarity matrix is shown as follows:

Table2. The link similarity matrix, B.

	V1	V2	V3	V4	V5
V1		0.904	0.904	0.730	0.913
V2	0.904		0.833	0.707	0.904
V3	0.904	0.833		0.707	0.707
V4	0.730	0.707	0.707		0.816
V5	0.913	0.904	0.707	0.816	

Then the modified Pagerank can be computed according to Equation 4, and we get the following Pagerank values:

Table3. The modified Pagerank Values.

	Value
V1	0.116
V2	0.099
V3	0.101
V4	0.093
V5	0.109

3.2 Second Experiment

Secondly, we apply our modified Pagerank method to one small network, which is the much-discussed “karate club” network of friendships between 34 members of a karate club at a US university, assembled by Zachary [5] by direct observation of the club’s members. This network is of particular interest because the club split in two during the course of Zachary’s observations as a result of an internal dispute between the director and the coach. In other words the network can be classified into two communities-one’s center is the director, the other’s center is the coach.

We select degree centrality and PageRank algorithm as baseline methods, which be identified as “degree” and “PageRank” respectively. And our proposed modified Pagerank method is identified as “M-Pagerank”. Table4 shows the result of the experiment. We select the nodes in the top 10 according to different methods.

Table4. The comparison of different methods

degree	Node ID	PageRank	Node ID	M-Pagerank	Node ID
0.515	34	0.101	34	0.093	34
0.485	1	0.097	1	0.089	1
0.364	33	0.072	33	0.065	33
0.303	3	0.057	3	0.050	3
0.273	2	0.053	2	0.036	2
0.182	32	0.037	32	0.028	32
0.182	4	0.036	4	0.017	14
0.152	24	0.032	24	0.016	4
0.152	9	0.030	9	0.014	9
0.152	24	0.030	14	0.014	31

From the above table, we can see that three methods appear to have many common results. However, degree can not distinguish nodes because some nodes have the equal value. For example, node 32 and node 4 have the common value according to degree method, which shows that degree method can not distinguish nodes well. M-Pagerank performs better than Pagerank in that M-Pagerank considers the link similarity of the nodes. For example, node 31 is on the boundary of two communities, which has much connection between the two communities.

According to M-PageRank, node 31 is important than node 24 obviously. However, PageRank rank node 24 higher than node 31, which is not reasonable.

4. CONCLUSION

In the paper, we have shown a new method to find important nodes in social works based on modified PageRank. The method is capable of incorporate the link similarity of the nodes via the link vector. The final goal for this model is to incorporate some features into original PageRank. The proposed method enables a new way of ranking the nodes in social works. We have analyzed the experiment results using test networks. In our future work, we will further discuss how to achieve better performance to detect importance nodes.

ACKNOWLEDGEMENT

This paper is supported by National Science Foundation for Post-doctoral Scientists of China under grant 2013M541938, and Shandong Province Postdoctoral special funds for innovative projects of China under grant 201302036.

REFERENCES

- [1] Khorasgani RR, Chen J, Zaiane OR. Top leaders community detection approach in information networks. KDD 2010, 1-9.
- [2] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. WWW7, 1998.
- [3] Bianchini M, Gori M, Scarselli F. Inside PageRank. ACM transactions on internet technology, 2006, 92-128.
- [4] Abbassi Z, Minokni VS. A recommender system based on local random walks and spectral methods. 9th Web KDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, 2007, 102-108
- [5] Zachary WW. An information flow model for conflict and fission in small groups, Journal of anthropological Research, 1977, 33: 452-473.

A SAT ENCODING FOR SOLVING GAMES WITH ENERGY OBJECTIVES

Raffaella Gentilini

Dip. di Matematica e Informatica, Universit`a di Perugia,
Via Vanvitelli 1, Perugia (IT)

ABSTRACT

Recently, a reduction from the problem of solving parity games to the satisfiability problem in propositional logic (SAT) have been proposed in [5], motivated by the success of SAT solvers in symbolic verification. With analogous motivations, we show how to exploit the notion of energy progress measure to devise a reduction from the problem of energy games to the satisfiability problem for formulas of propositional logic in conjunctive normal form.

1. INTRODUCTION

Energy games (EG) are two-players games played on weighted graphs, where the integer weight associated to each edge represents the corresponding energy gain/loss. The arenas of energy games are endowed of two types of vertices: in player 0 (resp. player 1) vertices, player 0 (resp. player 1) chooses the successor vertex from the set of outgoing edges and the game results in an infinite path through the graph. Given an initial credit of energy c , the objective of player 0 is to maintain the sum of the weights (the energy level) positive. The decision problem for EG asks, given a weighted game graph with initial vertex v_0 , if there exists an initial credit for which player 0 wins from v_0 .

Energy games have been introduced in [3, 2] to model the synthesis problem within the design of reactive systems that work in resource-constrained environments. Beside their applicability to the modeling of quantitative problems for computer aided design, EG have tight connections with important problems in game theory and logic. For instance, they are log-space equivalent to mean-payoff games (MPG) [2], another kind of quantitative two-player game very well studied both in economics and in computer science. The latter are characterized by a theoretically engaging complexity status, being one of the few inhabitants of the complexity class $NP \cap coNP$ (for which the inclusion in P is still an open problem). Moreover, parity games [4, 6]—notoriously known as poly-time equivalent to the model-checking problem for the modal mu-calculus—are in turn poly-time reducible to MPG and EG. It is a long-standing open question to know whether the model-checking problem for the modal mu-calculus is in P.

The algorithm with the currently best (pseudopolynomial) complexity for solving EG (and MPG via log-space reduction) is based on the so-called notion of *energy progress measure* [7].

Progress measures for weighted graphs are functions that impose local conditions to ensure global properties of the graph. A notion of *parity* progress measure [6] was previously exploited in [6] for the algorithmic analysis of parity games and reconsidered in [5] to devise a SAT encoding of the corresponding games, motivated by the considerable success that using SAT solvers has had in symbolic verification. As a matter of fact, clever heuristics implemented in nowadays SAT solvers can result in algorithms that are very efficient in practice. Furthermore, there are fragments of SAT that can be solved in polynomial time. Hence, the reduction in [5] opens up a new possibility for showing inclusion of parity games in P.

Motivated by analogous reasons, in this paper we show how to exploit the notion of energy progress measure to devise a reduction from the problem of energy games to the satisfiability problem for formulas of propositional logic in conjunctive normal form. Tight upper bounds on the sizes of our reductions are also reported.

The paper is organized as follows. We recall the notions of energy games and energy progress measure in Section 2. Section 3 and Section 4 develop the reductions from energy games to difference logic and pure SAT, respectively, reporting tight bounds on the sizes of the corresponding reductions.

2. PRELIMINARIES

Game graphs A *game graph* is a tuple $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ where $G^\Gamma = (V, E, v_0, w)$ is a weighted graph with weight function $w : E \rightarrow \mathbb{Z}$ and $\langle V_0, V_1 \rangle$ is a partition of V into the set V_0 of player-0 vertices and the set V_1 of player-1 vertices. An *infinite game* on Γ is played for infinitely many rounds by two players moving a pebble along the edges of the weighted graph G^Γ . In the first round, the pebble is on some vertex $v \in V$. In each round, if the pebble is on a vertex $v \in V_i$ ($i = 0, 1$), then player i chooses an edge $(v, v') \in E$ and the next round starts with the pebble on v' . A *play* in the game graph Γ is an infinite sequence $p = v_0 v_1 \dots v_n \dots$ such that $(v_i, v_{i+1}) \in E$ for all $i \geq 0$. A *strategy* for player i ($i = 0, 1$) is a function $\sigma : V^* \cdot V_i \rightarrow V$, such that for all finite paths $v_0 v_1 \dots v_n$ with $v_n \in V_i$, we have $(v_n, \sigma(v_0 v_1 \dots v_n)) \in E$. We denote by Σ_i ($i = 0, 1$) the set of strategies for player i . A strategy σ for player i is *memoryless* if $\sigma(p) = \sigma(p')$ for all sequences $p = v_0 v_1 \dots v_n$ and $p' = v'_0 v'_1 \dots v'_m$ such that $v_n = v'_m$. We denote by Σ_i^M the set of memoryless strategies of player i . A play $v_0 v_1 \dots v_n \dots$ is *consistent* with a strategy σ for player i if $v_{j+1} = \sigma(v_0 v_1 \dots v_j)$ for all positions $j \geq 0$ such that $v_j \in V_i$. Given an initial vertex $v \in V$, the *outcome* of two strategies $\sigma_1 \in \Sigma_1$ and $\sigma_2 \in \Sigma_2$ in v is the (unique) play $\text{outcome}^\Gamma(v, \sigma_0, \sigma_1)$ that starts in v and is consistent with both σ_0 and σ_1 . Given a memoryless strategy π_i for player i in the game Γ , we denote by $G^\Gamma(\pi_i) = (V, E_{\pi_i}, w)$ the weighted graph obtained by removing from G^Γ all edges (v, v') such that $v \in V_i$ and $v' \neq \pi_i(v)$.

Energy Games [3, 2] An *energy game* (EG) is an infinite game on the game graph Γ , where the goal of player 0 is to construct an infinite play $v_0 v_1 \dots v_n \dots$ such that for some *initial credit* $c \in \mathbb{N}$:

$$c + \sum_{i=0}^j w(v_i, v_{i+1}) \geq 0 \text{ for all } j \geq 0 \quad (1)$$

The quantity $c + \sum_{i=0}^{j-1} w(v_i, v_{i+1})$ is called the *energy level* of the play prefix $v_0 v_1 \dots v_j$. Given a credit c , a play $p = v_0 v_1 \dots$ is *winning* for player 0 if it satisfies (1), otherwise it is winning for player 1. A vertex $v \in V$ is *winning* for player i if there exists an initial credit c and a winning strategy for player i from v for credit c . In the sequel, we denote by W_i the set of winning states for player i . Energy games are memoryless determined [2], i.e. for all $v \in V$, either v is winning for player 0, or v is winning for player 1, and memoryless strategies are sufficient.

Theorem 1 ([2]). *Let $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ be an EG, for all $v \in V$, the following four statements are equivalent:*

- $\exists \sigma_0 \in \Sigma_0 \cdot \forall \sigma_1 \in \Sigma_1 \cdot \text{outcome}^\Gamma(v, \sigma_0, \sigma_1)$ is winning for player 0;
- $\forall \sigma_1 \in \Sigma_1 \cdot \exists \sigma_0 \in \Sigma_0 \cdot \text{outcome}^\Gamma(v, \sigma_0, \sigma_1)$ is winning for player 0;
- $\exists \pi_0 \in \Sigma_0^M \cdot \forall \pi_1 \in \Sigma_1^M \cdot \text{outcome}^\Gamma(v, \pi_0, \pi_1)$ is winning for player 0;
- $\forall \pi_1 \in \Sigma_1^M \cdot \exists \pi_0 \in \Sigma_0^M \cdot \text{outcome}^\Gamma(v, \pi_0, \pi_1)$ is winning for player 0;

Using the memoryless determinacy of energy games, the authors of [7] derived the next characterization lemma for EG winning strategies.

Lemma 1 ([7]). *Let $\Gamma = (V, E, w, \langle V_0, V_1 \rangle)$ be an EG. For all vertices $v \in V$, for all memoryless strategies $\pi_0 \in \Sigma_0^M$ for player 0, the strategy π_0 is winning from v if and only if all cycles reachable from v in the weighted graph $G^\Gamma(\pi_0)$ are nonnegative.*

Given the energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$, the EG decision problem asks whether v_0 is winning for player 0. Such a problem is polynomially equivalent to the corresponding decision problem for so-called meanpayoff games [2, 1].

The algorithm with the currently best (pseudopolynomial) complexity for solving energy games is based on the so-called notion of small energy progress measure [7]. Intuitively, the latter is a condition locally defined on the vertices of the given game graph, tailored to witness the global absence of negative cycles within the subgame induced by a proper strategy for player 0 (cfr. the characterization lemma 1). Formally, the notion of small progress measure is recalled in Definition 1 (below) and relies on the following notation. Given $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$, denote by \mathcal{C}_Γ the following set:

$$\mathcal{C}_\Gamma = \{n \in \mathbb{N} \mid n \leq \mathcal{M}_{G^\Gamma}\} \cup \{\top\}.$$

where:

$$\mathcal{M}_{G^\Gamma} = \sum_{v \in V} \max(\{0\} \cup \{-w(v, v') \mid (v, v') \in E\})$$

Moreover, denote by \preceq the total order on \mathcal{C}_Γ defined by $x \preceq y$ if and only if either $y = \top$ or $x \leq y \leq \mathcal{M}_{G^\Gamma}$. Finally, let $\ominus : \mathcal{C}_\Gamma \times \mathbb{Z} \rightarrow \mathcal{C}_\Gamma$ be the operator such that for all $a \in \mathcal{C}_\Gamma$ and $b \in \mathbb{Z}$:

$$a \ominus b = \begin{cases} \max(0, a - b) & \text{if } a \neq \top \text{ and } a - b \leq \mathcal{M}_{G^r} \\ \top & \text{otherwise} \end{cases}$$

Definition 1 ([7]). *Let $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ be an EG. A function $f : V \rightarrow \mathcal{C}_\Gamma$ is a small energy progress measure for Γ if and only if the following conditions hold:*

- if $v \in V_0$, then $f(v) \succeq f(v') \ominus w(v, v')$ for some $(v, v') \in E$;
- if $v \in V_1$, then $f(v) \succeq f(v') \ominus w(v, v')$ for all $(v, v') \in E$.

Given a small energy progress measure f for the game graph $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$, we denote by V_f the set of states $V_f = \{v \mid f(v) \neq \top\}$. A memoryless strategy $\pi_0^f : V_0 \rightarrow V$ for player 0 is called *compatible with f* whenever for all $v \in V_0$, if $\pi_0^f(v) = v'$ then $f(v) \succeq f(v') \ominus w(v, v')$. The following property holds [7]: if π_0^f is a strategy for player 0 compatible with the energy progress measure f , then π_0^f is a winning strategy for player 0 from all vertices in V_f . Formally:

Theorem 2 ([7]). *Let $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ be an EG. For all small energy progress measures f for Γ , if π_0^f is a strategy for player 0 compatible with f , then π_0^f is a winning strategy for player 0 from all vertices $v \in V_f$, i.e. $V_f \subseteq W_0$. Moreover, Γ admits a small energy progress measure f such that $V_f = W_0$.*

2.1 Difference Logic

Let $\mathcal{B} = \{b_1, \dots, b_n\}$ be a set of boolean variables and $\mathcal{X} = \{x_1, \dots, x_n\}$ be a set of integer variables. The set of atomic formulas of difference logic consists of the boolean variables in \mathcal{B} and integer constraints of the form $x_i - x_j \geq c$, $c \in \mathbb{Z}$.

The set \mathcal{F} of difference logic formulas is the smallest set containing the atomic formulas which is closed under negation and conjunction (the boolean connectives $\vee, \rightarrow, \leftrightarrow$ are defined in the usual way in terms of the operators of negation and conjunction \wedge, \neg). A $(\mathcal{B}, \mathcal{X})$ valuation consists of two functions (overloaded with the name α), $\alpha : \mathcal{B} \rightarrow \{1, 0\}, \alpha : \mathcal{X} \rightarrow \mathbb{Z}$. The valuation is extended to all difference logic formulas by letting $\alpha(x_i - x_j \geq c) = 1$ if and only if $\alpha(x_i) - \alpha(x_j) \geq c$ and applying the obvious rules for boolean connectives. A difference logic formula ϕ is satisfied by a valuation α if and only if $\alpha(\phi) = 1$. A formula ϕ is satisfiable if it admits a satisfying valuation. The satisfiability problem for difference logic is NP-complete [8].

3. ENCODING EG WINNING STRATEGIES IN DIFFERENCE LOGIC

In this section we show how to derive a difference logic formula ϕ_Γ from a given energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ such that ϕ_Γ is satisfiable if and only if player 0 has a winning strategy on Γ .

In particular, the difference logic formula ϕ_Γ uses the set of $|E|$ integer constants $\{w_{(v,z)} \mid (v, z) \in E\}$ and ranges over the following set of boolean and integer variables:

- for each $v \in V$, there is a boolean variable n_v and an integer variable c_v
- for each edge $(v, z) \in E$, there is a boolean variable $m_{(v,z)}$

Given the above variables, $\phi_\Gamma \equiv n_{v_0} \wedge \phi_0 \wedge \phi_1 \wedge \phi_\sigma \wedge \phi_e$ is the conjunction of five subformulas, where $\phi_0, \phi_1, \phi_\sigma, \phi_e$ are defined as follows:

- $\phi_0 \equiv \bigwedge_{v \in V_0} (n_v \rightarrow \bigvee_{(v,z) \in E} m_{(v,z)})$
- $\phi_1 \equiv \bigwedge_{v \in V_0} (n_v \rightarrow \bigwedge_{(v,z) \in E} m_{(v,z)})$
- $\phi_\rho \equiv \bigvee_{\substack{v \in V \\ v \neq v_0}} ((\bigvee_{(v,z) \in E} m_{(v,z)}) \rightarrow n_z)$
- $\phi_e \equiv \bigvee_{(v,z) \in E} (m_{(v,z)} \rightarrow \psi_{(v,z)})$
- $\psi_{(v,z)} \equiv c_v + w_{(v,z)} \geq c_z$

Theorem 3. *Player 0 has a winning strategy in the energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ if and only if the difference logic formula ϕ_Γ is satisfiable.*

Proof. (\Rightarrow) Let $G_\Gamma(\pi)$ be the graph induced by a winning strategy π for player 0 on the energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$. Consider the assignment α to the variables of ϕ_Γ defined as follows: for each boolean variable n_v (resp. $m_{(v,z)}$) let $\alpha(n_v) = 1$ (resp. $\alpha(m_{(v,z)}) = 1$) if and only if v is a node (resp. (v, z) is an edge) of $G_\Gamma(\pi)$. By definition of $G_\Gamma(\pi)$, the assignment α satisfies $n_{v_0} \wedge \phi_0 \wedge \phi_1$. By Theorem 2, $G_\Gamma(\pi)$ admits a small progress measure function $f : W \rightarrow \mathcal{M}_{G_\Gamma(\pi)}$, where W is the set of vertices of $G_\Gamma(\pi)$. For each integer variable c_v in ϕ_Γ , define $\alpha(c_v) = f(v)$ if $v \in W$. Since π is a winning strategy on Γ for player 0, the assignment α satisfies also the last conjunct ϕ_σ in ϕ_Γ . Therefore, $\alpha \models \phi_\Gamma$.

(\Leftarrow) Suppose that α is a satisfying variable assignment of ϕ_Γ . Define the following game $\Gamma' = (V', E', v_0, w', \langle V'_0, V'_1 \rangle)$: $v \in V'$ (resp. $(v, z) \in E'$) if and only if $\alpha(n_v) = 1$ (resp. $\alpha(m_{(v,z)}) = 1$) and for each $(v, z) \in E'$ let $w'(v, z) = w_{(v,z)}$. Since α satisfies $n_{v_0} \wedge \phi_0 \wedge \phi_1 \wedge \phi_\sigma$, we derive that Γ' is a non empty subgame of Γ . Hence, since α satisfies also ϕ_e , by Theorem 2 we deduce that $V' \subseteq W_0$ and Γ' induces a winning strategy for player 0 on Γ .

Theorem 4. *Given an energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$, the size of the difference logic formula ϕ_Γ is $\mathcal{O}(|E|)$, even if ϕ_Γ is required to be in CNF.*

Proof. Each subformula $\phi_0 \wedge \phi_1, \phi_\sigma, \phi_e$ has size $\mathcal{O}(|E|)$, while the remaining conjunct n_{v_0} in ϕ_Γ has size 1. ϕ_Γ can be rewritten in CNF with a constant blow up by reformulating the conjuncts $\phi_0, \phi_1, \phi_\sigma$ and ϕ_e using the boolean equivalences:

$$\chi \rightarrow (\phi \wedge \psi) \equiv (\chi \rightarrow \phi) \wedge (\chi \rightarrow \psi)$$

$$(\phi \vee \psi) \rightarrow \chi \equiv (\phi \rightarrow \chi) \wedge (\psi \rightarrow \chi)$$

4. SOLVING ENERGY GAMES BY A REDUCTION TO SAT

In this section, we present an encoding for the difference logic formula ϕ_Γ associated to a given energy game Γ into propositional logic, i.e. the subset of difference logic with boolean variables only. Clearly, all that remains to be done is to translate the integer variables and the constraints on them of the form $c_v + w_{(v,z)} \geq c_z$ inside the conjunct ϕ_e in ϕ_Γ .

Let $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ be the energy game underlying ϕ_Γ . By Theorem 2 the domain of the integer variables in ϕ_Γ can be bounded by $\mathcal{M}_{G_\Gamma} \leq V \cdot W$, where W is the maximum absolute weight in Γ . Let $k = \lceil \log(\mathcal{M}_{G_\Gamma} + W) \rceil$ be the number of bits necessary to code $\mathcal{M}_{G_\Gamma}, W$.

For each edge $(v, z) \in E$, let $\bar{w}_{(v,z)} = w_1 \dots w_k$ be the boolean encoding of $|w_{(v,z)}|$ (using k boolean variables), let $e_1^v, \dots, e_k^v, e_1^z, \dots, e_k^z, s_1^{(v,z)} \dots s_k^{(v,z)}, r_0^{(v,z)} \dots r_k^{(v,z)}$, be further boolean variables and consider the following propositional formulas:

- If $w_{(v,z)} \geq 0$:

- $\text{CURRY}(v, z, k) \equiv \neg r_k^{(v,z)}$
- for $i = k \dots 1$:

$$\begin{aligned} \text{SUM}(v, z, i) &\equiv s_i^{(v,z)} \Leftrightarrow (\neg e_i^v \wedge \neg w_i \wedge \neg r_i^{(v,z)}) \vee (\neg e_i^v \wedge w_i \wedge \neg r_i^{(v,z)}) \\ &\quad \vee (e_i^v \wedge \neg w_i \wedge \neg r_i^{(v,z)}) \vee (e_i^v \wedge w_i \wedge r_i^{(v,z)}) \\ \text{CURRY}(v, z, i-1) &\equiv r_{i-1}^{(v,z)} \Leftrightarrow (\neg e_i^v \wedge w_i \wedge r_i^{(v,z)}) \vee (e_i^v \wedge \neg w_i \wedge r_i^{(v,z)}) \\ &\quad \vee (e_i^v \wedge w_i \wedge \neg r_i^{(v,z)}) \vee (e_i^v \wedge w_i \wedge r_i^{(v,z)}) \end{aligned}$$

- $\text{CURRY}(v, z, 0) \equiv \neg r_0^{(v,z)}$
- $\text{GEQ}(v, z, 1) \equiv s_1^{(v,z)} \Rightarrow e_1^z$
- for $i = k \dots 1$:
 $\text{GEQ}(v, z, i) \equiv (s_i^{(v,z)} \Rightarrow e_i^z) \wedge ((s_i^{(v,z)} \vee \neg e_i^z) \Rightarrow \text{GEQ}(v, z, i-1))$

- If $w_{(v,z)} < 0$:

- $\text{CURRY}(v, z, k) \equiv \neg r_k^{(v,z)}$
- for $i = k \dots 1$:

$$\begin{aligned} \text{SUM}(v, z, i) &\equiv s_i^{(v,z)} \Leftrightarrow (\neg e_i^z \wedge \neg w_i \wedge \neg r_i^{(v,z)}) \vee (\neg e_i^z \wedge w_i \wedge \neg r_i^{(v,z)}) \\ &\quad \vee (e_i^z \wedge \neg w_i \wedge \neg r_i^{(v,z)}) \vee (e_i^z \wedge w_i \wedge r_i^{(v,z)}) \\ \text{CURRY}(v, z, i-1) &\equiv r_{i-1}^{(v,z)} \Leftrightarrow (\neg e_i^z \wedge w_i \wedge r_i^{(v,z)}) \vee (e_i^z \wedge \neg w_i \wedge r_i^{(v,z)}) \\ &\quad \vee (e_i^z \wedge w_i \wedge \neg r_i^{(v,z)}) \vee (e_i^z \wedge w_i \wedge r_i^{(v,z)}) \end{aligned}$$

- $\text{CURRY}(v, z, 0) \equiv \neg r_0^{(v,z)}$
- $\text{GEQ}(v, z, 1) \equiv e_1^v \Rightarrow s_1^{(v,z)}$
- for $i = k \dots 1$:
 $\text{GEQ}(v, z, i) \equiv (e_i^v \Rightarrow s_i^{(v,z)}) \wedge ((e_i^v \vee \neg s_i^{(v,z)}) \Rightarrow \text{GEQ}(v, z, i-1))$

Let ϕ'_Γ be the propositional logic formula obtained by replacing each integer constraint in ϕ_Γ of the form $c_v + w_{(v,z)} \geq c_z$ by the propositional formula $\text{GEQ}(v, z, k)$

Theorem 5. *Player 0 has a winning strategy in the energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$ if and only if the propositional logic formula ϕ'_Γ is satisfiable.*

Proof. (\Rightarrow) Let $G_\Gamma(\pi)$ be the graph induced by a winning strategy π for player 0 on the energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$. Consider the assignment α to the variables of ϕ_Γ defined as follows: for each boolean variable n_v (resp. $m_{(v,z)}$) let $\alpha(n_v) = 1$ (resp. $\alpha(m_{(v,z)}) = 1$) if and only if v is a node (resp. (v, z) is an edge) of $G_\Gamma(\pi)$. By Theorem 3, the assignment α satisfies $n_{v_0} \wedge \phi_0 \wedge \phi_1$. By Theorem 2, $G_\Gamma(\pi)$ admits a small progress measure function $f : W \rightarrow \mathcal{M}_{G_\Gamma(\pi)}$, where W is the set of vertices of $G_\Gamma(\pi)$. For each $(v, z) \in E$ such that $w(v, z) \geq 0$ (resp. $w(v, z) < 0$):

- let $\alpha(e_1^v), \dots, \alpha(e_k^v)$ be the boolean code of $f(v)$
- let $\alpha(e_1^z), \dots, \alpha(e_k^z)$ be the boolean code of $f(z)$
- let $\alpha(s_1^{(v,z)}), \dots, \alpha(s_k^{(v,z)}), \alpha(r_0^{(v,z)}), \dots, \alpha(r_k^{(v,z)})$ be the boolean code of the sum $f(v) + w(v, z)$ (resp. $f(z) + (-w(v, z))$) and the corresponding carry bits.

Since π is a winning strategy on Γ for player 0, the assignment α satisfies the propositional formula $\text{GEQ}(v, z, k)$. Therefore, $\alpha \models \phi'_\Gamma$.

(\Leftarrow) Suppose that α is a satisfying variable assignment of ϕ'_Γ . Define the following game $\Gamma' = (V', E', v_0, w', \langle V'_0, V'_1 \rangle)$: $v \in V'$ (resp. $(v, z) \in E'$) if and only if $\alpha(n_v) = 1$ (resp. $\alpha(m_{(v,z)}) = 1$) and for each $(v, z) \in E'$ let $w'(v, z) = w_{(v,z)}$. Since α satisfies $n_{v_0} \wedge \phi_0 \wedge \phi_1 \wedge \phi_\sigma$, we derive that Γ' is a non empty subgame of Γ . Hence, since α satisfies also ϕ_e , by Theorem 2 we deduce that $V' \subseteq W_0$ and Γ' induces a winning strategy for player 0 on Γ .

Theorem 6. *Given an energy game $\Gamma = (V, E, v_0, w, \langle V_0, V_1 \rangle)$, the size of the propositional logic formula ϕ'_Γ is $\mathcal{O}(|E| \cdot \lceil \log((V + 1) \cdot W) \rceil)$, even if ϕ'_Γ is required to be in CNF.*

5. CONCLUSIONS

We devise efficient encodings of the energy games problem into the satisfiability problem for formulas of difference logic and pure propositional logic in conjunctive normal form. Tight upper bounds on the sizes of the given reductions are also reported. Due to the success of nowadays SAT solvers in symbolic verification, the proposed encodings could result in algorithms that are very efficient in practice. Furthermore, they could open up new possibilities for devising tight bounds on the complexity of the energy games problem, as there are fragments of SAT that can be solved in polynomial time.

REFERENCES

- [1] A. Ehrenfeucht and J. Mycielski. International journal of game theory. *Positional Strategies for Mean-Payoff Games*, 8:109–113, 1979.
- [2] P. Bouyer, U. Fahrenberg, K. G. Larsen, N. Markey, and J. Srba. Infinite runs in weighted timed automata with energy constraints. In Proc. of FORMATS: *Formal Modeling and Analysis of Timed Systems*, LNCS 5215, pages 33–47. Springer, 2008.
- [3] A. Chakrabarti, L. de Alfaro, T. A. Henzinger, and M. Stoelinga. Resource interfaces. In *Proc. of EMSOFT: Embedded Software*, LNCS 2855, pages 117–133. Springer, 2003.
- [4] Y. Gurevich and L. Harrington. Trees, automata, and games. In Proc. of STOC: *Symposium on Theory of Computing*, pages 60–65. ACM, 1982.
- [5] Keijo Heljanko, Misa Keinänen, Martin Lange, and Ilkka Niemelä. Solving parity games by a reduction to sat. *J. Comput. Syst. Sci.*, 78(2):430–440, March 2012.
- [6] M. Jurdzinski. Small progress measures for solving parity games. In *Proceedings of STACS: Theoretical Aspects of Computer Science*, LNCS 1770, pages 290–301. Springer, 2000.
- [7] L. Brim, J. Chaloupka, L. Doyen, R. Gentilini, and J-F. Raskin. Faster algorithms for mean payoff games. *Formal Methods in System Design*, 38(2):97–118, 2011.
- [8] Moez Mahfoudh, Peter Niebert, Eugene Asarin, and Oded Maler. A satisfiability checker for difference logic. In *5-th Int. Symp. on the Theory and Applications of Satisfiability Testing*, 2002.

A CONTENT BASED WATERMARKING SCHEME USING RADIAL SYMMETRY TRANSFORM AND SINGULAR VALUE DECOMPOSITION

Lakehal Elkhamssa¹ and Benmohammed Mohamed²

¹LAMIE laboratory, Department of Computer Engineering,
Batna University, Algeria
lakehal_elkhamssa@yahoo.fr

²LIRE laboratory, Department of Computer Engineering,
Constantine University, Algeria
ben_moh123@yahoo.com

ABSTRACT

The Watermarking techniques represent actually a very important issue in digital multimedia content distribution. To protect digital multimedia content we embed an invisible watermark into images which facilitate the detection of different manipulations, duplication, illegitimate distributions of these images. In this paper we present an approach to embedding invisible watermarks into color images using a robust transform of images that is the Radial symmetry transform. The watermark is inserted in blocs of eight pixels large of the blue channel using the Singular Value Decomposition (SVD) of these blocs and those of the radial symmetry transform. The insertion in the blue channel is justified when we know that many works states that the human visual system is less sensible to perturbation in the blue channel of the image. Results obtained after tests show that the imperceptibility of the watermark using this approach is good and its robustness face to different attacks leads to think that the proposed approach is a very promising one.

KEYWORDS

Image watermarking, Singular value decomposition, Radial symmetry transform, Invisible watermarking

1. INTRODUCTION

Nowadays, multimedia content is largely used, due to the progressive development of new imaging techniques and devices. Users of these techniques and devices share generally these contents over communication channels that can be unsecure. This wide transmission of image content make it easily modified. These modifications disregard generally author's ownership that is gradually in danger because of this situation. So the multimedia content protection became a rigorous need to defend the author's copyright and to protect multimedia content from different illegal manipulations.

As a response to this need of protection come the digital watermarking of multimedia contents [1], [2]; which is no other than information hidden in the multimedia content such that a slight modification on the content results in a modification of the information hidden which is considered as a sign of unauthorized content's manipulation. A survey of watermarking techniques can be found in [3].

2. RELATED WORK

The watermarking method presented here belongs to the second generation ones. In such approaches image content is used in digital watermarking through the insertion of the watermark in specific points of image like points of interest, edges, corners or other features [4]. A feature-based watermarking scheme was first proposed in [5] where the authors use the Mexican Hat wavelet scale interaction to extract features in the image that can resist a series of attacks which makes them suitable to be used as emplacement to insert and to extract the watermark. Another content-based approach in [6] uses a reworked copy of the traditional Harris corner detector. The reworked copy calculates the corner response function within a circular window originate from the image centre and covers the largest area of it to resist image centre based rotation attacks. The new detector gives geometrically significant points that can detect possible geometric attacks. We find in [7] the use of interest points from the Harris corner detector to watermark synchronization before the extraction to recover the watermark positions that can be changed during a geometrical attack. To do, authors generate points of interest using the Harris detector after scale normalization in order to get the most stable points in the image by avoiding the sensitive character of the Harris detector face to scale changes. Then they search within a circular region around the detected points whether the detector response of a selected point reaches a local maximum or not. If it is the case they consider the point, otherwise they neglect it. After that they profit from the characteristic of Pseudo Zernike Moments magnitude which is invariant to rotation to design the watermark to embed in a rotation invariant pattern.

Another work [8] uses the robustness of interest's points to select the positions of the insertion which strengthen the relationship between the watermark and the image content. The detection of interest's points makes possible the creation of triangular partitions of the image and afterward the insertion of the watermark in each triangle. We note that image content is also used in watermarking systems in order to resist other geometric attacks like scale changing and translation in [9].

The watermarking system proposed in this paper uses image content through singular value decomposition (SVD). The use of SVD in watermarking schemes is not recent; we find in literature many watermarking schemes based on SVD decomposition [10], [11]. The SVD decomposition may be used in spatial or frequency domain. In frequency domain, the SVD decomposition can be combined with lifting wavelet transform like in [12]; it is also possible to combine it with the discrete cosines transform (DCT) as in [13] or many other transformations. The contribution of this work is the use of an SVD decomposition of a strong content transform (radial symmetry transform) to insert in the blue channel where the red and the green channels are just used to synchronize the detection of the watermark.

3. SCHEME OVERVIEW

3.1. Imperceptibility and Robustness

To be invisible to human visual eye, first, the mark is inserted in the blue channel, which is the one to which the human visual system is less sensitive [14]. Second, the watermark bits are

inserted in points of high interest in the image; those points with high luminance values make the modifications in the image imperceptible. That is the property of imperceptibility which marks any consistent watermarking scheme; the other property is the robustness of the system [15]. The robustness of our scheme is achieved by using robust transform of image information. This transform is called radial symmetry transform and it presents very good robustness against image transformations and noising. The radial symmetry transform is originally used to detect image features [16], [17].

3.2. The Radial Symmetry Transform

Author in [18] details an algorithm to calculate the radial symmetry transform of an image. With this approach a symmetry score is calculated from votes of one pixel to surrounding pixels. The transform is calculated in one or more radii n . the value of the transform at radius n indicates the contribution to radial symmetry of the gradients a distance n away from each point.

At each radius n , an orientation projection image O_n and a magnitude projection image M_n are formed. These images are generated by examining the gradient g at each point p from which a corresponding positively-affected pixel $p_{+ve}(p)$ and negatively-affected pixel $p_{-ve}(p)$ are determined, as shown in Fig. 1. The positively-affected pixel is defined as the pixel that the gradient vector $g(p)$ is pointing to, a distance n away from p , and the negatively-affected pixel is the pixel a distance n away that the gradient is pointing directly away from.

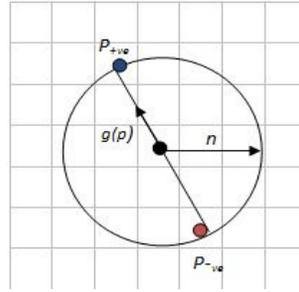


Figure 1. Positively and negatively affected pixels

Coordinates of the positively-affected pixel are given by:

$$p_{+ve} = p + ROUND \left(n \cdot \frac{g(p)}{\|g(p)\|} \right) \quad (1)$$

Coordinates of the negatively-affected pixel are given by:

$$p_{-ve} = p - ROUND \left(n \cdot \frac{g(p)}{\|g(p)\|} \right) \quad (2)$$

The orientation and magnitude projection images are initially zero. For each pair of affected pixels, the corresponding point p_{+ve} in the orientation projection image O_n and magnitude projection image M_n is incremented by 1 and $\|g(p)\|$, respectively, while the point corresponding to p_{-ve} is decremented by these same quantities in each image. That is:

$$O_n(p_{+ve}(p)) = O_n(p_{+ve}(p)) + 1 \quad (3)$$

$$O_n(p_{-ve}(p)) = O_n(p_{-ve}(p)) - 1 \quad (4)$$

$$M_n(p_{+ve}(p)) = M_n(p_{+ve}(p)) + \|g(p)\| \quad (5)$$

$$M_n(p_{-ve}(p)) = M_n(p_{-ve}(p)) - \|g(p)\| \quad (6)$$

The radial symmetry contribution at radius n is defined as the convolution:

$$S_n = F_n * A_n \quad (7)$$

$$\text{Where: } F_n(p) = \frac{M_n(p)}{k_n} \left(\frac{|\tilde{O}_n(p)|}{k_n} \right)^\alpha \quad (8)$$

$$\text{And } \tilde{O}_n(p) = \begin{cases} O_n(p) & \text{if } O_n < kn \\ k_n & \text{else} \end{cases} \quad (9)$$

A_n is a two-dimensional Gaussian, α is the radial strictness parameter, and k_n is a scaling factor that normalizes M_n and O_n across different radii.

The full transform is defined as the average of the symmetry contributions over all the radii considered:

$$S = \frac{1}{|N|} \sum_{n \in N} S_n \quad (10)$$

4. THE PROPOSED SCHEME

The proposed algorithm inserts the watermark bits into the blue channel of the image based on an SVD decomposition of this channel. The SVD decomposition used here is a bloc based one which necessitates the decomposition of the blue channel into blocs of 8 pixels large in different locations in the image the locations are objects centres obtained after a local maxima search over the radial symmetry transform.

At the same time we should calculate the symmetry transform using the red and green channel of the original image. The symmetry transform is also divided into blocks in the same locations as the blue channel and with the same size.

4.1. Insertion of the watermark

To insert the pixel of the binary watermark we need to calculate the *svd* of both, the blue block i and the symmetry transform of the block i as indicated in Figure 2.

The insertion mechanism uses the formula:

$$\sigma_{bk} = W_i * \sigma_{sk} * \alpha \quad (11)$$

Where k is the index of the singular value which holds the watermark.

α is the embedding strength (an adjustment parameter between the quality and the robustness). It is chosen experimentally.

σ_{bk} , σ_{sk} are respectively singular values from the range of big singular values of the blue bloc and those of the radial symmetry transform bloc.

4.2. Extraction of the watermark

Figure 3 presents the extraction scheme; it is based on the decomposition of blue image channel and radial symmetry transform into blocs, then the decomposition of each bloc to its singular values.

To extract the watermark we use the singular values following the formula:

$$\tilde{W}_i = \sigma_{bkw} / \sigma_{skw} / \beta \tag{12}$$

Where β is $1/2\alpha$, σ_{bkw} and σ_{skw} are respectively singular values from high singular values of the blue blocs and those of the radial symmetry transform blocs of watermarked image

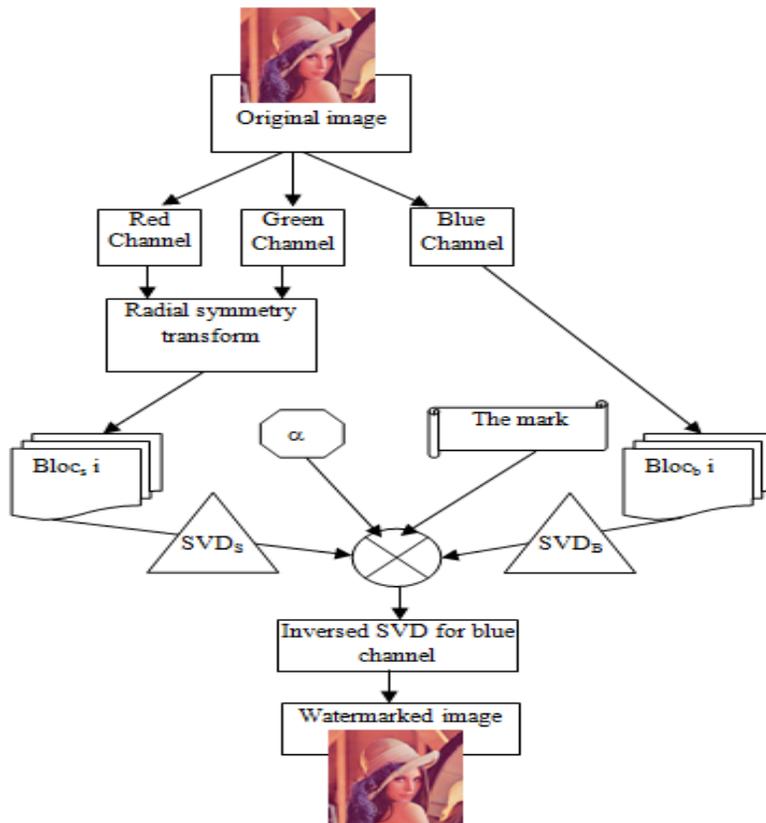


Figure 2. The mark insertion scheme

5. RESULTS AND DISCUSSION

In this section we present results of tests of imperceptibility and robustness for our scheme. To test our watermarking scheme we use an image database composed of color images of size 512 x 512 pixels: Lena, baboon, air plane, peppers. The mark used is a binary image of size 8 x 8 pixels.

Figure 4 presents the original images used in tests with the watermark image to the right bottom of the figure.

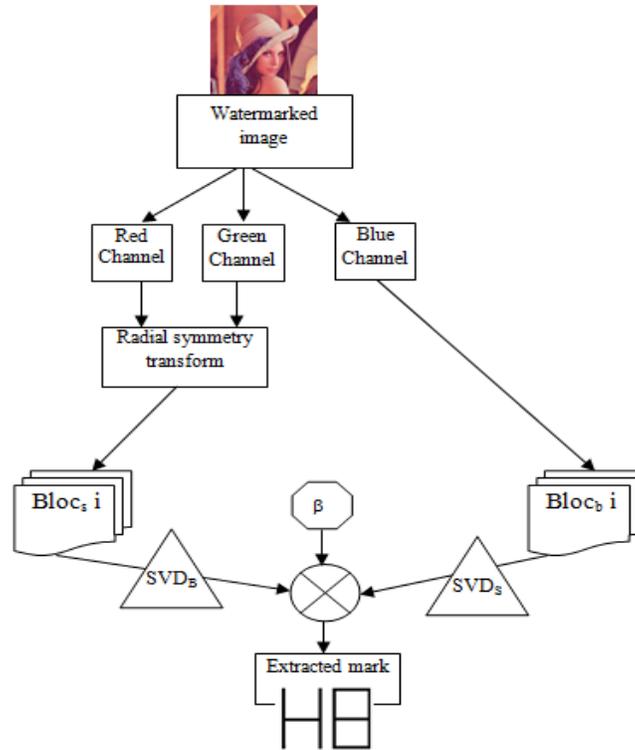


Figure 3. The mark extraction scheme



Figure 4. Test images with the watermark used

The table 1 presents the imperceptibility results according to three objective measures: PSNR (Peak Signal to Noise Ratio) calculated between the original image and the watermarked one, NC (Normal Correlation) and BCR (Bit Correct Ratio) calculated between the original watermark and the extracted one.

We note that the results obtained with the PSNR measure between original image and the watermarked one are very good when we know that authors considered a PSNR rate as good when it is greater than 30dB. With tested images we have PSNR from 42 to 46dB.

When investigating table 1 we find that the NC measure is between 0.97 and 1 when the BCR measure is between 98.44% and 100%. These are very interesting rates which indicates that the proposed algorithm makes a high-quality extraction of the watermark.

To evaluate the robustness of our algorithm we should test its performance face to different attacks. The attack chosen are common in watermarking algorithm evaluation and includes low pass filtering, noise, jpeg compression and cropping. In this paper, we have used attacks with different parameters to check as deeply as possible the robustness of the proposed algorithm. Attacks used to test our watermarking system are given in table 2.

In this paper we have not tested the capacity of our proposed scheme, but we claim that our algorithm has a very good capacity since it inserts watermark bits into radial symmetry transform local maxima.

Table 1. imperceptibility and robustness of the watermark in absence of attacks.

Image	PSNR	NC	BCR%
Lena	+ 46.13 dB	1	100
Baboon	+ 42.76 dB	0.97	98.44
Air Plane	+ 45.42 dB	0.97	98.44
Peppers	+ 45.42 dB	1	100

Table 2. Attacks used in different tests.

	Attacks	Parameters
SP1	Salt & Pepper noise	Density 0.008
SP2	Salt & Pepper noise	Density 0.002
SP3	Salt & Pepper noise	Density 0.05
GN1	Gaussian noise	M=0.0 V=0.001
GN2	Gaussian noise	M=0.1 V=0.001
GF	Gaussian filter	3x3
SH	Sharpening	3x3
HE	Histogram equalization	
MF	Median filter	3x3
JC1	JPEG Compression	Quality 80
JC2	JPEG Compression	Quality 60
CR	Cropping	1/8 of image
AF	Average filter	3x3
LF	Laplacian filter	3x3
Rot1	Rotation	0.2°

Table 3 presents the NC and BCR measure between the watermark inserted and the one extracted using our extraction scheme with different attacks over four test images.

By observing the interval of these two objective measures over all attacks (NC between 1 and 0.62, BCR between 100% and 71.88%) we can deduce that the proposed algorithm perform well face to these attacks. Furthermore, the robustness of the method is related to the tested image, thus to the content. Then, enhancing the content transform may enhance the extraction. The robustness of the algorithm face to the rotation attack can be reinforced using a robust rotational radial symmetry transform.

In the next figures (5-9) we present comparison graphics of NC rates between watermarks extracted with our algorithm and those presented in reference [12].

Algorithm in [12] uses gray scale images of 256 x 256 pixels where the watermark image is a binary image of 32 x 32 pixels which represents the letter "A".

Table 3. Robustness of the proposed scheme (NC and BCR) over different attacks

N°	Lena		Baboon		Air plane		Peppers	
	NC	BCR%	NC	BCR%	NC	BCR%	NC	BCR%
SP1	0.88	89.06	0.97	98.44	1.00	100	1.00	100
SP2	1.00	100	0.97	98.44	0.97	98.44	1.00	100
SP3	0.73	71.88	0.90	78.13	0.90	79.69	0.73	75.00
GN1	1.00	100	0.97	98.44	1.00	100	0.97	98.44
GN2	1.00	100	0.97	96.88	0.97	98.44	0.97	98.44
GF	1.00	100	0.97	98.44	0.97	98.44	1.00	100
SH	1.00	93.75	0.97	98.44	1.00	100	0.97	98.44
HE	1.00	100	0.97	98.44	0.91	95.31	0.84	92.19
MF	0.90	95.31	0.82	90.63	0.77	87.50	0.71	82.81
JC1	1.00	100	0.93	96.88	1.00	100	0.84	92.19
JC2	0.81	90.63	0.73	87.50	0.94	96.88	0.63	81.25
CR	0.93	95.31	0.98	98.44	0.93	95.31	1.00	100
AF	0.85	92.19	0.73	84.38	0.66	79.69	0.62	78.13
LF	0.71	85.94	0.90	95.31	0.90	95.31	0.80	90.63
Rot1	0.90	93.31	0.84	92.19	0.85	92.19	0.81	90.63

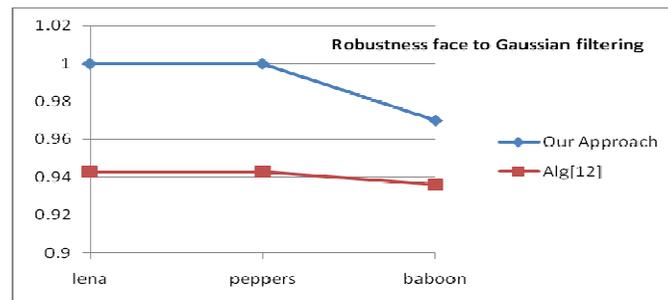


Figure 5. Comparison of NC rates face to gaussian filtering

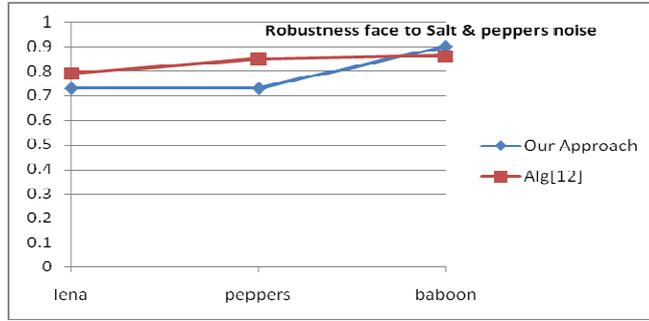


Figure 6. Comparison of NC rates face to salt & peppers noise (density 0.05)

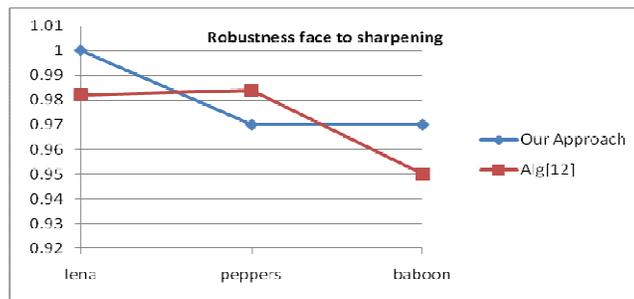


Figure 7. Comparison of NC rates face to sharpening attack

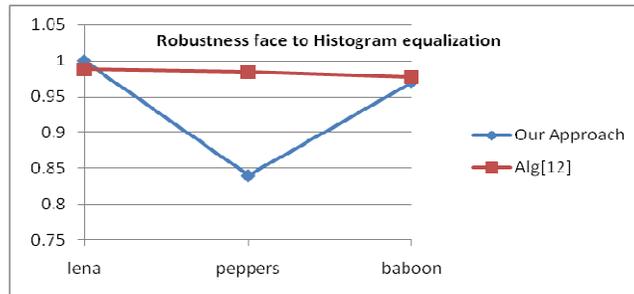


Figure 8. Comparison NC rates face to histogram equalization attack

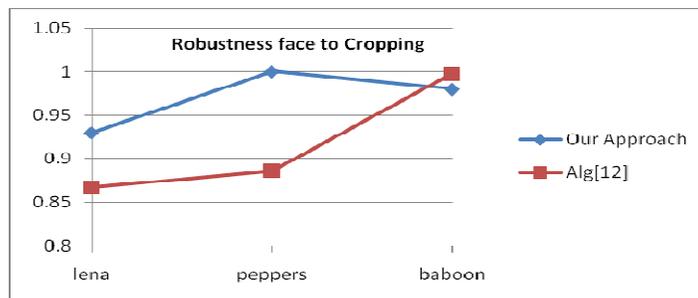


Figure 9. Comparison of NC rates face to cropping attack

6. CONCLUSION AND FURTHER WORKS

The watermarking algorithm presented in this paper is a promised one since it presents good results in term of robustness and imperceptibility. However, the approach necessitates to be compared to other works in order to be judged correctly. To this day there are no works which evaluate different content based algorithms face to different attacks. So the comparison of this work with other works should be based on the same attacks and considering the same metrics. Since it is the case, we resolve, in our future work, to evaluate and to compare the present work with other content based watermarking techniques using a unified testing benchmark.

ACKNOWLEDGEMENTS

Lakehal elkhamssa thanks Pr. Noui L. for his inestimable aid and considers his contribution to success this work.

REFERENCES

- [1] I. Cox, M. Miller, and J. Bloom. "Digital Watermarking: Principles & Practices". Morgan Kaufmann Publisher, San Francisco, USA, 2002.
- [2] C. I. Podilchuk and E. J. Delp, "Digital Watermarking: Algorithms and Applications," IEEE Signal Processing Magazine, pp. 33-46, 2001.
- [3] V. M. Potdar, S. Han and E. Chang , "A Survey of Digital Image Watermarking Techniques", 3rd international conference on industrial informatics (INDIN), 2005.
- [4] M. Kutter, S. K. Bhattacharjee, T. Ebrahimi, "Towards second generation watermarking schemes", international conference on image processing, 1999.
- [5] C. W. Tang and H. M. Hang, "A Feature-Based Robust Digital Image Watermarking Scheme", IEEE Transactions on Signal Processing, Vol. 51, pp. 950-959, 2003.
- [6] X. Qi and J. Qi, "A robust content-based digital image watermarking scheme," Signal Processing , Vol. 87, pp. 1264-1280, 2007.
- [7] L.D. Li, B. L. Guo AND L. Guo, "Combining Interest Point and Invariant Moment for Geometrically Robust Image Watermarking", Journal of information science and engineering, vol. 25, pp. 499-515, 2009.
- [8] P. Bas, J.M. Chassery, B. Macq, "Toward a content-based watermarking scheme", Journal of signal processing [journal de Traitement du Signal], vol. 19, pp. 11-17, 2002.
- [9] D. Simitopoulos, D.E. Koutsonanos, M.G. Strintzis, "Robust Image Watermarking Based on Generalized Radon Transformations", IEEE Trans. on Circuits and Systems for Video Technology, IEEE, vol. 13, pp. 732-745, 2003.
- [10] R. Liu and T. Tan, "A svd-based watermarking scheme for protecting right-ful ownership", IEEE Transactions on Multimedia, vol. 4, pp. 121-128, 2002.
- [11] C. Bergman, J. Davidson, "Unitary Embedding for Data Hiding with the SVD", Security, Steganography, and Watermarking of Multimedia Contents VII, SPIE vol. 5681, 2005.
- [12] K. Loukhaoukha, "Tatouage numérique des images dans le domaine des ondelettes basé sur la décomposition en valeurs singulières et l'optimisation multi-objective", Phd thesis, Laval University , pp. 115-126, 2010.
- [13] F. Huang, Z. H. Guan, "A hybrid SVD-DCT watermarking method based on LPSNR", Pattern Recognition Letters, vol. 25, pp. 1769-1775, 2004.
- [14] R. W. G. Hunt, The reproduction of color. Fountain press, Tolworth, England, 1988.
- [15] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques," Proceedings of the IEEE , vol. 87, pp. 1079-1107, 1999.
- [16] G. Heidemann, "Focus of Attention from Local Color Symmetries" IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 26, pp. 817-830, 2004.
- [17] A. Rebai, A. Joly, N. Boujemaa, "Constant Tangential Angle Elected Interest Points," in Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR), pp. 203-211, 2006.
- [18] A. Zelinsky, G. Loy, "Fast Radial Symmetry for Detecting Point of Interest," IEEE Trans. on Pattern Analysis and Machine Intelligence, pp. 959-973, vol. 27, August 2003.

AUTHOR

Lakehal elkhamssa is currently a researcher in the LAMIE laboratory, Batna University, Algeria. She received his Magister degree in computer science in the University of Batna in 2009. Her main research interests are: image watermarking, image understanding and colour constancy.



INTENTIONAL BLANK

ANALYSIS OF INTRADAY TRADING OF INDEX OPTION IN KOREAN OPTION MARKET

Young-Hoon Ko

Department of Computer Engineering, HyupSung University, South Korea

ABSTRACT

The option market in South Korea began on 7 July 1997. After then, the amount of option market has increased steeply. In these days, average daily payments is beyond 1 trillion won.

It is impossible to predict the market. But using the statistics, investors can get a profit steadily.

The open interest contracts of index future has increased over 4000 after start time of a day and decrease down to about 0 when closing time.

As for this characteristics of index future, Ko^[1] suggested the volatility strategy and brought the result of simulation with the profit of 1.07 % per a day. This profit comes to real if an investor finds a brokerage firm with low commissions.

This paper suggests another strategy. The price of options consists of time value and intrinsic value. And the fall of index future is faster than rising. Therefore velocity of moving index cause the price of options. The simulation results give a fascinating fact that put option tends to increase in the morning and call option tends to increase in the afternoon.

With this velocity strategy, investors get the profit 1.4% per a day except commissions of 0.15% per one trade.

KEYWORDS

Automatic Trading System, Volatility Strategy, Velocity Strategy, Open interest Contract

1. INTRODUCTION

The option market in South Korea began on 7 July 1997. An underlying asset of option market is KOSPI200 index which consist of 200 superior symbols in KOSPI. Stock index future began on 3 May 1996.

According to the statistics of Korea exchange(KRX), the average daily payments of index options was just 50 billion won in January 2000, it had increased and reached up 500 billion won in March 2002. And then it reached 1 trillion won in June 2007. It was up to 2.5 trillion won in August 2011. In these days it is about 1~1.5 trillion won

Monthly average daily trading payments is shown in the following graph.

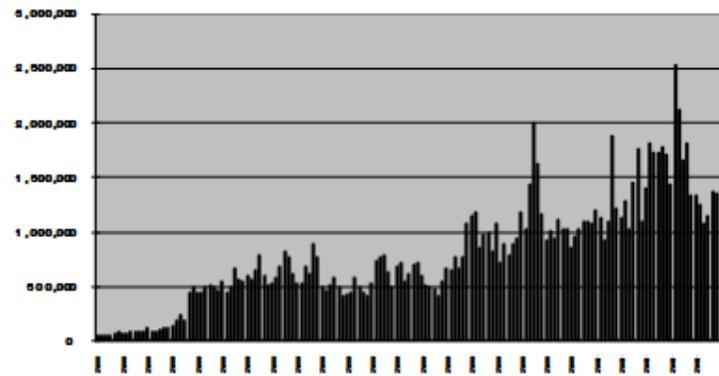


Figure 1. Average daily trading payments (million won)

Call-put ratio is shown in the following graph. The volume of call option is more than put option's in early 2000 decade. But after 2003, there is a trend that the volume of put option has increased more than call option's. Especially, in the months with a big fall of index, the volume of put option is 50% more than call option's.

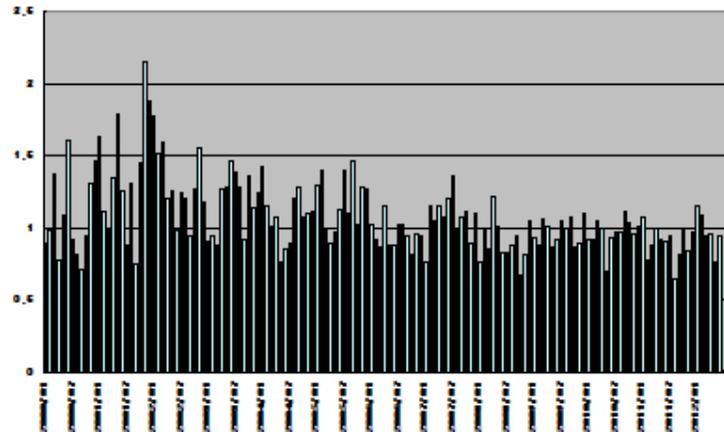


Figure 2. Call-put ratio by average daily trading contracts

Option is in the derivative market, It is not to trade stock such as real asset but it makes contracts mutually by abstract index of stocks. This type of contracts must be liquidated. The amount of contracts which is not yet liquidated is called open interest contracts.

It is impossible to predict the market. If the investors can predict the future market, the market cannot be established because everyone wants to trade one side.

However, in stock market there is sure one thing. It is that every contract has its own expiration date.

2. OPEN INTEREST CONTRACTS

Since 2008, nearly two year average of open interest contracts of index futures showed as a followed graph.

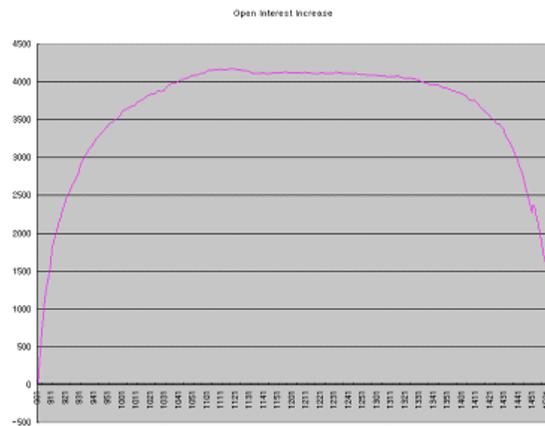


Figure 3. OID(average open interest contracts difference from the start time) (2-year)

As for minute candles, the difference between the close value of a candle and the close value of first candle at 9:00 is considered as the y-axis value. OID is used as abbreviation for open interest contracts difference from the start time of a day.

The OID at 9:00 is 0 and it is increasing over 4000 by around 10:40 and it reduces down to about 0 until closing time at 15:00. This curve appears due to the nature of the intraday trading of stock index futures.

Stock index futures and options trading is held from 9:00 to 15:15. From 15:05 to 15:15 is progressed simultaneously and investors can not know the asking and bid price, purchase orders and sell orders determine the closing price at 15:15.

Stocks that is an underlying asset of futures are trading from 9:00 to 15:00.

Therefore, the trade after 15:00 until 9:00 the following day will not be able. If some event happens during this period, it will make a big gap at start of the next trade time. the event of disaster can result in huge losses to position investors who sold options and held it overnight.

For this reason, a lot of unpredictability, most investors are trying to avoid the risk of overnight. Theoretically, option sell, especially in the case of possible loss of the margins, often wearing heavy losses occur.

Thus, investors in order to avoid overnight risk of the day tend to focus on daily marketing, which open interest contracts of index futures increase and decrease is illustrated look.

Next graph shows OID with the longer average duration of five years. The shape is almost same but a little difference occurs as the table below. Second graph include late two years, so it shows trend of recent characteristics.

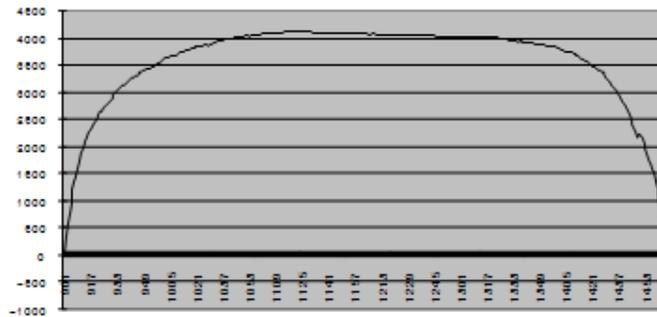


Figure 4. OID (5-year)

The times of accrossing above 4000 are the same. They are 10:40. The times of accrossing down below 4000 are 13:33 and 13:28 respectively. It means that nowadays investors tend to follow the market faster. The times of maximum open interest are the same which is 11:22 and the maximum contracts are 4138.55 both. This means the power of super investors who lead the market is not changed. But the investors who participate the market acclimate themselves to the market.

Table 1. Comparison between two OID curves

Average duration	2-year	5-year
last date	10-05-31	12-04-13
Time Above 4000	10:40	10:40
Time Under 4000	13:33	13:28
Time Maximum	11:22	11:22
Max Contracts	4138.55	4138.55

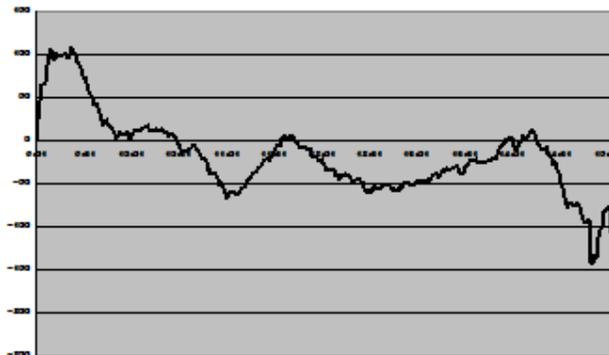


Figure 5. Difference between two open interest curves

The chart above shows the difference between two-year average graph and five-year average graph. In other words, this graph shows the recent trend of investors. From 9:00 to 10:00 the difference is rising up to 100. This means that the super investors who lead the market try to handle faster. And when about 11:00 and 13:00, the difference decrease down to 50. This means the weak investors tend to acclimate themselves to the market. After 14:30 the difference fall to 200. This means investors come to waive nearly the closing time.

3. VOLATILITY STRATEGY

Volatility strategy is based on the hypothesis that OID is relate to volatility. This strategy was suggested by Ko^[1] in 2010.

If the OID is increasing before the noon, the volatility of option market is also increased. And then if the OID is reducing after the noon, the volatility is also reduced.

With this properties, the two strategies are proposed. Buythensell strategy makes long straddle/strangle at 9:00 and liquidates it at 12:00. Sellthenbuy strategy makes short straddle/strangle at 12:00 and then also liquidates it near closing time, 14:40. For avoiding the volatility of near closing time, the time of liquidation of sellthenbuy strategy is selected at 14:40.



Figure 6. Strategy charts using Signals

To ensure the profitability of Buythensell, Sellthenbuy strategies, from 23 March 2009 to 31 May 2010 for 301-trading days are selected for an experiment.

These strategies are very simple. Buythensell strategy is buy both call and put option at 9:00 and liquidates both options at 12:00. Sellthenbuy strategy sell both call and put option at 12:00 and liquidates both options at 14:40. So after 15:00 there is no contract left and this features guarantee avoiding the high overnight risk.

Table 2. Profits of volatility

	Volatility strategy		
	total	BuythenSell	SellthenBuy
ATM	1.07%	0.51%	0.56%
OTM 2.5	0.94%	0.41%	0.53%
OTM 5	0.68%	0.21%	0.47%
OTM 7.5	0.53%	0.15%	0.38%
OTM 10	0.44%	0.14%	0.30%

As for the result of exercise price in above table, the ATM has the biggest profit. This means that the option at the money which is near the index is most influenced by super investors. In Korea, most of the trading in ATM is by foreigners which are regarded as super investors.

4. VELOCITY STRATEGY

Volatility strategy is due to the hypothesis that the OID is related to the volatility. But if you review the real data, you would know the fact that the call and put options show some different characteristics. So this paper is worked up more in detail about this.

It is clear that the OID influence the volatility of options. But something that is more influenced by OID may exist.

Another hypothesis is that the OID influence the velocity of index. If the OID increase, the super investors try to move index into a certain level. And weak investors resist this pressure. And if the OID decrease, the leading investors win this try or find compromise index. And then the weak investors waive and adapt themselves to market.

The move of index has two direction. Because of the characteristics of stock, rising of index is slower but falling of index is faster relatively.

This characteristics make call option stagnate because of slow rising of index. But this makes put option rise rapidly.

As for the average data of long days enough, call options tend to decrease and put options tend to increase when the OID increase. This characteristics can be confirmed by simulation of real data.

When the OID decrease, put option which has risen more than its value finds the appropriate price. And call option which has been compressed is easy to increase slightly.

By considering this characteristics, this paper suggest velocity strategy which consists of Morning strategy and Afternoon strategy.

Morning strategy is that buying put option and selling call option at 9:00. And they are liquidated by 12:00.

Afternoon strategy is that selling put option and buying call option at 12:00. And they are liquidated by 14:40.

Table 3. Profits of velocity strategies

	Velocity strategy		
	total profit	Morning strategy	Afternoon strategy
ATM	2.64%	1.11%	1.53%
OTM 2.5	1.95%	0.73%	1.22%
OTM 5	1.26%	0.41%	0.85%
OTM 7.5	1.08%	0.35%	0.73%
OTM 10	0.64%	0.18%	0.46%

Table 3 shows the result of simulation of velocity strategy. In ATM, the profit is 2.64% per a day. This amount is incredibly big because cumulative profit is enormous.

5. EXPERIMENTS AND RESULTS

Ko^[1] brought the results of his simulation from 23 March 2009 to 31 May 2010 . This duration is a little over one year and 301 trading days. Symbols for experiments are based on the ATM, OTM of 2.5 points, OTM of 5 points, OTM of 7.5 points, OTM of 10 points.

In this paper the simulation duration is expended. From 23 March 2009 to 13 April 2012 is selected for the experiment. This duration is 769 trading days.

The straddle/strangle trade consists of both call and put options. But this paper simulates individual option separately. So when straddle get profit, the result shows which part of call and put options brings the profit.

Table 4. Simulation result of Call and Put respectively

	strategy		type	trades	profit	Rate profitable	MDD
ATM	BuyThenSell	09:00-12:00	Call	769	-1,321,000	42	-2,210,000
			Put	769	4,650,000	47	-1,754,000
		12:00-14:40	Call	769	1,662,000	46.42	-1,261,000
			Put	769	-6,549,000	37.45	-7,283,000
	SellThenBuy	09:00-12:00	Call	769	1,321,000	55.79	-1,816,000
			Put	769	-4,650,000	51.11	-5,417,000
12:00-14:40		Call	769	-1,662,000	50.2	-2,510,000	
		Put	769	6,549,000	60.08	-1,940,000	
OTM1	BuyThenSell	09:00-12:00	Call	769	-563,000	39.14	-1,508,000
			Put	769	3,380,000	45.38	-1,529,000
		12:00-14:40	Call	769	1,107,000	44.47	-863,000
			Put	769	-5,445,000	36.28	-6,013,000
	SellThenBuy	09:00-12:00	Call	769	563,000	58.91	-1,600,000
			Put	769	-3,380,000	52.28	-4,189,000
12:00-14:40		Call	769	-1,107,000	51.89	-1,780,000	
		Put	769	5,445,000	60.47	-1,816,000	
OTM2	BuyThenSell	09:00-12:00	Call	769	-84,000	38.1	-1,078,000
			Put	769	2,129,000	43.95	-1,325,000
		12:00-14:40	Call	769	429,000	41.87	-757,000
			Put	769	-4,135,000	34.07	-4,553,000
	SellThenBuy	09:00-12:00	Call	769	84,000	59.04	-1,426,000
			Put	769	-2,129,000	53.45	-3,062,000
12:00-14:40		Call	769	-429,000	51.24	-1,124,000	
		Put	769	4,135,000	60.86	-1,575,000	
OTM3	BuyThenSell	09:00-12:00	Call	769	-138,000	36.54	-872,000
			Put	767	1,766,000	41.98	-1,125,000
		12:00-14:40	Call	769	508,000	39.4	-604,000
			Put	767	-3,399,000	34.29	-3,693,000
	SellThenBuy	09:00-12:00	Call	769	138,000	57.35	-1,024,000
			Put	767	-1,766,000	52.93	-2,508,000
12:00-14:40		Call	769	-508,000	51.11	-1,011,000	
		Put	767	3,399,000	59.97	-1,328,000	
OTM4	BuyThenSell	09:00-12:00	Call	768	-255,000	33.85	-757,000
			Put	766	723,000	40.08	-911,000
		12:00-14:40	Call	768	615,000	36.33	-500,000

	SellThenBuy	09:00-12:00	Put	766	-1,837,000	33.03	-2,118,000
			Call	768	255,000	53.91	-732,000
			Put	766	-723,000	53.39	-1,412,000
		12:00-14:40	Call	768	-615,000	47.4	-970,000
			Put	766	1,837,000	57.96	-1,215,000

The simulation result shows that put options tend to increase in the morning and decrease in the afternoon. In contrast, call options tend to decrease in the morning and increase in the afternoon.

Table 5. Profits of volatility strategies (old)

	BuythenSell strategy			SellthenBuy strategy		
	total profit	average daily profit	daily profit rate	total profit	average daily profit	daily profit rate
ATM	1,071,000	3,558	0.51%	1,244,000	4,133	0.56%
OTM 2.5	63,000	2,867	0.41%	1,167,000	3,877	0.53%
OTM 5	39,000	1,458	0.21%	1,040,000	3,455	0.47%
OTM 7.5	21,000	1,066	0.15%	827,000	2,748	0.38%
OTM 10	87,000	953	0.14%	650,000	2,159	0.30%

Table 6. Profits of volatility strategies (updated)

	BuythenSell strategy			SellthenBuy strategy		
	total profit	average daily profit	daily profit rate	total profit	average daily profit	daily profit rate
ATM	3,329,000	4,329	0.62%	4,887,000	6,355	0.91%
OTM 2.5	2,817,000	3,663	0.53%	4,338,000	5,641	0.81%
OTM 5	2,045,000	2,659	0.38%	3,706,000	4,819	0.69%
OTM 7.5	1,628,000	2,117	0.30%	2,891,000	3,759	0.54%
OTM 10	468,000	608	0.09%	1,222,000	1,589	0.23%

Above two tables show a difference of their simulation duration. Table 2 shows the recent trend. The difference is clear. In recent days, ATM has more influence rather than other OTMs. So the profit of ATM becomes bigger than other OTMs.

Table 7. Profits of velocity strategies

	Morning strategy			Afternoon strategy		
	total profit	average daily profit	daily profit rate	total profit	average daily profit	daily profit rate
ATM	5,971,000	7,765	1.11%	8,211,000	10,677	1.53%
OTM 2.5	3,943,000	5,127	0.73%	6,552,000	8,520	1.22%
OTM 5	2,213,000	2,877	0.41%	4,564,000	6,777	0.85%
OTM 7.5	1,904,000	2,476	0.35%	3,907,000	5,811	0.73%
OTM 10	978,000	1,271	0.18%	2,452,000	3,430	0.46%

In velocity strategy, BuythenSell is changed to Morning strategy and SellthenBuy is changed to Afternoon strategy. The initial time and liquidated time are the same. But selling is changed to buying in only call option.

6. CONCLUSION

The derivative market is made for hedging of stock market. For the purpose, it has a big leverage. This fact gives the investors chances to get profit. However it also makes speculative investors become bankrupt.

The prediction of market is not possible. If it is possible, all the investors go to one side that the trade is not established.

But the only fact of derivative products is that every product has its own expiration date. All the symbols including future, various options have its own expiration date when the open interest contracts must be liquidated.

An average curve of open interest contracts difference from the start shows that it has a regular pattern. This pattern is the fact and truth.

Ko suggested his paper that this pattern relate to the volatility. And suggested strategy get profit about 1.07% a day. But this amount of profit is just same to commission of 8 trades a day. If you have luck to find commission about 0.1% a trade, You can get profit.

This paper suggest another strategy. The pattern relate to the velocity. Time concept is very important to options. Because the price of option consist of time value and intrinsic value. The moving of stock index is differencet as rising and fall. The fall of stock index is faster than rising. This characteristics give call and put options a little bit different moving. In long time average, this characteristic is clear and investors can get profit with this characteristics.

As a simulations in the 517 trading day, The result gives that the amount of profit ATM is 1.44% except commissions of 0.15% a day.

And further research, it is very interesting to find a selection method of day which apply this strategy. This selection method gives surely more profit over 1.44% a day.

REFERENCES

- [1] Ko Young Hoon, Analysis of Straddle trading strategy for KOSPI200 Stock index, Pan-Pacific Journal of Business Research, Vol 1. No 2. 2010.
- [2] Balsara, Nauzer, Money Management Strategies for Futures Traders, John Wkley & Sons. NewYork, 1998, pp. 276.
- [3] Ko Young Hoon, Kim Yoon Sang, "Study on the performance analysis of push-pull strategy by Multicharts' Portpolio", journal on IWIT, Dec. 2010, 317-324
- [4] Ko Young Hoon, Kim Yoon Sang, "A design of automatic trading system by dynamic symbol using global variables", journal on KSDIM, Sep. 2010, 211~219.
- [5] Ko Young Hoon, Kim Yoon Sang, "The profit analysis of straddle sell by entry-time and delta at system trading", journal on KSDIM, May 2010, 151~157.
- [6] Ko Young Hoon, "MultiCharts multi-entry strategy for a portfolio of signal conversion system design", Software Engineering Institute of Society, Vol. 22, No. 1, 2009, pp. 44~52.
- [7] Kang Suchul, Kim HeeChul, Investra - System Trading Strategies, Bumhan Books, 2004.
- [8] Lukac and Brorsen, "The Usefulness of Historical Data in Selecting Parameters for Technical Trading Systems", The Journal of Futures Markets, John Wiley & Sons, 1993. pp. 55~59.

- [9] Kim Jungyoung, System trading by technical index, Truth Search, 2001.
[10] Balsara, Nauzer, Money Management Strategies for Futures Traders, John Wkley & Sons. 1998

AUTHOR

Young Hoon Ko He obtained B.S., M.S., and Ph.D. degrees in Electronics Engineering from Yonsei University, Seoul, South Korea, in 1991, 1993, and 1997, respectively. He was a member of the Invited professors of Chungbuk University from 1997 to 1999. Since March 1999, he has been a Professor at the department of Computer Engineering, Hyupsung University, Hwasung, South Korea.



QUALITY OF SERVICE MANAGEMENT IN DISTRIBUTED FEEDBACK CONTROL SCHEDULING ARCHITECTURE USING DIFFERENT REPLICATION POLICIES

Malek Ben Salem¹, Emna Bouazizi¹, Rafik Bouaziz¹, Claude Duvalet²

¹MIRACL, HI of Computer Science and Multimedia, Sfax University, Tunisia

²LITIS, Université du Havre, 25 rue Philippe Lebon, BP 540, F-76058 ,
Le Havre Cedex, France

ABSTRACT

In our days, there are many real-time applications that use data which are geographically dispersed. The distributed real-time database management systems (DRTDBMS) have been used to manage large amount of distributed data while meeting the stringent temporal requirements in real-time applications. Providing Quality of Service (QoS) guarantees in DRTDBMSs is a challenging task. To address this problem, different research works are often based on distributed feedback control real-time scheduling architecture (DFCSA). Data replication is an efficient method to help DRTDBMS meeting the stringent temporal requirements of real-time applications. In literature, many works have designed algorithms that provide QoS guarantees in distributed real-time databases using only full temporal data replication policy.

In this paper, we have applied two data replication policies in a distributed feedback control scheduling architecture to manage a QoS performance for DRTDBMS. The first proposed data replication policy is called semi-total replication, and the second is called partial replication policy.

KEYWORDS

DRTDBMS, Quality of Service, Partial Replication, Semi-Total Replication.

1. INTRODUCTION

Distributed real-time database systems have a wide range of applications such as mobile and wireless communications, distance education, e-commerce treatment and industrial control applications as well as the control of nuclear reactions.

The DRTDBMS include a collection of sites interconnected via communication networks for distributed transactions processing. Each site includes one RTDBMS. In distributed environment, the presence of many sites raises issues that are not present in centralized systems.

In an application involving the use of DRTDBMS, user requests arrive at varying frequencies. When the frequency increases, DRTDBMS balance will be affected. During overload periods, DRTDBMS will potentially have fewer resources, and then real-time transactions will miss their

deadlines. To address this problem, more studies focus on feedback control techniques have been proposed [1,5] to provide better QoS guarantees in replicated environment.

Data replication consists on replicating a data on participating nodes. Given that this technique increases the data availability, it can help DRTDBMS to meet the stringent temporal requirements. In literature, two data replication policies are presented: the full data replication policy and the partial data replication policy [8].

Wei et al [8] have proposed a QoS management algorithm in distributed real-time databases with full temporal data replication. They proposed a replication model in which only real-time data are replicated and the updating replicas data is propagate at the same time to all replicas in each other nodes. In replicated environment, user operations on data replicas are often read operations [8]. In this case, to guarantee the replicas data freshness, many efficient replica management algorithms have been added in literature. All proposed algorithms aim to meet deadlines of transactions and data freshness guarantees.

This proposed solution in [8] is shown to be appropriate for a fixed number of participating nodes, e.g., 8 nodes. However, in the presence of a high number of nodes (more than 8 nodes), using a full data replication policy is inefficient in these systems, and it may have many limitations. Those issues are related to communication costs between nodes, highly message loss and periodically collected data performance.

To address this problem, we proposed in this work a new data replication policy called a semi-total data replication policy and, we have applied it in feedback control scheduling architecture in replicated environment. Furthermore, we have applied the partial replication policy in this architecture, and we have presented a comparison between obtained results with these data replication policies and the existing results with full data replication policy proposed in previous work. Compared to previous work [8], we increment the number of nodes. Also, in our replication model, we proposed to replicate both types of data; the classical data and the temporal data. In this model, temporal replicas data are updating transparently, and the classical replicas data are updating according to the RT-RCP policy presented in [3].

The main objective of our work is to limit the miss ratio of the arrived transactions to the system. At the same time, our work aims to tolerate a certain imprecise value of replicas data and to control timely transactions, which must guarantee access only to fresh replicas data, even in the presence of unpredictable workloads.

In this article, we begin by presenting a distributed real-time database model. In the section 3, we present the related works on which, we base our approach. In section 4, we present the proposed QoS management approach based on DFCS architecture to provide QoS guarantees in DRTDBMS. Section 5 shows the details of the simulation and the evaluation results. We conclude this article by discussing this work and by focusing on its major perspectives.

2. RELATED WORK

In this section, we present the QoS management architecture proposed in [3], on which we based our work. We describe, also, an overview of the data replication policies presented in literature, which is used for the QoS enhancement.

2.1. Architecture for QoS management in DRTDBMS

The QoS can be considered as a metric which measures the overall system performance. In fact, QoS is a collective measure of the service level provided to the customer. It can be evaluated by different performance criteria that include basic availability, error rate, response time, the rate of successful transactions before their deadline, etc.

The DRTDBMS as RTDBMS have to maintain both the logical consistency of the database and its temporal consistency. Since it seems to be difficult to reach these two goals simultaneously because of the lack of predictability of such systems, some researchers have designed new techniques to manage real-time transactions. These techniques use feedback control scheduling theory in order to provide an acceptable DRTDBMS behaviour. They also attempt to provide a fixed QoS by considering the data and the transactions behaviour.

In RTDBMS, many works are presented to provide a QoS guarantees. Those works consist of the applicability of most of the management techniques of the real-time transactions and/or real-time data [1,2,9,11].

A significant contribution on QoS guarantees for real-time data services in mid-size scale of DRTDBMS is presented by Wei et al. [8] (cf. Figure 1). The authors have designed an architecture, which we base our work, that provides QoS guarantees in DRTDBMS with full replication of only temporal data with small number of nodes (8 nodes). This architecture, called Distributed FCSA (DFCSA), consists of heuristic Feedback-based local controllers and global load balancers (GLB) working at each site.

The general outline of the DFCSA is shown in Figure 1. In what follows, we give a brief description of its basic components.

The admission controller is used to regulate the system workload in order to prevent its overloading. Its functioning is based on the estimated CPU utilization and the target utilization set point. For each admitted transaction, its estimated execution time is added to the estimated CPU utilization. Therefore, transactions will be discarded from the system if the estimated CPU utilization is higher than the target utilization set by the local controller.

The transaction manager handles the execution of transactions. It is composed of a concurrency controller (CC), a freshness manager (FM), a data manager (DM) and a replica manager (RM).

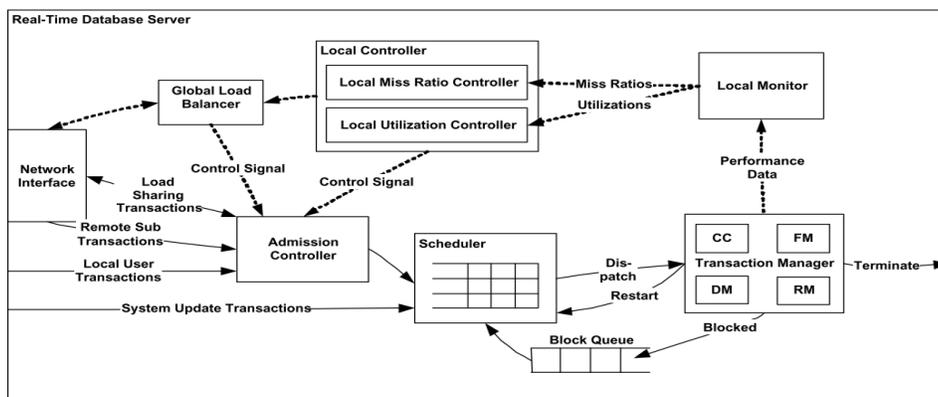


Figure 1 FCS architecture for QoS guarantee in DRTDBMS [8].

2.2. Distributed real-time database model

In this section, we present the distributed real-time database model on which we base our work. This model is issued from many works about QoS management in DRTDBMS [6]. The main difference is the applicability of different data replication policies. In our works, DRTDBMS model is defined by the interconnection between many centralized RTDBMS. We consider a main memory database model on each site, in which the CPU is the main system resource taken into account.

2.2.1. Data model

In this model, data objects are classified into either real-time or non real-time data. In our model, we consider both types of data objects. Non real-time data are classical data found in conventional databases, whereas real-time data have a validity interval beyond which they become stale. These data may change continuously to reflect the real world state (e.g., the current temperature value). Each real-time data has a timestamps which define its last observation in the real world state [9], a validity interval and a value.

2.2.2. Data replication model

In DRTDBMS, data replication is very attractive in order to increase the chance of distributed real-time transaction to meet its deadline, system throughput and provide fault-tolerance. However, it is a challenge to keep data copies consistent. For that purpose, two types of data replication policies have been developed in literature.

The first data replication policy is the full data replication; the entire database at each node is replicated to all other sites which admit transactions that can use their replicas data. Therefore, all the data of this database will be available in different sites which facilitates access by the various local and remote transactions.

The second policy data replication is the partial data replication policy, which is based on access history of all transactions in each node. In fact, for each node, if the number of access on one data object by current transactions reach a threshold then current node request to replicate locally this data object. Therefore, this policy consist of replicate the most accessed data object of the most accessed nodes to satisfy various user requests.

In this paper, we present the third policy. We called it the "Semi-total data replication". Wei et al. [8] have proposed a replication model that only temporal data had replicated and the updating replicas data is propagate at the same time to all replicas in each other nodes. In the replicated environment, the frequently user operations on data replicas are read operations. In this case, to guarantee the replicas data freshness, many efficient replica management algorithms was added in literature to manage the data replicas at every node that supported data replication. All proposed algorithms aims to meet transactions deadlines and data freshness guarantees.

Here, we discuss the difference between the two data replication policies. The full replication is characterized by a maximum number of replicated data. Then, full database replication means that all sites store copies of all data items. An analytical study in [12] has shown the scalability limits of full replication. Therefore, the time needed for updating replicated data is quite important. In some protocols update transactions are executed to preserve all the databases consistency. By taking into consideration the time of transmission of messages between sites, the chance to respect distributed real-time transactions decreases. However, partial data replication policy only assigns copies of a frequently accessed data item. When there is a high update workload, full

replication has too much overhead to keep all copies consistent and the individual replicas have little resources left to execute read operations. In contrast, with partial replication, updating protocols for replica data only has to execute the updates for mostly accessed replica data items, and thus, it has more potential to execute read operations.

In the next section, we present our QoS management algorithms based on a different data replication policy.

2.2.3. Transaction model

In this model, we use the firm transaction model, in which tardy transactions are aborted because they can't meet their deadlines. In fact, transactions are classified into two classes: update transactions and user transactions. Update transactions are used to update the values of real-time data in order to reflect the real world state. These transactions are executed periodically and have only to write real-time data. User transactions, representing user requests, arrive aperiodically. They may read real-time data, and read or write non real-time data.

Furthermore, each update transaction is composed only by one write operation on real-time data. User transactions are composed of a set of sub-transactions which are executed at local node or at remote nodes that participate in the execution of the global system.

We consider that a user transaction may arrive at any node of the global system and define its data needs. If all data needed by the transaction exist at the current site, then the transaction is executed locally. Otherwise, the transaction, called real-time distributed user transaction, is split into sub-transactions according to the location of their data. Those sub-transactions are transferred and executed at corresponding nodes.

There are two sub-types of real-time distributed user transaction: remote and local [3]. Remote transactions are executed at more than one node, whereas the local transactions are executed only at one node. There is one process called coordinator which is executed at the site where the transaction is submitted (master node). There is also a set of other processes called cohorts that execute on behalf of the transaction at other sites accessed by the transaction (cohort node). The transaction is an atomic unit of work, which is either entirely complete or not at all. Hence, a distributed commit protocol is needed to guarantee uniform commitment of distributed transaction execution. The commit operation implies that the transaction is successful, and hence all of its updates should be incorporated into the database permanently. An abort operation indicates that the transaction has failed, and hence requires the database management system to cancel or abolish all of its effects in the database system. In short, a transaction is an all or nothing unit of execution.

2.3. Performance metrics

The main performance metric, we consider in our model, is the *Success Ratio (SR)*. It is a QoS parameter which measures the percentage of transactions that meet their deadlines. It is defined as follows:

$$SR = 100 \times \frac{\# Timely}{\# Timely + \# Tardy} (\%)$$

Where #timely and #tardy represent, respectively, the number of transactions that have met and missed their deadlines.

3. QOS MANAGEMENT APPROACHES USING DIFFERENT DATA REPLICATION POLICIES

In this paper, the number of nodes is larger than the number used in experiments in [8] for full data replication policy. We have also proposed to apply other data replication policies that dedicated to mid-size scale system.

Our work consists of a new approach to enhance the QoS in DRTDBMS. We apply two data replication policies in conventional DFCS architecture; the semi-total replication and partial replication, on the DFCS architecture using a greater number of nodes (16 nodes) of overall system than classical DFCS architecture. Then, we have compared obtained results with those both data replication policies and the result obtained with full data replication policy used in [8]. Moreover, in our replication model, we proposed to replicate both types of data; the classical data and the temporal data. In this model, temporal replicas data are updating transparently, and the classical replicas data are updating as RT-RCP protocol presented by Haj Said et al. [3].

3.1 Approach using semi-total replication policy

In the first part, we propose a new algorithm replication called semi-total replication of real-time and non real-time data (cf. Algorithm 1). In this proposed data replication policy, the system execution start running without any database replication. Distributed real-time transactions require data located in local or in remote nodes. Each node define an access counter for transactions which require data located on remote sites which is define by an access counter of mostly accessed sites. In case where the number of accessed remote data at each node n_i reaches a maximum threshold, then, the current node request the full replication of the database from node n_i . In other case, if the number of accessed remote data at each node n_i reaches a minimum threshold, then the current node decide to remove the full replicated database from node n_i . The maximum and the minimum threshold parameters are defined as input parameter to the system.

Algorithm 1: The semi-total data replication policy

nbAccNode: number of accessed remote node by all transactions from current node
nbAccDataN_i: number of accessed remote data by all transactions from current node N_i
MaxAccNodeThres: maximum number of accessed remote node by all transactions.
MinAccNodeThres: minimum number of accessed remote node by all transactions.
 N_i : Node i.

```

begin
  for i from 1 to nbAccNode do
    if nbAccDataNi >= MaxAccNodeThres then
      Current node decide to replicate the full database of the Ni
    end
    if nbAccDataNi <= MinAccNodeThres then
      Current node decide to remove the replicated full database of the Ni
    end
  end
end

```

3.2 Approach using partial replication policy

In the second part of our work, we propose to apply a partial data replication policy of real-time and non real-time data (cf. Algorithm 2) in DFCS architecture. This technique may provide good management of database storage by reducing the updating replica data of the overfull system.

The principle of this algorithm is to start running system execution without any real-time and non real-time data replication. User's transactions require data located in local or in remote nodes. We use an accumulator of accessed remote data on other nodes by all transactions of the current node. In this policy, the access counter is calculated for each data, this means that it uses two accumulators; the first consists of the number of the frequently accessed remote nodes, and the second consists of the number of the accessed remote data on frequently accessed nodes. In the first case, if the number of accessed remote data at each node n_i reaches a maximum threshold, then the current node requests the data replication from node n_i . In another case, if the number of accessed remote data at each node n_i reaches a minimum threshold, then the current node decide to remove the current replicated data from node n_i . The maximum and the minimum threshold parameters are defined as input parameter to the system.

The both of those algorithms do not overload the system any more by updating replica data workload compared with full data replication policy. Each one of them has advantages to enhancing the QoS performance in DRTDBMS. This assertion for both algorithms is validated through a set of simulations.

Algorithm 2: The partial data replication policy

nbAccNode: number of accessed remote node by all transactions from current node

nbAccDataN_i: number of accessed remote data by all transactions from current node N_i

nbAccDataOcc: number of occurrences of accessed remote data from current node N_i

MaxAccNodeThres: maximum number of accessed remote node by all transactions.

MinAccNodeThres: minimum number of accessed remote node by all transactions.

MaxAccDataThres: maximum number of accessed remote data from current node by all transactions.

MinAccDataThres: minimum number of accessed remote data from current node by all transactions.

N_i : Node i.

```

begin
  for i from 1 to nbAccNode do
    if nbAccDataNi ≥ MaxAccNodeThres then
      for j from 1 to nbAccDataNi do
        if nbAccDataOcc ≥ MaxAccDataThres then
          Current node decide to replicate the current data of the Ni
        end
      end
    end
    if nbAccDataNi ≤ MinAccNodeThres then
      for j from 1 to nbAccDataNi do
        if nbAccDataOcc < MinAccDataThres then
          Current node decide to remove the current data of the Ni
        end
      end
    end
  end
end

```

4. SIMULATIONS AND RESULTS

To validate our QoS management approach, we have developed a simulator. In this section, we describe the overall architecture of the simulator, and we present and comments the obtained results.

4.1. Simulation model

Our simulator is composed of:

- **Database:** it consists of a data generator that generates data randomly avoiding duplicate information. The consistency in the database is provided by update transactions. In our simulator, the database contains real-time data and classic data.
- **Generator of transactions:** it is composed of two parts.
 - User transactions generator: it generates user transactions using a random distribution, taking into account their unpredictable arrival.

- Update transactions generator: it generates update transactions according to an arrival process that respects the periodicity of transactions.
- **Precision controller:** it rejects update transactions when the data to be updated are sufficiently representative of the real world, based on the value of MDE.
- **Scheduler:** it is used to schedule transactions according to their priorities.
- **Freshness manager:** it checks the freshness of data that will be accessed by transactions. If the accessed data object is fresh, then the transaction can be executed, and it will be sent to the transactions handler. Otherwise, it will be sent to the block queue. Then, it will be reinserted in the ready queue when the update of the accessed data object is successfully executed.
- **Concurrency controller:** it is responsible for the resolution of data access conflicts. Based on priorities (deadlines) of transactions, it determines which one can continue its execution and which one should be blocked, aborted or restarted.
- **Distributed commit protocol:** it manages the global distributed real-time transaction to ensure that all participating nodes agree with the final transaction result (validation or abortion) [9,10]. In our simulator, we use the commit protocol called “Permits Reading Of Modified Prepared-data for Timeliness” (PROMPT) defined in [4] which allows transactions accessing data not committed (optimistic protocol).
- **Admission controller:** the main role of this component is to filter the arrival of user transactions, depending on the system workload and respecting deadlines of transactions. Then, accepted transactions will be sent to the ready queue.
- **Handler:** it is used to execute transactions. If a conflict between transactions appears, it calls the concurrency controller.
- **Global load balancers (GLB):** it ensures that the load on each node does not affect the functioning of other nodes. So the GLB is used to balance the system load by transferring transactions from overloaded nodes to less loaded nodes in order to maintain QoS.

The input interface of our simulator allows choosing parameter values by which each simulation runs. It includes general parameters of the system, parameters for generating database, generating update transactions and generating user transactions. It also enables to choose the scheduling algorithm, the concurrency control protocol and the real-time distributed validation protocol. Once the simulation is well finished, results are saved in an output file. We can also show results in the form of curves, pie charts or histograms.

4.2. Simulation settings

The performance evaluation of the DRTDBMS is achieved by a set of simulation experiments, in which we varied some of the parameter values. Table 1 summarizes the general system parameters settings used in our simulations. The DRTDBMS is composed of 16 nodes. In each node we have 200 real-time data objects and 10000 classic data objects. Validity values of real-time data are distributed between 500 milliseconds and 1000 milliseconds. For real-time data, the value of MDE varies by 1 between 1 and 10. We choose the universal method to update real-time data. Within each queue in our system, transactions are scheduled according to the EDF algorithm. We use the 2PL-HP protocol for concurrency control. For the validation of

transactions, the distributed algorithm PROMPT is chosen. Update transactions are generated according to the number of real-time data in the database.

Table 1. System parameter settings.

Parameter	Value
Simulation time	3000 ms
Number of nodes	16
Number of real-time data	200/node
Number of classic data	10000/node
Validity of real-time data	[500,1000]
MDE value	[1,10]
Method to update real-time data	Universal
Scheduling algorithm	EDF
Concurrency control algorithm	2PL-HP
Distributed validation algorithm	PROMPT

Parameter settings for user transactions are defined in Table 2. User transactions are a set of read and write operations (from 1 to 4). Read operations (from 0 to 2) can access real-time data objects and classic data objects, however, write operations (from 1 to 2) access only classic data objects. The time of one read operations is used to be 1 millisecond, and for one write operation is set to 2 milliseconds. We set the slack factor of transactions to 10. The remote data ratio, which represents the ratio of the number of remote data operations to that of all data operations, is set to 20%. We note that user transactions are generated at arrival times calculated according to the "Poisson" process, which uses the value of the lambda parameter to vary the number of transactions.

Table 2. User transactions parameter settings.

Parameter	Value
Number of write operations	[1,2]
Number of read operations	[0,2]
Write operation time (ms)	2
Read operation time (ms)	1
Slack factor	10
Remote data ratio	20 %

Parameter settings for the real-time and non real-time data replication parameter settings are given in Table 3. To simulate using the partial data replication policy, we have to fix a minimum and a maximum threshold for nodes supporting replication which are set respectively to 0.2 and 0.5. We have to define, also, the value of the minimum and the maximum threshold of the number of accessed remote data at each node, which are set to 0.1 and 0.2. In case of a simulation with semi-total replication, only maximum values of the threshold of accessed remote nodes to replicate their database have to be fixed.

Table 3. Data replication parameter settings.

Parameter	Value
Maximum threshold of nodes supporting replication	0.5
Maximum threshold of data to be replicated	0.2
Minimum threshold of nodes supporting replication	0.2
Minimum threshold of data to be replicated	0.1

4.3. Simulation principle

To evaluate the performance of the proposed QoS management approach to enhance the performance of the overall system, we conducted a series of simulations by varying values of some parameters. Each transaction, whatever its type (user or update), undergoes a series of tests since its creation to its execution. In experiments, the workload distribution is initially balanced between all participating nodes.

4.3.1. Simulation using semi-total data replication policy

The first set of experiments evaluates the QoS management using semi-total data replication policy. For each node, an accumulator counter is used to calculate the mostly accessed remote nodes which the current node requests to replicate fully their databases.

4.3.2. Simulation using partial data replication policy

In this set of simulations, we evaluate the QoS management using partial data replication policy. For each node, as with semi-total replication, an accumulator counter is used to calculate the mostly accessed remote nodes. For each frequently accessed node, an accumulator counter is used to calculate the mostly accessed data which are requested to be replicated in the current node.

4.4. Results and discussions

As shown in Figure 2, the transactions success ratio is not affected by the increase of incoming user transactions. In fact, the system workload remains balanced and the system behaviour is maintained in stable state. Also, it is shown that the number of successful transactions with partial and semi-total data replication policies is greater than with full data replication policy. Indeed, the QoS guarantees is defined by increasing the number of transactions that meet their deadlines using fresh data.

We can say that the use of semi-total and partial data replication policy is suitable when increasing the number of participating nodes in DRTDBMS. By this way, the applicability of these data replication policies provides an optimistic management of the database storage while using fresh data by reducing the time of updating replicas.

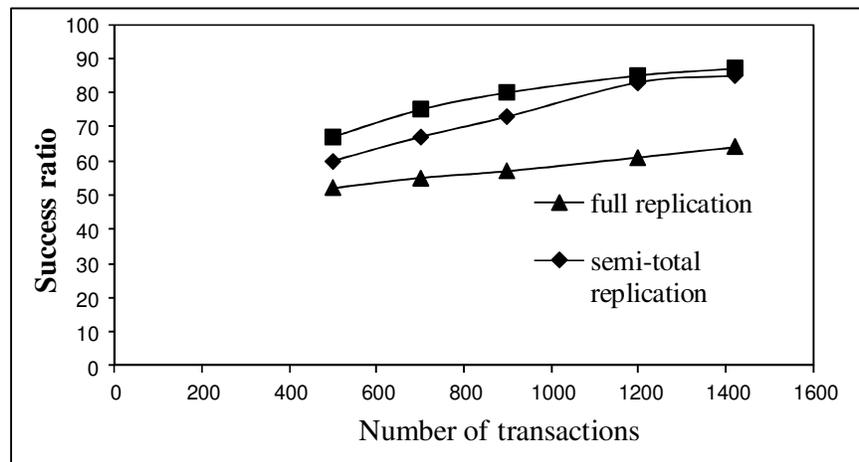


Figure 2. Simulation results for user transactions

The proposed QoS management, using semi-total and partial data replication policies, provides a better QoS guarantees than full data replication policy, ensuring stability and robustness of DRTDBMS.

5. CONCLUSION AND FUTURE WORK

In this article, we presented our QoS management approach to provide QoS guarantees in DRTDBMS. This approach is an extension work of the DFCS architecture proposed by Wei et al. [8]. It consists on applying two data replication policies, semi-total replication and partial replication of both classical data and real-time data on the conventional DFCS architecture, in order to make DRTDBMS more robust and stable. Indeed, the proposed approach is defined by a set of modules for data and transaction management in distributed real-time environment. The proposed approach helps to establish a compromise between real-time and data storage requirements by applying different data replication policies.

In future work, we will propose an approach for QoS enhancement in DRTDBMS using multi-versions data with both semi-total and partial data replication policies.

REFERENCES

- [1] Amirijoo, M. & Hansson J. & Son, S.H., (2003) « Specification and Management of QoS in Imprecise Real-Time Databases », Proceedings of International Database Engineering and Applications Symposium (IDEAS).
- [2] Amirijoo, M. & Hansson J. & Son, S.H., (2003) « Error-Driven QoS Management in Imprecise Real-Time Databases », Proceedings 15th Euromicro Conference on Real-Time Systems.
- [3] Haj Said, A. & Amanton, L. & Ayeb, B., (2007) « Contrôle de la réplication dans les SGBD temps réel distribués », Schedae, prépublication n°13, fascicule n°2, pages 41-49.
- [4] Haritsa, J. & Ramamritham, K. & Gupta, R., (2000) « The PROMPT Real-Time Commit Protocol », IEEE Transactions on Parallel and Distributed Systems, Vol 11, No 2, pp 160-181.
- [5] Kang, K. & Son, S. & Stankovic, J., (2002) « Service Differentiation in Real-Time Main Memory Databases », 5th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC02).
- [6] Ramamritham, K. & Son, S. & Dipippo, L., (2004) « Real-Time Databases and Data Services », Real-Time Systems journal, Vol 28, pp 179-215.

- [7] Sadeg, B., (2004) « Contributions à la gestion des transactions dans les SGBD temps réel », University of Havre.
- [8] Wei, Y. & Son, S.H. & Stankovic, J.A. & Kang, K.D., (2003) « QoS Management in Replicated Real Time Databases », Proceedings of the IEEE RTSS, pp 86-97.
- [9] Shanker, U. & Misra, M. & Sarje, A.K., (2008) « Distributed real time database systems: background and literature review », Springer Science + Business Media, LLC.
- [10] Jayanta Singh, J. & Mehrotra, Suresh C., (2009) « An Analysis of Real-Time Distributed System under Different Priority Policies », World Academy of Science, Engineering and Technology.
- [11] Lu, C. & Stankovic, J.A. & Tao, G. & Son, S.H., (2002) « Feedback Control Real-Time Scheduling: Framework, Modeling and Algorithms », Journal of Real- Time Systems, Vol 23, No ½.
- [12] Serrano, D. & Patino-Martinez, M. & Jimenez-Peris, R. and Kemme, B., (2007) « Boosting Database Replication Scalability through Partial Replication and 1-Copy-Snapshot-Isolation », IEEE Pacific Rim Int. Symp. on Dependable Computing (PRDC).

INTENTIONAL BLANK

EMOTION TEACHING INTERFACE FOR FINGER BRAILLE EMOTION TEACHING SYSTEM

Yasuhiro Matsuda¹ and Tsuneshi Isomura¹

¹Department of Robotics and Mechatronics, Kanagawa Institute of Technology,
Atsugi, Japan
yasuhiro@rm.kanagawa-it.ac.jp

ABSTRACT

Finger Braille is one of the tactual communication methods utilized by deafblind individuals. Deafblind people who are skilled in Finger Braille can catch up with speech conversation and express various emotions by changing dotting strength and speed. In this paper, we designed the emotion teaching interface in order to express joy, sadness, anger and neutral for the Finger Braille emotion teaching system. We changed the previous background color (beige) of the teaching interface into 17 different colors. We also designed 8 kinds of dot patterns with different horizontal width and vertical length. The experiment to select the most suitable emotion teaching interfaces for joy, sadness, anger and neutral was conducted. The results showed that the dot patterns 6 (the wide and middle length pattern) or 1 (the small circle) with the lime, dark orange or yellow background colors are suitable for joy; the dot patterns 7 (the narrow and long pattern) or 4 (the narrow and middle length pattern) with the lavender, navy or blue background colors are suitable for sadness; the dot patterns 9 (the large circle) or 8 (the middle width and long pattern) with the red background color are suitable for anger; the dot pattern 5 (the middle circle) with the previous, honeydew, saddle brown or white background colors are suitable for neutral.

KEYWORDS

Emotional communication, Finger Braille, deafblind & emotional color

1. INTRODUCTION

One of the most important factors of communication is emotion. Emotion is communicated in facial expressions, body movements, eye communication and paralanguage. Emotion is also affected by content of communication, particular objects and colors of environment. Colors are associated with particular emotions for human beings.

Deafblindness is a combination of varying degrees of both hearing and visual impairment. Deafblind people have difficulties with both verbal and emotional (nonverbal) communication. Deafblind people use many different communication media, depending on the age of onset of hearing and visual impairment and the available resources. Tactual communication is an important form of verbal and emotional communications for deafblind people.

“*Yubi-Tenji*” (Finger Braille) is one of the tactual communication media utilized by deafblind individuals (see Figure 1). In two-handed Finger Braille, the sender’s index finger, middle finger and ring finger of both hands function like the keys of a Braille typewriter. The sender dots Braille code on the fingers of the receiver as if typing on a Braille typewriter. The receiver is assumed to recognize the Braille code. In one-handed Finger Braille, the sender first dots the left column of Braille code on the distal interphalangeal (DIP) joints of the three fingers of the receiver, and then the sender dots the right column of Braille code on the proximal interphalangeal (PIP) joints. Deafblind people who are skilled in Finger Braille communicate words and express various emotions because of the prosody (intonation) of Finger Braille [1]. The prosody of Finger Braille is a kind of paralanguage that helps the receiver recognize the dotted Braille code. The following were the features of the prosody: (1) the sender dots long at the end of the clause; (2) the sender pauses long after the end of clause; (3) the sender also dots long and strongly at the end of the sentence. However, non-disabled people generally are not skilled in Finger Braille. Consequently, deafblind people can communicate only through interpreters.



Figure 1. Two-handed Finger Braille (left) and one-handed Finger Braille (right)

Various Finger Braille input devices were developed. Amemiya et al. developed oboe-like Braille input interface (keyboard) [2]. An et al. developed Braille input gloves [3]. Fukumoto et al. developed a wearable input device with accelerometers mounted on the top of rings [4]. Hoshino et al. developed a Finger Braille input system that mounted accelerometers on the middle phalanges [5]. With these devices, deafblind people are burdened with wearing sensors, and they must master a new communication system.

We have recently been developing a Finger Braille support device which employs tactual communication. Figure 2 shows the concept of the Finger Braille support device. The advantages of this support device are as follows: (1) both deafblind people and non-disabled people unskilled in Finger Braille can communicate using conventional Finger Braille; (2) because the non-disabled people operate the support device and wear all of the sensors, deafblind people are not encumbered by the support device; (3) the intent of the support device is to assist both verbal and emotional communication.

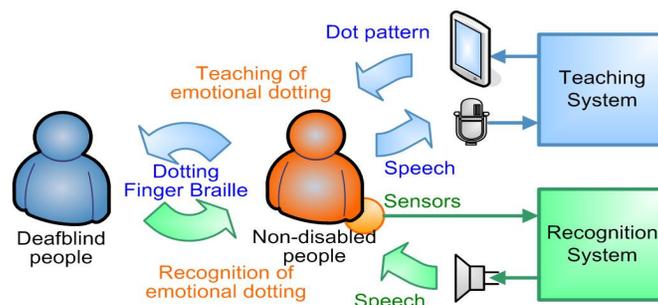


Figure 2. Concept of Finger Braille support device

The support device consists of a Finger Braille (emotion) teaching system and a Finger Braille (emotion) recognition system. The teaching system recognizes the speech of a non-disabled person and displays the associated dot pattern of Finger Braille. The non-disabled person can then dot Finger Braille on the fingers of the deafblind person by observing the displayed dot pattern [6]-[7]. The emotion teaching system also teaches non-disabled person to express emotions. We have developed the teaching method of emotional expression using sentences about the impression of emotional expression [8].

The recognition system recognizes the dotting of Finger Braille by the deafblind person and synthesizes the associated speech for the non-disabled person [9]. We have developed the emotion recognition system which recognizes the emotions dotted by the deafblind person and presents the information for the non-disabled person [10].

The objective of this study is development of the emotion teaching interfaces to express joy, sadness, anger and neutral. In this paper we first designed the emotion teaching interfaces for the emotion teaching system. Then, an experiment to select the most suitable emotion teaching interfaces for joy, sadness, anger and neutral was conducted.

2. DESIGN OF EMOTION TEACHING INTERFACE

2.1. Configuration of Teaching System

The configuration of the teaching system was shown in Figure 3 [6]-[7]. First, a speech recognition (SR) engine recognizes the speech by the sender. Second, the teaching system converts the Kana script to Braille code by using the results of the speech recognition. Third, the teaching system retrieves the clause information by parsing the Braille code and segments the Braille code into clauses. Finally, the teaching system displays the associated dot pattern of the Braille code. The teaching system was developed on a tablet PC (Dell Latitude XT, CPU Core 2 Duo 1.33 GHz, RAM 2 GB, 12.1 inch WXGA LCD with touch screen). The operating system was Microsoft Windows XP. The programming languages were Microsoft Visual Basic 6 and LPA WIN-PROLOG 4.500. The speech recognition engine was Microsoft Speech SDK (SAPI5.1). If the Braille code was not grammatically because of misrecognition of SR, the Braille code parser could not parse it. As a backup of the Braille code parser, we used Microsoft Global IME (Japanese) (IMM API).

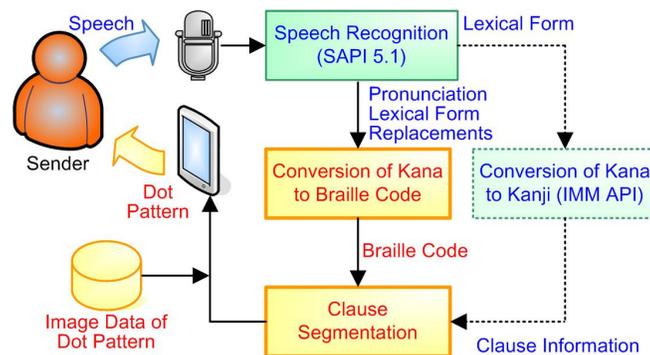


Figure 3. Configuration of the teaching system.

Figure 4 shows the previous teaching interfaces of the teaching system. The Braille code is displayed in the upper text box. The dot pattern is displayed in eighteen picture boxes (three columns and six rows). Clauses are displayed in the order of top to bottom of the first column, top

to bottom of the second column, and top to bottom of the third column, according to the length of the clause. Sometimes a long clause is continued on the next page. If a clause has fewer than six characters, the next clause appears in the next column or the next page. In this way, the dot pattern of the clauses is displayed explicitly by the columns.

The red pattern in Figure 4 indicates the dotting of the DIP joints and the blue pattern indicates the dotting of the PIP joints. We designed three kinds of teaching interfaces for the sender. Teaching interface 1 displays the dot pattern illustrated on the fingers; teaching interface 2 displays only the dot pattern; teaching interface 3 displays the dot pattern with long and short arrows to indicate the duration of dotting. Teaching interface 1 is more symbolic and easier for beginners because they can see the dotting fingers. Teaching interface 2 has the most simplified signing and is suitable for the experienced senders. Teaching interface 3 teaches the duration of dotting to realize the non-disabled sender's prosodic dotting.

The buttons of speech recognition, edit, restatement, previous page and next page are located on the lower part of the display. The sender can touch the LCD directly to operate the teaching system and edit the Braille code.

In this study, we adopted one-handed Finger Braille using the right hand as the communication medium, because one-handed Finger Braille is easy to dot for non-disabled people and most human populations are right-handed (see Figure 5).



Figure 4. Three kinds of teaching interfaces: teaching interface 1 (left), teaching interface 2 (center) and teaching interface 3 (right). The displayed dot pattern is “Ohayo- / gozaimasu” (“Good morning”).

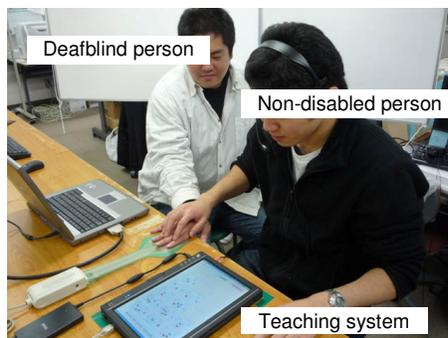


Figure 5. One-handed Finger Braille communication supported by the teaching system.

2.2. Teaching Method of Emotional Expression using Sentences

In the previous study, we analyzed the features of emotional expression by Finger Braille interpreters [1]. The features were as follows: (1) the duration of the code of joy was significantly shorter than those of the other emotions (neutral, anger and sadness); (2) the duration of the code of sadness was significantly longer than those of the other emotions; (3) the finger load of anger was significantly larger than those of the other emotions; (4) the finger load of joy was significantly larger than those of sadness and neutral; (5) the duration of the code of anger was significantly shorter than those of sadness and neutral. We also analyzed the effectiveness of emotional expression and emotional communication between people unskilled in Finger Braille. The results indicate that the features and the impression of emotional expression by the unskilled people were very similar to those by the interpreters.

We noted the similarities of the impression of emotional expression between the interpreters and the unskilled people. We developed the teaching method using the sentences about the impression of emotional expression [8]. The teaching sentences of emotional expression that we developed were as follows.

Joy: Dot rhythmically.

Sadness: Dot weakly and slowly.

Anger: Dot strongly and little bit quickly.

Neutral: Dot politely and slowly with a constant rhythm.

The unskilled people read these sentences about the emotion which he/she want to express and then express the impression of emotional expression in the dotting of Finger Braille. The results of the evaluation experiment showed that the non-disabled subjects could express emotions better than the subjects who were not taught the features of emotional expression.

2.3. Design of Emotion Teaching Interface

Colors are associated with particular emotions for human beings, such as semantic words (“warm - cool”, “heavy - light”, “active - passive”, etc.) or actual emotions [11]-[12]. In the present study, we designed the emotion teaching interface in order to express joy, sadness, anger and neutral, in addition to the teaching sentences of emotional expression. The concept of design is as follows: “the background color of the teaching interface will be associated with the emotion to express.” We changed the previous background color (beige) of teaching interface 2 (see Figure 4) into 17 different colors. The designed teaching interfaces with 18 background colors are presented in Figure 9. The RGB triplets of the designed background colors are listed in Table 1. The color names are pursuant to the HTML color names.

Next, we modified the horizontal width and vertical length of the dot pattern of teaching interface 2. The concepts of modification are as follows: (1) “the wide dot pattern will be associated with the strong dotting and the narrow dot pattern will be associated with the weak dotting;” (2) “the long dot pattern will be associated with the long dotting duration and the short dot pattern will be associated with the short dotting duration.” We designed 8 kinds of dot patterns with different horizontal width and vertical length (see Figure 10). Dot pattern 5 is the previous dot pattern (the middle circle). Dot pattern 1 is the small circle. Dot pattern 2 is the middle width and short pattern. Dot pattern 3 is the wide and short pattern. Dot pattern 4 is the narrow and middle length pattern. Dot pattern 6 is the wide and middle length pattern. Dot pattern 7 is the narrow and long pattern. Dot pattern 8 is the middle width and long pattern. Dot pattern 9 is the large circle.

In the present study, we selected the most suitable combinations of the background color and dot pattern to express joy, sadness, anger and neutral through an experiment, as we discussed below.

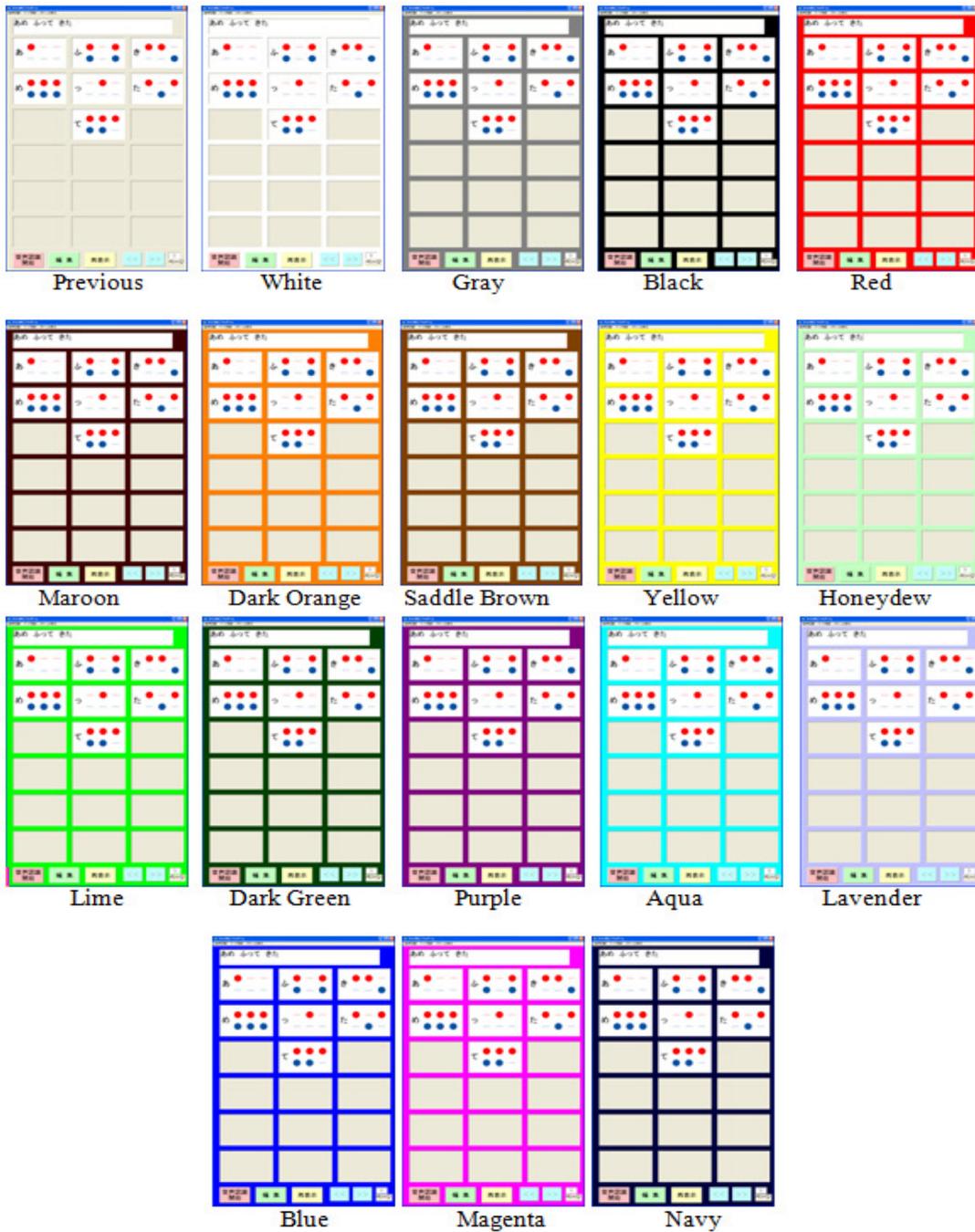


Figure 6. Emotion Teaching interfaces with 18 background colors. The displayed dot pattern is “Ame / futte / kita” (“Rain has fallen”).

Table 1. RGB triplets (hexadecimal) of designed background colors. (Color names with * are darker than the associated HTML color names)

No.	Color Name	R	G	B
1	Previous (Beige *)	EC	E9	D8
2	White	FF	FF	FF
3	Gray	80	80	80
4	Black	00	00	00
5	Red	FF	00	00
6	Maroon *	40	00	00
7	Dark Orange *	FF	80	00
8	Saddle Brown *	80	40	00
9	Yellow	FF	FF	00
10	Honeydew *	C0	FF	C0
11	Lime	00	FF	00
12	Dark Green *	00	40	00
13	Purple	80	00	80
14	Aqua	00	FF	FF
15	Lavender *	C0	C0	FF
16	Blue	00	00	FF
17	Magenta	FF	00	FF
18	Navy *	00	00	40

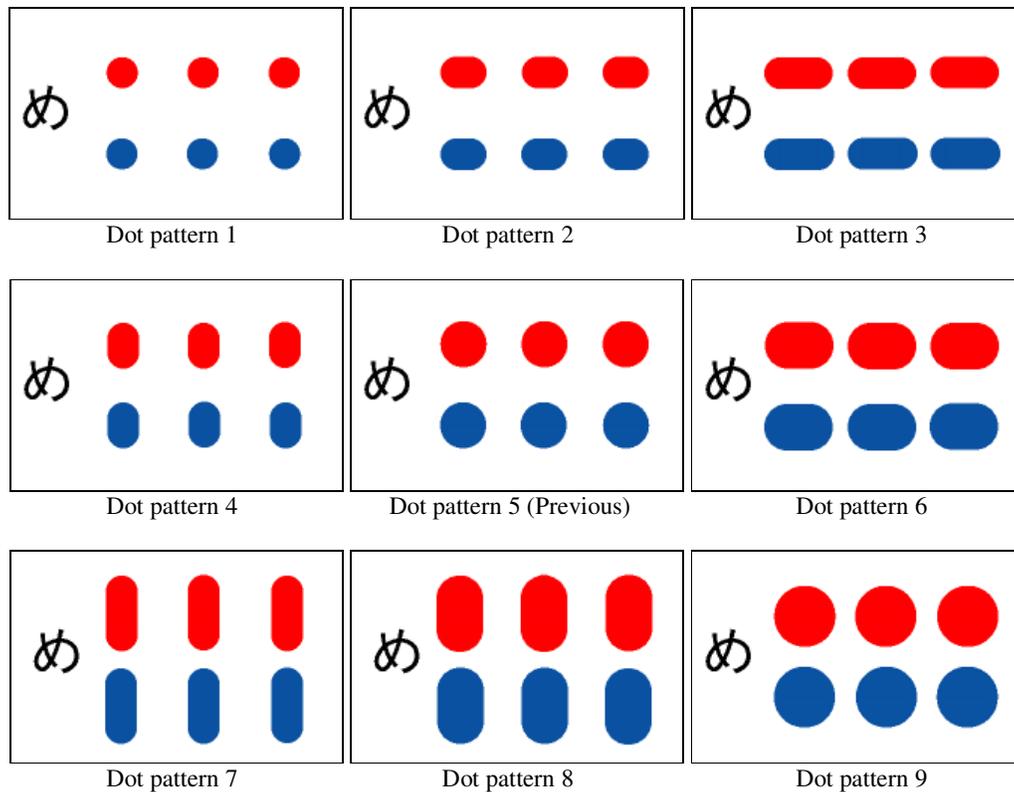


Figure 7. Dot patterns with different width and length.

3. SELECTION EXPERIMENT

3.1. Methods

To select the most suitable emotion teaching interfaces for joy, sadness, anger and neutral, a selection experiment was conducted.

The subjects were 14 male and 1 female college students (mean age = 22.6, S.D. = 1.2). The subjects have no experience of Finger Braille. At the beginning of the experiment, a tester described about Finger Braille, the teaching system and the objectives of this experiment. All subjects gave their informed consent after hearing a description of the study.

Two experimental sessions were conducted. In the session 1, the tester displayed one of the 18 emotion teaching interfaces of Figure 6. The dot pattern was the previous pattern (dot pattern 5 of Figure 7). By observing the displayed emotion teaching interface, the subject responded an associated emotion from “joy”, “anger”, “sadness”, “fear”, “disgust”, “surprise” and “not applicable (NA)”. These six emotions are the fundamental emotions of human being. The tester repeated displaying 18 emotion teaching interfaces with a predetermined random order.

In the session 2, the tester displayed one of the emotion teaching interfaces with the 9 dot patterns of Figure 7. The background color of the teaching interface was the previous color (beige). By observing the emotion teaching interface, the subject responded the impression about the dotting strength from “strongly”, “weakly” and “not applicable (NA)”; and the impression about dotting duration from “long”, “shortly” and “not applicable (NA)”. The tester repeated displaying 9 emotion teaching interfaces with a predetermined random order.

The emotion teaching interfaces were displayed on an external 14 inches LCD (ThinkVision LT1421, Lenovo) placed in front of the subject. The tester operated a note PC (ProBook 4515s/CT, HP) to display the teaching interfaces (see Figure 8).

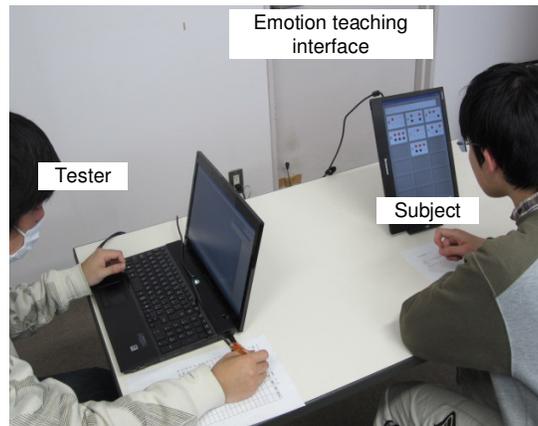


Figure 8. An experiment.

3.2. Results

Figure 9 shows the response ratio of emotions as a function of the background color of the emotion teaching interface. As a result, the associations of the lime, dark orange and yellow background colors with joy were common (73%, 60% and 60%, respectively). As for the red

background color, 60% of subjects responded anger. The responses for the lavender, navy and blue background colors were very similar with sad (60%, 53% and 47%, respectively). About 47% of subjects regarded the black background color as fear. Almost all subjects (87%) responded that the previous background color (beige) was not applicable (NA). A number of responses for the honeydew, saddle brown and white background colors were also NA (53%, 47% and 47%, respectively).

Figure 10 shows the response ratio of dotting strength and duration as a function of the number of the dot pattern of the emotion teaching interface. As a result, a number of responses for dot patterns 1 and 2 were weakly (73% and 53%, respectively) and shortly (80% and 53%, respectively). Dot patterns 4 and 7 were mostly associated with weakly (53% and 67%, respectively) and long (60% and 87%, respectively). A number of responses for dot patterns 3, 6 and 8 were strongly (73%, 80% and 60%, respectively) and long (67%, 53% and 67%, respectively). Dot pattern 9 was mostly associated with strongly (67%) and shortly (47%). A number of responses for dot pattern 5 were NA (47% and 53%).

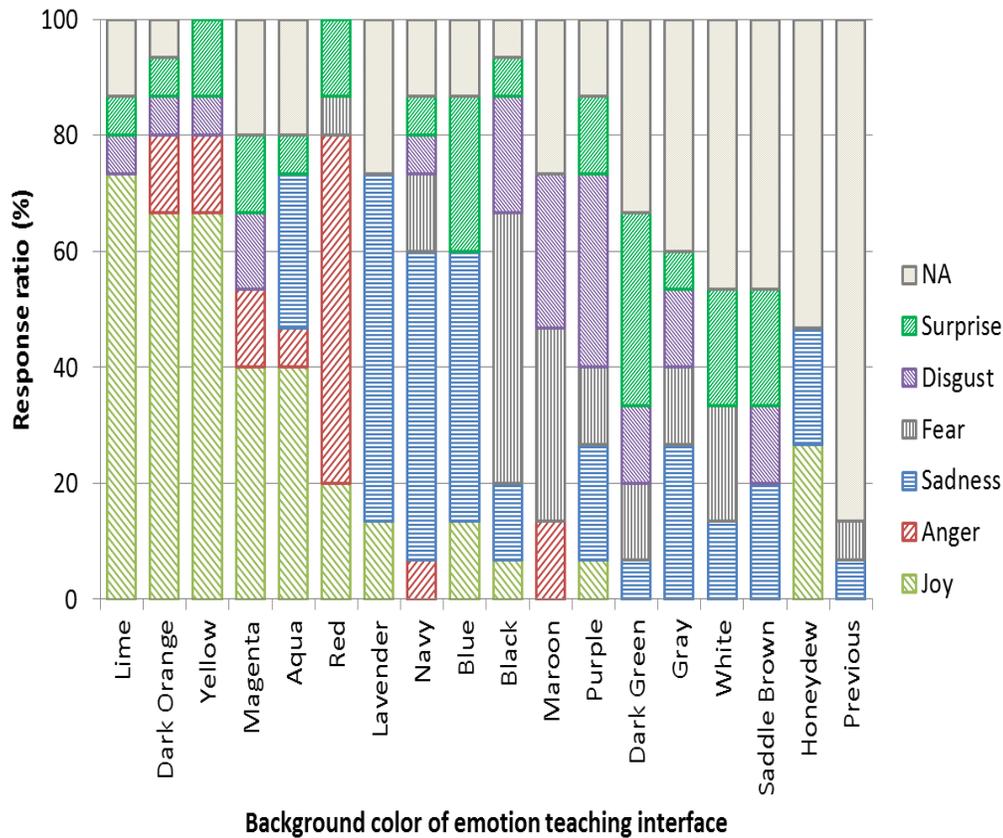


Figure 9. Response ratio of emotions as a function of the background color of the emotion teaching interface.

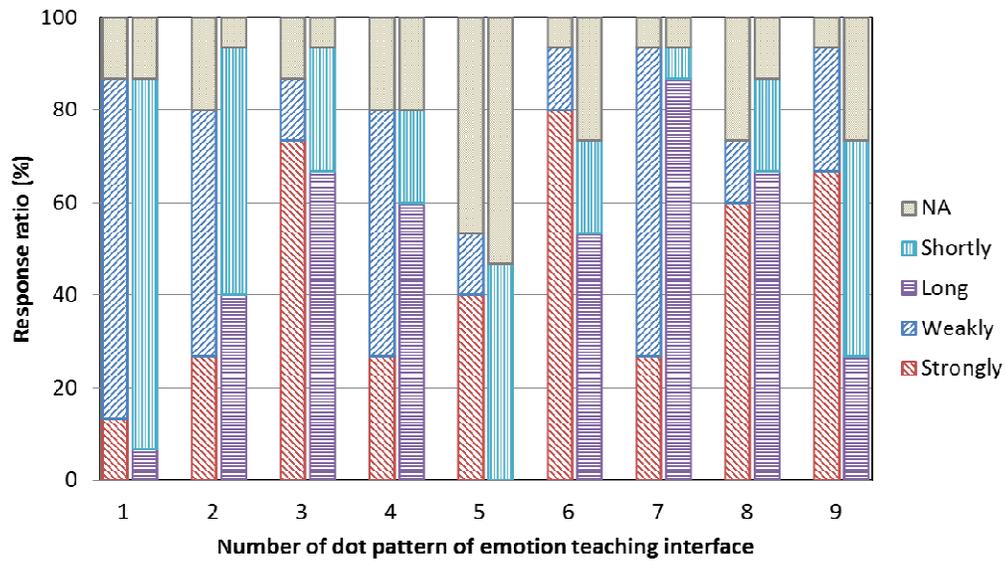


Figure 10. Response ratio of dotting strength and duration as a function of the dot pattern of the emotion teaching interface.

3.3. Discussion

According to the results of the session 1, the lime, dark orange and yellow background colors were associated with joy; the red background color was associated with anger; the lavender, navy and blue background colors were associated with sadness; the honeydew, saddle brown and white background colors were not associated with any emotions.

Clarke et al. investigated the relationship between colors and emotions [11]. They revealed that orange and yellow are associated with joy; red is associated with anger; blue is associated with sadness; gray is not associated with any emotions. These results were similar to our experimental results.

According to the results of the session 2, dot patterns 1 and 2 were associated with weak and short dotting. Dot patterns 4 and 7 were associated with weak and long dotting. Dot patterns 3, 6 and 8 were associated with strong and long dotting. Dot pattern 9 was associated with strong and short dotting. As for the concept of design of the dot patterns, dot pattern 3 and 6 should be associated with strong and short dotting. The other results were similar to the concept of the design.

As mentioned above, joy was characterized by little bit strong and short dotting. Sadness was characterized by weak and long dotting. Anger was characterized by strong and little bit short dotting. Neutral was characterized constant dotting without emotion. Thus, we conclude that the most suitable teaching interfaces for joy, sadness, anger and neutral are as follows.

Joy: Dot patterns 6 or 1 with the lime, dark orange or yellow background colors.

Sadness: Dot patterns 7 or 4 with the lavender, navy or blue background colors.

Anger: Dot patterns 9 or 8 with the red background color.

Neutral: Dot pattern 5 with the previous, honeydew, saddle brown or white background colors.

4. CONCLUSIONS

In this paper, we designed the emotion teaching interface in order to express joy, sadness, anger and neutral for the Finger Braille emotion teaching system. We changed the previous background color (beige) of the teaching interface into 17 different colors. We also designed 8 kinds of dot patterns with different horizontal width and vertical length. The experiment to select the most suitable emotion teaching interfaces for joy, sadness, anger neutral and was conducted. The results showed that dot patterns 6 (the wide and middle length pattern) or 1 (the small circle) with the lime, dark orange or yellow background colors are suitable for joy; dot patterns 7 (the narrow and long pattern) or 4 (the narrow and middle length pattern) with the lavender, navy or blue background colors are suitable for sadness; dot patterns 9 (the large circle) or 8 (the middle width and long pattern) with the red background color is suitable for anger; dot pattern 5 (the middle circle) with the previous, honeydew, saddle brown or white background colors are suitable for neutral.

Our future plans are: (1) verification of the impression of the combinations of the background color and dot pattern; (2) evaluation of the emotional expression using these emotion teaching interfaces.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 25350683.

REFERENCES

- [1] Y. Matsuda, I. Sakuma, Y. Jimbo, E. Kobayashi, T. Arafune & T. Isomura, (2010) "Emotional Communication in Finger Braille", *Advances in Human-Computer Interaction*, Vol. 2010, Article ID 830759, 23 pages.
- [2] S.S. An, J.W. Jeon, S. Lee, H. Choi & H.G. Choi, (2004) "A Pair of Wireless Braille-Based Chording Gloves", *Proceedings of 9th International Conference on Computers Helping People with Special Needs*, pp.490-497.
- [3] T. Amemiya, K. Hirota & M. Hirose, (2004) "OBOE: Oboe-Like Braille Interface for Outdoor Environment", *Proceedings of 9th International Conference on Computers Helping People with Special Needs*, pp.498-505.
- [4] M. Fukumoto & Y. Tonomura, (1997) "Body Coupled FingerRing: Wireless Wearable Keyboard", *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp.147-154.
- [5] T. Hoshino, T. Otake & Y. Yonezawa, (2002) "A Study on a Finger-Braille Input System Based on Acceleration of Finger Movements", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol.J85-A, no.3, pp.380-388.
- [6] Y. Matsuda & T. Isomura, (2010) "Finger Braille Teaching System", *Character Recognition, Sciyo*, pp.173-188.
- [7] Y. Matsuda & T. Isomura, (2011) "Improvement of Interfaces of Finger Braille Teaching System", *Journal of Computer and Information Technology*, Vol. 1, No. 2, pp.19-28.
- [8] Y. Matsuda & T. Isomura, (2013) "Development of Teaching Method of Emotional Expression using Finger Braille", *Journal of Communication and Computer*, Vol. 10, No. 4, pp.567-577.
- [9] Y. Matsuda & T. Isomura, (2012) "Finger Braille Recognition System", *Advances in Character Recognition, InTech*, pp.193-210.
- [10] Y. Matsuda & T. Isomura, (2012) "Emotion Recognition System of Finger Braille", *International Review on Computers and Software (I.RE.CO.S)*, Vol. 7, No. 7, pp.3494-3501.
- [11] T. Clarke & A. Costall, (2008) "The Emotional Connotations of Color: A Qualitative Investigation", *Color Research and Application*, Vol. 33, No. 5, pp.406-410.
- [12] M. Solli & R. Lenz, (2011) "Color Emotions for Multi-Colored Images", *Color Research and Application*, Vol. 36, No. 3, pp.210-217.

AUTHOR**Yasuhiro Matsuda**

Professor in the Department of Robotics and Mechatronics, Kanagawa Institute of Technology. He obtained Ph.D. degree in environmental studies from the University of Tokyo in 2007. His research interests include assistive technology, human interface and emotional communication. He is a member of IEEE Engineering in Medicine and Biology Society.



DISTRIBUTED SYSTEM FOR 3D REMOTE MONITORING USING KINECT DEPTH CAMERAS

M. Martínez-Zarzuela¹, M. Pedraza-Hueso, F.J. Díaz-Pernas,
D. González-Ortega, M. Antón-Rodríguez

¹Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática
University of Valladolid
marmar@tel.uva.es

ABSTRACT

This article describes the design and development of a system for remote indoor 3D monitoring using an undetermined number of Microsoft® Kinect sensors. In the proposed client-server system, the Kinect cameras can be connected to different computers, addressing this way the hardware limitation of one sensor per USB controller. The reason behind this limitation is the high bandwidth needed by the sensor, which becomes also an issue for the distributed system TCP/IP communications. Since traffic volume is too high, 3D data has to be compressed before it can be sent over the network. The solution consists in self-coding the Kinect data into RGB images and then using a standard multimedia codec to compress color maps. Information from different sources is collected into a central client computer, where point clouds are transformed to reconstruct the scene in 3D. An algorithm is proposed to conveniently merge the skeletons detected locally by each Kinect, so that monitoring of people is robust to self and inter-user occlusions. Final skeletons are labeled and trajectories of every joint can be saved for event reconstruction or further analysis.

KEYWORDS

3D Monitoring, Kinect, OpenNI, PCL, CORBA, VPX, H264

1. INTRODUCTION

A system for remote people monitoring can be employed in a large amount of useful applications, such as those related to security and surveillance[1], human behavior analysis[2] and elderly people or patient health care[3][4]. Due to their significance, human body tracking and monitoring are study fields in computer vision that have always attracted the interest of researchers[5][6]. As a result, many technologies and methods have been proposed. Computer vision techniques are becoming increasingly sophisticated, aided by new acquisition devices and low-cost hardware data processing capabilities.

The complexity of the proposed methods can significantly depend on the way the scene is acquired. An important requirement is to achieve fine human silhouette segmentation. State-of-art technologies are really good at this task. Apart from the techniques that use markers attached to the human body, tracking operations are carried out mainly in two ways, from 2D information or

3D information [7][8]. On the one hand, 2D body tracking is presented as the classic solution; a region of interest is detected within a 2D image and processed. Because of the use of silhouettes, this method suffers from occlusions. On the other hand, advanced body tracking and pose estimation is currently being carried out by means of 3D cameras, such as binocular, Time-of-Flight (ToF) or consumer depth-cameras like Microsoft(R) Kinect[9]. The introduction of low-cost depth sensors has pushed up the development of new systems based on robust segmentation and tracking of human skeletons. The number of applications built on top of depth-sensor devices is rapidly increasing. However, most of these new systems are aimed to track only one or two people thus have only direct application on videogames or human-computer interfaces.

There are some limitations to address in order to build a remote space monitoring system using consumer depth-cameras, and only a few separate efforts have been done to address these limitations. Even so, those developments do not pursue building a remote monitoring system, but covering part of the limitations in which we are also interested for our system. On the one hand, Kinect devices can capture only a quite small area, covering accurately distances only up to 3.5 meters[9]. There are proposals which allow to make a 3D reconstruction of spaces and objects using Kinect[10], but in themevery capturing device has to be connected to the same computer. Apart from that, these solutions cannot merge skeletons information from different Kinects. The first limitation is significant, since only two or three devices can be connected to a single computer, due to the high USB bandwidth consumption of these cameras. There is another proposal that allows to send data over a network[11]. However, this application uses Microsoft SDK [9], so it only works under Windows operating system.

The 3D monitoring system presented in this paper addresses these limitations and allows using an undetermined number of Microsoft® Kinect cameras, connected to an undetermined number of computers running any operating system (Windows, Linux, Mac), to monitor people in a large space remotely. The system codes the 3D information (point clouds representing the scene, human skeletons and silhouettes) acquired by each camera, so that bandwidth requirements for real-time monitoring are met. The information coming from different devices is synchronized. Point clouds are combined to reconstruct the scene in 3D and human skeletons and silhouettes information coming from different cameras are merged conveniently to build a system robust to self-user or inter-user occlusions. The proposed system uses low cost hardware and open source software libraries, which makes its deployment affordable for many applications under different circumstances.

Section 2 of this paper includes a general description of the tools and methods employed to develop the system, Section 3 describes the proposed system, describing important details about the 3D information coding strategy and the algorithm proposed to merge different skeletons information. Section 4 describes performance evaluation tests that were conducted in order to measure the robustness of the system. Finally, Section 5 draws the main conclusions and comments on future research tasks.

2. TOOLS AND METHODS

2.1 Consumer depth-cameras

For 3D scene acquisition, a number of devices can be used. For computer vision techniques, we can distinguish among passive and active cameras. The first include stereo devices, simulating the left and right eye in human vision: the images coming from each camera in the device are combined to generate a disparity map and reconstruct depth information[7]. In this category, some other proposals in which several passive cameras are disposed around the person or object to be reconstructed can be included. The second option consists in using an active device such as

a TOF (Time of Flight) camera or newer consumer-depth cameras like Microsoft® Kinect or ASUS® Xtion. Despite their high depth precision, TOF cameras are expensive and provide very low resolutions. On the other side, consumer depth-cameras provide resolutions starting at 640x480 px and 30 fps at very affordable prices.

We would pay special interest to Kinect cameras, since they are the chosen devices for the proposed system. Microsoft® Kinect emits a structured infrared pattern of points over its field of view, which is then captured by a sensor and employed to estimate the depth of every projected point in the scene. Although Kinect was initially devised only to computer games, the interest of the computer vision community rapidly made it possible to use the device for general purpose from a computer, even before the Microsoft® official Kinect SDK was available[9]. There is a wide variety of tools to work with Kinect. A commonly used framework for creating applications is OpenNI[12], which has been developed to be compatible with any commodity depth camera and, in combination with NiTE middleware, is able to automate tasks for user identifying, feature detection, and basic gesture recognition[13].

2.2 Data compression

Consumer depth cameras generate a large volume of data. This is an important issue, since one of the objectives of the system is the transmission of this information over a network. Therefore, data compression is necessary before sending data to a central computer. There are different ways to compress data. If the data to compress is not multimedia, we can use a zip encoder, which provides lossless compression, but generates large output data and is computationally expensive. For multimedia compression, there are picture encoders like jpeg, which do not use temporal redundancy. To compress video, there are many encoders like H.264 or VP8. These encoders are able to compress data taking advantage of the temporal redundancy, thus compressed information is suitable to send over the network. However, there are not extended codecs to compress depth maps yet. One type of compression codecs used for 3D images, are those used to transmit the 3D television signal, but they are based on the compression of two images (right and left) [14], thus are not useful for our system, where 3D information is directly acquired using an active infrared device.

2.3 CORBA

A distributed application based on the client-server paradigm does not need to be developed using low level sockets. For the proposed system, a much more convenient approach is using a middleware such as TAO CORBA, a standard defined by OMG (Object Management Group). This middleware allows using a naming service[15], that avoids the central client to know about the addresses of each one of the servers. The aim of CORBA is to hide to the programmer the low level complexity algorithms for data transmission over the network. It is object-oriented and supports C++, Python, Java, XML, Visual Basic, Ada, C, COBOL, CORBA-Scripting-Language, Lisp, PL/I, Smalltalk and C#. Besides, this middleware is chosen because it is independent of the programming language, so servers could be programmed in Java and a client in C++, for example. It represents a clear advantage over RMI, which can only be programmed in Java. CORBA is also cross platform, so clients and servers can be running on different operating systems. In the proposed system, the servers may be running on Windows computers and the client in a Linux computer or in the opposite way.

2.4 PCL: Point Cloud Library

PCL 'Point Cloud Library'[16], is a C++ free open source computer vision library to work with 3D information that can be used in many areas such as robotics. PCL is being developed by a

group of researchers and engineers from around the world. There are also many companies such as Toyota or Nvidia working to develop this powerful library [17]. The library contains algorithms for filtering, feature estimating, point cloud registration and segmentation.

Point clouds can be obtained and stored in 3D raw data files, read from 3D models or 3D cameras such as Kinect. The combination of both technologies, PCL and Kinect, is very convenient for our purpose of monitoring a space with 3D information. The library is comprised of the following modules: filters, features, keypoints, registration, kd-tree, octree, segmentation, simple consensus, surface, range image, IO, visualization, common and search.

3. PROPOSED SYSTEM

The proposed system key feature is the fusion of 3D information coming from multiple Kinect devices, including depth information and detected skeletons. This takes place under the client-server model, where servers are computers with attached devices and the client is the central computer responsible for information fusion, tracking and visualization.

3.1 General description

Figure 1 depicts the scheme of the proposed system. A server is a computer where one or more Kinect cameras are connected. The different servers, deployed in a remote space are responsible of capturing the information coming from different regions of the scene. This information is conveniently processed and then sent to a system central computer. The large amount of information acquired by Kinect devices has to be compressed using different strategies before it can be sent over the network. The central client is in charge of reconstructing the remote space in 3D using PCL library and includes a robust algorithm for multiple detected skeletons merging. The computer interface can be used to monitor the scene in 3D in real time, label people within it and record specific users movements for further analysis[18] The system is fully scalable to any number of servers and clients, thus any number of acquiring devices and locations.

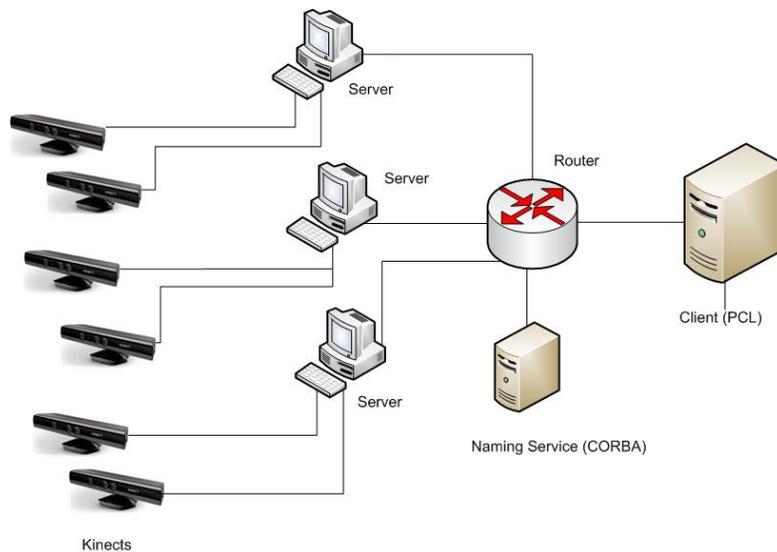


Figure 1. General scheme of the proposed system.

3.2 Data acquisition

Kinect devices present certain limitations. First, its view range covers depth precisely only between 0.5 m and 3.5 m[9].

For this reason, one of the motivations of our system is to expand the covered view by adding several Kinects to the scenario. To add more Kinect devices to the scene, the naïve solution is to try to connect multiple cameras to the same computer. However, due to the large volume of data generated by each camera, a USB controller is needed to handle the bandwidth emitted by each one. As a consequence, that is not a valid solution, since most computers only support a limited number of USB controllers, usually two or three. The solution adopted for our system was to develop a distributed application with multiple cameras connected to multiple computers. In such a configuration, another important issue has to be taken into account. The infrared pattern emitted by different Kinects can interfere with each other, causing 'holes' in the acquired point clouds. Therefore, we must be careful in the placement of the devices and avoid placing a camera right in front of other one.

The data provided by each Kinect in which we are interested in are: a three-channel *RGB image* of the scene, captured at a resolution of 640x480 px; a *depth map*, which is a texture of the same resolution in which each pixel takes a value that indicates the distance between the infrared pattern and the sensor; a texture of *user labels* with same resolution, in which each pixel takes the value of the user id in front of the camera or a zero value; and the *skeletons* of the users in the scene, which are formed by joints representing the parts of the body (head, elbow, shoulders ...) and include both xyz position and rotation from an initial pose position.

3.3 Data coding

Before the information captured by the remote devices can be sent, it has to be encoded. Kinect data to be processed includes RGB images, depth maps, user labels and skeleton joints. Skeleton joints information is sent without any compression, since the volume of data needed to store and transmit the position of all the joints is negligible compared to the volume of image or depth information.

To encode the RGB image we use a video compressor. It would be meaningless to use an image codec such as JPEG, since it only uses spatial information at the time of compression and thus the bit rate needed is much higher. For video compression, the proposed system uses the cross-platform library FFMPEG, which provides many audio and video codecs. The RGB image compression is done using the VP8 codec developed by Google TM [19] that needs a YUV 420 image format. Although VP8 codec introduces quality losses during compression and decompression, its balance between final quality and performance makes it adequate for our purposes. Additionally, the loss in quality remains quite low and the human eyes, acting as filters, are not able to appreciate it.

As it has been commented before, there is no compression codec to encode depth or a combination of RGB and depth information. In the proposed system, the compression of the depthmap has to be done in a tricky way, based on the scheme proposed by Pece[20]. Basically, one depth channel has to be converted into a three-channel image, and then a specific codec H264 is used to compress the result. The H264 codec is more computationally expensive than VP8 and it also needs more bandwidth. However, the final results obtained for the particular case of depth-information are much better than using VP8.

Finally, for labels codification, the chosen codec was VP8. Using this codec, quality losses that could result in user misidentifications in the remote computer, can be expected. In order to prevent these situations, the following strategy is proposed. Since encoders usually join together colors being too close, we propose spacing them before codification. The values 0 to 15 of user labels are translated into values from 0 to 255, preventing the encoder to mix up them. In Figure 2 the conversion equivalences are shown. With these new values, labels are stored into a luminance channel and then compressed.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	17	34	51	68	85	102	119	136	153	170	187	204	221	238	255

Figure 2. Correspondence of user labels to colors to avoid misidentifications after data compression.

Due to the computation requirements of the proposed system, in order to code and send the 3D information, it has been designed to process video sequences in parallel using threads. The RGB image, depth map, user labels and skeletons are acquired at the same time. Each type of data is then coded separately in parallel using the 'Boost' library.

3.4 Data transmission

System servers are registered in a CORBA naming service after starting, so that the system's central computer can find them, without needing to know their IP addresses. When the central computer establishes a communication and asks for data, the server collects the information from every local attached Kinect, encodes it and sends it continuously to the remote client. The client is constantly receiving data sent from each server, but it may not use all information that arrives to the client. The system is designed to decode only the information that is to be used. To this end, mutual exclusion techniques are employed.

Compressed information is stored into CORBA data arrays. Then, the server sends data by invoking remote methods in each client. These methods receive input arguments containing the compressed RGB image, depth map, user labels and uncompressed skeletons. The information is sent only to the clients who have previously registered on the server.

3.5 Point cloud fusion

Once the Kinect cameras have been installed in the location to be monitored, a first system calibration has to be performed. The goal of calibration is for the central client computer to find the proper transformation matrices to align and fuse the received point clouds. One of the devices is chosen to be the center of the coordinate system and then rotations are calculated from the other cameras. Given each pair of point clouds, the objective is to calculate a 4x4 rotation and translation matrices by solving the system of equations $\mathbf{B} = \mathbf{R}\mathbf{A} + \mathbf{t}$, where \mathbf{A} and \mathbf{B} are three-component points, \mathbf{R} is a 3x3 rotation matrix and \mathbf{t} is a three-component column translation vector.

Within the system interface, the calibration step will prompt the user to check the correspondence of at least 3 common points in different clouds of points. This calibration clouds are not yet compressed for better results. For this purpose, it is useful to place an object into the intersection area of different infrared patterns. Figure 3 shows this process using points belonging to a chair and a box on top of it. Taking the marked common points, the system can approximate an initial calibration, which serves to rotate the point clouds and apply the algorithm ICP (Iterative Closest Point), which refines the calibration. These rotation matrices have only to be computed the first

time the system is deployed and they are later used to rotate all information coming from the different cameras, including user skeletons, in different executions.



Figure 3. Initial calibration to determine rotation and translation matrices. These matrices are used to fuse 3D information coming from different Kinect cameras.

3.6 Skeleton merging

We distinguish between Kinect input skeletons and system's final output skeletons. Each output skeleton is computed dynamically from a linked list of input skeletons, which are merged and averaged together.

Figure 4 depicts the process of output skeletons computation. To merge the input skeletons, the first step is to apply rotation matrices to the detected joints. Once all the skeletons from all the cameras are in the same reference system, the algorithm for skeleton merging can be applied. The first step is to check for changes in the previous linked lists of skeletons, which contain the correspondence among similar input skeletons. These lists include the camera identifiers, the input skeletons identifiers and the output skeletons identifiers. Every time a remote camera considers a Kinect skeleton to have disappeared, it is removed from its linked list. Accordingly, every time a camera provides information of a new skeleton, the system tries to add it into a linked list. To this end, it compares the distance between the joints representing the two skeleton heads. This strategy has been used in other human skeleton tracking proposals[21]. If the distance is less than 15 cm, the system considers both input skeletons to be the same output skeleton. Evaluating skeleton matching during 25 consecutive frames strengthens the robustness of the system. In case no correspondence can be found to include the new skeleton into any existing linked list, then it is considered as a new output skeleton and a new linked list is built up. The final joints are calculated by averaging all joints from different cameras. If the confidence of a specific joint within a skeleton is less than 0.5, its position is not used to calculate final joint. The advantage of this design is that in case one camera cannot detect a given joint, its position can be determined from the information given of other camera. The probability of having all joints describing the skeleton available, and with accurate positions, grows with the number of cameras detecting the skeleton.

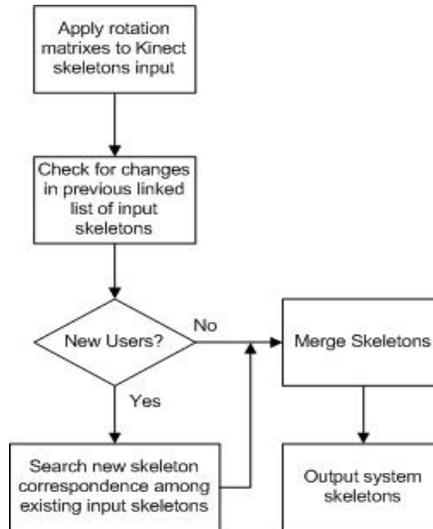


Figure 4. Skeleton merging algorithm.

3.7 Data visualization and skeleton tracking

The central client interface uses PCL to visualize the final reconstructed space and allows real-time tracking of labeled people inside the area covered by the cameras. Figure 5 shows a labeled skeleton being monitored in real-time by 5 cameras. The final scene can be rotated and analyzed from any point of view. However, there is a limitation on the available frame rate due to the VTK rendering methods employed by the system. When the number of points in the final point cloud grows, the frame rate is reduced. This is not a problem related to the compression/decompression computational cost, but the visualization methods included in PCL. Future releases of PCL are said to address this problem by adding native OpenGL rendering [17]. In order to guarantee usability of the system, despite of this problem, the user interface allows subsampling the number of points to visualize. Test and results section gives some figures of performance with 5 device cameras.

Finally, the system is designed to store skeleton information of people within the tracking area, associated to their labeled output skeletons. Once recording has started, all the user joints are stored in a raw file that can be further used to reproduce any situation occurred or serve as an input for another application for further situation analysis (i.e. movement recognition application). For applications that require human activity registration, the needed storage space is much smaller than in conventional 2D video systems, since only the skeletons may need to be stored.



Figure 5. Labeled skeleton and associated joints obtained from the combination of 5 cameras information.

4. TESTS AND RESULTS

The system has been tested using 5 Kinect cameras connected to three personal computers: two desktop computers equipped with an Intel i5-2400 CPU running at 3.10 GHz and a laptop equipped with an Intel i7-3610QM CPU running at 2.30GHz. The client computer was an Intel Xeon X5650 with 24 cores running at 2.66 GHz and equipped with aNvidia GeForce GTX580 GPU. The server computers were running Windows 7 operating system, while the central client was running Linux Fedora 16.

The aim of these tests was to check the performance of the final system in real conditions. The first tests conducted included data transmission, coding/decoding and visualization measurements. Using RGB input at 640x480 px resolution and coding depth information to 320x240 px color maps, the theoretical limit on the number of cameras that can be connected over a Gigabit Ethernet is higher than 50 for a framerate of 30 fps. These numbers do not consider the overhead of TCP connection. In our experimental tests, performed with up to 5 cameras (the maximum number of cameras we managed to have), the obtained framerate was actually 30 fps. However, in our tests, we detected that even having every server transmitting at 30 fps and the client computer decoding all cameras information at the same framerate, the final scene rendering was affected by VTK visualization limitations of PCL. As explained above, the achieved framerate depends on the number of points in the cloud. Table 1 shows how visualizing a cloud of points constructed from 5 cameras renders only at 7 fps if every point is drawn onto the screen. Subsampling the number of points by 16, which actually still provides a very nice representation of the scene, improves performance to 29 fps.

FPS	Rendered points
7	$\approx 5 \cdot 307200 = 1.536.000$
13	$\approx 5 \cdot 307200 / 4 = 384000$
22	$\approx 5 \cdot 307200 / 9 = 170666$
29	$\approx 5 \cdot 307200 / 16 = 96000$

Table 1. Frame rate obtained during visualization using VTK for 5 cameras 3D reconstruction. This is a limitation of VTK, not the system itself.

The second battery of tests conducted included situations to measure the behavior of the system with different people in the scene and measure the robustness to self-user and inter-user occlusions. The first test consisted in a user placed in the center of the scene. Meanwhile, another user revolves around him or her, so that some cameras can see the first user and some others cannot. The goal is to test the robustness of the system when different cameras detect and lost Kinect skeletons over and over again. The test was conducted ten times using combinations of different height users and the obtained result was always successful in every situation, since the system did not confuse users or incorrectly merged their skeletons. The second test consisted in users sitting and getting up from chairs in an office space. This test measured the robustness of the system to some skeletons joint occlusions, since some of the cameras are not able to provide accurate positions for body parts behind tables or chairs. The test was repeated for ten different people sitting in front of the four tables in the scene in Figure 6 and again the system worked perfectly. The third test consisted in covering and uncovering one by one the different cameras in the scene while 5 people were being tracked in the scene. The goal was to test what happens when

multiple Kinect input skeletons are removed and detected at the same time. The result was again satisfactory and every computed output skeletons in the scene kept being tracked consistently.

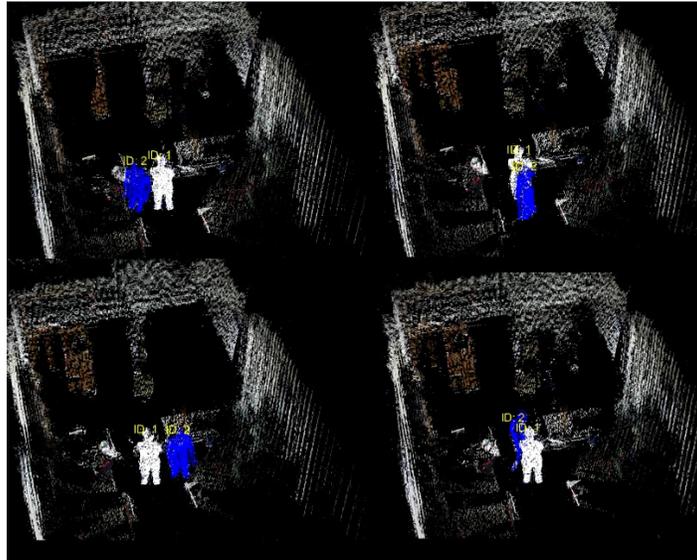


Figure 6 Inter-user occlusion test within a space monitored by 5 cameras.

5. CONCLUSIONS

This article describes a distributed CORBA system for remote space 3D monitoring using Kinect consumer depth cameras. Due to the high bandwidth needs of these cameras, the maximum number of cameras that can be connected to a single computer is usually two. The solution provided in this paper includes a client-server application that can handle the information acquired by any number of cameras connected to any number of computer servers. Since one Kinect camera can only detect precisely the depth information within a field of view of 3.5 meters, the proposed system solves, at the same time, the limitation on the size of the location that can be monitored precisely. A central client computer can be used to monitor the reconstructed 3D space in real time and track the movements of people within it.

In the central client computer, a skeleton-merging algorithm is used to combine the information of skeletons belonging to the same person, but generated by different Kinect cameras, into a single output skeleton. The tests conducted showed that this algorithm is robust under several situations, avoiding unwanted duplication of skeletons when new people enter the scene or under camera or inter-user occlusions. Moreover, the algorithm combines the information coming from each skeleton joint independently, so the 3D location of joints in the final generated skeleton is more precise, having been averaged among all the cameras detecting that joint. In case a self-user or a inter-user occlusion causes one joint not to be detected by one or more of the cameras, its position is reconstructed using the information coming from cameras in which the joint has been detected with enough confidence. Output skeleton movements can be stored in raw files for further analysis of situations.

This system provides a very precise and convenient way of monitoring a 3D space at an affordable price. People activity in the scene can be registered for further analysis and the storage needs to keep track of human behavior under different circumstances can be much lower than for

conventional 2D systems, if only the skeletons are needed. Future research tasks will include designing a top activity recognition layer that could monitor people behavior and interactions.

ACKNOWLEDGEMENTS

This research was supported in part by the Ministerio de Ciencia e Innovación under project TIN2010-20529 and Junta de Castilla y León under project VA171A11-2.

REFERENCES

- [1] Sage, K. & Young S. (1999) Security Applications of Computer Vision, Aerospace and Electronic Systems Magazine, IEEE 14(4):19-29.
- [2] Boutaina, H., Rachid, O.H.T. & Mohammed, E.R.T. (2013) Tracking multiple people in real time based on their trajectory, in Proc. of Intelligent Systems: Theories and Applications (SITA), 8th International Conference on, pp.1-5, Rabat, Morocco.
- [3] Strbac, M., Markoviu, M., Rigolin, L. & Popoviu, D.B. (2012) Kinect in Neurorehabilitation: Computer vision system for real time and object detection and instance estimation, in Proc. Neural Network Applications in Electrical Engineering (NEUREL), 11th Symposium on, pp. 127-132, Belgrade, Serbia.
- [4] Martín Moreno, J., Ruiz Fernandez, D., Soriano Paya, A. & Berenguer Miralles, V. (2008) Monitoring 3D movements for the rehabilitation of joints in Physiotherapy, in Proc. Engineering in Medicine and Biology Society (EMBS), 30th Annual International Conference of the IEEE, pp. 4836-4839, Vancouver, Canada.
- [5] Ukita, N. & Matsuyama, T. (2002) Real-Time Cooperative Multi-Target Tracking by Communicating Active Vision Agents, in Proc. Pattern Recognition, 16th International Conference on, vol.2, pp.14-19.
- [6] RBodor, R., Jackson, B. & Papanikolopoulos, N. (2003) Vision-Based Human Tracking and Activity Recognition, in Proc. of the 11th Mediterranean Conf. on Control and Automation, pp. 18-20.
- [7] Poppe R. (2007). Vision-based human motion analysis: An overview, Computer Vision and Image Understanding 108:4-18.
- [8] Weinland, D., Ronfard, R. & Boyer, E. (2011) A survey of vision-based methods for action representation, segmentation and recognition, Computer Vision and Image Understanding 115(2):224-241.
- [9] Kinect for Windows (2013) Kinect for Windows. Retrieved from <http://www.microsoft.com/en-us/kinectforwindows/develop/overview.aspx>, last visited on July 2013.
- [10] Burrus, N. (2013) RGBDemo. Retrieved from <http://labs.manctl.com/rgbdemo/>, last visited on July 2013.
- [11] KinectTCP (2013) KinectTCP. Retrieved from <https://sites.google.com/a/temple.edu/kinecttcp/>, last visited on July 2013
- [12] OpenNI (2013) OpenNI. Plataform to promote interoperability among devices, applications and Natural Interaction (NI) middleware. Retrieved from <http://www.openni.org>, last visited on July 2013.
- [13] Prime Sense (2013) Prime Sense. Retrieved from <http://www.primesense.com/>, last visited June 2013.
- [14] Martínez Rach, M.O., Piñol, P., López Granado, O. & Malumbres, M.P. (2012) Fast zerotree wavelet depth map encoder for very low bitrate, in Actas de las XXIII Jornadas de Paralelismo, Elche, Spain.
- [15] Joon-Heup, K., Moon-Sang J. & Jong-Tae, P. (2001) An Efficient Naming Service for CORBA-based Network Management, in Integrated Network Management Proceedings, IEEE/IFIP International Symposium on, pp.765-778, Seattle, USA
- [16] PCL (2013) Point Cloud Library. Retrieved from <http://pointclouds.org/>, last visited on July 2013.
- [17] PCL Developers Blog (2013) PCL Developers Blog. Retrieved from <http://pointclouds.org/blog/>, last visited on July 2013.
- [18] Parajuli, M., Tran, D.; Wanli, Ma; Sharma, D. (2012) Senior health monitoring using Kinect, in Proc. Communications and Electronics (ICCE), Fourth International Conference on, pp. 309-312, Hue, Vietnam.
- [19] The WebM Project (2013) The WebM Project. Retrieved from <http://www.webmproject.org>, last visited on July 2013.

- [20] Pece, F., Kautz, J. & Weyrich, T. (2011) Adapting Standard Video Codecs for Depth Streaming, in Proc. of the 17th Eurographics conference on Virtual Environments & Third Joint Virtual Reality (EGVE - JVRC), Sabine Coquillart, Anthony Steed, and Greg Welch (Eds.), pp. 59-66, Aire-la-Ville, Switzerland.
- [21] García, J., Gardel, A., Bravo I., Lázaro, J.L., Martínez, M. & Rodríguez, D. (2013). Directional People Counter Based on Head Tracking, IEEE Transactions on Industrial Electronics 60(9): 3991-4000.

Authors

Mario Martínez-Zarzuela was born in Valladolid, Spain; in 1979. He received the M.S. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2004 and 2009, respectively. Since 2005 he has been an assistant professor in the School of Telecommunication Engineering and a researcher in the Imaging & Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. His research interests include parallel processing on GPUs, computer vision, artificial intelligence, augmented and virtual reality and natural human-computer interfaces.



Miguel PedrazaHueso was born in Salamanca, Spain, in 1990. He received his title in Telecommunication Engineering from the University of Valladolid, Spain, in 2013.



Since 2011, he has collaborated with Imaging & Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. His research interests include applications with Kinect and augmented reality.

Francisco Javier Diaz-Pernas was born in Burgos, Spain, in 1962. He received the Ph.D. degree in industrial engineering from Valladolid University, Valladolid, Spain, in 1993. From 1988 to 1995, he joined the Department of System Engineering and Automatics, Valladolid University, Spain, where he has worked in artificial vision systems for industry applications as quality control for manufacturing. Since 1996, he has been a professor in the School of Telecommunication Engineering and a Senior Researcher in Imaging & Telematics Group of the Department of Signal Theory, Communications, and Telematics Engineering. His main research interests are applications on the Web, in intelligent transportation system, and neural networks for artificial vision.



David González-Ortega was born in Burgos, Spain, in 1972. He received his M.S. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2002 and 2009, respectively. Since 2003 he has been a researcher in the Imaging and Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. Since 2005, he has been an assistant professor in the School of Telecommunication Engineering, University of Valladolid. His research interests include computer vision, image analysis, pattern recognition, neural networks and real-time applications.



Míriam Anton-Rodríguez was born in Zamora, Spain, in 1976. She received her M.S. and Ph.D. degrees in telecommunication engineering from the University of Valladolid, Spain, in 2003 and 2008, respectively. Since 2004, she is an assistant professor in the School of Telecommunication Engineering and a researcher in the Imaging & Telematics Group of the Department of Signal Theory, Communications and Telematics Engineering. Her teaching and research interests include applications on the Web and mobile apps, bio-inspired algorithms for data mining, and neural networks for artificial vision.



COLOR SATELLITE IMAGE COMPRESSION USING THE EVIDENCE THEORY AND HUFFMAN CODING

Khaled SAHNOUN and Noureddine BENABADJI

Laboratory of Analysis and Application of Radiation (LAAR)
Department of Physics, University of Sciences and Technology of Oran
(USTOMB)
B.P. 1505, El M'nouar, 31024, ORAN, Algeria

sahnounkhaled@hotmail.fr and benanour2000@yahoo.com

ABSTRACT

The color satellite image compression technique by vector quantization can be improved either by acting directly on the step of constructing the dictionary or by acting on the quantization step of the input vectors. In this paper, an improvement of the second step has been proposed. The k -nearest neighbor algorithm was used on each axis separately. The three classifications, considered as three independent sources of information, are combined in the framework of the evidence theory. The best code vector is then selected, after the image is quantized, Huffman schemes compression is applied for encoding and decoding.

KEYWORDS

Vector quantization, compression, k -nearest neighbor, Huffman coding, evidence theory.

1. INTRODUCTION

By definition, the compression with vector quantization [1] accepts an input vector \vec{x} of n dimension and replaced by a vector \vec{y} of the same size belonging to a dictionary which is a finite set of code vectors $(w_j)_{j \in [1, \dots, N]}$ also called classes, or centroids, since they are calculated by an average iterative of vectors \vec{x} . The quantization step based on the nearest neighbor rule: vector \vec{x} to classify is assigned to one class of $(w_j)_{j \in [1, \dots, N]}$ under the condition that this assignment produces the smallest distortion. Such assignment rule may be too drastic in cases where the distances between the vector \vec{x} and two centroids are very close. A possible improvement to avoid this hard decision would be to consider each color components independently to obtain a classification by component. In this work, components R, G and B are considered as three independent information sources. In a first step, the K -nearest neighbor rule is applied to all three components, then generating three sets of potential classes. This step, taking into account K - neighbors and not one, allows considering uncertainty according to each of color components and to push decision making. Finally, the decision of the assignment final class of \vec{x} is done after combining these three classifications. This technique refers to methods of data fusion. Among all the tools available in this domain, we decide to use the evidence theory [2]

which makes it possible firstly to process uncertain information and secondly to combine information from several sources. In the framework of this theory, several decision rules are defined to enable us selecting the final class of \vec{x} .

2. USE THE EVIDENCE THEORY

2.1. Basic principle

Let $\Omega = \{w_1, \dots, w_N\}$ a set of all possible classes for \vec{x} , called the frame of discernment and corresponding to the dictionary in our application, the evidence theory extends over the entire power of Ω , noted 2^Ω . We define an initial mass function m of 2^Ω in $[0,1]$ which satisfies the following conditions:

$$\sum_{A \subset \Omega} m(A) = 1 \text{ And } m(\emptyset) = 0 \quad (1)$$

Where \emptyset is the empty set $m(\emptyset)$, $m(A)$ quantifies the belief that be given to the fact that the class belongs to the subset A of Ω . Subsets A like $m(A) > 0$ are called focal elements. Two initial mass functions m_1, m_2 representing the respective information from two different sources, can be combined according to Dempster rule [3]:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{1 - k} \quad \forall A \in 2^\Omega, A \neq \emptyset \quad (2)$$

k is called the conflict factor and represents the detuning between the two sources, Note that the combination of Dempster also called orthogonal sum and denoted $m = m_1 \oplus m_2$.

After combination, a decision on the most likely element of Ω must be taken. Several decision rules are possible, but one of the most used is the ‘‘maximum of Pignistic probability’’ Presented by Smets [4] which uses the Pignistic transformation, and allows to evenly distributing the weight associated with a subset of Ω , on each of its elements:

$$BandP(\omega) = \sum_{\omega \in A \subset \Omega} \frac{m(A)}{|A|}, \forall \omega \in \Omega \quad (3)$$

$|A|$ is the Cardinal of A . The decision then goes to the element of Ω , where the value is the largest.

$$\omega^* = Arg \{max_{\omega \in \Omega} [BandP(\omega)]\} \quad (4)$$

2.2. Application to the vector quantization

All compression methods based on the Discrete Cosine Transform (DCT) following three steps, transformation of the difference image or predicted, quantizing and encoding the transformed coefficients [3]. The transformation is applied on blocks of 8×8 (pixels) and is defined by:

$$f(u, v) = \frac{2c(u)c(v)}{n} \sum_{x=0}^{n-1} \sum_{y=0}^{n-1} f(x, y) \cos(2x + 1) \frac{u\pi}{2n} \cos(2y + 1) \frac{v\pi}{2n}$$

$$\text{With } c(0) = \frac{1}{\sqrt{2}} \text{ and } c(u) = 1 \text{ if } u \neq 0. \quad (1)$$

N is the block size ($N = 8$ is selected), x, y are the coordinates in the spatial domain and u, v coordinates in the frequency domain. Each block is composed of 64 coefficients. The coefficient

(0.0) is denoted by DC and represents the average intensity of the block, others are denoted AC. The coefficients with the global information of the image are located in low frequency while those in the higher frequencies are often similar to noise [4].

2.3. The quantization

We propose to represent the information provided by each independent classification according to each of the three components (R, G, B) by a function of initial mass. These three functions $(m)_i, i \in \{R, G, B\}$ are created after calculating the K-nearest neighbor and before the final decision they allow to take into account the uncertainty associated with each axis. Thus, classes that are very close to each other on the same axis are grouped in the same focal element, and the decision is made only after having combined the results of the other two projections. For each axis identifies the most significant elements "k" by a distance d_i on this axis. The initial mass function constructed according to the axis i has three focal elements $A_i, \overline{A_i}, \Omega$. $\overline{A_i}$ is the complement of $A_i, i \in \Omega$. We construct:

$$A_i = \{\omega \in \Omega, \omega = \text{class}(\vec{x}) \mid d_i(\vec{x}, \vec{x}^*) \leq \varepsilon_i(\vec{x}_1, \vec{x}^*), \forall \vec{x}\} \quad (5)$$

$i \in \{R, G, B\}$. \vec{x}^* Is the vector for classifying and \vec{x}_1 its nearest neighbor depending on d_i . ε_i is a constant greater than 1 to take into account the sensitivity of the human visual system according to the axis i , if $\varepsilon_i = 1$ then $A_i = \{\text{class}(\vec{x}_1)\}$ is a singleton corresponding to the nearest neighbor. The masses are then assigned to the sets A_i taking into account the distribution of elements in the set A_i which is represented by the average distance between two of these elements. The initial mass function for the i axis then:

$$m_i(A_i) = \alpha_i e^{-\beta_i \overline{d}} \quad (6)$$

$$m_i(\overline{A_i}) = 1 - m_i(A_i) - m_i(\Omega) \quad (7)$$

$$m_i(\Omega) = 0,01 \quad (8)$$

α_i Is a constant and $\beta_i = 1/d_{max} \cdot d_{max}$ is the maximal distance between \vec{x}^* and of A_i elements in (R,G,B) space, \overline{d} is the average distance between each elements of A_i . Thus, More \overline{d} is large more the mass of A_i is small. m_R, m_G, m_B The three functions of initial mass from projections R, G and B respectively. Mass function resulting from the combination of three functions is obtained from equation (02):

$$m = m_R \oplus m_G \oplus m_B \quad (9)$$

Finally, assignment class of \vec{x} is selects from m equation (09) on the basis of maximum of Pignistic probability equation (03)

3. HUFFMAN COMPRESSION FOR R, G AND B

In the proposed compression method, before applying Huffman compression, quantization is applied as it will give better results. In quantization, compression is achieved by compressing a range of values to a single quantum value. When the given number of discrete symbols in a given stream is reduced, the stream becomes more compressible. After the image is quantized, Huffman compression is applied. The Huffman has used a variable-length code table for the encoding of each character of an image where the variable-length code table is derived from the estimated probability of occurrence for each possible value of the source symbol. Huffman has used a

particular method for choosing the representation for each symbol which has resulted in a prefix codes. These prefix codes expresses the most common source symbols using shorter strings of bits than are used for less common source symbols. In this way, we have achieved a compressed image. [5]

4. EXPERIMENTAL RESULTS AND DISCUSSION

Four different color images have been taken for experimental purpose. Simulation results for different images are given in Table 1. For measuring the originality of the compressed image Peak Signal to Noise Ratio (PSNR) is used, which is calculated using the formula

$$PSNR(db) = 10 \log_{10} 255^2 / MSE \quad (10)$$

Where MSE is the mean squared error between the original image f_{ij} and the reconstructed compressed image f'_{ij} of the size MN, which is calculated by the equation[6]

$$MSE = \frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N [f'_{ij} - f_{ij}]^2 \quad (11)$$

The algorithm realized in Builder C++ to code and to decode the satellite image, but all these Images are resized into a resolution of 256 X 256.



Figure 1. Original Satellite image (1)

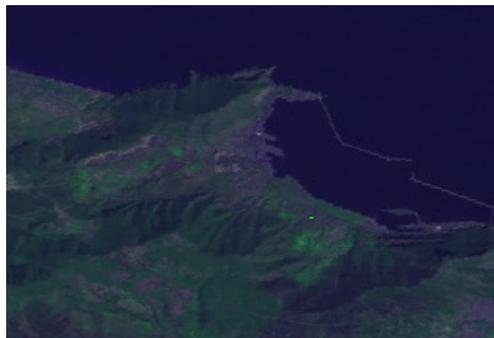


Figure 2. Reconstructed Satellite image (1)



Figure 3. Original image 'Lena'



Figure 4. Reconstructed image 'Lena'



Figure 5. Original Satellite image (2)



Figure 6. Reconstructed Satellite image (2)

Image	PSNR(db)	CR
Lena	32.92	13.66
Satellite image (1)	34.68	15.89
Satellite image (2)	33.89	15.55

Table 1. The compression ratios and PSNR values derived for imageries

It can be seen from the Table 1 that for all the images, the PSNR values are greater than 32, the compression ratios achievable different. It is clearly evident from the table that for two types of images with reasonably good PSNR values clearly indicate that the compression ratio achievable for satellite imageries is much higher compared to the standard Lena image.

5. CONCLUSIONS

To improve the quantization step of the input vectors according to the code vectors present in a dictionary. Using the evidence theory has obtaining promising results. In our study setting, a vector quantization was performed on each of the three colors R, G and B, according to the color dispersion of the K-nearest neighbor. The results show that the use of evidence theory during the quantization step and Huffman coding is an improvement of the quality of the reconstructed images

REFERENCES

- [1] A. Gersho and R. M. Gray. Vector Quantization and Signal Compression. Kluwer Academic Publishers, 1991.
- [2] G. Shafer. A mathematical theory of evidence. Princeton University Press, 1976.
- [3] A. Dempster. Upper and Lower Probabilities Induced by Multivalued Mapping. Ann. Math. Statist , 38:325– 339, 1967.
- [4] P. Smets. Constructing the pignistic probability function in a context of uncertainty. Uncertainty in Artificial Intelligence, 5 :29–39, 1990. Elsevier Science Publishers.
- [5] H.B.Kekre, Tanuja K Sarode, Sanjay R Sange “ Image reconstruction using Fast Inverse Halftone & Huffman coding Technique”, IJCA, volume 27-No 6, pp.34-40, 2011
- [6] Varun Setia, Vinod Kumar, “Coding of DWT Coefficients using Run-length coding and Huffman Coding for the purpose of Color Image Compression”, IJCA, volume 27-No 6, pp.696-699, 2012.

ON-BOARD SATELLITE IMAGE COMPRESSION USING THE FOURIER TRANSFORM AND HUFFMAN CODING

Khaled SAHNOUN and Noureddine BENABADJI

Laboratory of Analysis and Application of Radiation (LAAR)
Department of Physics, University of Sciences and Technology of Oran
(USTOMB)
B.P. 1505, El M'nouar, 31024, ORAN, Algeria

sahnounkhaled@hotmail.fr, benanour2000@yahoo.com

ABSTRACT

The need to transmit or store satellite images is growing rapidly with the development of modern communications and new imaging systems. The goal of compression is to facilitate the storage and transmission of large images on the ground with high compression ratios and minimum distortion. In this work, we present a new coding scheme for satellite images. At first, the image will be downloaded followed by a fast Fourier transform FFT. The result obtained after FFT processing undergoes a scalar quantization (SQ). The results obtained after the quantization phase are encoded using entropy encoding. This approach has been tested on satellite image and Lena picture. After decompression, the images were reconstructed faithfully and memory space required for storage has been reduced by more than 80%

KEYWORDS

Compression, Encoding Entropy, FFT, Scalar Quantization, Satellite

1. INTRODUCTION

Compression of satellite images is a set of techniques and methods used to reduce the volume of data without losing important information. The reduction will take place either by lossless algorithm which the original data will be found, either lossy algorithm where the retrieved data after compression are reasonable reconstruction of the original data. [1] Reduce the amount of data used to store more information on a single media or take less time for data transmission to the ground. [2] In some cases the volume of data is such that it would be almost impossible to manage without using a compression operation with the best possible compromise between compression ratio and the quality of reproduction of images. [3] In this paper we propose a technique based on the Fourier transform and scalar quantization (SQ) to drastically increase the compression ratio, while maintaining a satisfactory quality of the reconstructed image. This paper is organized as follows Fourier transforms is illustrated in section two. The proposed scheme is presented in part three. Simulation results are given in section four and finally a conclusion in Part Five.

2. FOURIER TRANSFORM

The Fourier transform, also known as frequency analysis or spectral involved in the implementation of many digital techniques for processing signals and images. [4] It is found in applications such as direct harmonic analysis of musical signals and vibrations, but also reduces the rate coding of speech and music (mp3), voice recognition, improving the quality of images, compression, and digital transmissions. Applying a Fourier transform give a complex image. [5] In general, we calculated the module F_m and the phase F_p of the source image, and we represent the module. These two images can be defined as following:

$$\text{Re}(F(f)(x, y)) = F_m(x, y) \cos(F_p(x, y)) \rightarrow (1)$$

$$\text{Im}(F(f)(x, y)) = F_m(x, y) \sin(F_p(x, y)) \rightarrow (2)$$

Where Re and Im denotes the real and imaginary parts. One then finds that F_p is not unique. In general, to represent the transform, it is only the module.

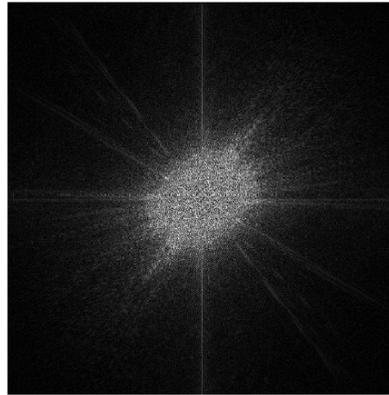


Figure 1. Module blue channel butterfly

Note that the image is square, we completed the butterfly image with black to make it square. In addition, we work more frequently with square images. [6]

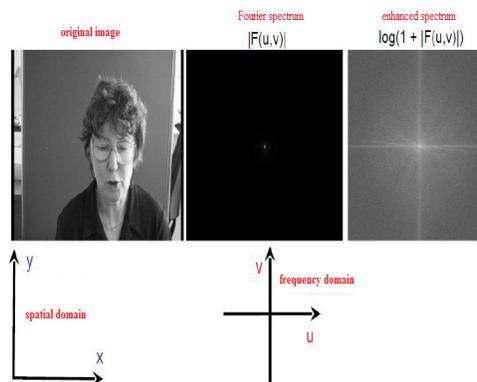


Figure 2. Fourier Transform

The Fourier transform of a real image can be expressed as follows:

$$F(u, v) = \frac{1}{mn} \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} f(x, y) \cdot e^{-2\pi i \left(\frac{ux}{m} + \frac{vy}{n} \right)}$$

With u and $v = 0..N-1 \rightarrow (3)$

Reverse:

$$f(x, y) = \sum_{u=0}^{m-1} \sum_{v=0}^{n-1} F(u, v) \cdot e^{2\pi i \left(\frac{ux}{m} + \frac{vy}{n} \right)}$$

With x and $u = 0..N-1 \rightarrow (4)$

The variables u and v used in the equation (3) are variable frequency (frequency domain), x and y used in the equation (4) are variable in the spatial domain. $F(u,v)$ Is often represented by its amplitude and phase, rather there are the real and imaginary parts, the formula is given by:

$$\text{amplitude}(F(u, v)) = \sqrt{R^2(u, v) + I^2(u, v)} \rightarrow (5)$$

$$\text{phase}(F(u, v)) = \tan^{-1} \left[\frac{I(u, v)}{R(u, v)} \right] \rightarrow (6)$$

3. PROPOSED APPROACH

The main philosophies of our image compression technique based on the fast Fourier transform (FFT). The general architecture of the system coding of our method relied primarily on the steps shown in the following figure:

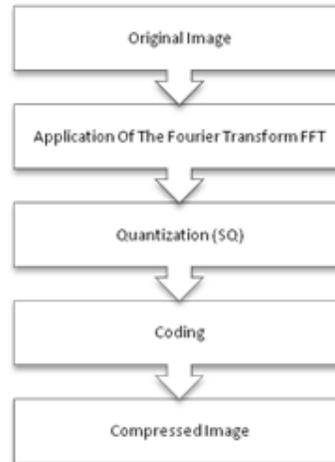


Figure 3. The steps of the proposed method

3.1. Read the image source

The input image is satellite picture, the size will be equal to 2^n , in our case it is 2^8 (256 * 256).

3.2. Calculate the Fourier transform (2D FFT) for this image

The type of data returned by the FFT is complex, which contains real and imaginary parts. The real part is the amplitude, and the imaginary part is the phase. In the proposed method, just the amplitude is concerned, which is the only party represented in the surface and the displays of the transformation results.

3.3. Quantization

The compression technique by quantization can be improved either by acting directly on the step of constructing the dictionary or by acting on the quantization step of the input pixels. In this method, an improvement of the second step has been proposed (input vector). The scalar quantization of each is the approximate value of the random signal $x(t)$ by a value that belongs to a finite set of codes $\{y_1, y_2, \dots, y_l\}$. At any amplitude x in the interval $[x_{i-1}, x_i]$, there corresponds a quantized value y_i situated in that Interval. [8]

3.4. Coding

In the coding phase we used the RLE encoding. It is a compression mode of the simplest and oldest, it is both easy to implement and fast execution. The algorithm is to identify and remove redundant information by encoding more compact form any sequence of bits or characters is replaced by the same number of occurrences of a couple, bit or character repeated. The image coding by RLE method coded the sequence of identical gray pixel values, assigning the three parameters, the position (x, y) of the first pixel in the sequence, the gray value of the first pixel and the length of the sequence. Finally the Huffman algorithm is applied which is a compression algorithm capable of generating variable length codes to a whole number of bits. This algorithm can achieve good results, but it should be kept the codebook used between the compression and decompression. [9]

4. EXPERIMENTAL RESULTS

4.1. Evaluation of compression and loss

Compression ratio (T):

$$Q = \frac{\text{initial size}}{\text{final size}} \rightarrow (7)$$

$$T = \frac{1}{Q} \rightarrow (8)$$

Compression gain:

$$G = \frac{\text{initial size} - \text{final size}}{\text{final size}} \rightarrow (9)$$

Mean Squared Error:

$$MSE = \frac{1}{N} \sum_{i=0}^{N-1} (n_{comp}(i) - n(i))^2 \rightarrow (10)$$

Peak Signal to Noise Ratio

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \text{ db} \rightarrow (11)$$

A powerful compression algorithm has a gain of maximum compression and a minimum mean square error. [7] The compression ratio, the mean squared error and PSNR are calculated by equations (8) and (6), (11). The proposed scheme has been tested by satellite image and Lena picture.

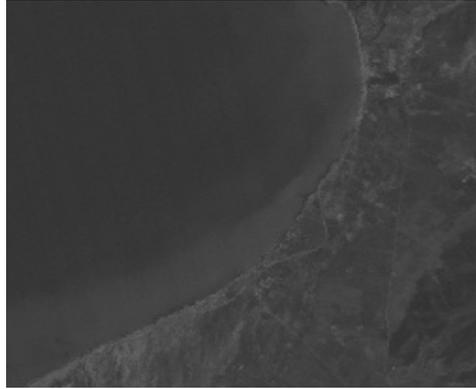


Figure 4. Original satellite image (1)

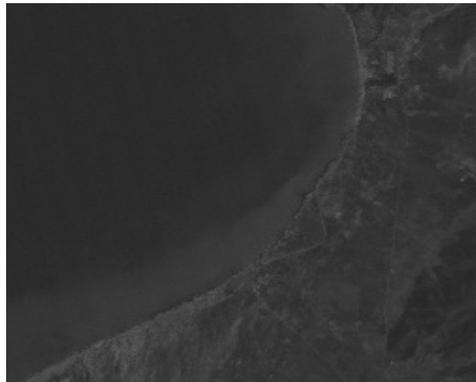


Figure 5. Satellite image reconstructed (2)



Figure 6. Original Lena image



Figure 7. Reconstructed Lena image



Figure 8. Original satellite image (2)



Figure 9. Satellite image reconstructed (2)

Table 1. Table of results with satellite image and Lena picture 256 * 256

Image	Lena	Satellite (1)	Satellite (1)
MSE	11.67	9.92	8.87
PSNR (db)	37.49	38.20	33.79
T (%)	65.52	87.27	76..3

5. CONCLUSIONS

In this paper we have presented an approach for still image compression based on the Fourier transform and scalar quantization (SQ) and also the entropy encoding. The compression and decompression algorithm that we have developed in this article is able to compress satellite images and grayscale picture with high compression ratio and return with a better quality. Thus, tests on Lena image and other images show the superiority of this algorithm with respect to other compression methods.

REFERENCES

- [1] LAHDIR MOURAD “new approach to image compression based on wavelet and fractal for Meteosat image application” PhD Thesis University Mouloud Mammeri UMMTO, Tizi-Ouzou Algeria 2011.
- [2] STEVEN PIGEON “Contributions to the Data Compression” PhD Thesis University of Montreal Canada 2001.
- [3] SYLVAIN ARGENTIERI ”Introduction to image compression” Institute Of Robotics And Intelligent Systems Paris 6 France 2009 .
- [4] BENJAMIN HARBELOT YOAN “Project mathematical Fourier Transform” University of Burgundy France 2010.
- [5] ERIC FAVIER. “Image analysis and processing, principles of computer vision”. ENISE France.
- [6] A. AMAAR, E.M. SAAD, I. ASHOUR AND M. ELZORKANY Electronics Department, National Telecommunication Institute (NTI) Image “Compression Using K-Space Transformation Technique” Recent Researches in Communications, Electronics, Signal Processing and Automatic Control 2012.
- [7] R.KUMAR, K.SINGH, R.KHANNA Dept. of ECE Thapar University, Patiala “Satellite Image Compression using Fractional Fourier Transform” International Journal of Computer Applications (0975 – 8887) Volume 50 – No.3, July 2012.
- [8] A. MOULAY LAKHDAR, M. KANDOUICI, B. BELGHEIT “Image compression by wavelet transforms and adaptive vector quantization” University Of Bechar and Department of Electronic faculty of Science & Engineering University Djilali Liabes. IMAGE conference Biskra Algeria 2009.
- [9] ERIC FAVIER “image compression Part H” National School of engineers Saint-Etienne France 2008.

INTENTIONAL BLANK

AMINO ACID INTERACTION NETWORK PREDICTION USING MULTI-OBJECTIVE OPTIMIZATION

Md. Shiplu Hawlader¹ and Saifuddin Md. Tareeq²

¹Department of Computer Science & Engineering,
University of Asia Pacific, Dhaka, Bangladesh
shiplu.cse@uap-bd.edu

²Department of Computer Science & Engineering,
University of Dhaka, Dhaka, Bangladesh
smtareeq@cse.univdhaka.edu

ABSTRACT

Protein can be represented by amino acid interaction network. This network is a graph whose vertices are the proteins amino acids and whose edges are the interactions between them. This interaction network is the first step of proteins three-dimensional structure prediction. In this paper we present a multi-objective evolutionary algorithm for interaction prediction and ant colony probabilistic optimization algorithm is used to confirm the interaction.

KEYWORDS

Protein Structure, Interaction Network, Multi-objective Optimization, Genetic Algorithm, Ant Colony Optimization

1. INTRODUCTION

Proteins are biological macromolecules performing a vast array of cellular functions within living organisms. The roles played by proteins are complex and varied from cell to cell and protein to protein. The best known role of proteins in a cell is performed as enzymes, which catalyze chemical reaction and increase speed several orders of magnitude, with a remarkable specificity. And the speed of multiple chemical reactions is essential to the organism survival like DNA replication, DNA repair and transcription. Proteins are storage house of a cell and transports small molecules or ions, control the passages of molecules through the cell membranes, and so forth. Hormone, another kind of protein, transmits information and allow the regulation of complex cellular processes.

Genome sequencing projects generate an ever increasing number of protein sequences. For example, the Human Genome Project has identified over 30,000 genes [1] which may encode about 100,000 proteins. One of the first tasks when annotating a new genome is to assign functions to the proteins produced by the genes. To fully understand the biological functions of proteins, the knowledge of their structure is essential.

Proteins are amino acids chain bonded together in peptide bonds, and naturally adopt a native compact three-dimensional form. The process of forming three-dimensional structure of a protein is called protein folding and this is not fully understood yet in System Biology. The process is a result of interaction between amino acids which form chemical bond to make protein structure.

In this paper, we proposed a new algorithm to predict a interaction network of amino acids using two new emerging optimization techniques, multi-objective optimization based on evolutionary clustering and ant colony optimization.

2. AMINO ACID INTERACTION NETWORK

Amino acids are the building blocks of proteins. Protein a sequences of amino acids linked by peptide bond. Each amino acid has the same fundamental structure, differing only in the side-chain, designated the R-group. The carbon atom to which the amino group, carboxyl group, and side chain (R-group) are attached is the alpha carbon ($C\alpha$). The alpha carbon is the common reference point for coordinates of an amino acid structure. Among the 20 amino acids some are acidic, some are basic, some are polar, some non-polar. To make a protein, these amino acids are joined together in a polypeptide chain through the formation of a peptide bond. The structure, function and general properties of a protein are all determined by the sequence of amino acids that makes up the primary sequence. The primary structure of a protein is the linear sequence of its amino acid structural units and it is a part of whole protein structure. The two torsion angles of the polypeptide chain, also called Ramachandran angles, describe the rotations of the polypeptide backbone around the bonds between $N - C\alpha$ (called Phi angle, ϕ) and $C\alpha - C$ (called Psi angle, ψ). Torsion angle is one of the most important parameter of protein structure and controls the protein folding. For each type of the secondary structure elements there is a characteristic range of torsion angle values, which can clearly be seen on the Ramachandran plot [2].

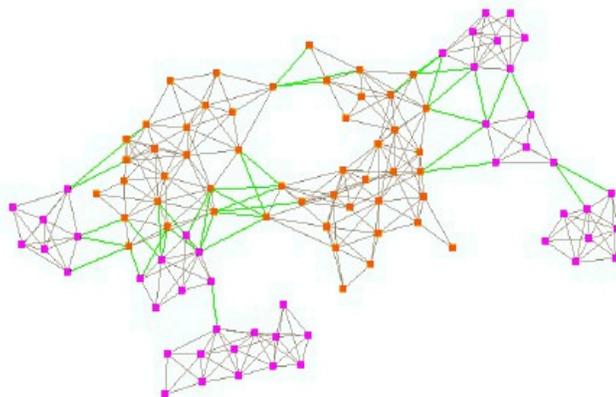


Figure 1: SSE-IN of 1DTP protein. Green edges are to be predicted by ant colony algorithm

Another important property of protein is hydrophobicity. Proteins tertiary structure's core are hydrophobic and the amino acids inside core part do not interact much as like their counterpart hydrophilic, those made the outer side of the protein structure. Many systems, both natural and artificial, can be represented by networks, that is by site or vertices connected by link or edges. Protein also be represented as a network of amino acid whose edges are the interactions or the protein functions between amino acids. The 3D structure of a protein is determined by the coordinates of its atoms. This information is available in Protein Data Bank (PDB) [3], which regroups all experimentally solved protein structures. Using the coordinates of two atoms, one can compute the distance between them. We define the distance between two amino acids as the distance between their $C\alpha$ atoms. Considering the $C\alpha$ atom as a center of the amino acid is an

approximation, but it works well enough for our purposes. Let us denote by N the number of amino acids in the protein. A contact map matrix is an $N \times N$, 0 - 1 matrix, whose element (i, j) is 1 if there is a contact between amino acids i and j and 0 otherwise. It provides useful information about the protein. For example, the secondary structure elements can be identified using this matrix. Indeed, α - helices spread along the main diagonal, while β - sheets appear as bands parallel or perpendicular to the main diagonal [4]. There are different ways to define the contact between two amino acids. In [5], the notion is based on spacial proximity, so that the contact map can consider non-covalent interactions. Gaci et al. in [5] says that two amino acids are in contact iff the distance between them is below a given threshold. A commonly used threshold is 7 Å. Consider a contact map graph with N vertices (each vertex corresponds to an amino acid) and the contact map matrix as incidence matrix. It is called contact map graph. The contact map graph is an abstract description of the protein structure taking into account only the interactions between the amino acids. Now let us consider the sub-graph induced by the set of amino acids participating in SSE, where SSE is secondary structure element of protein like alpha helix, beta sheet etc. We call this graph SSE interaction network (SSE - IN). The reason of ignoring the amino acids not participating in SSE is simple. Evolution tends to preserve the structural core of proteins composed from SSE. In the other hand, the loops (regions between SSE) are not so important to the structure and hence, are subject to more mutations. That is why homologous proteins tend to have relatively preserved structural cores and variable loop regions. Thus, the structure determining interactions are those between amino acids belonging to the same SSE on local level and between different SSEs on global level. In [6] and [7] the authors rely on similar models of amino acid interaction networks to study some of their properties, in particular concerning the role played by certain nodes or comparing the graph to general interaction networks models. Thanks to this point of view the protein folding problem can be tackled by graph theory approaches.

Gaci et al. in [8], has described the topological properties of a network and compared them with some *All alpha* and *beta* to prove that a protein can be treat as a network of amino acids. According to the diameter value, average mean degree and clustering coefficient shown in the experiment in [8], we can say a protein is a network of amino acids.

3. PREDICT AMINO ACID INTERACTION NETWORK

We can define the problem as prediction of a graph G consist of N vertices V and E edges. If two amino acids interact with each other in protein we mention it as an edge $(u, v) \in E, u \in V, v \in V$ of the graph. A SSE-IN is a highly dense sub-graph G_{SSE-IN} with edge set E_{SSE-IN} . Probability of the edge $(u, v) \in E_{SSE-INA}, u \in V_{SSE-INA}, v \in V_{SSE-INA}$ is very high and probability of the edge $(u, v) \notin E_{SSE-INA}, u \in V_{SSE-INA}, v \in V_{SSE-INB}$ is very low where $V_{SSE-INA}$ and $V_{SSE-INB}$ are respectively the vertex set of SSE-IN A and SSE-IN B. SCOP and CATH are the two databases generally accepted as the two main authorities in the world of fold classification. According to SCOP there are 1393 different folds. To predict the network we have to solve three problems, as i) find a associate SCOP protein family from the given protein sequence ii) predict a network of amino acid secondary structure element (SSE) from the known SCOP protein family and iii) Predict interactions between amino acids in the network, including internal edges of SSE-IN and external edges.

We are going to avoid the description the first problem, because it can be solve using a good sequence alignment algorithm like BLAST as discussed in [8]. In this paper we are going to solve the second and third problem individually with multi-objective optimization using genetic algorithm and ant colony optimization respectively.

Gaci et al. in [9], described a solution to the prediction of amino acid interaction network. He used a genetic algorithm with single objective as the distance between to amino acid in protein atom. But it is very difficult to define real world problems like amino acid interaction problem in terms of a single objective. A multi-objective optimization problem deals with more than one objective functions that are to be minimized or maximized. These objectives can be conflicting, subject to certain constraints and often lead to choosing the best trade-off among them. As we have described before, the interaction between amino acids in protein depends not only distance between two amino acids but also the torsion angles and hydrophobic property of the amino acid. So to get more accurate interaction network of amino acid we have to consider it is as a multi-objective problem rather than single objective.

4. ALGORITHM

As we have mentioned before, we can solve the amino acid interaction network prediction problem as well as the protein folding problem using two new and emerging algorithms. The multi-objective optimization algorithm will predict structural motifs of a protein and will give a network or graph of secondary structural element (SSE) of the protein. On the other hand, the ant colony optimization (ACO) algorithm will find the interactions between amino acids including the intra-SSE-IN and inter-SSE-IN interactions. In our algorithm we have considered a folded protein in the PDB as an unknown sequence if it has no SCOP v1.73 family classification. According to [8], we can associate the most compatible and best fit structural family based on topological criteria like average diameter, average mean distance etc.

4.1. Prediction of SSE interaction network using Multi-objective Optimization

There are several ways to solve multi-objective optimization problem. In this research we have decided to use Genetic Algorithm (GA) as multi-objective optimization. The GA has to predict the adjacency matrix of unknown sequence when it is represented by chromosome.

In this paper we proposed a evolutionary clustering algorithm to predict the SSE-IN, which is a modified algorithm of the second version of strength pareto evolutionary algorithm (SPEA2) in [10]. SPEA2 preserve better solutions than NSGA-II [11] and its diversity mechanism is better than the others, this is the reason to choose SPEA2 to implement the evolutionary clustering algorithm.

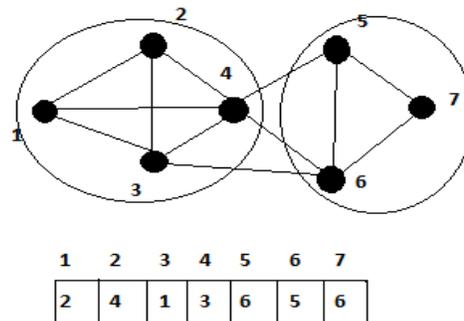


Figure 2: Network of 7 nodes clustered into 1,2,3,4 and 5,6,7 and their genetic representation

As proposed in [12], we are using a local-based adjacency representation. In this representation an individual of the population consist of N genes g_1, \dots, g_N , where N is the number of nodes. Each gene can hold allele value in the range $1, \dots, N$. Genes and alleles represents nodes in the graph $G =$

(V, E) modelling a network N . A value j assigned in i -th gene interpreted as a link between node i and j and in clustering node i and j will be in the same cluster as in Figure 2. In decoding step all the components are identified and nodes participating in the same component are assigned to the same cluster.

Algorithm 1 Multiobjective genetic algorithm to predict SSE interaction algorithm

- 1: **Input:** A protein sequence , T = total time steps, N_E = Archive size, N_p = Population size
- 2: **Output :** A predicted incident matrix M and clustering for each network N^i of N
- 3: Use BLAST to find a associate protein family of the given sequence from PDB
- 4: Generate initial cluster $\mathcal{C}_1 = \{C_1^1, \dots, C_k^1\}$ of the network N^i with number of vertex equal to number of SSE of the associate protein family
- 5: **for** $t = 2$ to T **do**
- 6: Create initial population of random individual P_0 and set $E_0 = \emptyset, i = 0$
- 7: **Loop**
- 8: Decode each individual of $P_i \cup E_i$
- 9: Evaluate each individual of $P_i \cup E_i$ to find rank and density value using equation 1 and 2
- 10: Assign fitness value to each individual, as the sum of rank and inverse of density value
- 11: Copy all no dominating solution to E_{i+1}
- 12: **if** $|E_{i+1}| > N_E$ **then**
- 13: truncate $|E_{i+1}| - N_E$ solutions according to topological property
- 14: **Else**
- 15: copy best $N_E - |E_{i+1}|$ dominated solution according to their fitness value and topological property
- 16: **end if**
- 17: **if** stopping criteria does not satisfies **then**
- 18: **return** non-dominated solution in $|E_{i+1}|$
- 19: **else**
- 20: Select some individual form— E_{i+1} for mating pool as parents using binary tournament with replacement
- 21: Apply crossover and mutation operators to the mating pool to the mating pool to create N_p offspring solution and copy to P_{i+1}
- 22: $i := i+1$
- 23: **end if**
- 24: **end loop**

- 25: From the returned solution in E take the best cluster according to the highest modularity value
- 26: **end for**

It takes a dynamic network $N = N_1, N_2, \dots, N_T$, the sequence of graphs $G = G_1, G_2, \dots, G_T$ and the number of timestamps T as input and gives a clustering of each network N_i of N as output.

In the amino acid interaction network, total number of gene is the number SSE in the associate protein family found from the first step and each SSE represents one gene or allele notably considering its size that is the number of amino acids which compose it, of the population. We represent a protein as an array of alleles. The position of an allele corresponds to the SSE position it represents in the sequence. At the same time, an incident matrix is associates for each genome.

For the first time-stamp of first input network there is no temporal relation with the previous network. The only objective function is snapshot quality or snapshot score. Thus we can apply any static clustering algorithm or trivial genetic algorithm to find the initial cluster. In this algorithm we used genetic algorithm to find the best cluster by maximizing the only objective function. As it is single objective algorithm we can find the single best cluster from this step.

As a first step in each time-stamp from 2nd time-stamp to T , it creates a population of random individuals. Each individual is a vector of length equal to number of nodes in the graph G^t . Genetic variant operators will be applied on this population for a fixed number of pass.

Each individual of the population and archive is decoded into component graph. As each individual gene is working as an adjacency list, if a node in x of graph is reachable from y by maintaining the edges in the individual, then x and y is in same cluster of component.

Give each individual chromosome of the population and chromosome in archive a rank value. Smaller the rank value better it is as fitness value. Each non-dominated individual gets the rank 0. After removing the 0 ranked individuals, give the rank 1 to the next non-dominated individuals and so on. After giving each individuals a rank value, sort the individuals according to the ascending rank.

$$r(x) = \sum_{x < y} s(y) \quad (1)$$

There could be many individuals of in same area of solution space or objective space. If we take all these solution into account, we could loss diversity in the population. To remain the population diverse, we are using distance of k -th nearest neighbour. The fitness value of each individual is the sum of its non-dominated rank and the inverse of the distance of k -th nearest neighbours distance. More the distance between solutions, better the fitness functions value.

$$m(x) = (\sigma_x^k + 1)^{-1} \quad (2)$$

where σ_x^k is the distance between individual x and its k -th nearest neighbour. To calculate the distance between chromosome, we have to take account the three objectives, atomic distance of amino acids, torsion angles and hydrophobicity.

After evaluating fitness values of each of the population and archive, the best individuals are selected as a new population. From the total individuals of population and archive population size individuals are selected as new population. From the rank 0 to the highest rank, all the individuals

are added if number of population of this rank is not exceeding the current population size. If it is exceeding, then some individuals are truncated according to the value of each individuals.

Table 1: Example of uniform crossover

Parent 1	4	3	2	2	6	5	6
Parent 2	3	3	1	5	4	7	6
Mask	0	1	1	0	0	1	1
Offspring	4	3	1	2	6	7	6

After selecting the new population, a mating pool is created of pool size from the new population to apply the genetic variation operators. To choose the mating pool, binary tournament with replacement has been used in this algorithm. According to binary tournament, two individuals are randomly selected from the new population and the better fitness valued individual is chosen for the mating pool.

4.1.1. Genetic Variation Operators

Genetic operators are used to create offspring from parent or mating pool. As other genetic algorithms, in this algorithm two widely used genetic variation operators have been used. These are crossover and mutation.

Crossover is the operator which is used to create offspring from two parents. The offspring bear the genes of each parent. As a genetic variation operator there is very high probability to crossover occurs other than mutation. In this algorithm we are using uniform crossover. A random bit vector of length of number of the node in the current graph is created. If *i-th* bit is 0 then the value of the *i-th* gene comes from the first parent otherwise it comes from the *i-th* gene of second parent. As each of the parents holding true adjacency information, the offspring will also hold it.

One of the most widely used variation operator in genetic algorithm, which perform the operation in a single individual is mutation. Though the probability of mutation is normally very low, but it is the best way to make small variation in the individual. To mutate and create a offspring, some position of the of the individuals are chosen randomly and changed to other values. But the value should be one of its neighbours in the current graph.

A topological operator is used to exclude incompatible population generated by the algorithm. We compute the diameter, the characteristic path length and the mean degree to evaluate the average topological properties of the family for the particular SSE number.

4.1. Ant Colony Optimization (ACO) to Predict Interactions

After predicting the SSE-IN network we have to identify the interactions involve between the amino acids in the folded protein. We have used an ant colony optimization (ACO) approach to select and predict the edges which link different SSE's, considering about the correction of the matrix of motifs previously predicted.

We have built a two steps algorithm as the hierarchical structure of the SSE-IN.

- In interaction, consider each pair of SSE's separately. This is the local step. We use an ant colony algorithm to identify the suitable interactions between amino acids belonging to these SSE's.
- A global ant colony algorithm is run to predict the interaction between amino acids from different SSE-IN.

4.2.1. Parameters for Interaction Network Prediction

To predict the interactions, firstly we have to know how many edges to be add in the network and which nodes we should consider in interactions. To find and evaluate these parameters, we incorporated the template proteins from the associate family.

We select some template proteins from the associate family whose SSE number is same as the sequence to predict the edge rate of the sequence and represent them as chromosome or array of alleles as in the multi-objective genetic algorithm. Thus, we build a comparative model to compute the edge ratio, which is used to fold the sequence SSE-IN.

We calculate the average chromosome from all the template proteins in associate protein family. Here we used the distance between two chromosomes as discussed in the previous section to compare the sequence with the average chromosome. We add up the distance allele by allele to obtain a distance between the sequence and the average family chromosome. After that, we calculate the cumulated size by adding up the chromosome cell values. If the distance is less than 20% of the sequence cumulated size and the average family chromosome then the sequence is closer to the template protein. Then we compute the average edge rate in the closer protein to add the initial edges in the disconnected network of the sequence. If we can't find a sequence closer to the template one, we add the sequence with the average family chromosome and start again the same procedure.

We do the same procedure to find the designation of the vertices, which vertices should interact with each other as they also use comparative model.

To define, which edges link two SSE's, we consider the following problem.

Let $X = x_1, x_2, \dots, x_n$ and $Y = y_1, y_2, \dots, y_m$ be two SSEs in interaction. We want to add e edges among the $n \times m$ possible combinations. For $i \in [1, n]$ and $j \in [1, m]$ the probability to interact the amino acid x_i with y_j , is correlated with the occurrence matrix of the predicted edges ratios, represented by $Q(x_i, y_j)$ and we can assume $s_{ij} \sim Q(x_i, y_j)$. To add approximately e edges, we need

$$\sum_{i=1}^n \sum_{j=1}^m S_{ij} = e \quad (3)$$

and

$$S_{ij} = \frac{eQ(x_i, y_j)}{\sum_{p=1}^n \sum_{q=1}^m Q(x_p, y_q)} \quad (4)$$

4.2.2. Ant Colony Algorithm

The prediction of interaction network consists of two approaches, local and global algorithm.

4.2.2.1. Local Algorithm

The local algorithm is used to predict the suitable shortcut edges between pair of SSEs in the network. Thus, we differentiate each pair of SSEs which have connection and build a graph where each vertex of the first SSE is connected to each vertex of the other SSE. The connection or the edges are weighted (S_{ij}). Then we used an ant colony approach consists of an ant number equals to the number of vertices in two SSE. The ant system has to reinforce the suitable edges between the SSEs. We use these edges in the global algorithm which is described in the next section.

The local ant colony algorithm first creates n ants which is total number of vertices in the two SSEs related in the search. For an ant to be positioned we choose a random vertex of to SSE involved and place it. All the n ants are positioned this way and two ants can share same vertex. An ant in vertex i will choose the vertex j with probability p_{ij} , defined as follows:

$$p_{ij} = \frac{\tau_{ij}^{\alpha} \cdot s_{ij}^{\beta}}{\sum_{k \in V(i)} \tau_{ik}^{\alpha} \cdot s_{ik}^{\beta}} \quad (5)$$

The weight s_{ij} also called heuristic vector, calculated before. If the vertices i and j are in the same SSE, then the edge between these two vertices has weight equal to the average weight of the shortcut edges:

$$\bar{s} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m s_{ij} \quad (6)$$

After each move of an ant we update the pheromone value on the inter-SSE edges using the formula,

$$\tau_{ij} = (1 - \rho)\tau_{ij} + n_{ij} \Delta \tau \quad (7)$$

where s_{ij} is the number of ants that moved on the edge (i, j) and $\Delta \tau$ is the quantity of pheromone dropped by each ant. As far as the edges belonging to the same SSE are concerned, we keep the pheromone rate on them equals to the average pheromone rate on the inter-SSE edges

$$\bar{\tau} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \tau_{ij} \quad (8)$$

Ants are move inside an SSE randomly, described above, on the other hand if they decide to change the SSE they are guided by the edge weight and the weight is guided by the pheromone value. The algorithm stops after a predefined number of iteration or the maximum pheromone rate is e time bigger than the average pheromone rate on the edge. After the execution of the algorithm we keep the edges whose pheromone quantity exceeds a threshold λ_{min} .

Algorithm 2 Local algorithm to find Inter-SSE edges

```

1:   Input: The predicted network from the multiobjective genetic algorithm
2:   Output: Predicted inter-SSE edges
3:   Create n ants, where n is the total number of nodes in process
4:   while stopping criteria does not meet do
5:     for all ant  $\alpha$  do
6:       moveAnt( $\alpha$ )
7:     end for
8:     updatePheromone()
9:   end while
10:  selectEdges( $\lambda_{min}$ )

```

Algorithm 3 Global algorithm to predict edges into SSEs

```

1:   Input: The network with predicted edges  $E_s$  from local algorithm and  $E_p$ , number edges to predict
2:   Output: The network with total  $E_p$  edges
3:   buildSSEIN( $E_s$ )
4:   create n ants
5:   while stopping criteria does not meet do
6:     for all ant  $a$  do
7:       moveAnt( $\alpha$ )
8:     end for
9:     updatePheromone()
10:  end while
11:  selectEdges( $E_p$ )

```

4.2.2.2. Global Algorithm

After the local algorithm execution, we get the SSE-IN composed of these specific inter-SSE edges. The global algorithm will keep the number of edges exactly E_p , which was predicted before. As the local one, the global algorithm uses the ant colony approach with the number of vertices equal to the SSE-IN vertex number. The ants movements contribute to emerge the specific shortcut who's only a number E_p is kept. We rank the shortcut edges as a function of the

pheromone quantity to extract the E_p final shortcuts. Finally, we measure the resulting SSE-INs by topological metrics to accept it or not.

We compute the diameter, the characteristic path length and the mean degree to evaluate the average topological properties of the family for a particular SSE number. Then, after we have built the sequence SSE-IN, we compare its topological properties with the template ones. We allow an error up to 20% to accept the built sequence SSE-IN. If the built SSE-IN is not compatible, it is rejected. We compare the predicted value, denoted E_p , with the real value, denoted E_R

$$AC = 1 - \frac{|E_R - E_p|}{E_p} \quad (9)$$

where AC is the accuracy of the prediction.

5. PERFORMANCE ANALYSIS

In this paper we have discussed two algorithms to predict the interaction network of amino acid. We are going to analysis each algorithm independently.

5.1. Analysis of Genetic Algorithm as Multi-objective Optimization

In order to test the performance of proposed multi-objective genetic algorithm, we randomly pick three chromosomes from the final population and we compare their associated matrices to the sequence SSE-IN adjacency matrix. To evaluate the difference between two matrices, we use an error rate defined as the number of wrong elements divided by the size of the matrix. The dataset we use is composed of 698 proteins belonging to the *All alpha* class and 413 proteins belonging to the *All beta* class. A structural family has been associated to this dataset as in [8].

All alpha class has an average error rate of 14.6% and for the *All beta* class it is 13.1% and the maximum error rate shown in the experiment is 22.9%. Though, the error rate depends on other criteria like the three objectives described before but according to the result we can firmly assert that the error rate is depends on the number of initial population, more the number of initial population less the error rate. With sufficient number of individuals in the initial population we can ensure the genetic diversity as well as the improved SSE-IN prediction. When the number of initial population is at least 15, the error rate is always less than 10%.

As compared to the work in [8] we can claim better and improved error rate in this part of SSE-IN prediction algorithm.

5.2. Analysis of Ant Colony Optimization

We have experimented and tested this part of our proposed method according to the associated family protein because the probability of adding edge is determined by the family occurrence matrix. We have used the same dataset of sequences whose family has been deduced.

For each protein, we have done 150 simulations and when the topological properties are become compatible to the template properties of the protein we accepted the built SSE-IN. The results are shown in Table 2. The score is the percentage of correctly predicted shortcut edges between the sequence SSE-IN and the SSE-IN we have reconstructed [8]. In most cases, the number of edges to add were accurate according to the Figure 3. From this we can percept that, global interaction scores depends on the local algorithm lead for each pair of SSEs in contact. The plot, in Figure 3

confirms this tendency, if the local algorithm select at least 80% of the correct shortcut edges, the global intersection score stays better than the 80% and evolve around 85% for the All alpha class and 73% for the All beta class.

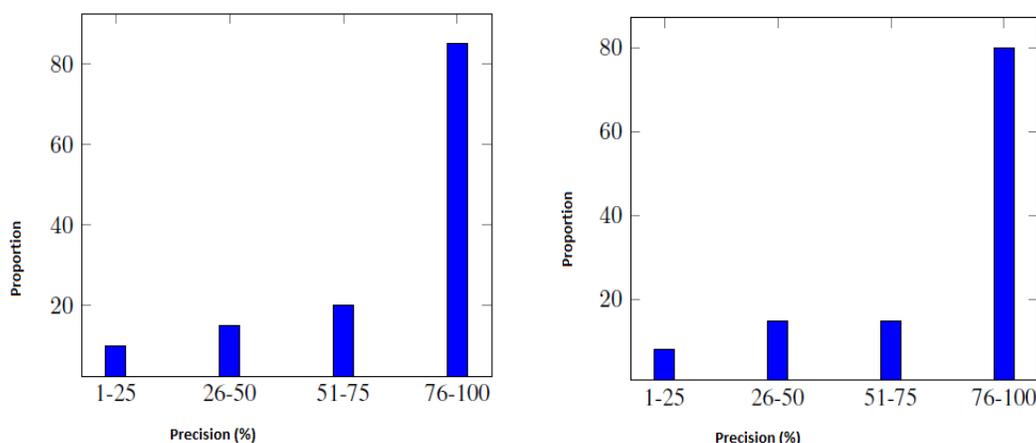


Figure 3: Precision in number of edges to be added in All Alpha (left) and All Beta (right).

After the discussion we can say that, though for the big protein of size more than 250 amino acids the average score decreases, but in an average the score remains for the global algorithm around 80%.

5.3. Algorithm Complexity

Our proposed algorithm is independent of specific time bound. Both the optimization algorithm used as multi-objective genetic algorithm and ant colony algorithm, is iteration based. We can stop the algorithm at any time. Though the result of the algorithm depends on the number of iteration but if we give sufficient amount of iteration it provides good result. In compare to other state of art algorithms, those uses exponential complexity algorithm, our is linear in terms of time and memory.

6. CONCLUSIONS

We have proposed an computational solution to an biological problem. We have described how we can formulate a biological problem like folding protein into optimization and graph theory problem. The formulation consists of finding the interactions between secondary structure element (SSE) network and interaction between amino acids of the protein. The first problem was solving by an multi-objective genetic algorithm and the second one solve by ant colony optimization approach.

As discussed before, we have given theoretical and statically proof that our proposed algorithm gives more accurate result in terms of accuracy and score to predict the amino acid interaction network. Though it can be furnished further with improved data structure and parallel algorithms.

Table 2: Folding a SSE-In by ant colony approach. The algorithm parameter values are : $\alpha = 25$, $\beta = 12$, $\rho = 0.7$, $\Delta\tau = 4000$, $e = 2$, $\lambda_{\min} = 0.8$.

Class	SCOP Family	Number of Proteins	Protein Size	Score	Average Deviation
All Alpha	46688	17	27-46	83.973	3.277
	47472	10	98-125	73.973	12.635
	46457	25	129-135	76.125	7.849
	48112	11	194-200	69.234	14.008
	48507	18	203-214	66.826	5.504
	46457	16	241-281	63.281	17.025
	48507	20	387-422	62.072	9.304
All Beta	50629	6	54-66	79.635	2.892
	50813	11	90-111	74.006	4.428
	48725	24	120-124	80.881	7.775
	50629	13	124-128	76.379	9.361
	50875	14	133-224	77.959	10.67

REFERENCES

- [1] E. Pennisi. A low number wins the genesweep pool. In *Science*, volume 300, page 1484, 2003.
- [2] K Madhusudan Reddy, Sunkara V Manorama, and A Ramachandra Reddy. Bandgap studies on anatase titanium dioxide nanoparticles. *Materials Chemistry and Physics*, 78(1):239–245, 2003.
- [3] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [4] Amit Ghosh, KV Brinda, and Saraswathi Vishveshwara. Dynamics of lysozyme structure network: probing the process of unfolding. *Biophysical journal*, 92(7):2523–2535, 2007.
- [5] Omar Gaci and Stefan Balev. The small-world model for amino acid interaction networks. In *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*, pages 902–907. IEEE, 2009.
- [6] Usha K Muppurala and Zhijun Li. A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. *Protein Engineering Design and Selection*, 19(6):265–275, 2006.
- [7] KV Brinda and Saraswathi Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophysical journal*, 89(6):4159–4170, 2005.
- [8] Gaci, Omar. Building a topological inference exploiting qualitative criteria. *Evolutionary Bioinformatics*, 2010.
- [9] Gaci, Omar. How to fold amino acid interaction networks by computational intelligence methods. In *Bioinformatics and BioEngineering (BIBE), 2010 IEEE International Conference on*, pages 150 – 155, 2010.
- [10] Marco Laumanns Eckart Zitzler and Lothar Thiele. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. Zurich, Switzerland, 2001.

- [11] Deb K, Agrawal S, Pratap A, and Meyarivan T. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In Proc. of 6th international conference on parallel problem solving on nature, Paris, France, September 2000. Springer.
- [12] Y.J. Park and M.S. Song. A genetic algorithm for clustering problems. In Proc. of 3rd Annual Conference on Genetic Algorithm, pages 2–9, 1989.

Authors

Md. Shiplu Hawlader, born 1987 in Dhaka, Bangladesh, graduated from the Faculty of Engineering of University of Dhaka in 2011 at the Computer Science and Engineering department. After the graduation he completed his M.Sc. degree at the same department. He is a lecturer of Department of Computer Science and Engineering, University of Asia Pacific.



Saifuddin Md. Tareeq, Dr., graduated from Department of Computer Science and Engineering, University of Dhaka and working as Associate Professor at the same Department.



FLIGHT TRAJECTORY RECREATION AND PLAYBACK SYSTEM OF AERIAL MISSION BASED ON OSSIMPLANET

Wu Wu¹, Jiulin Hu², Xiaofang Huang³, Huijie Chen⁴, Bo Sun⁵

College of Information Science and Technology,
Beijing Normal University, Beijing, China
tosunbo@bnu.edu.cn

ABSTRACT

Recreation of flight trajectory is important among research areas. The design of a flight trajectory recreation and playback system is presented in this paper. Rather than transferring the flight data to diagram, graph and table, flight data is visualized on the 3D global of ossimPlanet. ossimPlanet is an open-source 3D global geo-spatial viewer and the system realization is based on analysis it. Users are allowed to choose their interested flight of aerial mission. The aerial photographs and corresponding configuration files in which flight data is included would be read in. And the flight statuses would be stored. The flight trajectory is then recreated. Users can view the photographs and flight trajectory marks on the correct positions of 3D global. The scene along flight trajectory is also simulated at the plane's eye point. This paper provides a more intuitive way for recreation of flight trajectory. The cost is decreased remarkably and security is ensured by secondary development on open-source platform.

KEYWORDS

flight trajectory, open-source platform, 3D global, ossimPlanet, KML

1. INTRODUCTION

Flight trajectory is important in the flight collision, flight planning, flight accidents investigation and flight simulation areas. Generally, the recreation of flight trajectory [1-2] is to transfer the flight data to the diagram, graph, table, etc. It is important in aircraft safety assessments, aircraft maintenance, accident investigation and event analysis. However, the results by these common solutions are not intuitive for the users [3]. The flight data, e.g., the flight angles, positions and m Accordingly, it is a better way to recreate the flight trajectory by visualizing the flight data.

The visual simulation technology makes it possible. Flight Viz [4] produced by SimAuthor Company could playback the flight trajectory in 3D visual effects, but its cost is very high. The 3D Flight Simulation System EasyFlight developed by China Academy of Civil Aviation Science and Technology is a professional aeronautic platform [5]. It's mainly used in flight simulation of major accidents, but rarely in common flight trajectory recreation. Yong Tang [6] presented a solution for 3D flight trajectory and 6-DOF flight simulation based on Google Earth [7]. But the research result relies much on the server of Google Earth, thus the users may concern about the security and the cost.

From the point of view of the cost and security, secondary development on open-source platform is a better choice. In this paper, ossimPlanet [8] is taken as the development platform. It is an accurate 3D global geo-spatial viewer that is built on the OSSIM [9], OpenSceneGraph [10], and Trolltech QT[11] open source software libraries [12]. It could provide accurate 3D global visualization and collaboration [13], and has the following three advantages: (1) It's open-source. It costs less than platform, e.g., Google Earth. Especially, we can realize more customized functions and ensure its security. (2) It's built on OSSIM, which has a powerful suite of geospatial libraries and applications to process imagery, maps, terrain, and vector data [9]. This paper focuses on the flight trajectory of aerial mission, which includes a lot of aerial photographs. Thus, OSSIM can provide strong support on image processing. (3) any other flight parameters could describe the spatial status of flight.

It's written in C++ and thus has higher performance than the platforms written in other languages, such as World Wind written in C#.

In this paper, a flight trajectory recreation and playback system of aerial mission will be implemented based on ossimPlanet. The system would recreate and playback the trajectory on 3D global, thus it will be more intuitive. The development on ossimPlanet ensures the security in a low cost and high performance. This paper is organized as follows. In Section II, the requirement analysis of the whole system is presented. The key problems and the corresponding solutions are given in Section III. The system realization is introduced in Section IV. The simulation results are shown in Section V. The conclusions are summarized in Section VI.

2. REQUIREMENT ANALYSIS

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

The system functions are shown in Fig.1, and their detailed descriptions are given below.

- Choose flight. Users are allowed to choose their interested flight, that is, to fix the local path where the aerial photographs and corresponding configuration files are located. Then these files would be read in, and the statuses of the plane are stored after the necessary processing on these input files.
- Observe photographs. Users are allowed to observe the input photographs. These photographs would be pasted on their correct positions which are set in the configuration files.
- Observe trajectory. Users are allowed to observe the flight trajectory on the 3D global of ossimPlanet. Both the input flight trajectory points and the interpolated trajectory points are marked on the 3D global.
- Observe simulation. Users could follow the plane's eye point to view the flight trajectory dynamically.

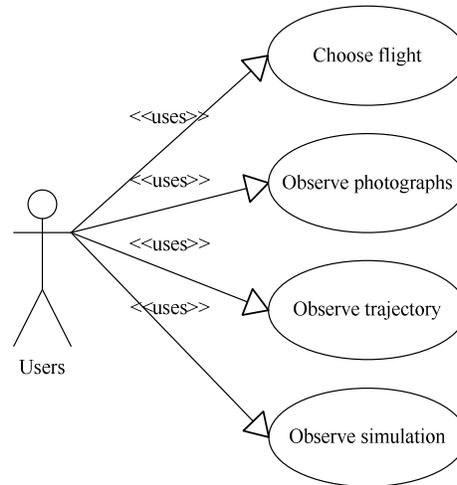


Fig. 1. Use case diagram.

3. KEY PROBLEMS

There are three key problems for the system realization. (1) Data processing. The input data includes the aerial photographs and configuration files. Every aerial photograph has a corresponding configuration file in which the flight data is included, e.g., the flight statuses, flight positions, pilot's operations, etc. The required flight data will be taken for interpolation. (2) Data display. It is to display the aerial photographs and mark flight trajectory points on the 3D global of ossimPlanet. (3) Flight trajectory playback. It is to playback the flight trajectory on the 3D global of ossimPlanet.

3.1. Data Processing

Without loss of generality, we make following assumptions on the motion of plane: (1) It's rigid body motion. (2) The translation is with the centroid and the rotation is around the centroid.

To describe the motion clearly, we should take proper flight data from the configuration files. Generally, 6 degree-of-freedom (DOF), i.e., 3 position coordinates (longitude, latitude and height) and 3 posture angles (heading angle, pitch angle and roll angle), is always used [6]. The posture angles are shown in Fig.2.

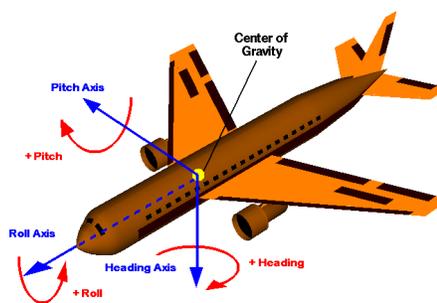


Fig. 2. The heading, roll and pitch angle of plane [14].

After we get the flight trajectory points contained the 6 DOF parameters, it's necessary to smooth the flight trajectory by interpolation. The interpolation on the position coordinates and posture angles will be done respectively. Since it's shown that the unstable heights may result in flight collision [15], the height is assumed invariable in position coordinates and not interpolated. De Boor's algorithm [16] will be used to interpolate the longitude and latitude. In the posture angles, the heading angle is changed with the stress of plane [17] and always very small. The changes of pitch angle in real aerial mission are generally less than 5 degrees and roll angle is always 0 degree [18]. In this paper, we use linear interpolation for the posture angles smoothing.

3.2. Data Display

3.2.1. Photograph Display

To display the photographs on the 3D global of ossimPlanet, the corresponding geometry files for photographs are required. A geometry file instance is given in Fig. 3. In which the projection type, datum, longitude, latitude and some other geographic parameters are set.

```

type: ossimEquDistCylProjection
origin_latitude:0.0
central_meridian: 0.0
pixel_scale_units: degrees
pixel_scale_xy: (.133, .133 )
datum: WGE
tie_point_units: degrees
tie_point_xy: (-180.0, 90.0)
pixel_type: area

```

Fig. 3. Geometry file instance.

In Fig. 3, the type defines the projection of the photograph, and the default projection of ossimPlanet is cylindrical equidistant projection [19]. The origin_latitude and the central_meridian are always 0 degree. The pixel_scale_xy is the actual scale of every pixel of the photograph and its unit is defined in the pixel_scale_units. The tie_point_xy is the coordinate of the photograph as (longitude, latitude) and its unit is defined in tie_point_units.

After the photographs and configuration files are read in, the corresponding geometry files are created according to the configuration files. Then by using ossimPlanet's API, the photographs could be displayed on the 3D global.

3.2.2. Flight Trajectory Display

Keyhole Markup Language (KML) [20] is a Markup Language to describe and store geographical information, such as point, line, surface, three-dimensional models, etc. A KML file instance is given in Fig.4.

Generally, a KML file includes 3 parts: (1) XML Header; (2) The definition of KML namespace; (3) The object of geographical indication [21]. In Fig.4, <Stlye> indicates a style may be used for objects and <Placemark> indicates a place mark. The KML file in Fig.4 indicates a point at (121.48844, 53.332649, 0), and it will be shown as an icon whose hyperlink is given in the referenced link.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2"
xmlns:gx="http://www.google.com/kml/ext/2.2">
<Document>
<Style id="style60" >
<IconStyle>
<Icon><href>reference link</href></Icon>
</IconStyle>
</Style>
<Placemark>
<styleUrl>#style60</styleUrl>
<Point>
<coordinates>121.48844,53.332649,0</coordinates>
</Point>
</Placemark>
</Document>
</kml>
```

Fig. 4. KML file example.

A KML file is created for the input trajectory points and interpolated trajectory points. Then by using ossimPlanet's API, the KML file could be loaded and thus the trajectory points could be marked on the 3D global.

3.3. Flight Trajectory Playback

To playback the flight trajectory, it's necessary to have the knowledge of the 3D world of ossimPlanet. (1) Coordinate Systems and Transformations. In the 3D world, the basic work is to confirm the coordinate systems and find out the coordinate transformations. (2) View Transformation. To playback flight trajectory is to change the eye point with the flight status. Thus the view transformation is important. (3) Rendering theory of ossimPlanet. The actual development work should be based on the rendering theory of ossimPlanet.

3.3.1. Coordinate Systems and Transformations

The Geographical Coordinate System, World Coordinate System and Local Coordinate System are briefly introduced as follows:

- a) Geographical Coordinate System. In this coordinate system, each point is determined by its longitude, latitude and the height above a WGS-84 reference ellipsoid [22].
- b) World Coordinate System. The world coordinate of ossimPlanet is Earth Centered Earth Fixed (ECEF) [23], as the XYZ coordinate system shown in Fig.5.
- c) Local Coordinate System. The local space reference (LSR) system of ossimPlanet is as UVW coordinate system shown in Fig.5.

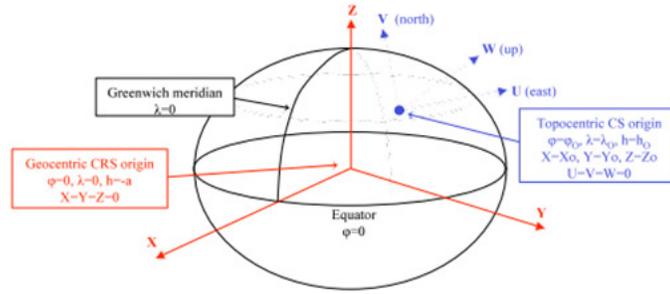


Fig. 5. Coordinate systems of ossimPlanet[24]. (λ is longitude, ϕ is latitude and h is the height)

Both the Geographical Coordinate and the Local Coordinate could be converted into the World Coordinate as follows [24].

a) From Geographical Coordinate to ECEF

$$\begin{aligned} X &= (v + h) \cos \phi \cos \lambda \\ Y &= (v + h) \cos \phi \sin \lambda \\ Z &= [(1 - e^2)v + h] \sin \phi \end{aligned} \quad (1)$$

b) From LSR to ECEF

$$\begin{aligned} X &= X_0 - U \sin \lambda_0 - V \sin \phi_0 \cos \lambda_0 + W \cos \phi_0 \cos \lambda_0 \\ Y &= Y_0 + U \cos \lambda_0 - V \sin \phi_0 \sin \lambda_0 + W \cos \phi_0 \sin \lambda_0 \\ Z &= Z_0 + V \cos \phi_0 + W \sin \phi_0, \end{aligned} \quad (2)$$

where v is normal vector of latitude ϕ and its value is $v = a / (1 - e^2 \sin^2 \phi)^{0.5}$, h is the height above the surface of ellipsoid, e is eccentricity and $e^2 = (a^2 - b^2) / a^2 = 2f - f^2$, ϕ is latitude and λ is longitude, a is semi-major axis, b is semi-short axis and f is flattening.

3.3.2. View transformation

From the general process of 3D graphics display [25], we can easily convert the World Coordinate to the View Coordinate as follows,

$$ViewCoord = WorldCoord * VM * PM * WM, \quad (3)$$

where VM is the view matrix, PM is the projection matrix and WM is the window matrix.

In ossimPlanet, PM and WM in (3) are fixed. Therefore, VM should be calculated for the view transformation. That is to place the eye point of ossimPlanet on the proper position in proper posture. The position of eye point is determined by the position of the plane given in the configuration files. Then we can convert the position coordinate in the Geographic Coordinate System $GeoEye(l, l, h)$ to that in the World Coordinate System $WorldEye(x_0, y_0, z_0)$ as in (1).

From the posture angle, we can get the rotation matrix of the eye point:

$$\text{RotateMatrix} = R_z(h) * R_y(p) * R_x(r), \quad (4)$$

where h is heading angle, p is pitch angle and r is roll angle. $R_z(h)$, $R_y(p)$ and $R_x(r)$ are the corresponding rotation matrixes [25].

Since RotateMatrix is in the Local Coordinate System, we should convert it to the World Coordinate System as in (2) and have the rotation matrix of the eye point in LSR,

$$\text{RotationLsrMatrix} = \text{RotateMatrix} * \text{LsrMatrix}, \quad (5)$$

then, VM is as follows:

$$\text{ViewMatrix} = \text{RotationLsrMatrix} * \text{WorldEye}, \quad (6)$$

3.3.3. Rendering Theory of ossimPlanet

The rendering circle [26] of ossimPlanet is to loop the frame() before the scene is finished. Every frame has the following three traversals:

- a) Event Traversal. This part is implemented in eventTraversal(), where the different kinds of events are handled. The events include the mouse events, keyboard events, windows, callbacks of cameras, etc.
- c) Updating Traversal. This part is implemented in updateTraversal(), where the updating callbacks are traversed and executed.
- d) Rendering Traversal. This part is implemented in renderingTraversal(), where the rendering work such as the Cull and the Draw are done.

The basic rendering progress in ossimPlanet is described below. (1) The events from GUI or the scene are caught and handled in Event Traversal. The corresponding scene parameters are also calculated. (2) The scene parameters calculated by the Event Traversal are updated in Updating Traversal. (3) The scene parameters updated in Updating Traversal would be shown by Rendering Traversal. Thus, the scene would change with the events through the cooperation of the 3 traversals.

In ossimPlanet, the rendering circle is finished in a component called Viewer. It's shown in Fig. 6 that the Viewer includes Manipulator, GUI Event Handler, Scene and Camera. The Manipulator is an instance for roaming in the scene of the Viewer. All of the events from the GUI or the scene will be collected in the Manipulator. And these event messages will be translated to and finally handled in Navigator. In Navigator, the calculations of scene parameters of Event Traversal are completed.

To implement playback of flight trajectory in ossimPlanet, we should add our own event handlers in Event Traversal. And the handlers would handle customized events and calculate the scene parameters as we design. The rest work could be then done by the other 2 traversals.

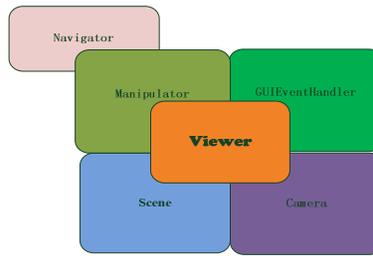


Fig. 6. Viewer of ossimPlanet.

4. SYSTEM REALIZATION

Based on the analysis in Section III, the system realization is mainly to overwrite the Manipulator and Navigator of ossimPlanet. The user interactions are through the GUI (ossimPlanet QtMainWindow) and the corresponding event handlers are added in Navigator following the theory of Event Traversal. To store the information of flight trajectory, data structures are designed.

4.1. Component Design

The component diagram of the whole system is shown in Fig.7 and the corresponding description is summarized below.

GUI (ossimPlanetQtMainWindow) provides 3 interfaces for user interactions. on_viewStartInputtingPath_triggered() is for inputting the interested flight of aerial mission, on_flieOpenKml_triggered() is for displaying the flight trajectory and on_viewShowThePath_triggered() is for simulating the flight trajectory playback. After users' operations, GUI will translate event messages to the Navigator which is bridged by the Manipulator. Then, Navigator will handle these events.

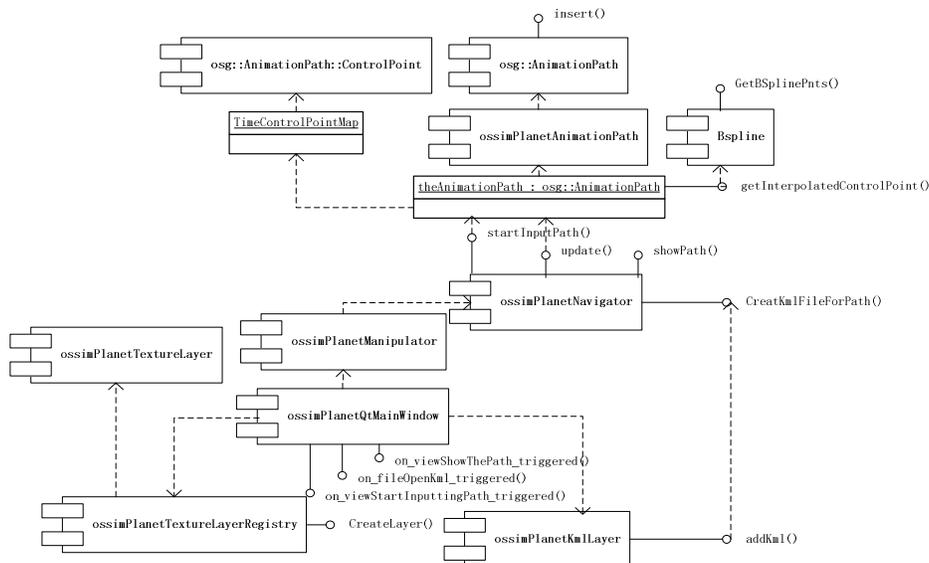


Fig. 7. Component diagram.

Interfaces (`startInputPath()`, `CreateKmlFileForPath()` and `showPath()`) are added to the Navigator. They are the responses to the GUI events. The `startInputPath()` is to receive the raw files (photographs and configuration files) and store them. The `CreateKmlFileForPath()` is to create a KML file and write the points of flight trajectory into it. The `showPath()` is to change the rendering mode of `ossimPlanet` to simulation mode. The calculation of scene parameters is done in the `update()`. The `ossimPlanetTextureLayer` is the layer of photographs and the `ossimPlanetKmlLayer` is the layer of the KML files. Every photograph or KML file shown on the 3D global of `ossimPlanet` corresponds to a layer.

4.2. Data Structures

The data structures used for storing the flight trajectory points are shown in Fig.8, and the corresponding description is given below. (1) Struct `PathPoint`. It includes 6 attributes corresponding to the 6 DOF parameters and denotes the input flight trajectory point. An array of the `PathPoint` denotes the flight trajectory points gotten from the configuration files. (2) Class `BSpline`. It is for the de Boor's interpolation algorithm and shown in Fig.8 (a). (3) Class `ControlPoint`. It denotes the flight trajectory point after interpolation. And it has 3 attributes, where the `_position` is the position of eye point in the world coordinate system, the `_rotation` is the rotation of the eye point and the `_scale` is scale factor. The view matrix could be set by them. It's shown in Fig.8 (b). (4) Class `ossimPlanetAnimationPath`. It is to store the flight trajectory points after interpolation and is shown in Fig.8 (c). In which `map<double,ControlPoint>` is to store the mapping relationship between the flight trajectory point and its relative time.

BSpline
#ShapePoints
#NodeVector
#MyControlPoints
#BSplinePoints
+BSpline()
+CalNodeVector()
+CalControlPnts()
+GetdeBoorValue()
+CalBSplinePnts()
+GetBSplinePnts()

Fig.8 (a). Class BSpline.

ControlPoint
#osg::Vec3d _position
-osg::Quat _rotation
-osg::Vec3d _scale
+ControlPoint()
+setPosition() : void
+getPosition()
+setRotation()
+getRotation()
+setScale()
+getScale()
+getMatrix()

Fig.8 (b).Class ControlPoint.

ossimPlanetAnimationPath
+ControlPoint
+map<double,ControlPoint> TimeControlPointMap
TimeControlPointMap _timeControlPointMap
+getMatrix() : bool
+getInterpolatedControlPoint() : bool
+insert() : void
+getFirstTime() : double
+getLastTime() : double
+getPeriod() : double
+getTimeControlPointMap()

Fig.8 (c). Class ossimPlanetAnimationPath.

Fig. 8.Main data structures.

4.3. Storage of Interested Flight

After users input the interested flight from GUI, the `on_viewStartInputtingPath_triggered()` is triggered. The cooperation diagram is shown in Fig.9.

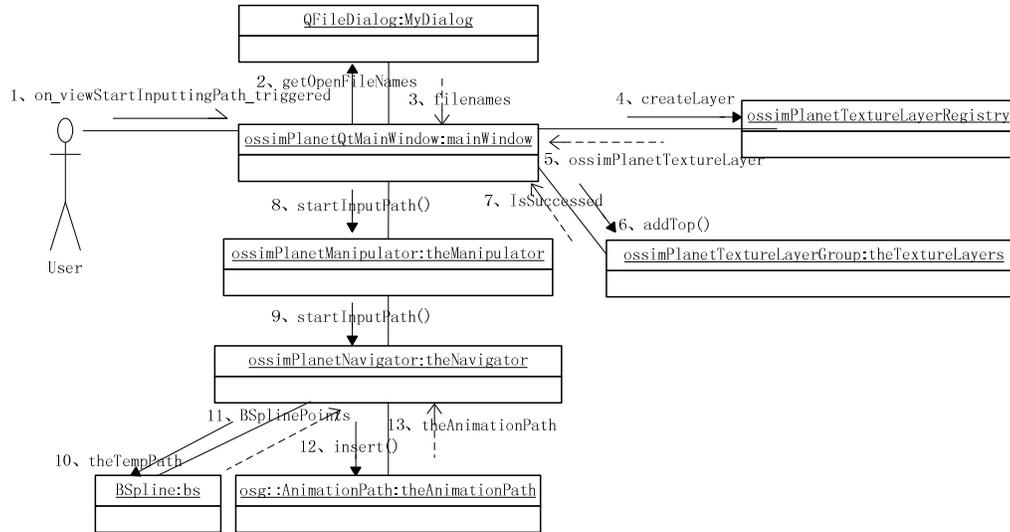


Fig. 9. Cooperation of the components after inputting interested flight.

The message translation and cooperation are summarized below.

- Users fix the path of the aerial photographs and configuration files. Then the point of flight trajectory is stored as `PathPoint` and the input trajectory points as an array of `PathPoint`. The geometry files are also created.
- `createLayer()` is invoked, and the texture layers corresponding to the photographs are created.
- `addTop()` is invoked to load these layers. The actual locations of the photographs are defined in the corresponding geometry files so that the photographs are displayed in the correct positions.
- GUI translates the event messages to the Navigator through the Manipulator.
- In the Navigator, the input flight trajectory points are interpolated and the results are stored as `ossimPlanetAnimationPath`.

4.4. Display of Flight Trajectory

After users choose to display the flight trajectory from GUI, the `on_flieOpenKml_triggered()` is triggered. The cooperation diagram is shown in Fig. 10, and the message translation and cooperation are summarized below.

- The users' option is translated to the Navigator through the Manipulator.
- In the Navigator, the `CreateKmlFilePath()` is invoked to create a KML file.
- Traverse the input trajectory points and export them into the KML file.
- Traverse the interpolated trajectory points and export them into the KML file.
- The `addkml()` in `ossimPlanetKmlLayer` is invoked to create a corresponding KML layer on `ossimPlanet` for display.

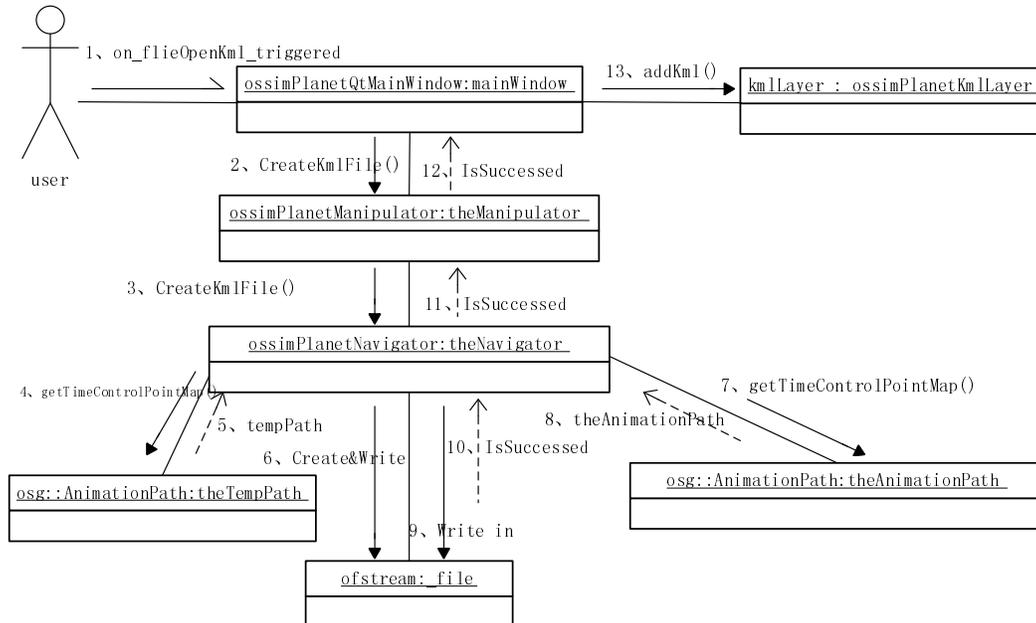


Fig. 10. Cooperation of the components after choosing to display trajectory.

4.5. Playback of Flight trajectory

After users choose to playback the flight trajectory, `on_viewShowThePath_triggered()` is triggered. The cooperation diagram is shown in Fig.10. The message translation and cooperation are summarized blow.

- The users' option is translated to the Navigator through the Manipulator.
- In the Navigator, the `showPath()` is invoked. Current rendering mode is set as the simulation mode.
- The rest work is done along with the Event Traversal. In every frame, the scene parameters are calculated in `update()`. And the calculation is done according to the ControlPoint in the map of `ossimPlanetAnimationPath`.
- Loop the third step with the refresh of `ossimPlanet` to playback the flight trajectory dynamically.

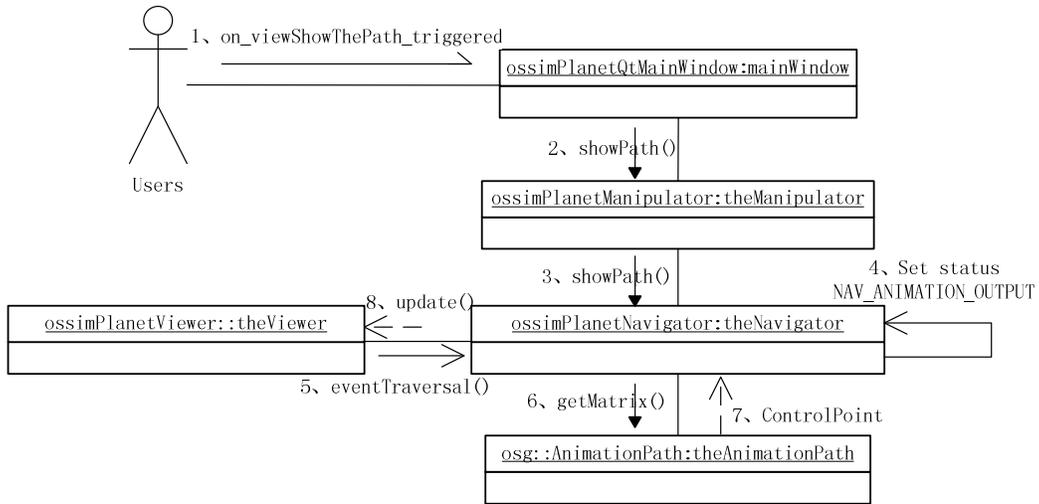


Fig. 11. Cooperation of the components after choosing to playback the flight trajectory.

5. SIMULATION RESULTS

The simulation is done on the ossimPlanet 1.8.4. The interested flight data includes ten photographs and their configuration files. The 10 red-dotted points are the input flight trajectory points as shown in Fig.12 and the corresponding photograph is pasted as shown in Fig.13. The green marks are the interpolated flight trajectory points.

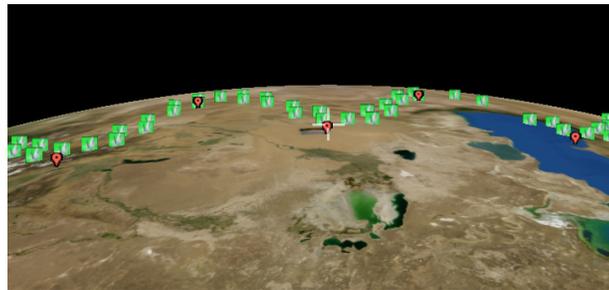


Fig. 12. Marks of flight trajectory.

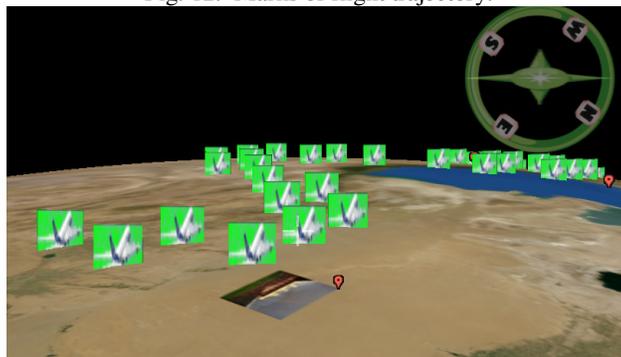


Fig. 13. Playback of flight trajectory.

During playback of flight trajectory, the eye point changes along the flight trajectory. As shown in Fig.13, users could view the photographs and the trajectory marks on the 3D global dynamically. The eye point will change with the plane when the ossimPlanet refreshes its scene.

6. CONCLUSIONS

In this paper, flight trajectory recreation and playback system of aerial mission is implemented based on open-source 3D global platform – ossimPlanet. Users can choose their interested flight of aerial mission. Then the aerial photographs would be displayed on the proper geographic positions of ossimPlanet. The flight trajectory also would be recreated and marked. In addition, the playback of the flight trajectory is simulated on ossimPlanet. These functions allow users to analyze their interested flight in a more institutive way.

The development on open-source platform ensures the security of system in a low cost and high performance. Especially, it allows developers to implement more customized functions. During the development, APIs of ossimPlanet about loading images, loading KML files and rendering frame are overwritten. This paper provides a general method for the development on ossimPlanet with its rendering theory.

REFERENCES

- [1] T.J. Chung and N.Y. Lee,“Trajectory Simulation of the Small Atmospheric Re-entry Module,”in SICE-ICASE International Joint Conference ,Bexco, Busan, Korea, Oct.2006, pp.18-21
- [2] C. Foster, “Trajectory Browser: An Online Tool for Interplanetary Trajectory Analysis and Visualization,”in IEEE Aerospace Conference, Big Sky,MT,Mar.2013,pp.1-6
- [3] X. Zhao, “Research on Flight Trajectory Recreation and Three-Dimensional Flight Playback,” Computer Engineering & Science,vol.34,no.7,2012 (in Chinese)
- [4] SimAuthor, The Leader in Flight Data Analysis & Visualization. [Online].Available: <http://www.simauthor.com/>
- [5] 3D Flight Simulation System-EasyFlight-China Academy of Civil Aviation Science and Technology. [Online]. Available: http://www.cast.c.org.cn/kjcg/hkaqlycg/2006yq1/201006/t20100624_537.html
- [6] Y. Tang, C. Liu and H. Wu, “3D Flight Track and 6-DOF Flight Simulation based on Google Earth, ” Journal of Computer Applications, vol.29,no.12,Dec.2009.(in Chinese)
- [7] Google Earth.[Online].Available:<http://www.google.com/earth/index.html>
- [8] ossimPlanet.[Online].Available:<http://trac.osgeo.org/ossim/wiki/OssimPlanet>
- [9] OSSIM.[Online]. Available:<http://trac.osgeo.org/ossim/>
- [10] OpenSceneGraph.[Online].Available:<http://www.openscenegraph.org/>
- [11] QT.[Online].Available:<http://qt.digia.com/>
- [12] ossimPlanet Users Manual.[Online]. Available:<http://download.osgeo.org/ossim/installers/windows/ossimPlanetUsers.pdf>
- [13] M.H. Tsou and J. Smith,Free and Open Source software for GIS education. [Online].Available:http://geoinfo.sdsu.edu/hightech/WhitePaper/tsou_free-GIS-for-educators-whitepaper.pdf
- [14] Flight dynamics.[Online].Available:<http://en.wikipedia.org/wiki/File:Rollpitchyawplain.png>
- [15] J. Yang,“Analysis of Human Factors in Flight Height Deviation,”Journal of China Civil Aviation Flying College, pp.16-19,Mar. 2000 (in Chinese)
- [16] D. Du,Y. Liu,X. Guo, K. Yamazaki and M. Fujishima, “An accurate adaptive NURBS curve interpolator with real-time flexible acceleration/deceleration control,”Robotics and Computer-Integrated Manufacturing,vol.26, pp.273–281,2010.
- [17] J. Zhao,G. Shan,B. Ji,H. Chen, “MM Tracking Algorithm with the Aid of Pose-Angle of Plane,”Electronics Optics& Control, vol. 18 no.3,Mar.2011 (in Chinese)
- [18] K. Wang and X. Ben, “Gait recognition using linear interpolation,” J. Huazhong University. of Science & Techology. (Natural Science Edition), vol.38,no.2,Feb. 2010 (in Chinese)
- [19] H. Thomas, “Map projections and airborne moving map displays,” in Digital Avionics Systems Conference, Proceedings, IEEE/AIAA 10th. Los Angeles, CA ,Oct.1991,pp.493 – 49

- [20] KML Overview.[Online].Available:<http://www.opengeospatial.org/standards/kml>
- [21] Y. Du,C. Yu and L. Jie, "A Study of GIS development based on KML and Google Earth,"in 5th International Joint Conference on INC, IMS and IDC, Seoul, Aug.2009,pp.1581-1585
- [22] T. Weiss,N. Kaempchen and K. Dietmayer, "Precise ego-localization in urban areas using Laserscanner and high accuracy feature maps," in Intelligent Vehicles Symposium,2005.Proceedings. IEEE, June 2005, pp.284-289
- [23] Y. Zhou, "Sensor Alignment with Earth-Centered Earth-Fixed (ECEF) Coordinate System,"IEEE Transactions on Aerospace and Electronic Systems, vol.35, no.2, Apr.1999
- [24] Coordinate Conversions and Transformation including Formulas, 2nd ed, OGP Surveying and Positioning Guidance Note,no.7,part 2,Jan.2012,pp.92-96.
[Online].Available:<http://www.epsg.org/guides/docs/G7-2.pdf>
- [25] D. Hearn and M.P. Baker, "3D Observation" in Computer Graphics with OpenGL,3rd ed,China:Publishing House of Electronics Industry,2004, pp.288-290
- [26] R. Wang,The Longest Frame.[Online].Available:
<http://bbs.osgchina.org/redirect.php?tid=696&goto=lastpost>

Authors

Wu Wu is a graduate student for master at the Research Center for Natural Computing and Software, College of Computer Science and Technology, Beijing Normal University. Her research interests include software engineering and information systems,3D GIS.Her email is midou@mail.bnu.edu.cn;



AUDIT MATURITY MODEL

Bhattacharya Uttam, Rahut Amit Kumar, De Sujoy

Cognizant Technology Solutions, Kolkata, India

uttam.bhattacharya@cognizant.com / amit.rahut@cognizant.com /
sujoy.de@cognizant.com

ABSTRACT

Today it is crucial for organizations to pay even greater attention on quality management as the importance of this function in achieving ultimate business objectives is increasingly becoming clearer. Importance of the Quality Management (QM) Function in achieving basic need by ensuring compliance with Capability Maturity Model Integrated (CMMI) / International Organization for Standardization (ISO) is a basic demand from business nowadays. However, QM Function and its processes need to be made much more mature to prevent delivery outages and to achieve business excellence through their review and auditing capability. Many organizations now face challenges in determining the maturity of the QM group along with the service offered by them and the right way to elevate the maturity of the same. The objective of this whitepaper is to propose a new model –the Audit Maturity Model (AMM) which will provide organizations with a measure of their maturity in quality management in the perspective of auditing, along with recommendations for preventing delivery outage, and identifying risk to achieve business excellence. This will enable organizations to assess QM maturity higher than basic hygiene and will also help them to identify gaps and to take corrective actions for achieving higher maturity levels. Hence the objective is to envisage a new auditing model as a part of organisation quality management function which can be a guide for them to achieve higher level of maturity and ultimately help to achieve delivery and business excellence.

KEYWORDS

Audit; Software Quality Assurance; Risk Management; Engagement Maturity; Business Excellence

1. INTRODUCTION

For any world class organization, quality compliance to its standard software process [1] is considered as a basic hygiene factor. ISO [2] and CMMI [3] are official certification/assessment for this which each business unit must ensure.

In today's business scenario, focus of the Quality Assurance (QA) function needs to be elevated from traditional compliance related aspects to more value added services to justify its presence to meet business objectives. Audit function, instead of ensuring mere compliance needs to be much more matured to prevent delivery outage and to achieve business excellence which are the call of the day for survival and to prove oneself best in class in the industry.

To keep the quality function as one of the essential business functions, the focus of Quality Assurance activities (audit, review etc.) should be elevated towards higher quality of deliverables and higher performance by strengthening process maturity and quality of data. That way,

prevention of delivery outage can be achieved through proactive identification of the risks associated with delivery management, product quality and process adherence. Furthermore, focusing on business excellence by business risk assessment along with management of client's expectation will help in reaching highest maturity.

2. AUDIT MATURITY MODEL (AMM)

Audit Maturity Model (AMM) framework will provide organizations with an assessment of the maturity of audit and review processes / capabilities in the perspective of auditing capability, along with recommendations for achieving higher levels of maturity. This will ensure assessment of not only the basic hygiene factors but also of engagement maturity and business excellence.

At the bottom level, audit / review activities are informal, chaotic and adhoc. Reviews and audits are carried out mainly on reactive basis to understand and correct burning project issues. Hence success of the reviews and audits depends on the skill of the people conducting the reviews & audits. There is no Software Quality Assurance (SQA) group defined to assess the audit process. This level can be called as Level 1 initial. There is no formal auditing team to meet the basic objective.

At level 2, localized standards of reviews and audits have been recognized, best practices for different reviews and audits are identified and software quality assurance group formed to make it more manageable. At this level, reviews and the audit activities are much more disciplined than level 1 and meet all basic need by focusing on setting up of a standard / compliant process. At this level, SQA Team exists and the objective of audits is to ensure verbatim compliance to meet all basic hygiene. This type of audit can be called as Disciplined Audit, and are carried out by members of the SQA group.

At the next level, the audit activities are completely standardized and consistent. Reviews and audits are now much more compliant to many international standards. The audit function now focuses on process maturity through repeatable results and increasing scope of audits. Sets of well-defined and documented standard processes are established and the auditing activities are now formal. The main objective of audits at this level is to ensure process maturity, and audits are carried out by experienced members of the SQA group.

Level 4 is much more matured and now the focus of audits shifts to proactive risk identification to ensure product quality and maturity. Delivery management with stable product quality and process adherences are key aspects to prevent delivery outage at this level. Audits here are carried out by senior members of the SQA team along with seasoned project and delivery managers.

At the level 5, there is a paradigm shift audits focus on business excellence rather than process maturity or delivery maturity. Assessment of business risks in the area of Finance, Customer Relations, Employee, Infrastructure, and Security are the main objective at this level. At this level, audits are carried out by senior management team members.

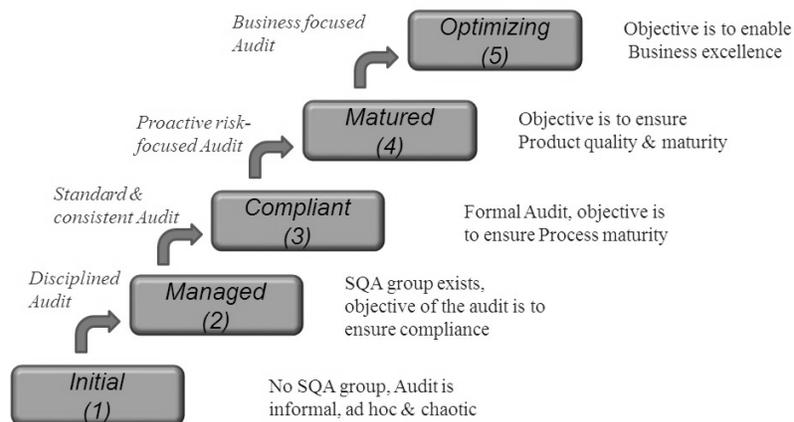


Fig. 1. Audit Maturity Model (AMM)

3. CHARACTERISTICS OF THE AUDIT MATURITY MODEL (AMM)

In Audit Maturity Model, lower levels of maturity form the basis of a higher maturity level. Hence, it is not possible to achieve higher maturity level if a lower level is skipped. Hence assessment of reviews / audit maturity can be achieved stage wise from level 2 to upwards. Followings are few characteristics of Audit Maturity Model:

- This audit model automatically helps to ensures process compliance. Organizations assessed at CMMI level 2 or certified in ISO, AMM helps to ensure compliance to the organization standard software process, thereby confirming basic hygiene.
- At lower maturity level, basic risks are identified and mitigation actions are planned so that the higher maturity level can focus on more vital aspects and identify more business-critical risks.
- Delivery management, product quality and process adherences risks are proactively identified till maturity level 4 which help in enhancing execution maturity.
- Maturity Level 5 reinforces client expectations by identifying and mitigating business risks in the area of Finance, Customer Relations, Employee, Infrastructure, and Security.

4. IMPLEMENTATION APPROACH OF AUDIT MATURITY MODEL (AMM)

The assessment of maturity reviews / audit activities is an examination of different goals defined at different levels by a trained team of professionals using Audit Maturity Model framework as a basis for determining strengths and weaknesses of an organization. This will help to identify gaps at different levels in the framework. Weaknesses can be analyzed and proper action items can be implemented to close the gaps and thus achieve maturity of a particular level, as also proceed to higher maturity levels.

The relationship between the different audits to be conducted and focus area of Audit Maturity Model (AMM) is demonstrated in the figure below. At the bottom of sharp end of V, there is no formal audit or risk assessment. At the next level, the audit is called Discipline Audit to check compliance of level 2 goals of focusing on process compliance and data quality. This can be done through desktop audit by auditing, collecting and analyzing the data for projects of the

organisation. In a mature organization, this can also be performed remotely by extracting necessary data from defined tools. The risk of non-compliance of process and data quality needs to be shared with the corresponding stakeholders to identify and implement further corrective and preventive actions.

	Focus Area	Audit Name
Level 5: Focusing on Finance, Customer Relations, Employee, Infrastructure, Security	Business Risk Assessment	Engagement Maturity Audit
Level 4: Preventing delivery outage	Delivery Management Risk Assessment	Execution Maturity Audit
Level 3: Focusing on quality of deliverables	Product Quality Risk Assessment	Process, Work Product & Delivery Audits
Level 2: Focusing on Process Maturity and Data Quality	Process Compliance Risk Assessment	Desktop Audit
Level 1: Reactive audit, adhoc and chaotic	Adhoc/ reactive audit/ no Risk Assessment	

Fig. 2. Implementation Approach of Audit Maturity Model (AMM)

At the next level, different types of audit are executed like Process Audit which focuses on process maturity, Work Product Audit which ensures quality of all deliverables; and finally Delivery Audit which controls quality of the delivered product or services. These standard and consistent audits can focus on quality of deliverables with process maturity by identifying risks of product quality.

Once the focus has shifted completely from process compliance to process maturity, and quality of deliverables are assured by level 2 and level 3 audit capability of AMM implementation, audits now need to focus on product quality and maturity by identifying proactive risks of delivery management. This Execution Maturity Audit includes product quality with delivery management aspects to prevent delivery outage.

At the highest level, the objective is to identify and assess business risks associated with financial performance, the relationship between various groups in the program / project, customer relationship, staffing, infrastructure, business continuity and security, etc. through Engagement Maturity Audit. At this level, execution maturity transforms to engagement maturity so as to achieve business excellence. The Quality Assurance function aided by senior management must also work proactively at this stage to align the vendors / suppliers, the organization and its customers.

The audit function must identify the aforesaid risks proactively and escalate through defined path to the stakeholders in coordination with project senior team members. The risks must be identified and mitigated proactively before they affect the business or customer. Detailed audit checklists can be made based on different goals and these can be used to dig to a granular level to make the audits more stringent. The appraisal process also needs to be mature enough to produce consistent results through these audits for elevating themselves to the next level.

When planning an audit of the AMM framework, the scope of the disciplines to be included needs to be determined. Other considerations include whether the audit team will consist of members

internal or external to the organization; individuals to be interviewed; and the type or class of maturity necessary.

5. BENEFIT

- A Maturity Level rating assessment of quality assurance function in the perspective of auditing capability will be available
- Helps to comply with basic hygiene factor like ISO and CMMI once audit maturity level 2 is achieved
- Findings that describe the strengths and weaknesses of organisation relative to the AMM
- Consensus regarding the organization's key quality management area.
- An appraisal database in quality assurance area that the organization can continue to use to monitor quality assurance process improvement progress and to support future appraisals
- A proactive risk identification and mitigation for all projects of organisation in the area of delivery management, process, product and business area
- Engagement to execution level maturity of organization
- Align the vendors / suppliers, the organization and its customers as part of a single to reap maximum efficiencies and thus achieve business excellence

6. CHALLENGES

Followings are identified challenges to implement Audit Maturity Model (AMM) framework:

- The commitment from higher management (required for conducting level 5 audits) will be a key challenge as they need to understand the maturity assessment value addition based on their business objective.
- Identifying each aspect of audit checklist for each level would be crucial as this is cost effective in terms of technology, resource and training.
- The level of manual expertise at the internal or external organization level would be crucial.
- Identified findings or risks logging will be a true challenge. Coordination and further risk mitigation, in all levels, need to be synchronized to meet the business objective.

7. CONCLUSION

The Audit Maturity Model (AMM) and its implementation is a new concept in the area of quality assurance to unveil maturity assessment at different levels. Here a lower maturity level forms the basis of the next higher maturity level and hence it is not possible to achieve maturity of a higher level if a lower level is skipped. Hence audit maturity can be achieved stage wise from level 2 upwards. This model strengthens the organization standard process compliance at level 2 with all basic hygiene of process compliance and data quality. Level 3 focuses on process maturity and quality of deliverables by unearthing risk of product quality. At the next level, delivery outage has been prevented by proactive risk identification of delivery management area and finally, at the top level, business risks in the area of finance, customer relations, employee, infrastructure, and security. Based on the impact of business risks, varied levels of rigor are also implemented to check aspects in bottom three levels. Hence, it is a synchronized pre-emptive method of enrichment from a conventional to more business focused state. Proper mitigation of these risks can ensure success of the project and ensures customer satisfaction. The benefits identified for this framework far outweighs the challenges identified.

REFERENCES

- [1] Richard H. Thayer, Merlin Dorfman, "Software Engineering, Volume 2, The Supporting Processes, 3rd Edition", ©2005, Wiley-IEEE Computer Society Pres, August 2005, pp.280-281.
- [2] David I. Levine, Michael W. Toffel, "Quality Management and Job Quality: How the ISO 9001 Standard for Quality Management Systems Affects Employees and Employers," Copyright Harvard Business School© 2008, 2009, 2010 IEEE, January 18, 2010, pp.3-18.
- [3] CMMI for development, version 1.2, CMMI-DEV, V1.2, Carnegie Mellon, Software Engineering Institute, 2006, pp. 116.

Authors

Bhattacharya Uttam is a Senior Consulting Manager of Cognizant Technology Solutions having 19 Years of experience in the field of strategic assessment, process definition, implementation and process improvement in CMMI, Six Sigma, and ISO 9001. Mr. Bhattacharya was born in Kolkata, India on 2nd August, 1970 and obtained his engineering graduation (Bachelor in Technology) in the year 1993 from Calcutta University, India. Mr. Bhattacharya has also completed his MBA (part time) from Calcutta University, India in 2001.



He had played the role of Quality manager for Cognizant and was responsible for ensuring quality of deliverables of the projects. He has implemented CMMI, Six Sigma, ISO 9001 framework, metrics definition for various business units in Cognizant. He has also led the CMMI assessment for Cognizant. He has wide experience in the field of consulting with direct interfacing with many clients for Strategic assessment, Process definition, implementation, improvement and maintaining their Quality Management System for the client organizations spread across geographies. He has also led a number of Six Sigma projects. He has wide experience in organization wide implementation of various processes in different types of projects and has an in-depth understanding of SDLC concepts, continual improvements and high maturity process areas.

Mr. Bhattacharya is a certified Project Management Professional (PMP®) from PMI, USA and has cleared the ITIL® version 3 Foundation Examination from Quint. He is also a certified Six Sigma Black Belt Certification from BMG, and is a certified internal auditor of ISO 9000. Mr. Bhattacharya is a certified Scrum master from Scrum Alliance and is a member of Project Management Institute (PMI), USA. He is also an eminent writer in the Cognizant Process Quality Consulting newsletter and is part of the editorial board.

Rahut Amit Kumar is a consultant of Cognizant Technology Solutions having 11 Years of experience in the field of process definition, implementation and process improvement with CMMI, Six Sigma, and ISO 9001 model. Mr. Rahut was born in Kolkata, India on 31st October, 1977 and became an engineering graduate (Bachelor in Technology) in the year 2002 from Calcutta University, India.



He has wide experience in the field of consulting with direct interfacing for many clients for process definition, implementation, and process improvement and maintaining their Quality Management System. He has implemented CMMI, Six Sigma, ISO 9001 framework, metrics definition for a client organization. He has worked as a Configuration Manager in the IT division of the largest private bank in Europe. He has experience in organization wide implementation of process management applications for application development and maintenance projects and has an in-depth understanding of SDLC concepts, continual improvements and high maturity process areas. He has worked as a Quality Lead for process benchmarking and implementation for a big manufacturing organization and had implemented Theory of Constraint project resulting in increased profitability.

Mr. Rahut is certified Project Management Professional (PMP®) from PMI, USA, A PRINCE2® Practitioner from APMG, UK and certified in ITIL® version 3 Foundation from APMG, UK. He is also an eminent writer in the Cognizant Process Quality Consulting newsletter and is part of the editorial board.

De Sujoy is a consultant of Cognizant Technology Solutions having 8 years of experience in various fields of Software Quality and Tool Implementation. Mr. De was born in Bankura, India on 28th of July, 1981 and received his engineering degree (Bachelor in Computer Science & Engineering) in the year 2004 from Burdwan University, India, and Diploma in Business Administration in the year 2009 from Pune University, India.



He has wide experience in various fields of software quality like Process definition & implementation, process improvement and maintaining the Quality Management System. He has also experience in CMMI Level 3 implementation, ISO 9001 framework and metrics definition. He has worked as a Configuration Manager for the IT division of one of the largest private banks in Europe. He has experience in organization wide implementation of process management applications for application development and maintenance projects and has an in-depth understanding of SDLC concepts, continual improvements and high maturity process areas. In his previous organization, he was instrumental in the organization's achieving the ISO 9001:2000 recertification and its preparation for ISO 140001 certification.

INTENTIONAL BLANK

A STRUCTURAL APPROACH TO IMPROVE SOFTWARE DESIGN REUSABILITY

Tawfig M. Abdelaziz, Yasmeen.N.Zada and Mohamed A. Hagal

University of Benghazi, Faculty of Information Technology,
Department of Software Engineering

tawfig@cs.uni-essen.de, yasmeen.zada@hotmail.com
and Mohamed.hagal@benghazi.edu.ly

ABSTRACT

Software reuse become a very promising area that has many benefits such as reducing costs, time, and most importantly, increasing quality of software. However, the concept of reuse is not only related to implementation level, in fact, it can be included in the earlier stages of the software development life cycle such as design stage. Adopting reuse at this stage provides many benefits such as increasing productivity, saving time and reducing cost of software development.

Accordingly, this paper presents the concept of reuse at design level in more details. As well as, it proposes an approach to improve the reusability of software design using the directed graph concept. This leads to produce a design to be considered as reusable components which can be adapted in many software systems.

KEYWORDS

Software Reusability, Software Component, Unified Modeling Language (UML), Parameterization, Directed Graph.

1. INTRODUCTION

Software reuse is a fundamental aspect of high quality software. Effective reuse of software products is increasing productivity, saving time and reducing cost of software development. However, as the concept of reusing software components is very clear at the code level, while the same concept becomes more difficult to address when discussed in the context of reusing designs. The problem with design reuse in Software Engineering is the shortage of guidelines or approaches that support and guide the designers to be useful from previous design components.

In response to this, some researches related to reuse at design stage presented approaches that aim to improve design reusability. Gui and Scott in [1] worked on measuring software reusability by applying coupling and cohesion metrics on java components, and also focused on reflecting the complexity of those components to be used in reusability activities. Kang, Cohen and Holibaugh

in [2] proposed the work based on the refinement of the software lifecycle to identify reuse activities. This work concentrated more on identification of reusable resources than constructing reusable resources. Price and Demorgian[3]measure object oriented design reusability focusing on abstraction concept using metrics to measure coupling and cohesion dependency relationships. Mishra's Misra's [4]used reverse engineering of legacy software to create reusable components as an attempt to understand the re-existing software by re-designing it. Johnson and Russo[5] described design techniques that support abstract classes and framework. It provided a way to express the design to customize it, developing frameworks and tools that facilitated the design reusability.

The work in this paper is motivated primarily by the possibility of improving and increasing the degree of reusability of design in any software system by discussing the concept of reuse associated with the level of design. It presented a structural approach that is mainly considered with the improvement of software design reusability, to result a design that is potentially reusable. The following section describes the steps of the proposed approach that is expected to improve the design reusability.

2. THE PROPOSED APPROACH

The work of this paper was motivated by the design reuse model represented and illustrated in [6], and the design for reuse process of that model was the base of the proposed approach introduced in this section.

The proposed approach consists of four activities as shown in Figure 1: Design classes, Refinement, Reusable packages and Documentation.

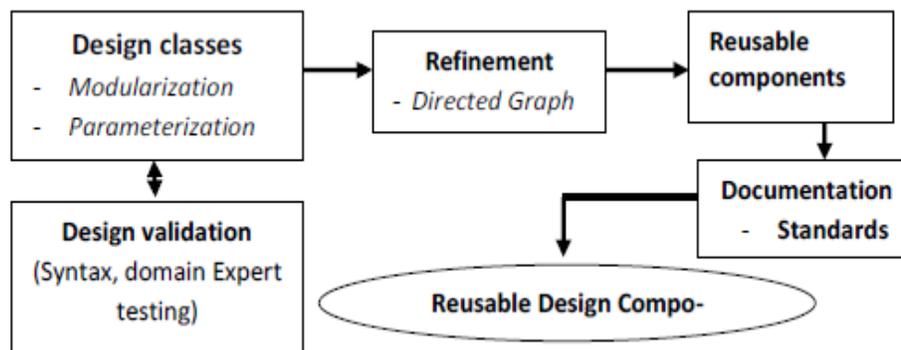


Fig. 1. Conceptual overview of the proposed approach

These activities will produce design components that are ready to be reused in many software systems. Figure 1 illustrates a conceptual overview of the proposed approach. The directions show the priority of the activities' steps. So, it must be considered that it cannot be move from Design classes step unless the design validation activity effectively achieved. This insures that mistakes are detected and handled early.

2.1 Design classes

The first activity in this approach is to decide major classes of the system design. The technique used to decide what classes are needed is to go through the software requirements and identify the nouns (entities) that can be considered as classes. After that, identify the characteristics (state and behavior) of each class, and then define the relationships between these classes. Finally, construct a class diagram as shown in figure 2. The structure of a system is represented using class diagrams. Therefore, modularization and the parameterization concepts must be taken into consideration in this process, due to their great impact on reusability of design. The use of modularity concept makes a system design to be considered as a set of smaller parts that should satisfy the quality concepts such as maintainability, testability and reusability. The effective modularity can be achieved by developing functional independence modules with single-minded function and refusal to excessive interaction with other modules.

Modularity and functional independence could be measured by two qualitative criteria: coupling and cohesion. Coupling is "a measure of inter-module connectivity, and is concerned with identifying the forms of connection that exist between modules" [7]. Cohesion, in its turn, provides "a measure of the extent to which the components of a module can be considered to be 'functionally related'. The ideal module is in which all the components can be considered as being solely present for one purpose"[7]. Furthermore, parameterizing methods of classes is a very important activity to improve design reusability [3]. Figure 2 illustrates an example of how a method (operation) can be parameterized in a class diagram at the design level. The Employee class on the left shows a duplication of method raise() to do same things with different values fivePercentRaise() and tenPercentRaise(). Where, the Employee class on the right shows the parameterization of the method by adding a parameter percentage to the raise() method which later prevents duplication, which will not require much effort in creating the program, and most importantly, makes it easier to reuse the design with no much information to be included.



Fig. 2. Parameterized method [11]

Another important note regarding with class attributes, a class with one or two attributes should be focused on, this may indicate that those attributes belong to another class related to the first class [9]. Therefore these attributes could be aggregated into one class. In this paper, Figure 3 illustrates an example of a small hospital system.

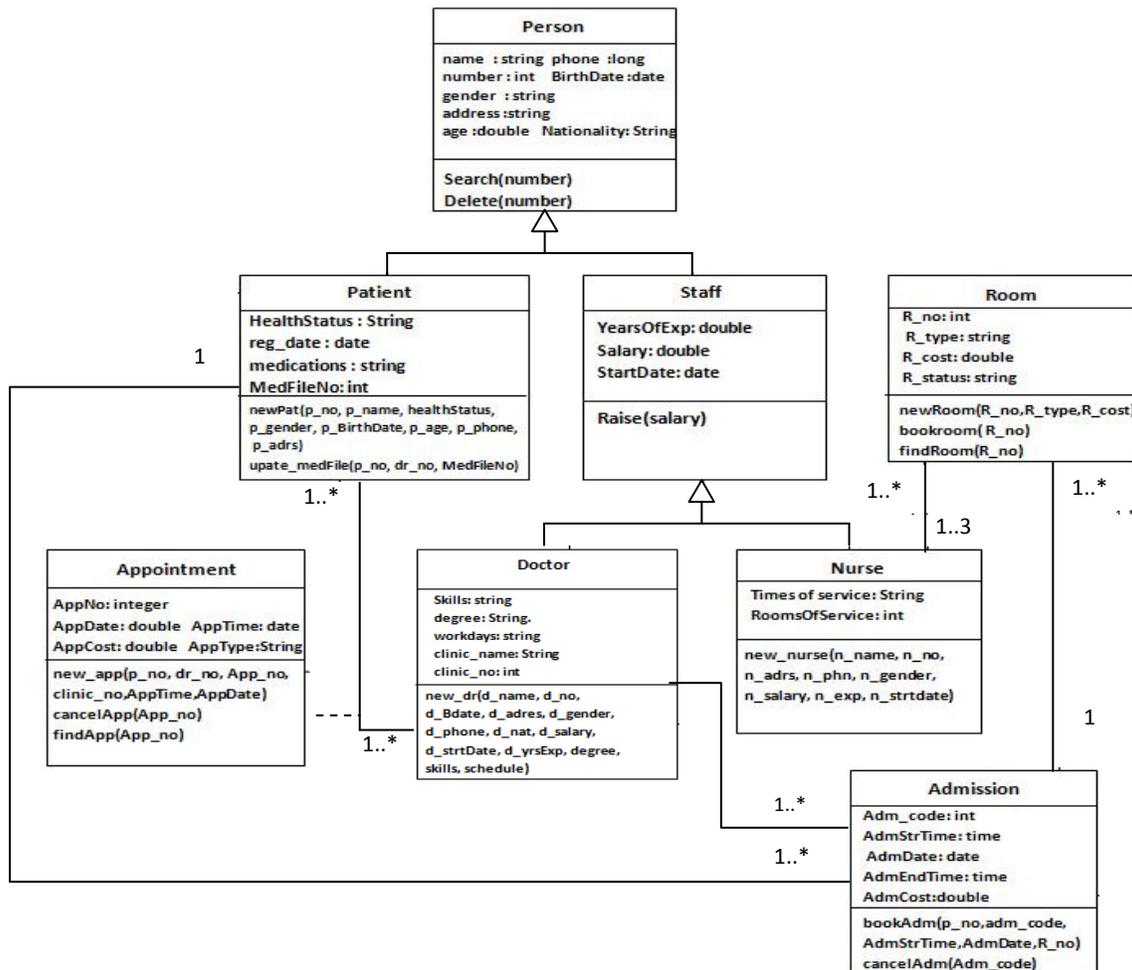


Fig. 3. Health- care UML Class diagram

It consists of the following classes: person, patient, staff, doctor, nurse, clinic, room, appointment, admission, surgery, and test. Modularity concept was represented by a class diagram and their methods are parameterized to increase reusability by providing more information.

2.2 Design Validation

To be able to reuse a component, and to insure that a component is reliable and ready to be reused, you have to make sure that this component does not contain any defects or errors.

Since the design validation work in this paper is based on the approach presented in [14] for class diagram testing. Therefore, UML class diagrams can be tested using some independent methods: Syntax Testing and Domain Expert Testing. Syntax testing is used to verify that the class diagram is correctly and properly constructed. Accordingly, three questions need to be answered: Is it complete? Is it correct? Is it consistent?. Then the domain expert testing is used to insure that the design is correct. Table 1 and Table 2, illustrate the syntax and domain expert testing for the chosen health-care systems.

Table 1. Syntax Testing of Health care system class diagram

1.	Does each class define attributes, methods, relationships, and cardinality?	✓
2.	Is each associations' and aggregations' cardinality correct?	✓
3.	Are all parameters explicit rather than being embedded in method names?	✓
4.	Do all subclasses implement the "is-a-kind-of" relationship properly?	✓
5.	In inheritance structures, are all attributes and methods pushed as high in the inheritance structure as is proper?	✓
6.	Does each association reflect a relationship that exists over the lives of the related objects?	✓
7.	Are each 0..* and 1..* relationships implemented ?	✓

Table 2. Domain Expert Testing of Health care system class diagram

1	Is each class named with a strong noun?	✓
2	Is each attribute defined within the proper class? Is it of the correct type?	✓
3	Is each method in the correct class?	✓
4	Are all method names strong verbs?	✓
5	Does each method take the correct input parameters and return the correct output parameter?	✓
6	Does each method implement one and only one behavior?	✓

2.3 Refinement

The main goal of the refinement step is to produce reusable components. This will be achieved by convert the class diagram into a directed graph. Each class is represented as a node in the graph and the directions of the edges is the direction of dependencies between classes as shown in figure 4. The idea behind this transformation is to introduce a technique to help increasing cohesion of design and improve reusability as a result.

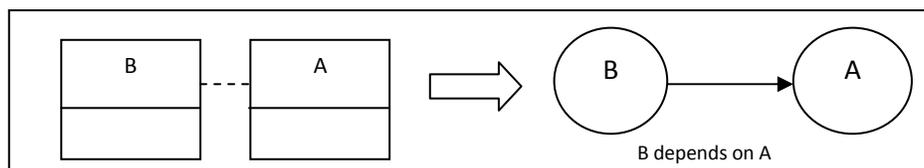


Fig. 4. Transforming classes to a graph

Attention should be paid to the inheritance relationships in a class diagram when transforming it into a directed graph. This relation is represented by a dotted edge between nodes in the graph. The nature of object oriented design with inheritance is to migrate more general information and operations up to the hierarchy where they can be reused by all descendants. Therefore when there is a need to reuse a child class it should be also reuse the parent class. However, when there is a need to reuse the parent class, it is not required to reuse its subclasses, since the parent class does not use members of its subclasses.

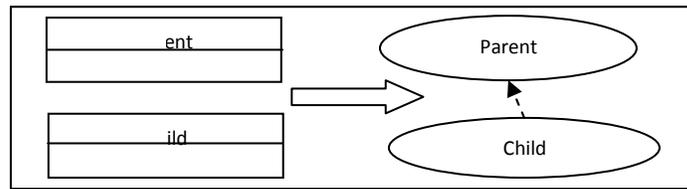


Fig. 5. Inheritance to directed graph

As a result, new systems can reuse the top portion of a hierarchy or the whole hierarchy, but they cannot reuse just a lower part of a hierarchy. Figure 5 shows an illustration of transforming inheritance relationship between classes to a directed graph.

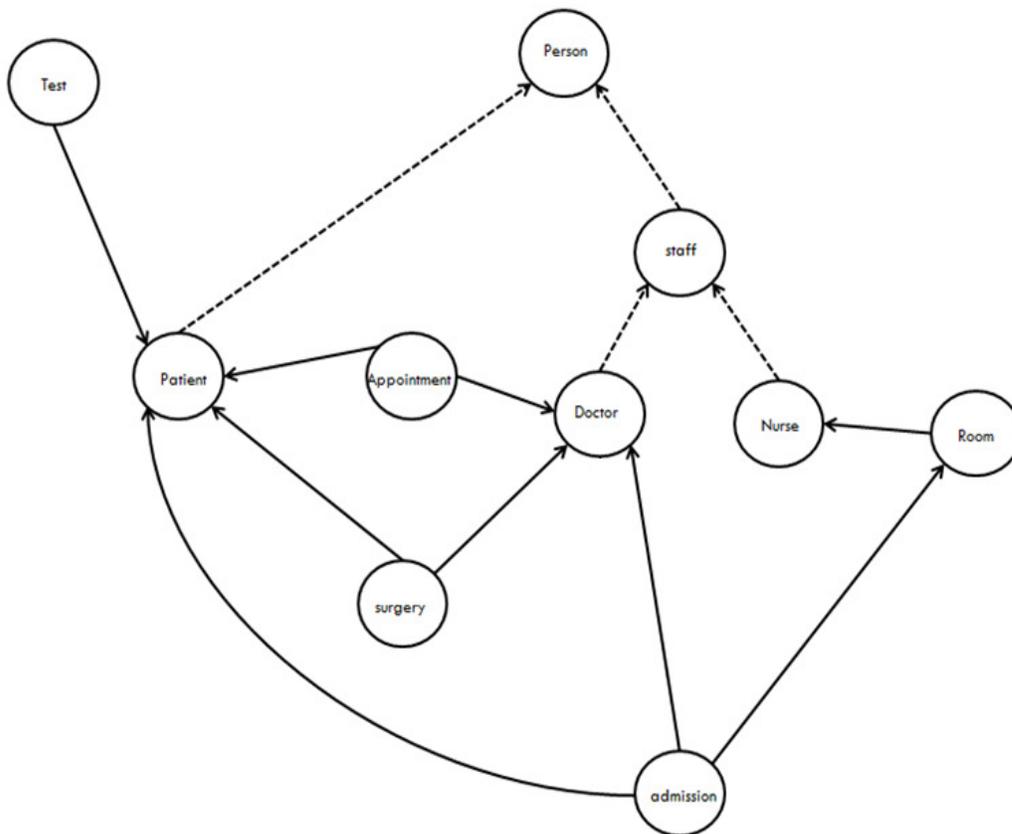


Fig. 6. Directed graph of a health-care system

Reusing just the top portion of a hierarchy is desirable in many cases, since the lower level classes are more specific classes, so with respect to their parents, they are less likely to be needed in other applications [3]. The following figure (Figure 6) illustrates the transformation of the figure 3 from class diagram into a full directed graph with taking into consideration the direction of dependencies and the inheritance relationships between classes.

2.4 Reusable Components

The purpose of this step is to extract the reusable components (packages) of the design. This will be achieved by grouping the nodes with the same directions, which was determined according to the relationships between classes of the system to deal with them as independent system components. So, every sub-graph (component) can be reused separately. Also, from another perspective, we can say that each component is a cohesive module (component), where every member is related with other members of the same module without the need to have dependencies on other modules. Figure 7 illustrates some examples of reusable components that can be generated from the directed graph illustrated in Figure 6.

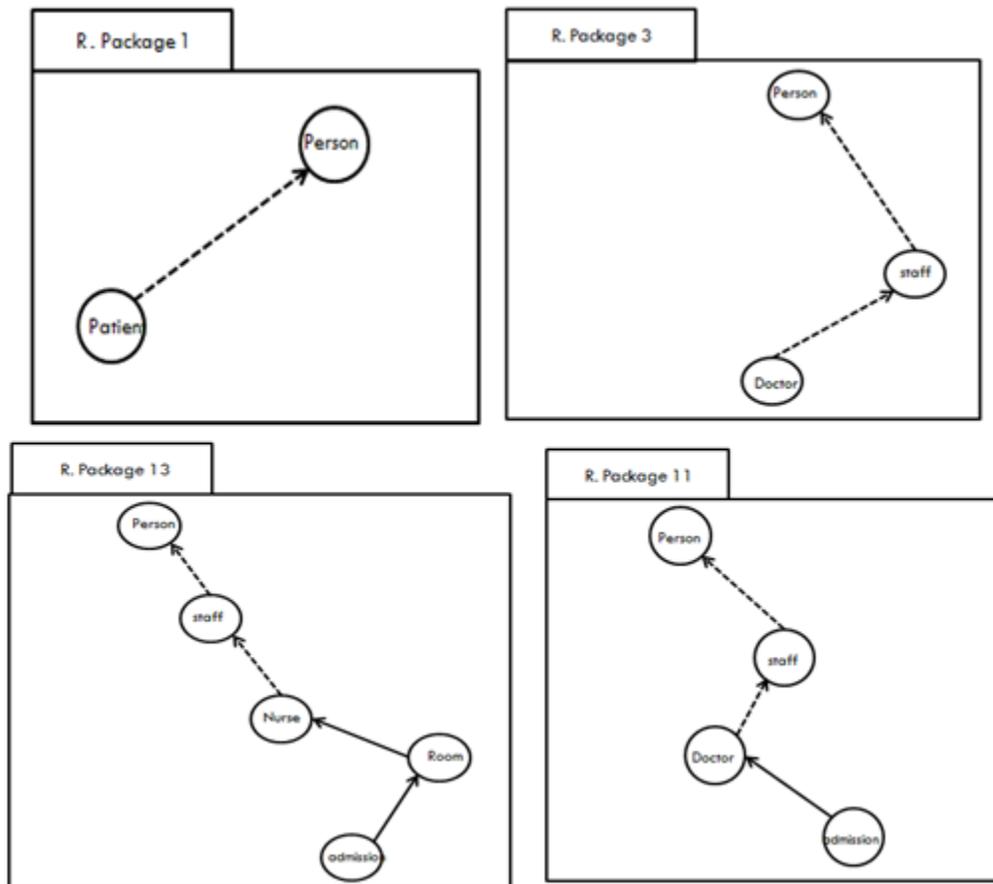


Fig. 7. Reusable packages

2.5 Documentation

The importance of documentation in software component reuse is critical. It needs accurate information about a component in order to state the component with a requirement by referencing the Software Design Document (SDD) [12][13]. Although good documentation of components is essential to software reusability, but in fact, Traditional documentation usually does not meet these needs, where there are some specific information that must be provided in order to indicate the reusability of a component whether it was a code component or a component from an early stages of the development such as requirements or design.

Some advice was provided in [10] about what should be included in a reuse document. Table 3 illustrates of documentation of two selected packages as an example from the given Health care system.

Table 3. Component documentation example

Component name	Identification	Specification	Technical restrictions	Commercial or legal restrictions	Problems	Recommended enhancements
R.Package1	Reusing this component is when the system needs to have patient information	Patient class that inherits from Person class, which contains all the properties and operations that need to be done for a patient	VB.Net programming language	This component is suitable only for health care systems	None (component is tested and bugs are detected and fixed)	This components can be integrated with other components to extend functionality
R.Package2	Reusing this component is when the system needs to have staff information	Staff class that inherits from Person class, which contains all the properties and operations that need to be done for a Staff member	VB.Net programming language	This component is suitable for health care or management systems	None (component is tested and bugs are detected and fixed)	This components can be integrated with other components to extend functionality

3. CONCLUSION

Reuse at design level is an important aspect that should be focused on during the software development life cycle. The proposed approach described how to improve design reusability in a structured way that can be applied on any software design in order to produce design of reusable components. The use of directed graphs helps the designers to understand how to extract the design components by putting them in a form of nodes and directed edges that can be grouped into packages (components) that could be reused in other systems.

As a future work, we are considering to develop a tool that performs the steps of the approach by providing the class diagram, and then the tool performs the refinement process of the class diagram to a directed graph and generates reusable packages of the produced.

REFERENCES

- [1] Gui, G & Scott, P.D. (2009), “Measuring software component reusability by coupling and cohesion metrics”, Journal of computers, Vol.4, No.9.
- [2] Kang, K.C, Cohen, S & Holibaug H.R. (1992), “Reuse-Based Software Development Methodology”, (Report No. SEI-92-SR-4). Application of Reusable Software Component Project.
- [3] Price, M.W & Demorgian, S.A. (1997), “Analyzing and measuring reusability in object oriented designs”, University of Connecticut, computer science and engineering department.
- [4] Mishra, S.K, Kushwaha, D.S & Misra, A.K. (2012), “Creating reusable software components from object oriented legacy system through reverse engineering”, Journal of object technology, pp 1-13.
- [5] Johnson, R.E & Russo, V.F. (1991), “Reusing object oriented designs”, Unpublished Manuscript, department of computer science, University of Illinois, West Lafayette.
- [6] Duffy, S.M, Duffy, AHB & MacCallum, K.J. (1995), “A Design Reuse Model”. International conference of engineering design (iced 95): Glasgow, Heurista, 490-495.
- [7] Budgen, D. (2003), “Software design”, England: Pearson education.
- [8] Price, M.W & Demorgian, S.A. (1997) “Analyzing and measuring reusability in object oriented designs”, University of Connecticut, computer science and engineering department.
- [9] Chidamber, S.R & Kemerer, C.F. (1994), “A Metrics Suite for Object Oriented Design”, IEEE Transaction on Software Engineering, Vol.20, No. 6.
- [10] Sametinger, J. (1996) “Reuse Documentation and Documentation Reuse”, A&M University, Texas, USA.
- [11] Sourcemaking.com/refactoring, Access: (January, 2013).
- [12] Jones, M & Mortensen, U. (1995), “Guide to the software detailed design and production phase”, ESA publications divisions, Vol.1: Paris, France.
- [13] Kuhns, R.D. (1998) “Strategies for designing and building reusable GIS application components”, Unpublished Manuscript, Convergent Group, Englewood, Colorado.

AUTHORS

Tawfig M. Abdelaziz

He is an assistant professor at the department of software engineering and a vice dean of faculty of Information Technology, University Of Benghazi, Libya. He is interesting in Agent systems, software Quality Assurance, software project management and Formal methods



Yasmeen.N.Zada

She is a graduate student at Department of Software Engineering, Faculty of Information Technology, University of Benghazi, Libya.



Mohamed Ali Hagal

He is a lecturer at the department of software engineering, faculty of Information Technology, Benghazi University-Libya. He is interesting in software requirements engineering, software design and software project management.



INTENTIONAL BLANK

QUALITY-AWARE APPROACH FOR ENGINEERING SELF-ADAPTIVE SOFTWARE SYSTEMS

Mohammed Abufouda

Department of Computer Science,
Technical University of Kaiserslautern, Kaiserslautern, Germany
abufouda@cs.uni-kl.de

ABSTRACT

Self-adaptivity allows software systems to autonomously adjust their behavior during run-time to reduce the cost complexities caused by manual maintenance. In this paper, an approach for building an external adaptation engine for self-adaptive software systems is proposed. In order to improve the quality of self-adaptive software systems, this research addresses two challenges in self-adaptive software systems. The first challenge is managing the complexity of the adaptation space efficiently and the second is handling the run-time uncertainty that hinders the adaptation process. This research utilizes Case-based Reasoning as an adaptation engine along with utility functions for realizing the managed system's requirements and handling uncertainty.

KEYWORDS

Software Quality, Model-Driven Software, Self-adaptive Software Systems, Case-based Reasoning, Run-time Uncertainty

1. INTRODUCTION

The majority of the existing work in the literature agrees [1][2] that *self-adaptivity* in software systems is the ability of a software system to adjust its behaviour during run time to handle software system's complexity and maintenance costs [3] while preserving the requirement of the system. This property dictates the presence of an adaptation mechanism in order to build the logic of self-adaptivity without human intervention. Developing a self-adaptive software system is subjected to many challenges like handling the complexity of the adaptation space of the managed system. This complexity is conceived when the number of the states that the managed system can run in is relatively large. Also, this complexity manifests itself when new states are needed to be inferred from previous one i.e. learning from past experience. Another challenge is the uncertainty that hinders the adaption process during run-time. This paper will address these challenges. More precisely, our framework is concerned with the following problems:

- *Adaptation responsible unit:* The majority of the existing work do not provide a modular separation between the adaptation engine and the managed system. Embedding the adaptation logic within the managed system components increases the complexity in the development process of a self-adaptive software system. This also limits the reusability of the work achieved in one application to other applications or domains.

- *Run-time uncertainty handling*: Uncertainty is a challenge that exists not only in self-adaptive software systems but also in the entire software engineering field on different levels. Therefore managing uncertainty is an essential issue in constructing a self-adaptive software system as uncertainty hinders the adaptation process if it is not handled and diminished.
- *Adaptation space*: The adaptation process raises a performance challenge if the adaptation space is relatively large, particularly when new adaptations are required to be inferred. This challenge requires an efficient mechanism that guarantees learning new adaptations as well as providing the adaptation with satisfactory performance. This means that the adaptation engine's response should be provided as soon as an adaptation is issued since late adaptations provided by the adaptation engine could be futile.

The rest of this paper is structured as follows: Section 2 lists the related work and the existing gaps in the literature. Section 3 shows the expected contributions of our research and Section 4 describes our proposed solution and its model. Section 5 and Section 6 contains the progress and the future of our research, in particular the evaluation. This paper concludes in Section 7.

2. RELATED WORK

The body of literature in the area of self-adaptivity has provided a plethora of frameworks, approaches and techniques to enhance self-adaptivity that is widespread in many fields. This section contains the related work to our research. In the following sections, we will present the related work categorized according to the mechanisms used to support self-adaptivity.

2.1 Learning based adaptation

Salehie and Tahvildari [2] proposed a framework for realizing the deciding process performed by an external adaptation engine. They used knowledge base to capture the managed system's information namely domain information, goals and utility information. This is used in the decision-making algorithm, as they name it, which is responsible for providing the adaptation decision. In [5], Kim and Park provided a reinforcement learning-based approach for architecture-based self-managed software using both on-line and off-line learning. FUSION [6], was proposed by Elkhodary et al. to solve the problem of foreseeing the changes in environment, which hinders the adaptation during run time for feature-based systems using a machine learning technique. In [7], Mohamed-Hedi et al. provided a self-healing approach to enhance the reliability of web services. A simple experiment was used to validate their approach without empirical evidence.

2.2 Architecture & model based adaptation

RAINBOW [8] is a famous contribution in the area of self-adaptation based on architectural infrastructures reuse. RAINBOW monitors the managed system using abstract architectural models to detect any constraints violation. GRAF [9] was proposed for engineering self-adaptive software systems. The communication between the managed system and GRAF framework is carried out via interfaces. This approach has a performance overhead because GRAF reproduces a new adaptable version of the managed system. Similar to GRAF [9] Vogel and Giese [10] assumed that adaptation can be performed in two ways, parameter adaptation and structural adaptation. They provided three steps to resolve structural adaptation and used a self-healing web application as an example. Morin et al. [11] presented an architectural based approach for realizing software adaptivity using model-driven and aspect oriented techniques. The aim of this approach was to reduce the complexities of the system by providing architectural adaptation based solution. They provided model-oriented architectures and aspect models for feature

designing and selection. Khakpour et al. [12] provided PobSAM, a model-based approach that is used to monitor, control and adapt the system behaviour using LTL to check the correctness of adaptation. Asadollahi et al. [13] presented StarMX framework for realizing self-management for Java-based applications. In their work they provided so called autonomic manager, which is an adaptation engine that encapsulates the adaptation logic. Adaptation logic was implemented by arbitrary policy-rule language. StarMX uses JMX and policy engines to enable self-management. Policies were used to represent the adaptation behaviour. This framework is restricted to Java-based application as the definition of processes is carried out by implementing certain Java interfaces in the policy manager. They evaluated their framework against some quality attribute. However, their evaluation for quality attributes was not quantified quantitatively. The work in [14] provided a new formal language for representing self-adaptivity for architecture-based self-adaptation. This language was used as an extension of the RAINBOW framework [8]. This work explains the use of this new language using an adaptation selection example that incorporate some stakeholders' interests in the selection process of the provided service which represents the adaptive service. Bontchev et al. [15] provides a software engine for adaptable process controlling and adaptable web-based delivered content. Their work reuses the functionality of the existing component in order to realize self-adaptivity in architecture-based systems. This work contains only the proposed solution and the implementation without experiment and evaluation.

2.3 Middleware based adaptation

In [16], a prototype for seat adaptation was provided. This prototype uses a middleware to support an adaptive behaviour. This approach was restricted to the seat adaptation which is controlled by a software system. Adapta framework [17] was presented as a middleware that enabled self-adaptivity for components in distributed applications. The monitoring service in Adapta monitored both hardware and software changes.

2.4 Fuzzy control based adaptation

Yang et al. [18] proposed a fuzzy-based self-adaptive software framework. The framework has three layers: (1) Adaptation logic layer, (2) Adaptable system layer, which is the managed system and (3) Software Bus. The adaptation logic layer represents the adaptation engine that includes the fuzzy rule-base, fuzzification and de-fuzzification components. This framework has a set of design steps in order to implement the adaptation. POISED [19] introduced a probabilistic approach for handling uncertainty in self-adaptive software systems by providing positive and negative impacts of uncertainty. An evaluation experiment had been applied which showed that POISED provided an accepted adaptation decision under uncertainty. The limitations of this approach are that it handles only internal uncertainty and does not memorize and utilize previous adaptation decisions.

2.5 Programming framework based adaptation

Narebdra et al. [20] proposed programming model and run time architecture for implementing adaptive service oriented. It was done via a middleware that solves the problem of static binding of services. The adaptation space in this work is limited to three situations that require adaptation of services. MOSES approach was proposed in the work [21] to provide self-adaptivity for SOA systems. The authors used linear programming problem for formulating and solving the adaptivity problem as a model-based framework. MOSES aimed to improve the QoS for SOA, and the work in [21] provides a numerical experiment to test their approach. QoS MOS [22] provided a tool-supported framework to improve the QoS for the service based systems in adaptive and predictive manner. The work in [23] provided an implementation of architecture-based self-adaptive software using aspect oriented programming. They used a web-based system as an experiment to

test their implementation. Their experiment showed that the response time of the self-adaptive implementation is better than the original implementation without a self-adaptivity mechanism. Liu and Parashar [24] provided Accord, which is a programming framework that facilitates realizing self-adaptivity in self-managed applications. The usage of this framework was illustrated using forest fire management application.

Table 1, which is similar to what proposed in [4], summarizes the related work done in this research. The table has two aspects of comparison (1) Research aspects and (2) Self-adaptivity aspect. The earlier aspect is important and represent an indication regarding the maturity and creditability of the research. The later aspect is related to the topic of this paper.

Table 1: Summary of related work

Covered literature categorization	Work	Research aspects						Self-adaptive software system aspects					
		Problem Statement	Contribution statement	Experiment	evaluation metrics	Limitations	Threats to validity	Adaptation Expediency	Adaptation remembrance	Uncertainty Handling	Adaptation Res. Time	Adaptation style	Adaptation engine
Learning based adaptation	[2]	√	√	X	X	X	X	X	√	X	X	Dynamic	External
	[5]	√	√	√	X	X	X	√	X	X	Dynamic	External	
	[6]	√	√	√	√	√	X	√	√	X	√	Dynamic	External
	[7]	X	X	√	X	X	X	X	X	X	X	Dynamic	External
Architecture & model based adaptation	[8]	√	√	√	√	X	X	X	X	√	√	Dynamic	External
	[9]	√	√	√	√	X	√	X	X	X	X	Dynamic	External
	[10]	√	√	√	X	X	X	X	X	X	X	Static	Internal
	[13]	X	X	√	X	X	X	√	X	X	X	Dynamic	External
	[11]	X	X	√	√	X	X	√	X	X	√	Dynamic	External
	[12]	√	√	X	X	X	X	X	X	X	X	Dynamic	Internal
	[14]	√	√	√	X	X	X	X	X	X	X	Static	External
Middleware based adaptation	[15]	√	√	X	X	X	X	X	√	X	X	Dynamic	External
	[16]	√	√	√	X	X	X	√	X	X	X	Static	Internal
Fuzzy control based adaptation	[17]	√	√	X	X	X	X	X	X	X	X	Dynamic	External
	[18]	√	√	X	X	X	X	X	X	X	X	Dynamic	External
Programming framework based adaptation	[19]	√	√	√	√	X	X	√	X	√	√	Dynamic	Internal
	[20]	X	X	√	√	X	X	X	X	X	X	Dynamic	External
	[21]	√	√	√	X	X	X	√	X	X	X	Dynamic	External
	[23]	√	√	√	√	X	X	√	X	X	√	Dynamic	Internal
	[24]	√	√	√	X	X	X	√	X	X	√	Dynamic	Internal

3. RESEARCH CONTRIBUTION

In this research, we realize self-adaptivity in software system by providing an external adaptation engine which reduces the changes in the managed system and subsequently in the entire self-adaptive system. Our approach utilizes Case-based Reasoning (CBR) [25] as an external adaptation engine in order to overcome the aforementioned challenges. Specifically, this research proposes a framework that we claim it addresses the following challenges:

- Separating the managed system and the adaptation engine in a modular fashion in order to overcome the drawbacks of embedding the self-adaptivity logic within the managed system. This idea is one of the key ideas in the IBM autonomic element [26] which suggests a modular separation between the managed system and the adaptation engine.
- Managing the complexity of adaptation space by remembering the previously achieved adaptations stored in a knowledge base, which improves the performance of the

adaptation process. The remembrance supports not only the complexity of the adaptation space, but also the performance of the adaptation engine. That is because recalling already existing adaptation is better than constructing it from scratch in terms of performance.

- Handling the run-time uncertainty that appears in the adaptation process due to the managed system's environment changes or our framework's internal model. We utilize and incorporate the probability theory and the utility functions as proposed in [27].

4. PROPOSED SOLUTION

In this section, an overview of our proposed solution will be presented. Figure 1 shows a reference model of our solution that will be described in the following sections.

4.1. External adaptation engine

The adaptation engine contains an *adaptation mediator*, which is responsible for:

- Monitoring the managed system by reading its attributes to decide whether an adaptation is required or not. We suppose that the managed system provides a service with overall utility U . If U is below or is approaching a predefined utility threshold i.e. " UT ", then the monitoring unit issues an adaptation process. The *adaptation request* is the set of the managed system attribute values at the time of issuing the adaptation. Consequently, the adaptation request is sent to the adaptation engine to perform the adaptation process.
- Executing the *adaptation response* received by the adaptation engine. The adaptation response is the result of the adaptation process performed by the adaptation engine, which is the corrective state to be applied on the managed system.

In addition to the adaptation mediator, the adaptation engine embraces a *Case-based reasoning engine*. Typically, CBR life cycle consists of four stages:

1. *Retrieve*: The CBR system retrieves the most similar case(s) from the Knowledge Base (KB) by applying the similarity measures on the request case.
2. *Reuse (Adapt)*: In this stage, CBR employs the information of the retrieved cases. If the retrieved cases are not sufficient in themselves to solve the query case, the CBR engine adapts this/these case/s to generate a new solution. Some of the common techniques for reusing and adapting the retrieved knowledge are introduced in [28]. Our approach uses *Generative Adaptation* [29], which requires some heuristics e.g. utility functions, to provide an efficient adaptation process.
3. *Revise*: A revision of the new solution is important to make sure that it satisfies the goals of the managed system. Revision process can be done by applying the adaptation response to real world, evaluate it by the domain expert or by simulation approaches.

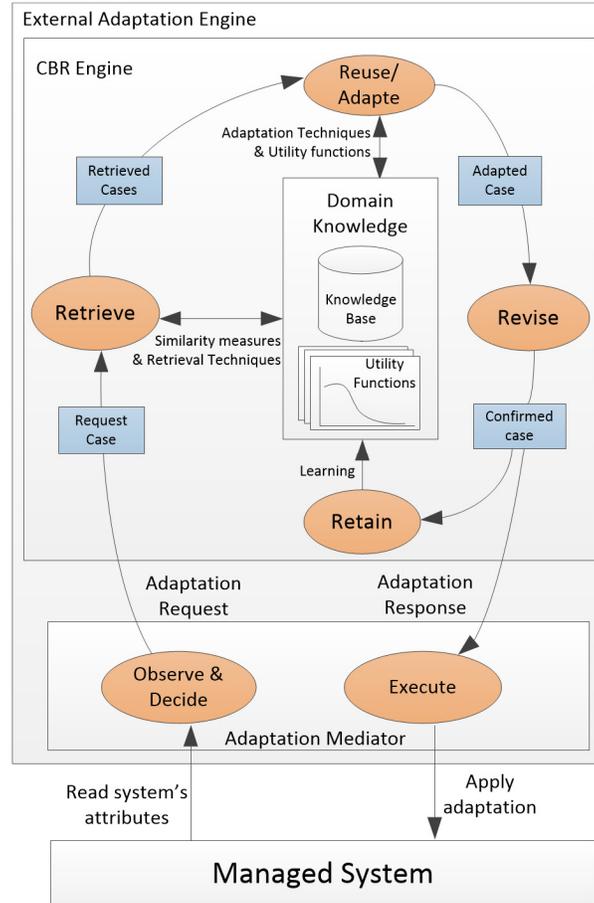


Figure 1: Reference model for the proposed solution

4. *Retain*: In this stage, the new generated cases are saved in the knowledge base. Case-Based Learning (CBL) have been introduced in [30] to provide algorithms and approaches for the retain process.

In our model, the case is a set of attributes that represents the attributes of the managed system. For example, if one attribute of the managed system causes a *UT* break, then the adaptation engine should alter the value of this attribute in order to provide the required utility. In our solution we incorporate the utility functions for capturing the requirements of the managed system. Also, utility function is used to judge the quality of the cases stored in the KB and generated by the adaptation engine.

Algorithm 1 abstracts the adaptation process of our solution, where β is a predefined level of the accepted similarity and QAF is the qualified adaptation frame, a set of cases that have the potential to be used directly as an adaptation response or as basis for adaptation. *Case Expediency* is a measure for the usefulness of a case, and this measure uses the similarity of the case beside its utility.

Algorithm 1 Adaptation process

Require: KB, A_{req}
Ensure: $Utility(A_{res}) > UT$

- 1: $List\ cases \leftarrow Retrieve(KB, A_{req})$
- 2: $List\ QAF$
- 3: $Case\ A_{res}$
- 4: **while** $Case\ c \leftarrow Iterate(cases)$ **do**
- 5: **if** $Sim(A_{req}, c) \in [\beta, 1]$ **then**
- 6: $QAF.add(c)$
- 7: **end if**
- 8: **end while**
- 9: **if** QAF is not *Empty* **then**
- 10: $A_{res} \leftarrow max(CaseExpediency(QAF))$
- 11: **Return** A_{res}
- 12: **else**
- 13: $A_{res} \leftarrow ConstructiveAdapt(A_{req})$
- 14: **Retain**(A_{res}, KB)
- 15: **Return** A_{res}
- 16: **end if**

Algorithm 1: Adaptation process algorithm

4.2. Uncertainty quantification

We follow the uncertainty quantification approach in [31], where uncertainty has three dimensions:

- The Location of uncertainty: Where the uncertainty manifests in the system.
- The Level of uncertainty: A variation between deterministic level and total ignorance. This means that uncertainty about one attribute of the system can take a value between one and zero.
- The Nature of uncertainty: Whether the cause of uncertainty is variability or lack of knowledge in the uncertainty meant attribute of the system.

Based on [32], uncertainty in self-adaptive software systems can be found in three places, namely: System requirement, system design and architecture, and run-time. In our solution, we focused on run-time uncertainty by quantifying it based on the aforesaid three dimensions.

5. PROGRESS AND CURRENT STATUS

A prototypical implementation of the solution has been done. This implementation includes the integration of the CBR engine with utility functions. The implementation also includes the generative adaptation of the adaptation responses. Moreover, uncertainty analysis and quantification are provided in this implementation paving the way for handling uncertainty during run-time. The three dimensions of the uncertainty [31] has been modelled and implemented.

6. FUTURE DIRECTION AND EVALUATION

For future direction, firstly, we will use a case study to empirically evaluate and validate our approach. The case study i.e. the managed system, should require the self-adaptivity mechanism that performs well under run-time uncertainty. Secondly, we will evaluate the results of the case study application. The evaluation will be based on software quality metrics and GQM [33]. We expect that the experimentation of our solution will provide a positive potential results for both handling the uncertainty and the complexity of adaptation space. However, we do not have a clue regarding the response time of the adaptation engine, the results will reveal this issue.

7. CONCLUSION

In this paper, we have presented our research for realizing self-adaptivity in software systems. We started by showing the gaps in the research and the expected contributions of the research. Also, we have presented details about the solution model and the used technology, Case-based reasoning. The progress of the work was presented along with the future directions. This paper ended with our vision of the evaluation process of the solution.

REFERENCES

- [1] B. H. Cheng and others, "Software Engineering for Self-Adaptive Systems," Springer-Verlag, 2009, pp. 1-26.
- [2] M. Salehie and L. Tahvildari, "A Quality-Driven Approach to Enable Decision-Making in Self-Adaptive Software," in Companion to the proceedings of the 29th International Conference on Software Engineering, 2007.
- [3] M. Salehie and L. Tahvildari, "Self-adaptive software: Landscape and research challenges," ACM Transactions on Autonomous and Adaptive Systems (TAAS) , 2009.
- [4] D. a. I. M. U. a. S. M. a. A. J. Weyns, "Claims and Supporting Evidence for Self-adaptive Systems : A Literature Study," in Software Engineering for Adaptive and Self-Managing Systems, SEAMS, 2012.
- [5] D. Kim and S. Park, "Reinforcement learning-based dynamic adaptation planning method for architecture-based self-managed software," in SEAMS '09. ICSE Workshop on, 2009.
- [6] A. Elkhodary, N. Esfahani and S. Malek, "FUSION: a framework for engineering self-tuning self-adaptive software systems," in Proceedings of the 18 ACM SIGSOFT international symposium, 2010.
- [7] M.-H. Karray, C. Ghedira and Z. Maamar, "Towards a Self-Healing Approach to Sustain Web Services Reliability," in AINA Workshops'11, 2011.
- [8] D. Garlan and others, "Rainbow: architecture-based self-adaptation with reusable infrastructure," Computer, pp. 46-54, 2004.
- [9] M. Derakhshanmanesh, M. Amoui, G. O'Grady, J. Ebert and L. Tahvildari, "GRAF: graph-based runtime adaptation framework," 2011.
- [10] T. Vogel and H. Giese, "Adaptation and abstract runtime models," in Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems, 2010.
- [11] B. Morin, O. Barais, J.-M. Jezequel, F. Fleurey and A. Solberg, "Models@ Run.time to Support Dynamic Adaptation," Computer, vol. 42, no. 10, pp. 44-51, #oct# 2009.
- [12] N. Khakpour, R. Khosravi, M. Sirjani and S. Jalili, "Formal analysis of policy-based self-adaptive systems," 2010.
- [13] R. Asadollahi, M. Salehie and L. Tahvildari, "StarMX: A framework for developing self-managing Java-based systems," in SEAMS, 2009.
- [14] S.-W. Cheng, D. Garlan and B. Schmerl, "Architecture-based self-adaptation in the presence of multiple objectives," in Proceedings of the 2006 international workshop on Self-adaptation and self-managing systems, 2006.
- [15] B. Bontchev, D. Vassileva, B. Chavkova and V. Mitev, "Architectural design of a software engine for adaptation control in the ADOPTA e-learning platform," 2009
- [16] G. Bertolotti, A. Cristiani, R. Lombardi, M. Ribarić, N. Tomašević and M. Stanojević, "Self-Adaptive Prototype for Seat Adaptation," in SASOW Fourth IEEE International Conference, 2010.

- [17] M. A. S. Sallem and F. J. da Silva e Silva, "Adapta: a framework for dynamic reconfiguration of distributed applications," in Proceedings (ARM '06), 2006.
- [18] Q. Yang and others, "Toward a fuzzy control-based approach to design of self-adaptive software," in Proce. of the 2nd Asia-Pacific Symposium on Internetware, 2010.
- [19] N. Esfahani, E. Kourosfar and S. Malek, "Taming uncertainty in self-adaptive software," in Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering, 2011.
- [20] N. C. Narendra and U. Bellur, "A middleware for adaptive service orientation in pervasive computing environments," 2010.
- [21] V. Cardellini, E. Casalicchio, V. Grassi, F. Lo Presti and R. Mirandola, "Qos-driven runtime adaptation of service oriented architectures," 2009.
- [22] R. Calinescu, L. Grunske, M. Kwiatkowska, R. Mirandola and G. Tamburrelli, "Dynamic QoS Management and Optimization in Service-Based Systems," IEEE Trans. Softw. Eng., vol. 37, no. 3, pp. 387-409, #may# 2011.
- [23] Y. Wu, Y. Wu, X. Peng and W. Zhao, "Implementing Self-Adaptive Software Architecture by Reflective Component Model and Dynamic AOP: A Case Study," in QSIC'10, 2010.
- [24] H. Liu, M. Parashar and S. Member, "Accord: A Programming Framework for Autonomic Applications," IEEE Transactions on Systems, Man and Cybernetics, Special Issue on Engineering Autonomic Systems, vol. 36, pp. 341-352, 2006.
- [25] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," AI COMMUNICATIONS, vol. 7, no. 1, pp. 39-59, 1994.
- [26] IBM, "An architectural blueprint for autonomic computing," IBM Corporation, 2005.
- [27] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 2 ed., Pearson Education, 2003.
- [28] W. Wilke and R. Bergmann, "Techniques and Knowledge Used for Adaptation During Case-Based Problem Solving," 1998.
- [29] E. Plaza and J. L. Arcos, "Constructive Adaptation," in Proceedings of the 6th European Conference on Advances in Case-Based Reasoning, 2002.
- [30] D. W. Aha, "Case-Based Learning Algorithms," 1991.
- [31] W. Walker and others, "Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support," Integrated Assessment, 2003.
- [32] A. a. J. A. a. C. B. H. C. Ramirez, "A taxonomy of uncertainty for dynamically adaptive systems," Software Engineering for Adaptive and Self-Managing Systems (SEAMS), 2012 ICSE, pp. 99-108, 2012.
- [33] R. van Solingen, V. Basili, G. Caldiera and H. D. Rombach, "Goal Question Metric (GQM) Approach," in Encyclopedia of Software Engineering, 2002.

AUTHOR

Mohammed Abufouda received the BSc. in Computer Engineering from Islamic University, Palestine in 2006 and the MSc. degree in computer science from Technical University of Kaiserslautern, Germany, in 2013. He is a PhD candidate at Technical University of Kaiserslautern in computer science department. His research interests includes software engineering and complex system analysis.

INTENTIONAL BLANK

EVALUATION OF THE SOFTWARE ARCHITECTURE STYLES FROM MAINTAINABILITY VIEWPOINT

Gholamreza ShahMohammadi¹

¹Department of Information Thechnology,
Olum Entezami University-Amin, Tehran, Iran
shah_mohammadi@yahoo.co.uk

ABSTRACT

In the process of software architecture design, different decisions are made that have system-wide impact. An important decision of design stage is the selection of a suitable software architecture style. Lack of investigation on the quantitative impact of architecture styles on software quality attributes is the main problem in using such styles. So, the use of architecture styles in designing is based on the intuition of software developers. The aim of this research is to quantify the impacts of architecture styles on software maintainability. In this study, architecture styles are evaluated based on coupling, complexity and cohesion metrics and ranked by analytic hierarchy process from maintainability viewpoint. The main contribution of this paper is quantification and ranking of software architecture styles from the perspective of maintainability quality attribute at stage of architectural style selection.

KEYWORDS

Maintainability Evaluation, Software Architecture, Architecture Style, Coupling, Complexity, Cohesion

1. INTRODUCTION

Functionality may be achieved using a number of possible structures [1], so software architecture styles (SASs) are selected based on the amount of their support from quality attributes. SASs present models to solve the problem of designing the software architecture in a way that each model describes its components, responsibilities of the components and the way they cooperate [2]. Architectural decisions made early in the design process are a critical factor in the successful development of the system. In particular, the selection of an appropriate architectural style has a significant impact on various system quality attributes [3]. Since quantitative impacts of SASs on quality attributes have not been studied so far [4], their applications are not systematic [5]. In other words, the current use of SASs in design is ad-hoc and based on the intuition of software developers.

A method has been shown in [6], to map an architectural style into a relational model that can be checked for various style properties such as consistency of style. In [7], two graph-based approaches have been shown and compared to the specification and modeling of dynamic software architectures. The impact of a distributed software system's architectural style on the

system's energy consumption has been estimated in [3]. A method for specifying the relation between six SASs and quality attributes such as maintainability has been proposed in [8]. The relationship between the quality attributes, design principles and some SASs has been specified in [8] using a tree-based framework. In [4], the impacts of SASs on quality attributes are determined based on the description of style in [2]. The methods offered in [4] and [8] are not able to determine the amount of style support from quality attributes, do not offer quantitative results about their maintainability, and are not precise. SASs are evaluated in [9] from maintainability viewpoint based on the scenario-based evaluation method that is less precise, less reliable and less analyzable as compared to the measurement-based evaluation method utilized in this paper.

In [10], the performance of three SASs has been investigated through simulation-based evaluation method. Implicit/invocation style has been verified in [11], by model checking method.

In this study, the quantitative impact of SASs on software maintainability, one of the important quality attributes required by all software, is determined based on the measurement-based evaluation of SASs. The SASs evaluated include Repository (PRS), Blackboard (BKB), Pipe and Filter (P/F), Layered (LYD), Implicit/Invocation (I/I), Client/Server(C/S), Broker (BRK) and Object-Oriented (OO), which have been introduced in [2], [12].

Software architecture evaluation methods include: 1) scenario-based evaluation, 2) simulation-based evaluation, 3) measurement-based evaluation and 4) mathematic model-based evaluation. Measurement-based evaluation method uses metrics to measure software architecture. Metrics evaluate internal attributes of software (e.g. coupling). External attributes (e.g. maintainability) reflect those properties that are desirable for the software user and usually are evaluated by internal attributes. It is believed that there is a relationship between internal and external quality attributes. This relationship is based on theoretical models and empirical study [13], [14]. There is a general agreement in software community that modularity has an influence on external attributes such as maintainability [15]. Therefore, in this paper, we use coupling, complexity and cohesion metrics to quantify the impact of SASs on software maintainability. These metrics are essential in evaluation of software design quality and their effects on maintainability have been extensively investigated [15]-[18].

The advantage of measurement-based evaluation as compared to scenario-based evaluation is that the evaluation would be easier and more precise, if there are appropriate metrics. In addition, it does not have the problems of scenario-based evaluation, namely the dependency of the results on the scenarios used, and extensive participation of the expert. As a result, the problem is evaluated more comprehensively.

Multi-criteria decision-making methods are used in the ranking problem of SASs. These methods are in three categories: 1) scoring, 2) compromise and 3) concordance [19]. Analytic hierarchy process (AHP) [20] is one of the most comprehensive multi-criteria decision making methods. It structures the problem as a hierarchy and provides a means of decomposing the problem into a hierarchy of sub problems that can more easily be comprehended and subjectively evaluated. AHP reflects the way people think and behave. It also considers different quantitative and qualitative criteria in the problem and provides sensitivity analysis on the criteria and sub-criteria. The AHP has been proven a theoretically sound, market-tested and accepted methodology. In this paper, to rank SASs based on the results of measurement-based evaluation of SASs, AHP method is used.

This paper is structured as follows: Section 2 discusses software maintainability and its measurement metrics. Section 3 explains the quantitative measurement of SASs. Section 4 deals with the ranking of SASs. Finally, Section 5 presents the conclusion.

2. SOFTWARE MAINTAINABILITY AND ITS MEASUREMENT METRICS

The main objective of any software is to offer desired services according to the predetermined quality level. There is a strong connection between many quality attributes and the software architecture of the software system. The architecture defines the overall potential that a software system possesses to fulfil certain quality attributes. Software are often redesigned not for the deficiency in the functionality, but due to difficulty in maintenance, port or scale [21].

Maintainability is the capability of the software product to be modified [22]. Modifications may include corrections, improvements or adaptations of the software to changes in the environment and in the requirements and functional specifications. The ease of software correction is determined through: 1) analyzability, 2) changeability, (3) stability and (4) testability [22].

A close look at software maintainability attributes reveals that provision of each characteristic depends on the amount of modularity of software design, design with low coupling among modules, low complexity of the modules and high cohesion of modules. Therefore, the less is the amount of coupling and complexity of the components and the more their cohesion, the easier will be the analyzability, changeability, stability and testability of the software. Various researches emphasize the impact of complexity, cohesion of components and coupling among components metrics on software maintainability [15]-[18].

2.1. Coupling Metric

High interaction of modules makes the understanding and modification of the modules more difficult [15]. The more independent the components, the easier their understanding, modification and maintainability [16]. Coupling is a complex concept that has been categorized by Yourdon and Constantine [23] as: 1) Data coupling, 2) Stamp coupling, 3) Control coupling, 4) Shared coupling and 5) Content coupling.

In this work, we generalize the “coupling among modules” concept to the coupling among software architecture components and use it to measure the amount of coupling of SASs. Components of SASs investigated in this work have three coupling types: data, stamp and shared quantified based on Table 1. In [24] also consecutive numeric values from 1 to 5 were used and the basis of such assignment was the experience from some software systems

Table 1.Type of Components Coupling

Row	Coupling type	Symbol	Weight
1	Data	w_1	1
2	Stamp	w_2	2
3	Common	w_3	4

designs. Regarding the coupling metric, SASs are investigated in terms of the type of coupling among the components and the number of components involved in the coupling. The more the strength of coupling among components and the more the number of components involved in the coupling, the less the understandability, correction and maintainability of the components [15]. To measure the coupling value of any SAS, (1) is used that is Euclidean norm, where n is the number of style components, SCP is the amount of SAS coupling and CCP_i is the amount of coupling of the i -th component. CCP_i is computed by (2), where NCT_j is the number of type j couplings, w_j is the weight of the corresponding coupling type and p is the number of coupling of the component i :

$$SCP = \sqrt{\sum_{i=1}^n CCP_i^2} \quad (1)$$

$$CCP_i = \sum_{j=1}^p NCT_j \cdot w_j \quad (2)$$

2.2. Complexity Metric

Complexity value of SASs is computed by (3) where, SCM is the complexity of SAS, n is the number of style components and CCM_i is the amount of complexity of the i-th component. CCM_i is computed by (4), using the module evaluation metric of Shepperd et al [25], where f_{in}(i) is the fan-in of component i and f_{out}(i) is the fan-out of component i.

$$SCM = \sqrt{\sum_{i=1}^n CCM_i^2} \quad (3)$$

$$CCM_i = [f_{in}(i) * f_{out}(i)]^2 \quad (4)$$

f_{in}(i) and f_{out}(i) are computed by (5) and (6). In (5), Nc_i is the sum of the number of invocations of component i by other components and Nr_i is the number of data that component i has retrieved from the repository. In (6), Nce_i is the number of other components called by component i and Nu_i is the number of the repository data updated by component i. A component that controls a lot of components usually performs various functions and so it will have a high complexity [15],[26].

$$f_{in}(i) = Nc_i + Nr_i \quad (5)$$

$$f_{out}(i) = Nce_i + Nu_i \quad (6)$$

2.3. Cohesion Metric

The cohesion of a module is the extent to which its individual components are needed to perform the same task. Types of cohesion are: 1) Coincidental, 2) Logical, 3) Temporal, 4) Procedural, 5) Communicational, 6) Sequential and 7) Functional [23]. The cohesion type of every component is computed based on the available information of functionality of each component of SASs regarding the definition of the type of component cohesion in Table 2 [26].

In this work, we generalize the “modules cohesion” concept to the cohesion of software architecture components and use it to measure the amount of cohesion of SASs. Our investigations showed that cohesion of SASs in component are of three types: Functional, Communicational and Logical, which are quantified based on Table 2.

Table 2. Type of Components Cohesion [26]

Cohesion type	Description	Symbol	Weight
Logical	Component performs multiple functions, and in each calling, one of them is executed	C ₁	1
Communicational	Component refers to the same data set and/or creates the same data set	C ₂	2
Functional	Component performs a single well-defined function	C ₃	3

Since every component i may have different type of cohesion (C_j), so the cohesion type of component i , CCH_i , is computed by (7). Finally, cohesion of SASs is computed by (8).

$$CCH_i = \arg \min_j C_j \quad (7)$$

$$SCH = \sqrt{\sum_{i=1}^n CCH_i^2} \quad (8)$$

3. Quantitative Measurement of SASs

In this section, SASs are measured from the viewpoint of maintainability based on coupling, complexity and cohesion metrics.

The effect of software size on SASs ranking is taken into account in the computations of this section. In object-oriented style, the number of objects (no) and in other SASs, the number of components (n) correspond the software size. So in the evaluations done in this section, the number of SASs components is considered as 3, 4, 5, 6, 7, 8 and 9 and the number of classes in object-oriented style is considered accordingly as 21, 28, 35, 42, 49, 56 and 63.

3.1. Measuring the Coupling of SASs

In this section, the coupling formula of every SAS is computed using (1) to (2).

A. Repository style. In this style, all components have common coupling with the repository. Therefore, any change in the repository affects them. If the number of components in the repository style is n , then the number of couplings in this style will be n as well. Thus coupling of repository style is obtained from $\sqrt{n} \cdot W_3$.

B. Blackboard style. The control component has a common coupling with the blackboard and has data coupling with the knowledge resources. Therefore, the control component has one common coupling and n data coupling while the knowledge resources have a common coupling with the blackboard. Thus, the coupling of the control component is $n \cdot w_1 + w_3$ and the coupling of each knowledge resource is w_3 . Then coupling of this style is obtained from $\sqrt{(n \cdot w_1 + w_3)^2 + n \cdot w_3^2}$.

C. Pipe and filter style Every filter (component) has a stamp coupling with the next filter while the last filter has no coupling with any other filter. The number of couplings is $n-1$, and regarding the coupling type, the coupling of this style is obtained from $\sqrt{n-1} W_2$

D. Layered style The coupling type of every layer (component) with its lower layer is data. Considering the fact that coupling is two way, the last and first layers have only one coupling while other layers have two couplings. So for n layer, the coupling of this style is obtained from $\sqrt{4(n-2) \cdot W_1^2 + 2 \cdot W_1^2}$.

E. Implicit invocation style. If, in average, $n/2$ of components publish the events that are favored by $n/2$ of the components, the coupling type of the event publisher component with the dispatcher component is data. If an event occurs, the dispatcher component invokes the interested components, so the coupling type of the dispatcher component with the interested components is

data, and the coupling of the dispatcher component will be $(n/2).w_1$. The coupling type of independent components $(n/2)$ is data, so coupling of this style is obtained from $\sqrt{((n/2).W_1)^2 + (n/2).W_1^2}$.

F. Client/server style. The coupling of the client with the server is data type. Supposing that, in average, the coupling of each server component is f and, since some server components are in transaction and usually the last component is related to repository, thus about r % of the components have just one connection with the repository. So, coupling of this style is obtained from $\sqrt{w_1^2 + (1-r)n(f.w_1)^2 + rnw_3^2}$.

G. Broker style. Coupling of all components is data type. Considering these facts: (1) the client component is related to the client side proxy, (2) the client is related to the broker in order to be informed of different services of the server, (3) the server side proxy is coupled with the broker, (4) the broker is coupled with the server side proxy and (5) the broker is coupled with the server for being informed of different type of services of the server, and also considering the similarity of the coupling of the server components to client/server style, the style coupling is obtained from $\sqrt{8w_1^2 + (1-r)n(f.w_1)^2 + r.n.w_3^2}$.

H. Object oriented style. In this style, the type of coupling is data. A case study done by Yu and Ramaswamy [28] on components dependency showed that 83% of the couplings between classes are of parameter (data) type. Coupling of each class with other classes is considered as f_o . So style coupling is obtained from $f_o \sqrt{n_o} \cdot W_1$.

Column 2 of Table 3 shows the coupling formulas of SASs. The third column shows the coupling value obtained by replacing the weight of coupling type based on Table 1.

Table 3. Coupling Formulas of SASs

Symbol	Coupling Formula	Coupling Value
RPS	$\sqrt{n}.W_3$	$3\sqrt{n}$
BKB	$\sqrt{(n.w_1 + w_3)^2 + n.w_3^2}$	$\sqrt{(n+3)^2 + 9n}$
P/F	$\sqrt{n-1} W_2$	$2\sqrt{n-1}$
LYD	$\sqrt{4(n-2).W_1^2 + 2.W_1^2}$	$\sqrt{4n-6}$
II	$\sqrt{((n/2).W_1)^2 + (n/2).W_1^2}$	$\sqrt{(n/2)^2 + (n/2)}$
C/S	$\sqrt{w_1^2 + (1-r)n(f.w_1)^2 + rnw_3^2}$	$\sqrt{1+(1-r)n.f^2 + 9.r.n}$
BRK	$\sqrt{8w_1^2 + (1-r)n(f.w_1)^2 + r.n.w_3^2}$	$\sqrt{8+(1-r)n.f^2 + 9.r.n}$
OO	$f_o \sqrt{n_o} \cdot W_1$	$f_o \sqrt{n_o}$

The coupling value of classes in object-oriented style (f_o) is related to the designing manner of the past software systems. This is true for the coupling value of server components (f) in the broker and client/server styles as well. Therefore, software designers determine the average value of coupling (i.e. f and f_o) by referring to the previous software design records. For displaying the relationship between coupling value and software size, it is necessary that first the values of f , r and f_o parameters are determined. Thus, documents of software design projects of a large and

valid software company in Iran is investigated. Accordingly, after computations, the values of these parameters become $f=1.65$ and $f_o=1.5$ and $r=0.2$. By setting the parameters of f , f_o and r to the designated formulas and parameters n and n_o , the coupling value of SASs is computed considering the software size (number of components), and its diagram is shown in figure 1. According to this diagram, the coupling value of SASs is increased by increasing of the software size.

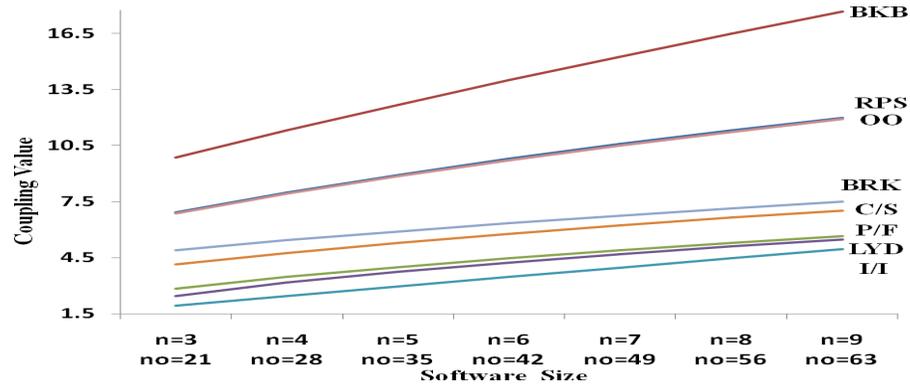


Figure 1. Coupling value of SASs based on the size of software

3.2. Measuring the Complexity of SASs

In this section, the complexity formula of every SASs is computed using (3) to (6).

A. Repository style. In this style, all components read from the data repository and modify it. Thus, both their fan-in and fan-out is equal to 1. Therefore, the total fan-out of each component, considering the writing in the repository and invoking the repository for this writing, is 3. Thus the complexity of independent components is 9 and the complexity of style is obtained from $9\sqrt{n}$.

B. Blackboard style. The fan-in of the control component is 1 (for examining the status of the blackboard) and its fan-out is 2 (for invoking the blackboard for reading its status and invoking the knowledge resources). The fan-in of knowledge resources is 2 (for invoking by the control component and reading from the blackboard) and its fan-out is 3 (for invocation of the blackboard for reading and writing into the blackboard). So the complexity of the control component is 2^2 , the complexity of each of the knowledge resource is 36, and complexity of style is obtained from $\sqrt{4^2 + 36^2 n}$.

C. Pipe and filter style. The first filter (component) has no input and the last filter does not have any output. Thus, their complexity is 0. The other filters have one input and one output. So the complexity of style is obtained from $\sqrt{n-2}$.

D. Layered style. In this style, the relation of lower layer to upper layer is response to the request of upper layer, so in computing of layer's fan-out, this relation is ignored, i.e. only upper layer invokes lower layer. Thus, each layer has fan-in and fan-out equal to 1. None of the layers does not invoke first layer and the last layer invokes no layer. So their complexity is 0 and the complexity of style is obtained from $\sqrt{n-2}$.

E. Implicit invocation style. With the occurrence of an event, the dispatcher component invokes the interested components. Therefore, the fan-in of dispatcher component is 1 (for occurrence of the event that led to the invoking of the interested component by the dispatcher component) and its fan-out is 1 (for invocation of the interested component, when an event occurs). Therefore, its complexity is 1. The complexity of event publisher due to lack of fan-in and the complexity of interested components due to lack of fan-out is 0. Therefore, the complexity of style is 1.

F. Client/server style. The client component invokes a procedure from the server, so fan-in of the server and fan-out of the client is equal to 1. Since the client is not invoked by the components and has no direct access to the repository, its fan-in is equal to 0 and its complexity is 0. The number of fan-ins and fan-outs of the server components, in average, is considered as f . So the complexity of each server component is f^4 and the complexity of style is $f^4 \sqrt{n}$.

G. Broker style. The client component gets informed of the services of the server through the method interface of the server that has been offered to the broker component, so both fan-in and fan-out of the server becomes 1. In addition, fan-in of the broker becomes 1 due to accessing the interface of the server services. The client invokes the client side proxy, thus its fan-out becomes 1 as well. The client side proxy sends a request to the broker component, therefore, both its fan-in and fan-out become 1. The broker component sends the request to the server side proxy. On the other hand, the broker invokes the server to get informed of the interface of the server services. Therefore, both fan-in and fan-out of the broker become 2. The server side proxy has the fan-in and fan-out equal to those of the client side proxy too. The complexity of server components is considered similar to that of the client/server style, thus style complexity is obtained from $\sqrt{274 + f^8 n}$.

H. Object-Oriented style. If, in average, the number of fan-in and fan-out of each class is considered as f_o , then the complexity of each class becomes f_o^4 and the complexity of style is $f_o^4 \sqrt{n_o}$.

Table 4 shows the complexity formulas of SASs. Values of f and f_o are considered as similar to those in the Section 3.1.

Table 4. Complexity Formulas of SASs

Symbol	Complexity Formula
RPS	$9\sqrt{n}$
BKB	$\sqrt{4^2 + 36n}$
P/F	$\sqrt{n-2}$
LYD	$\sqrt{n-2}$
I/I	1
C/S	$f^4 \sqrt{n}$
BRK	$\sqrt{274 + f^8 n}$
OO	$f_o^4 \sqrt{n_o}$

By setting the parameters n and n_o , the complexity value of SASs is computed considering the software size and its diagram is shown in figure 2. According to this diagram, the complexity value of most SASs is increased by increasing of the software size.

3.3. Measuring the Cohesion of SASs

In this section, the cohesion formula of every SAS is computed using (7) to (8).

A. Repository style. Each component processes the same set of data, so their cohesion type is communicational. The repository component performs various functions on the data, and in each calling, one of the functions is performed. So its cohesion type is logical, and the cohesion of style is $\sqrt{nC_2^2 + C_1^2}$.

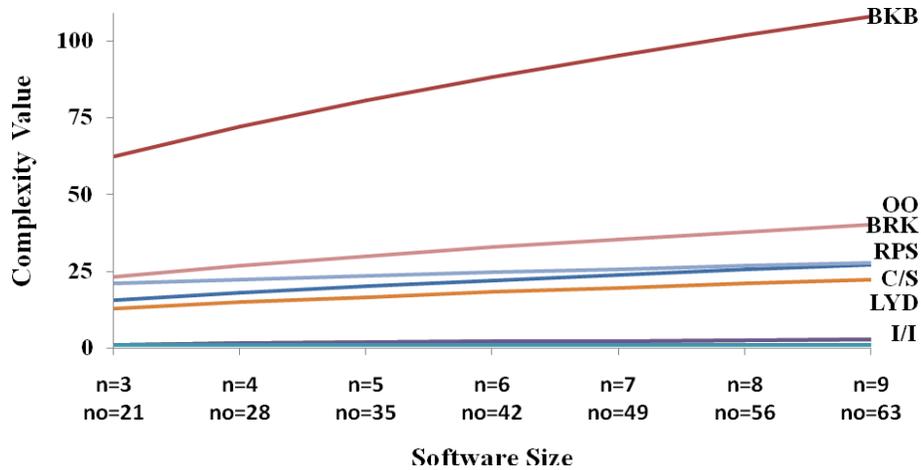


Figure 2. Complexity value of SASs based on size of software

B. Blackboard style. Each knowledge resource processes the same set of data, so their cohesion type is communicational. The control component invokes the knowledge resources based on the status of the blackboard. Therefore, its cohesion type is logical. The blackboard component performs various functions and, in each invocation, one of these functions is performed. So, its cohesion type is logical, and the cohesion of style is $\sqrt{nC_2^2 + 2C_1^2}$.

C. Pipe and filter style. Each filter processes the same set of data, so its cohesion type is communicational and the cohesion of style is $\sqrt{n} \cdot C_2$.

D. Layered style. Each layer contains some components; regarding the invoking of upper layer, one of components of the lower layer is performed, so the cohesion type of each layer is logical and the cohesion of style is $\sqrt{n} \cdot C_1$.

E. Implicit invocation style. Since the components are publisher or interested in the event, their cohesion type is communicational. The dispatcher component performs various functions and, in each invocation, one of them is performed. Thus, its cohesion type is logical and the cohesion of style is $\sqrt{C_1^2 + nC_2^2}$.

F. Client/Server style. The server provides various services for the client by its components, and in each invocation, one or some of the server components are performed so that each one works on the same data. Accordingly, their cohesion type is communicational. The client component performs a specific function, so its cohesion type is functional. The repository component

performs various functions and in each calling, one of them is performed. So its cohesion type is logical and the cohesion of style is $\sqrt{C_1^2 + nC_2^2 + C_3^2}$.

G. Broker style. The client side proxy, server side proxy, broker and server components perform various functions and in each invocation, just one of the functions is performed, so their cohesion type is logical. The client component performs a specific function, thus its cohesion type is functional. The repository component performs various functions and in each calling, one of them is performed. Therefore, its cohesion type is logical. Cohesion of the server components is considered similar to that of the client/server style. Thus, the cohesion of style is $\sqrt{4C_1^2 + nC_2^2 + C_3^2}$.

H. Object-Oriented style. The classes in this style define the data of an entity and its related functions, so, the cohesion type of each class is communicational and the cohesion of style is $\sqrt{n_o} \cdot C_2$.

Column 2 of Table 5 represents the cohesion formulas of SASs. The third column shows the cohesion value obtained by replacing the weight of cohesion type based on Table 2. By setting the parameters n and no, the cohesion value of SASs is computed considering the software size and its diagram is shown in figure 3. According to this diagram, the cohesion value of SASs is increased by increasing of the software size and the amount of increase is higher in the object-oriented style relative to the other styles.

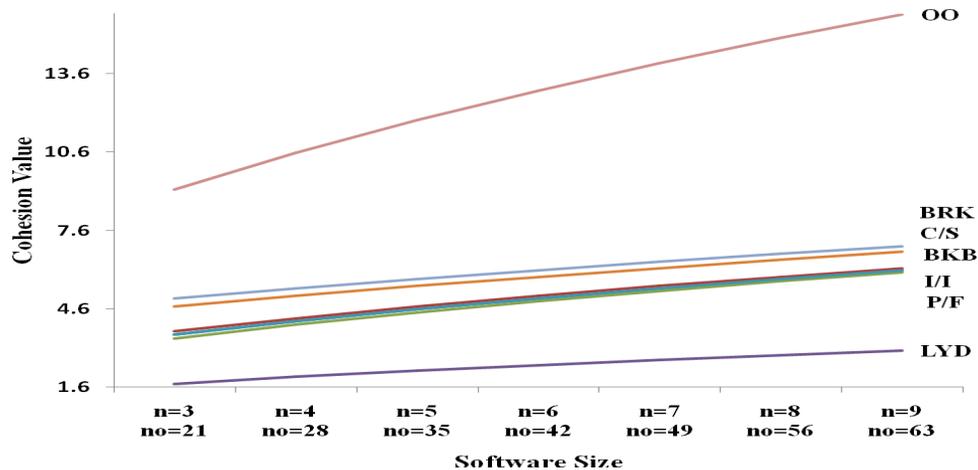


Figure 3. Cohesion value of SASs based on the software size

Table 5. Cohesion Formulas of SASs

Symbol	Cohesion Formula	Cohesion Value
RPS	$\sqrt{nC_2^2 + C_1^2}$	$\sqrt{4n + 1}$
BKB	$\sqrt{nC_2^2 + 2C_1^2}$	$\sqrt{4n + 2}$
P/F	$\sqrt{n} \cdot C_2$	$2\sqrt{n}$
LYD	$\sqrt{n} \cdot C_1$	\sqrt{n}
I/I	$\sqrt{C_1^2 + nC_2^2}$	$\sqrt{1 + 4n}$
C/S	$\sqrt{C_1^2 + nC_2^2 + C_3^2}$	$\sqrt{10 + 4n}$
BRK	$\sqrt{4C_1^2 + nC_2^2 + C_3^2}$	$\sqrt{13 + 4n}$
OO	$\sqrt{n_o} \cdot C_2$	$2\sqrt{n_o}$

4. COMPUTATION OF THE RANK OF SASs

In this section, the ranking of SASs is performed based on the results of measurement coupling, complexity and cohesion of SASs using AHP method.

4.1. Organizing Ranking Problem of SASs

In SASs ranking problem, aim is in the first level, metrics are in the second level and SASs are in the third levels of the structure.

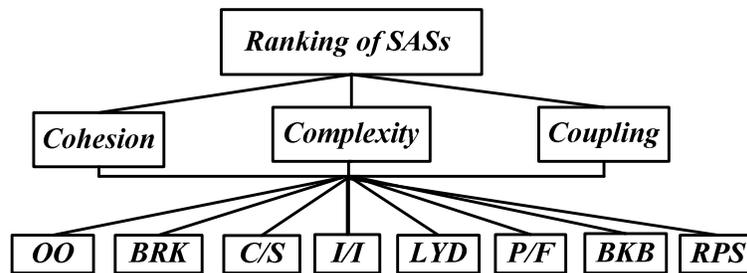


Figure 4. Hierarchical structure of SASs ranking

4.2. Computation of Priority of Metrics and the Relative Rank of SASs

In this stage, comparison matrix of the metrics and comparison matrices of SASs for the metrics are formed. The complexity and cohesion values of a component do not affect on the other components of SAS. However, the coupling value of a component affects the related components. Accordingly and due to the emphasis of researches on the importance of coupling [13], [15] the preference of coupling metric is considered more important than (1.6) the other metrics, and the preferences of other metrics are considered equal. Then the relative priority of metrics is computed by AHP method, the relative priority of coupling becomes 0.444 and that of the other metrics become 0.278.

To determine the relative rank of SASs for each metric, comparison matrices of SASs for each metric is formed. To set cell (i, j) of the comparison matrix of metric k, for the style x in row i with the style y in column j, if there is a direct relation between the metric k and maintainability,

the ratio of the metric value of style x to the metric value of style y is set to cell (i,j) , otherwise the inverse of the ratio is set to cell (i,j) . After setting of the comparison matrices based on the described procedure, the relative rank of SASs for each metric is computed by AHP method. Investigation of the consistency using the Expertchoice tool, tool of AHP method, showed that consistency index is zero, so there is no inconsistency between the comparisons.

4.3. Computing the Final Rank of SASs

The final rank of SASs is computed regarding the priority of metrics and the relative ranks of SASs. Table 6 shows the final rank of SASs. Based on the values of this Table, the Implicit/Invocation (I/I), Pipe and Filter (P/F), and Layered (LYD) styles provide the highest support for maintainability, respectively.

Figure 5 shows the changes in maintainability value of SASs based on the changes of software size. With the increasing of software size, the rank of some styles such as Pipe and Filter (P/F) and Layered (LYD) are decreased, and the rank of some styles such as Implicit Invocation (I/I) are increased while the rank of some styles such as Blackboard (BKB) are not changed considerably.

Table 6. Rank of SASs from the maintainability viewpoint

Symbol	n=3	n=4	n=5	n=6	n=7	n=8	n=9
	no=21	no=28	no=35	no=42	no=49	no=56	no=63
RPS	64	67	69	69	70	70	70
BKB	52	54	54	55	55	55	55
P/F	187	176	170	166	163	160	158
LYD	185	169	161	155	151	149	146
I/I	223	238	246	251	255	257	260
C/S	95	97	97	98	99	99	99
BRK	87	89	91	92	94	94	95
OO	107	110	112	114	115	116	116

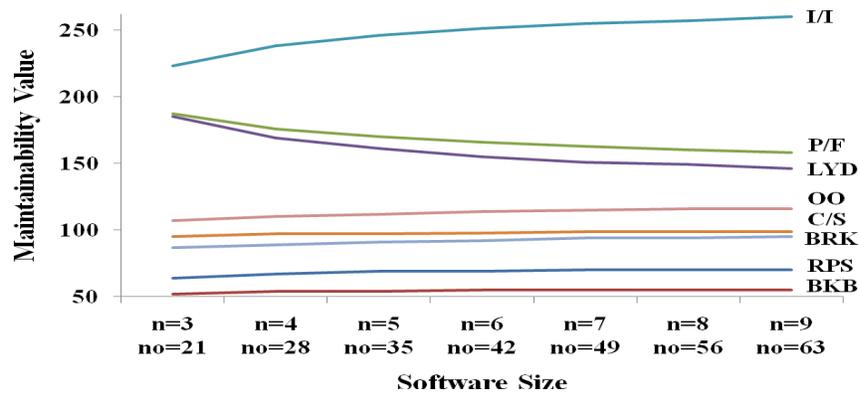


Figure 5. Maintainability value of SASs based on the changes of software size

Figure 6 shows the diagram of styles ranks based on the relative priority of metrics. It is known as sensitivity analysis diagram, which is drawn by Expertchoice. In this diagram, the vertical lines show the relative priority of metrics and the horizontal lines show the rank of SASs based on the metrics. The final rank of SASs is determined by the "OVERALL" label based on the vertical line (figure 6). The coupling metric accords with the y-axes and after that are complexity, cohesion and combination of the three metrics.

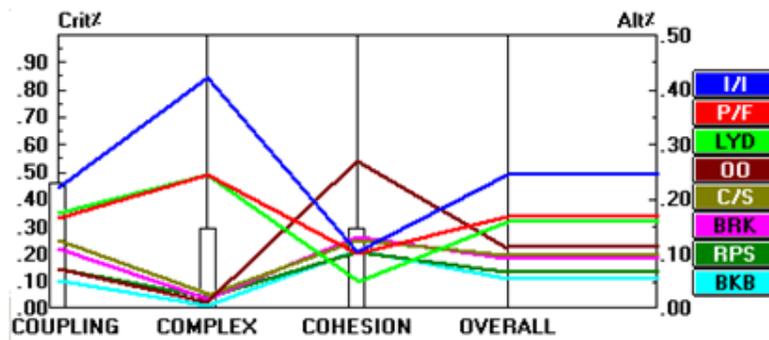


Figure 6. Diagram of styles rank regarding the relative priority of metrics

4.4. Analyzing the Rank of SAS

Here, by changing the values of some parameters, the effects of these changes on the rank of SASs are investigated.

- For the values of coupling types (Section 2.A), other values were used besides the values mentioned in table 1 (for twelve values in the ranges $1 \leq w1 \leq 1.5$, $1.5 \leq w2 \leq 2.5$ and $2.5 \leq w3 \leq 3.5$), but they did not lead to any changes in the rank position of the SASs' maintainability.
- For the values of cohesion types (Section 2.C), other values were used besides the values mentioned in table 2 (for twelve values in the ranges $1 \leq c1 \leq 1.5$, $1.5 \leq c2 \leq 2.5$ and $3 \leq c3 \leq 3.5$), but they did not lead to any changes in the rank position of the SASs' maintainability.
- By changing the f parameter (coupling of the server components in Section 3.A) in the range of $1.65 \leq f \leq 2.8$ at the Client/Server (C/S) style, the change in the rank position of this style was checked. It was found that only for $f \geq 2$, the rank position of this style is placed after the Broker (BRK) style and no other change in the rank position of other styles was seen.
- For determining the relative priority of metrics (In Section 4.B), in addition to 1.6 (the relative priority of coupling metric compared to that of the other metric), the ten values in the range of 1.3 to 2.2 were used. The results showed no changes in the rank position of the styles from maintainability viewpoint.

5. CONCLUSION

In this study, a model was offered to analyze the impact of SASs on software maintainability according to the measurement-based evaluation of SASs. In this model, first, the formulas were presented to compute the coupling, complexity and cohesion values of each SAS. Next, the coupling, complexity and cohesion values of SASs were computed quantitatively using the presented formulas. Then, the relative rank of each SAS was determined regarding the coupling, complexity and cohesion values of SASs. Afterward, the priority of metrics was determined. Subsequently, the final rank of SASs maintainability was determined using AHP method.

The analyses done showed that our proposed method had stability regarding the value of coupling types, different values of f parameter, value of cohesion types and preference of coupling metric to the other metrics.

Since the evaluation of this paper is based on measurement as compared to the method used in [9], which uses scenario-based evaluation and the quality of its results is dependent on the used scenarios and also on the extensive expert participation, the results of our proposed model is more precise, more reliable and more analyzable.

The proposed method gives formulas to determine the values of 1) coupling, 2) complexity and 3) cohesion of each SAS, while this has not been done in previous methods.

As compared to [4], [8], both the proposed method and the method used in [9] give the quantitative results about the maintainability of SASs that is basis of the systematic recommendation and selection of SAS.

Finally, only the proposed method examines the effect of the software size on the maintainability rank of SASs.

The methods given [6], [7], [11] use the mathematic model-based evaluation and the method used in [10] uses the simulation-based evaluation. These methods verify specific features such as consistency and satisfaction of some properties by SASs that are different from the quality attributes required in this paper. The above points and table 8 clearly show the position of the proposed method as compared to the methods of [4], [8] and [9].

It is worth noting that the ranking of SASs based on our proposed method is consistent with the priorities of SASs from the viewpoint of maintainability in the methods used in [2], [12], which are based on experimental studies.

Table 7. Comparison of the proposed method with the related methods

Criteria \ Method	Proposed Method	Method [4]	Method [8]	Method [9]
Base	Measurement	Tree	Unsystematic	Scenario
Offering the Quantitative Results about the Maintainability of SASs	•			•
Total SASs that were Investigated	8		6	8
Considering the Effect of Software Size on the Rank of SASs	•			

REFERENCES

- [1] Len Bass, Paul Clements. & Rick Kazman(2003) Software Architecture in Practice (2nd Edition), Addison-Wesley, p 89.
- [2] F. Buschmann, R. Meunier, H. Rohnert, P. Sornmerlad, & M. Stal,(1996) Pattern-Oriented Software Architecture- A system of Patterns" John Wiley & Sons, p. 394.
- [3] C. Seo, G. Edwards, S. Malek, & N. Medvidovic,(2009) "A Framework for Estimating the Impact of a Distributed Software System's Architectural Style on Its Energy Consumption", 7th Working IEEE/IFIP Conf. on Software Architecture, pp. 277-280.
- [4] B. Harrison, & P. Avgeriou,(2007) "Leveraging Architecture Patterns to Satisfy Quality Attributes", 1st European Conf. on Software Architecture, Springer, pp. 263-270.
- [5] P. Avgeriou P, & U. Zdun, (2005) "Architectural Patterns Revisited: A Pattern Language", Proc. of 10th European Conf. on Pattern Languages of Programs, pp.1-39.
- [6] J.S Kim, and D. Garlan, (2006) "Analyzing Architectural Styles with alloy", Proc. of the ISSTA 2006 workshop on Role of Software Architecture for Testing and Analysis, pp. 70-80.
- [7] R. Bruni, A. Bucchiarone, A. Gnesi, D. Hirsch, & A.L. Lafuente, (2008) "Graph-based Design and Analysis of Dynamic Software Architectures", LNCS 5065, pp. 37-56,.

- [8] H. Reza, & E. Grant, (2005) "Quality-Oriented Software Architecture", the IEEE Int. Conf on Information Technology, pp. 140 – 145.
- [9] Gholamreza Shahmohammadi, & Saeed Jalili, (2009) "Scenario-Based Quantitative Evaluation of Software Architecture Style from Maintainability Viewpoint", 14th Annual of CSI Computer Conference (CSICC 2009), Iran, Amirkabir University.
- [10] H. Grahm, & J. Bosch, (1998) "Some Initial Performance Characteristics of Three Architectural Styles", Proc. of Int. Workshop on Software and Performance.
- [11] D. Garlan, & S. Khersonsky, (2000) "Model Checking Implicit Invocation Systems", 10th Int. Workshop On Software Specification and Design.
- [12] M. Shaw & D. Garlan, (1996) Software Architecture: Perspectives Discipline on an Emerging Discipline, Prentice Hall.
- [13] L. Briand, S. Morasca, & V. Basili, (1996) "Property Based Software Engineering Measurement", IEEE Trans on Software Eng., vol. 22, no. 1, pp. 68-86.
- [14] L. Briand, J. Wust, & H. Lounis, (1999) "Using Coupling Measurement for Impact Analysis in Object-Oriented Systems", IEEE Int. Conf. on Software Maintenance.
- [15] S.L. Pfleeger, & J.M. Atlee, (2006) "Software Engineering, Theory and Practice", 3rd Edition, Prentice Hall.
- [16] P. Yu, T. Systs, & H. Muller, (2002) "Predicting Fault Proneness using OO Metrics. An Industrial Case Study," 6th European Conf. on Software Maintenance and Reengineering, pp.99 – 107.
- [17] M. Alshayeb, and L. Wei, (2003) "An Empirical Validation of Object-Oriented Metrics in Two Different Iterative Software Processes," IEEE Trans on Software Engineering, vol. 29 (11), pp. 1043 – 1049.
- [18] F. Bachmann, L. Bass, M. Klein, M. & C. Shelton, (2005) "Designing Software Architectures to Achieve Quality Attribute Requirements", IEE Proc. of Software, Vol. 152, No 4, pp. 153- 165.
- [19] C.L. Hwang, K. Yoon, (1981) "Multiple Attribute-Decision Making", Springer-Verlag.
- [20] T. L. Saaty, & L. G. Vargas, (2001) "Models, Methods, Concepts & Applications of the Analytic Hierarchy Process", Kluwer Academic Publisher.
- [21] L. Bass, P. Clements, & R. Kazman, (1998) Software Architecture in Practice, Addison-Wesley, p. 17.
- [22] ISO, (2001), International Organization for Standardization, "ISO 9126-1:2001, Software Engineering – Product quality, Part 1: Quality model".
- [23] E. Yourdon, & L. Constantine, (1978) Structured Design, Englewood Cliff, NJ, prentice Hall.
- [24] N. Fenton, & A. Melton, (1990) "Deriving Structurally Based Software Measures", Journal of Systems and Software 12(3), pp. 177-187.
- [25] M. J. Shepperd, & D.C. Ince, (1990) "The use of metrics in the early detection of design errors", Proc.of the European Software Engineering Conf, pp.67-85.
- [26] NE. Fenton, & SL. Pfleeger, (1997) "Software Metrics: A Rigorous and Practical Approach", (2nd Edition), International Thomson Computer PRESS.
- [27] S. Chidamber, & C. Kemerer, (1994) "A Metrics Suite for Object Oriented Design", IEEE Trans on Software Engineering, vol. 20, pp. 476-493.
- [28] L. Yu, & S. Ramaswamy,(2007) "Component Dependency in Object-Oriented Software", Journal of Computer Science and Technology, 22(3), pp. 379-386.
- [29] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

AUTHORS

Gholamreza Shahmohammadi received his Ph.D. degree from Tarbiat Modares University (TMU, Tehran, Iran) in 2009 and his M.Sc. degree in Computer Engineering from TMU in 2001. Since 2010, he has been Assistant Professor at the Olum Entezami University-Amin(Tehran, Iran). His main research interests are software engineering, quantitative evaluation of software architecture, software metrics and software cost estimation.



E-mail: Shah_mohammadi@yahoo.co.uk

INTENTIONAL BLANK

A SIMILARITY MEASURE FOR CATEGORIZING THE DEVELOPERS PROFILE IN A SOFTWARE PROCESS

Hamid Khemissa¹, Mohamed Ahmed-nacer² and
Abdelkader Belkhir³

Computer Systems Laboratory, Computer Science Institute, USTHB
University,
EL ALIA BP N°32, BAB EZZOUAR ALGERIA. TEL/FAX (00)213 21247917
¹hkhemissa@usthb.dz, ²anacer@cerist.dz and
³kaderbelkhir@hotmail.com

ABSTRACT

Software development processes need to have an integrated environment that fulfills specific developer needs. In this context, this paper describes the modeling approach SAGM ((Similarity for Adaptive Guidance Model) that provides adaptive recursive guidance for software processes, and specifically tailored regarding the profile of developers. A profile is defined from a model of developers, through their roles, their qualifications, and through the relationships between the context of the current activity and the model of the activities. This approach presents a similarity measure that evaluates the similarities between the profiles created from the model of developers and those of the development team involved in the execution of a software process. This is to identify the profiles classification and to deduce the appropriate type of assistance (that can be corrective, constructive or specific) to developers.

KEYWORDS

Software process modeling, Process engineering, Adaptive guidance profile, Similarity measure.

1. INTRODUCTION

Improving quality and productivity of software development requires assisting developers at both methodology level and consistency results level [1]. A guidance model in software engineering should combine the needed features to build the support system [2, 3].

Several PSEEs (**Process-Centered Software Engineering Environments**) [2, 4] deal the assistance aspect in the support of the software product development. Some PSEEs use an assistance description structured in steps like prescribing systems or proactive systems to control the operations carried out by the developer. The main limitations of these PSEEs are:

- The human actor has a central role in the progress of the development process regardless of his profile (qualifications and behavior).

- The basic guidance is defined as a global orientation core whatever the profiles of both the activity and the developer.
- The selection of the appropriate type of guidance is often more intuitive and not suitable.

To respond to these limits, several studies [2, 3, 5, 6] try to offer more flexibility in the language of software process modeling and a more adapted base of support and control. This tendency aims to define interventions of direct and adaptive assistance during the software process progress [7]. The following PSEEs included in the M1 level are as:

ADELE/APEL is based on a reactive database. It proposes a global assistance of proscriptive type and automates part of the development process using triggers [8, 9].

RHODES/PBOOL+ uses an explicit description of a development process. The software processes are modeled in PBOOL language [10]. The activities are associated to a guidance system with various scenarios of possible realization.

ADDD/ALADYN provides process automation and control the impact in a concrete system. The task hierarchy is used to organize the process descriptions, called policies. Several aspects are grouped and treated in a policy. A policy can be instantiated for several tasks. The instantiated triggers are rules of the form event-condition-action (ECA) and used to implement a reactive behavior [11].

On the M2 level of Meta model, SPEM [12] introduced the concept of "Guidance" in the "*Managed Content*" package by defining the stereotype "Guidance". According to SPEM, "the Guidance is a describable element which provides additional information to define the describable elements of a modeling. It also offers, through the stereotype "Guidance_kind" different types of guidance such as: Template, Guidelines, Checklists, etc. ..

However, the selection of guidance types remains defined in a manual and in an intuitive way. It depends on the experience and on the informal personality of the project manager. In addition, the proposed guidance is not adaptive to the actor's profile (role, qualifications and behavior). In considering the principal limitations of PSEEs and essential characteristics of our approach as the context adaptation aspect and the abstraction levels, a comparative table of the studied Meta models is as follows:

Table 1. A Comparative table of the studied Meta models

Meta model Criteria	ADELE /APEL	RHODES / PBOOL+	ADDD / ALADYN	SPEM
Global guidance core	Global	Global	Customized for each task	Global
Human performer profile oriented guidance	Not adapted	considered strategy Model	Not adapted	Not adapted
Context development Guidance	Not adapted	Adapted	Adapted	Not adapted
Guidance types	Not invoked	associated with a specific guide system	Not invoked	Intuitive selection

Explicit activity abstraction	Explicit abstraction	Implicit abstraction	Implicit abstraction	Explicit abstraction
Explicit task abstraction	Implicit abstraction	Not invoked	Explicit abstraction	Explicit abstraction
Process Modeling Language(PML)	APEL With predefined primitives	PBOOL+ With explicit primitives	ALADYN Not explicitly mentioned	UML Profile With explicit primitive

The current tendency is that developers would like to have integrated environments that are suitable to specific needs according to the role and the characteristics of each developer and closed to the context of the underway task. However, the provided efforts to develop such environments remain an insufficient contribution.

In this context, our conceptual model is based on the conventional reasoning of software processes enriched by the "Adaptive Guidance" element which supervises the running of the activities. It also provides adaptive support to the actor. It is described by the following figure:

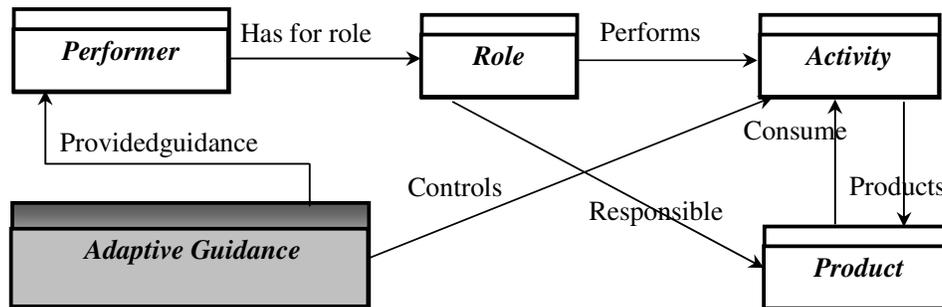


Figure 1. : Conceptual Model with Adaptive Guidance

This tendency of adaptive guidance environments is yet a subject of much research focusing on defining the concepts and objectives of process modeling software-based adaptive guidance [2, 3, 6].

For the sake of productivity and development time, our goal is to establish an optimal relationship between profile of guidance type and the adapted developer's profile to the context of development. The context is defined by the activity model, the developer and team development [2, 3, 13]. It is interesting to have an operator to assess the similarities within the handled data; this operator is the similarity function [13, 14, 15, 16]. The numerical similarity measures turn out to be extremely flexible employment. They are able to work on a broad spectrum of data types and it is fairly easy to introduce into the calculation, (statistical approximations if the underlying information is complex). In addition, similarities quantification by a continuous value implies that it is always possible and easy to compare pairs of objects. This is not the case for the symbolic similarity; treatment of numerical values is often done in an unsatisfactory way by rewriting these values in symbolic form [14, 16].

Our approach operates in the optimization of profiles classes in relation to the semantics of data manipulation. It defines a system for processing the similarity index and classification of guidance profiles. For this, we have to design a classifier in order to facilitate the analysis of our

population and the type of assistance offered to appropriate developers involved in a software process.

The second section summarizes the technique of assessing similarity. Section three presents our approach (to model similarity with the software process), in addition to the implementation and practical evaluation of our approach by giving algorithms and related results. The last section concludes and presents future works perspectives.

2. SIMILARITY MEASURE

A similarity measure is defined on the set N (developers, documents, websites ...). Each object is described by m features. Each feature can be present or absent in every object. A measure of similarity, denoted by s , (between the elements of N is a specific application of $N \times N$ in R and satisfying some properties [14, 15]).

Examples of the use of similarity techniques are described in cases of heterogeneous binary data [17]. To transform a direct measure of similarity s into a dissimilarity measure d , we can apply the following formula: $d(x, y) = \text{smax} - s(x, y)$.

Thus, each element x is associated to a binary vector (x_1, x_2, \dots, x_m) such that:

$$x_i = \begin{cases} 1 & \text{If the feature } i \text{ is present in the object } x \\ 0 & \text{Else} \end{cases} \quad \text{For } i \in \{1, 2, \dots, m\}.$$

The m characteristics are considered of equal importance and each object has at least one feature present. Note by:

- ▶ **a: The number of common characteristics between x and y .**
- ▶ **b: The number of features present in x but not y .**
- ▶ **c: The number of features present in y but not x .**
- ▶ **d: The number of missing features in x and y .**

Thus, the similarity measure s is given by the following formula:

$$\forall x, y \in N : s(x, y) = \frac{2a + b + c}{2(a + b + c)}$$

We can deduce the measure of dissimilarity from the following formula:

$$\forall x, y \in N : d(x, y) = 1 - \frac{2a + b + c}{2(a + b + c)} = \frac{b + c}{2(a + b + c)}$$

Besides this general form, there are also other forms of similarity measures such as binary data [14, 15, 17] as those given in table 2.

Table 2. Different forms of similarity measures

<i>Author</i>	<i>Definition</i>
Russel, Rao	$S = \frac{a}{m}$
Kendall, Sokal-Michener	$S = \frac{a + d}{m}$
Rogers, Tanimoto	$S = \frac{a + d}{m + b + c}$
Hamann	$S = \frac{a + d - b - c}{m}$
Sokal, Sneath	$S = \frac{b + c}{a + d}$

To review and evaluate the effect of binary similarity measures, we will illustrate all the similarity measures in an example as follows: Consider a set of 16 objects (A1, A2, . . . , A16), each object has thirteen (13) features present or absent (C1, C2, ..., C13), all these data are illustrated in the following table:

Table 3. Different forms of similarity measures

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	C ₉	C ₁₀	C ₁₁	C ₁₂	C ₁₃
A ₁	1	0	0	1	0	0	0	0	1	0	0	1	0
A ₂	1	0	0	1	0	0	0	0	1	0	0	0	0
A ₃	1	0	0	1	0	0	0	0	1	0	0	0	1
A ₄	1	0	0	1	0	0	0	0	1	0	0	1	1
A ₅	1	0	0	1	0	0	0	0	1	1	0	1	0
A ₆	1	0	0	1	0	0	0	0	1	1	0	1	0
A ₇	0	1	0	1	0	0	0	0	1	1	0	1	0
A ₈	0	1	0	0	1	1	0	0	0	1	0	0	0
A ₉	0	1	0	0	1	1	0	0	0	0	1	0	0
A ₁₀	0	1	0	0	1	1	0	1	0	1	1	0	0
A ₁₁	1	0	0	0	1	1	0	0	0	1	0	0	0
A ₁₂	0	0	1	0	1	1	0	0	0	1	1	0	0
A ₁₃	0	0	1	0	1	1	0	1	0	1	1	0	0
A ₁₄	0	0	1	0	1	1	1	1	0	0	1	0	0
A ₁₅	0	0	1	0	1	1	1	1	0	0	1	0	0
A ₁₆	0	0	1	0	1	1	1	0	0	0	0	0	0

In applying the measures of similarities between some pairs of objects, we obtain a rate of similarity on the similarity function used:

Table 4. A rate of similarity on the similarity function used

	S (A ₁ , A ₂)	S (A ₁ , A ₁₂)	S (A ₁₂ , A ₁₆)	S (A ₁ , A ₁₆)
Russel, Rao	0.23	0.00	0.23	0.00
Kendall, Sokal-Michener	0.92	0.31	0.77	0.38
Rogers, Tanimoto	0.86	0.18	0.63	0.24
Hamann	0.85	-0.38	0.54	-0.23
Sokal, Sneath	0.96	0.47	0.87	0.56

The previous results give relations that may exist between different data from a sample objects and deduce statistical information for describing more condensed key information contained in these data. We also seek to classify data into different subgroups that are similar. According to

the area, the nature of this knowledge is different in data analysis; information will be taken into account instead of a statistical nature.

3. SIMILARITIES IN THE SOFTWARE PROCESS

The proposed guidance system [3] addresses multiple views providing assistance to stakeholders. Our approach aims to optimize the profile classes. To be adaptive to both the context and identified needs, our model of adaptive guidance covers two levels of abstraction. It is based on a set of task and activity model, the model developer and development team, as well as the selection criteria specified by the mode of access for responding the objects of support by the defined assistance interventions (Figure 2.). The instantiation of this system is through rules of assistance detailed with the requirements for initiating appropriate actions to support a particular context [2, 3, 5].

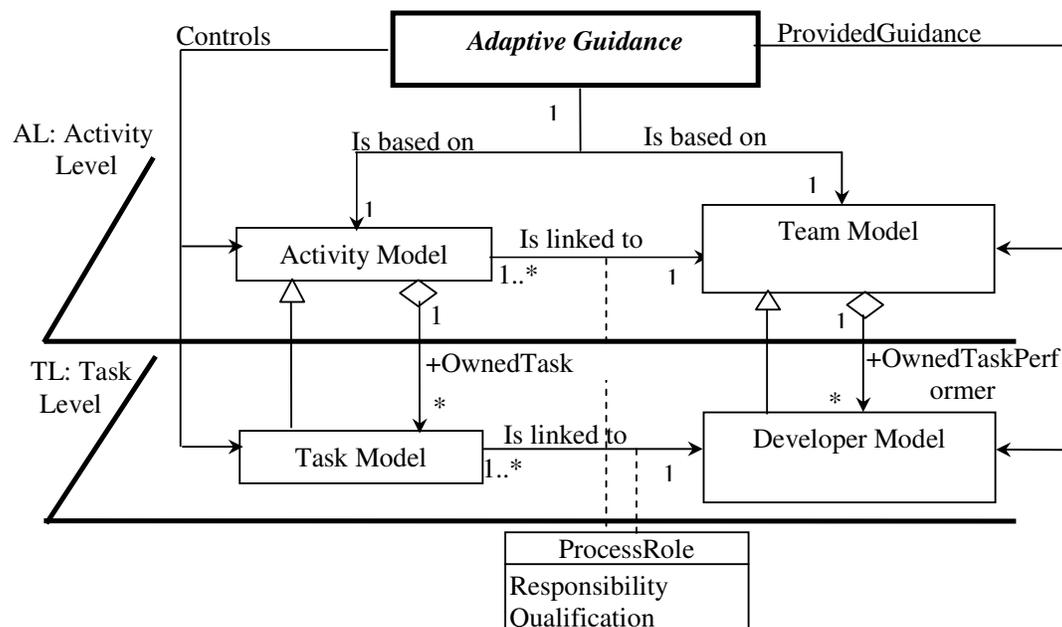


Figure 2. : Adaptive Guidance Model.

3.1. The Adaptive guidance model

This assistance system is based on the major models: the activity model, the developer model and the team development model.

a) *The activity model:* models the workflow, it is defined by:

- A hierarchical list of tasks,
- A mode of progression in the activity ensuring that all tasks can be performed under control in a preset order established by the designer,
- A temporal mode of progression specifying deadlines for completion.

The aspects of the activity model are useful for the assistance system to provide assistance on contextual growth in activity.

b) The developer model: defines the specific properties of each user. It allows our model to make adaptation according to these properties while maintaining the activity model. These properties can be either static or dynamic.

- *The static aspect refers to the user characteristics:*
 - his expertise in the field,
 - his familiarity with the software process model or with the software process,
 - his role in the activity.
- The dynamic aspect refers to the behavior of using the assistance system, the assistance system must be interpreted during the use of the process or system software support, for example :
 - the fact to execute, to define or to complete resource of software process,
 - the workload of an activity,
 - his reaction to a message of support.

c) The development team model: development environments allow exchanges and collaborative work. The assistance system can then construct a development team model that represents elements of the team. Example: trace of the various activities of the team as well as different interactions allow the developer to have a script about his own progress in the activity and the progression of the team. The properties of this model can be static or dynamic order.

- The static dimension references skills and team performance in the field of collaboration and distribution of task.
- The dynamic dimension deals with the behavior of the development team. It describes the actions taken by the team during the course of software process.

These data constitute indications that can be interpreted on the use of the assistance by the developer.

3.2. The assistance intervention

During the construction or interpretation of a software process model, the proposed model for assistance allows the developer to choose various support functions, namely:

- 3.2.1.** Controlling and taking corrective initiative: protect the user of his own initiatives when they are inappropriate, inadequate initiative under progress.
- 3.2.2.** Controlling and taking constructive initiative: the ability to take positive initiatives, executing and combining the performance of operations without user intervention.
- 3.2.3.** Specific assistance: analyze the impact projection to avoid deadlocks or delays.

3.3. The profiles categorization

For the sake of productivity and optimal lead time, we were led to define an effective process for allocating appropriate guidance's. This efficiency is based on the process of maximizing the number of profiles classes to be considered in a development system. We will present our analysis of similarity and classification of our population.

The conception of the processing system will be done through various algorithms. They process both similarity index and hierarchical classification threshold for different profiles. To avoid an important dissemination of similarity, this classification will be ordered by level of similarity

index. This classification will serve as the basis for the selection and assignment of appropriate types of assistance.

To reach an objective comparison between the profiles, an operator should be used for calculating similarity based on the instantaneous evaluation of selected features and associated weights. Despite the fact that this evaluation is not formal, it remains a crucial step for the classification. For this, we use the notion of symbolic learning for instant evaluation of some attributes of the profile as the behavior of the developer or development team.

3.3.1. Algorithm for computing Similarity Index

The evaluation of characteristics is based on the evolution of developer productivity. The weight value of each feature indicates the degree of its importance. The approach used for the evaluation of characteristics is based on "COCOMO II" work [19, 20]. The table of weights could be refined as soon as we have more data.

Consider two people profiles symbolized by:

$$\mathbf{X} = (x_1, x_2, x_3, \dots, x_n)$$

$$\mathbf{Y} = (y_1, y_2, y_3, \dots, y_n)$$

Each characteristic is related to a weight representing its impact in the degree of similarity and symbolized by: $\mathbf{W}_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{in})$

Example: a family of two profiles is semantically described in table 4 and table 5. The semantics evaluation and the weighting are determined by the project manager on an ongoing project [3, 19, 20].

Table 5. the profiles evaluation.

	Features	Evaluation of profile 1	Evaluation of profile 2
<i>Model of activity</i>	<i>Density of tasks in the activity</i>	High	Medium
	<i>Complexity level</i>	Medium	Low
	<i>Activity Type</i>	Tolerance zero	Margin Free
<i>Model of developer</i>	<i>Role</i>	Critical	Classic
	<i>Competence</i>	High	Medium
	<i>Familiarity with Process Software</i>	Medium	Low
	<i>Behavior for assistance</i>	Adequate	Acceptable
<i>Model of development team</i>	<i>Skill Area Collaboration</i>	High	Medium
	<i>Behavior for assistance</i>	Acceptable	Adequate

To scan the semantics evaluation, we associate the weight corresponding to the consideration according to each attribute.

Table 6. Table of weights

W [1]	P2
W [2]	P2
W [3]	P3
W [4]	P2
W [5]	P2
W [6]	P2
W [7]	P2
W [8]	P2
W [9]	P4

With $[i] \in [1, 5]$. Where P_i represents the computing value

The algorithm processing is as follows: For any feature, whether it is identical to the profiles, we increment the similarity index by the weight of this feature, otherwise, if the difference between the two characteristics is $< 1/2$, we add half the weight of it, otherwise, we move to the next feature.

The value $1/2$ represents the average distance between two successive levels of an attribute evaluation.

After all iterations, the similarity function obtained is formalized as follows:

$$S(x, y) = \frac{A(x, y)}{\sum W[i]}$$

With:

- X and Y represent the characteristics of the two profiles.
- W [] represents the weighting of each feature.
- A (x, y) represents the sum of the weights between the two profiles, it is included between 0 and $\sum W[i]$.

The similarity function developed verifies the properties of a similarity measure.

$\forall x, y$ two profiles $\in N$

- **If the profiles (x, y) are identical**
Then $A(x, y) = A(x, x) = \sum W[i]$ with $i = 1..9 \Rightarrow S(x, y) = A(x, y) / \sum W[i] = 1$
- **If the profiles (x, y) are totally different (the values are all above features $> 1/2$). Then for**
- **any characteristic** $A(x, y) = 0 \Rightarrow S(x, y) = A(x, y) / \sum W[i] = 0$.
- **If any profiles are neither identical nor different then " It is appropriate to consider three subsets possible through the following 03 cases":**

This allowed us to affirm that:

- ✓ $\forall x, y$ two profiles $\in N$, The similarity function $S(x, y) \in [0, 1]$.
- ✓ $\forall x, y$ two profiles $\in N$, then $S(x, x) = S(y, y) \geq S(x, y)$.

- (1) $A(x, y) = 0$ for features with a difference $> 1/2$
- (2) $A(x, y) = \sum 1/2 W[i]$ for the characteristics with a difference $< 1/2$
- (3) $A(x, y) = \sum W[i]$ for completely identical characteristics.

Finally for all characteristics:

$$A(x, y) = A(x, y)_{(1)} + A(x, y)_{(2)} + A(x, y)_{(3)}$$

$$= 0_{(1)} + 1/2 \sum W[i]_{(2)} + \sum W[i]_{(3)}$$

Then

$$S(x, y) = A(x, y) / \sum W[i] = (0_{(1)} + 1/2 \sum W[i]_{(2)} + \sum W[i]_{(3)}) / (\sum W[i]_{(1)} +$$

Input: The Profiles of Two Developers X and Y. - Table Weights W [].

Output: The value of similarity index $S(x, y) \in [0, 1]$.

Begin:

$A(x, y) = 0$; // Similarity index between X and Y

For all characteristics X_i and Y_i Do

If both characteristics are identical

Then $A(x, y) = A(x, y) + W[i]$;

Else

If $(|X_i - Y_i| < 1/2)$

Then $A(x, y) = A(x, y) + 1/2 W[i]$;

End if;

End if;

End.

Figure 3. : Algorithm for calculating similarity index.

Example: based on the assessing approach of the COCOMO model, the quantification of each characteristic of a profile P is on the data range] 0, 2 [. It is usually done through three steps, described by high, medium or low levels contribution, applying the following rules:

- 1: impact of middle order.
- <1: positive impact.
- >1: negative impact.

Table 7. The profiles evaluation

	Features	profile 1	profile 2	profile 3	profile 4
Model of activity	<i>Density of tasks in the activity</i>	1.65	1.20	1.10	1.65
	<i>Complexity level</i>	1.00	0.60	1.00	1.00
	<i>Activity Type</i>	1.70	1.20	1.20	1.70
Model of developer	<i>Role</i>	1.15	0.70	1.15	1.60
	<i>Competence</i>	0.55	1.00	1.00	0.55
	<i>Familiarity with Process Software</i>	0.40	0.80	0.35	0.40
	<i>Behavior for assistance</i>	0.40	0.60	0.40	0.40
Model of development	<i>Skill Area Collaboration</i>	0.70	0.70	0.70	0.70
	<i>Behavior for assistance</i>	0.95	0.10	0.95	0.95

The weight value of each feature indicates the degree of its importance. The project manager associates the value correspondence table of weights, for example.

Table 8. Table of weights

W [1]	1
W [2]	1
W [3]	2
W [4]	1
W [5]	1
W [6]	1
W [7]	1
W [8]	1
W [9]	2

Based on our approach, the calculation of the similarity value between profiles is given by:

Table 9. The similarity values

	Similarity value
S (P ₁ , P ₂)	0.36
S (P ₁ , P ₃)	0.63
S (P ₁ , P ₄)	0.95
S (P ₂ , P ₃)	0.59
S (P ₂ , P ₄)	0.31
S (P ₃ , P ₄)	0.59

3.3.2. Ascending Hierarchical Classification Algorithm for (addressing) and similarity threshold

Classifying is grouping objects together according to similar criteria. There are two main families of classification techniques:

- The non-hierarchical classification or partitioning leads to the decomposition of the set of all individuals in m disjoint sets or equivalence classes, the number of classes is fixed for m .
- The hierarchical classification for a given accuracy, two individuals may be confused in the same group, whereas in a higher level of accuracy, they will be separated and belong to two different subgroups.

We opted for the hierarchical classification in increments of similarity that led to construct a classification tree showing the transition profiles to the group through a series of consolidation.

The obtained classification is related to the variables selected to describe individuals, in our case the developers. They are called the active variables, which will be based on the classification of individuals. For this, and to avoid dispersion of profiles similarity, the user must set the level of similarity describing each time the similarity values to consider and the level of precision represents the similarity threshold to be applied on the profiles of guidance to classify.

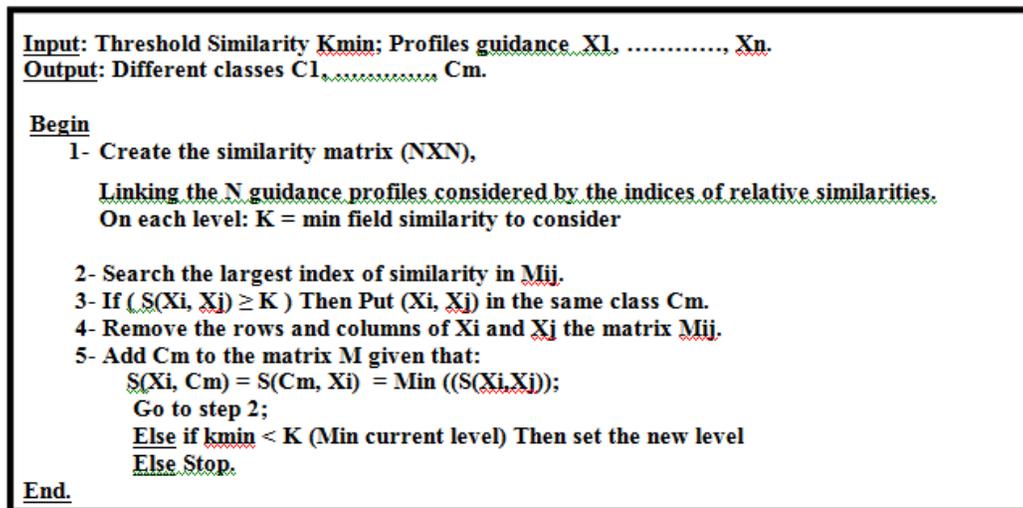


Figure 4. : Ascending Hierarchical Classification Algorithm

3.3.3. Processing of the algorithm on a sample guidance profiles

We have the similarity threshold set and profiles guidance X_1, X_2, \dots, X_n as input. Our algorithm will create a square matrix of size (NXN) considering the number of profiles to classify and index of similarity between profiles. See the following graph of similarity:

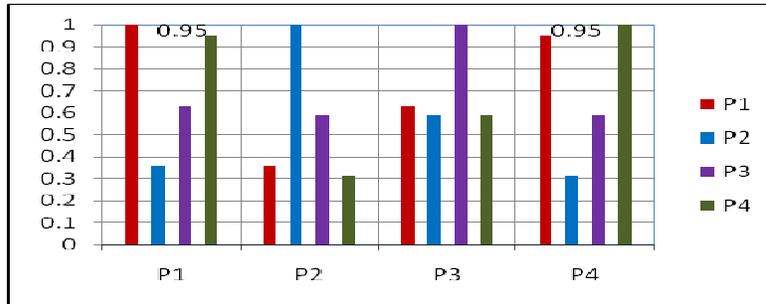


Figure 5. : The graph of similarity

For a level of 0.2 and a minimum similarity threshold of 0.50, set initially by the user, which fixes K to 0.80, the application of this algorithm is illustrated as follows:

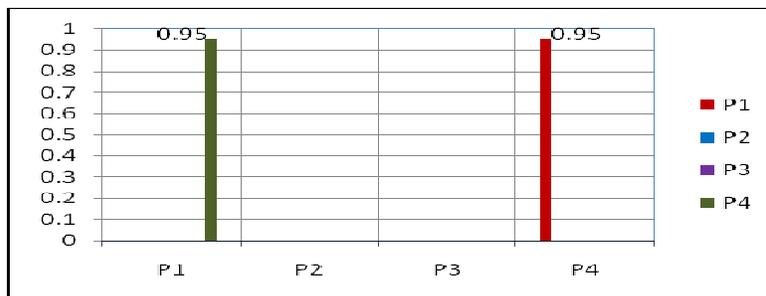


Figure 6. : Illustration of the algorithm

The maximum value of similarity in this table is 0.9, it is the index of similarity between two profiles P1 and P4, and these profiles will be aggregated to the first group C1.

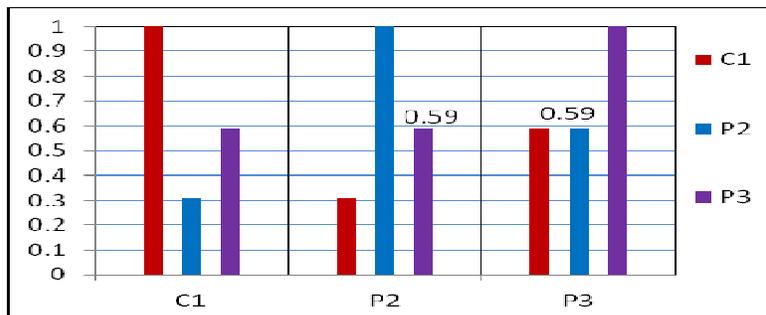


Figure 7. : Classification (first set)

At this level, we find that the similarity indices are all below the minimum level of similarity of the first set.

- Since K_{min} is less than K , we fix the next level, for this example, the new K will be set at 0.50. It repeats the previous steps until all the indices similarities are below the new threshold of similarity.

- The maximum value of similarity in the new similarity matrix is 0.59, it is the index of similarity between two profiles P2 and P3, and the two profiles will be aggregated to the first group C2.

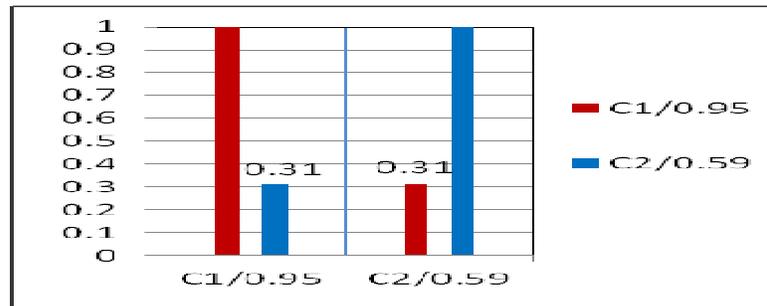


Figure 8. : The major classification

The application of our approach to similarity index and hierarchical clustering allowed us to deduce of the initial profiles number, two major classes C1, C2. This will provide the basis for optimal allocation of the appropriate guidance.

The guidance profile (GP) associated to each profile (Px) class is based on the following formula:

$$GP(Px) = \sum A_i W_i / 2 * \sum W_i \text{ avec } i=1 \text{ to } 9$$

With: A_i : the characteristic value and W_i : the associated weight. In our case, the guidance profile of each class based on the smallest similarity value is given by:

Table 10. The associate guidance value

	C1	C2
GP	<i>0.50</i>	<i>0.37</i>

It should be noted that the value of GP is ranged from 0 to 1. The range associated with each type of guidance is defined by the project manager. For instance, if the range of the corrective guidance is bounded between 0 and 0.40 and the range of the constructive guidance is between 0.41 and 0.70, we automatically associate a corrective guidance to the class C2 and a constructive guidance to the class C1.

4. CONCLUSION

The system presented in this paper is an approach based on similarity to process guidance model in the software process. It allows the profiles optimization, ie: classes presented through the semantics description of handled data, the definition of a system for processing the similarity index and classification of guidance profiles. The aim of this work is to facilitate the analysis of our population using the adaptive development context involved in the execution of a software process.

The system has been designed and implemented and its practical assessment seems to be promising with a significant impact on the productivity of software process development.

In perspective and in order to improve this approach, it would be interesting to develop a similarity measure that takes into account the partial knowledge of profile characteristics. This allows the selection of a profile as the "best effort".

REFERENCES

- [1] Kirk, D.C, MacDonell, S.G., & Tempero, E. 2009 Modeling software processes - a focus on objectives, in Proceedings of the Onward, 2009. USA, ACM Press, pp.941-948.
- [2] Ivan Garcia and Carla Pacheco. 2009. Toward Automated Support for Software Process Improvement Initiatives in Small and Medium Size Enterprises. Book chapter. Software Engineering Research, Management and Applications 2009. Volume 253, pp. 51–58. Springer-Verlag Berlin Heidelberg. ISBN: 978-3-642-05440-2.
- [3] Hamid Khemissa, Mohamed Ahmed-Nacer, Mourad Daoudi, 2008. A Generic assistance system of software process. In proceeding of the IASTED International Conference on Software Engineering: Software Engineering. (SE '08), ACTA Press, Anaheim, CA, USA, 237-242 ©2008 February., Austria.
- [4] Dadam, P. and Reichert, M. 2009. The ADEPT Project: A Decade of Research and Development for Robust and Flexible Process Support – Challenges and Achievements,' Springer, Computer Science - Research and Development, 2009. Vol. 23, No. 2, pp. 81-97.
- [5] MALGOUYRES H., MOTET G. 2006. A UML Consistency Verification Approach Based on Meta modeling Formalization. Symposium on Applied Computing, Dijon, France, ACM publishers.
- [6] C. EyssautierBavay, 2008. Modèles, langage et outils pour la réutilisation de profils d'apprenants", Thèse de doctorat de l'Université Joseph Fourier Grenoble 1, 26 Mai 2008.
- [7] CHU09 CHUNG-FOO-WO Daniel. 2009. Adaptation dynamique par tissage d'aspects de l'optimisation. Thèse de doctorat, Université de Nice - Sophia Antipolis, 5 Mars 2009, Equipe RAINBOW, Pôle GLC.
- [8] S. dami et al, 1998. APEL: a Graphical Yet Executable Formalism for Process Modeling. Kowler Academic Publisher, pp. 60-96, Boston, January 1998.
- [9] Borislava I. Simidchieva, Lori A. Clarke , Leon J. Osterweil. 2007. Representing process variation with a process family. ICSP'07 Proceedings of the 2007 international conference on Software process. Springer-Verlag Berlin, Heidelberg © 2007 ISBN: 978-3-540-72425-4.
- [10] Coulette Bernard., Crégut Xavier. et al, 2000. RHODES, a Process-centered Software Engineering Environment, in Proc. of ICEIS 2000, Stafford, pp 253-260, 2000.
- [11] Bradshaw J., Editor,. 2000. Handbook of Agent Technology. MIT Press, 2000.
- [12] OMG. Inc. 2008. Software Engineering Meta-Model Specification version 2.0: Formal/2008-04- 01.
- [13] Vivien Robinet, Gilles Bisson, Mirta B. Gordon, Benoît Lemaire. 2007. Induction of High-level Behaviors from Problem-solving Traces using Machine Learning Tools. Published in "IEEE Intelligent Systems 22, 4 (2007) 22".
- [14] Ahmed Belkhirat, Abdelghani Bouras, Abdelkader Belkhir. 2009. A New Similarity Measure for the Anomaly Intrusion Detection. In Third International Conference on Network and System Security (NSS).
- [15] Ahmed Belkhirat, Abdelkader Belkhir, Abdelghani Bouras 2011 A New Similarity Measure for the Profiles Management, UKSIM '11: Proceedings of the Tenth International Conference on Computer Modeling and Simulation, Cambridge England.
- [16] Xavier Aimé, Frédéric Furst, Pascale Kuntz, Francky Trichet. 2009. «SEMIOSEM: A Semiotic-Based Similarity Measure". On the Move to Meaningful Internet Systems (OTM 2009), International Workshop on Ontology content and evaluation in Enterprise (OntoContent'2009), Lecture Notes in Computer Science-LNCS 5872, pp. 584-593. Springer-Verlag (Berlin Heidelberg). ISBN 978-3-642-05289-7. Villamoura, Portugal.
- [17] S. Boriah, V. Chandola, and V. Kumar. 2008. Similarity measures for categorical data: A comparative evaluation. In SDM 2008: Proceedings of the eighth SIAM International Conference on Data Mining, pages 243-254, 2008.
- [18] Fernando Lourenco, Victor Lobo, Fernando 2004. Bacao: Binary-based similarity measures for categorical data and their application in self-organizing maps. JOCLAD - XI Days of classification and analysis of data, April 1-3, Lisbon.

- [19] Barry W. Boehm, Chris Abts, A. Winsor Brown, Sunita Chulani, Bradford K. Clark, Ellis Horowitz, Ray Madachy, Donald J. Reifer, Bert Steece. 2009 Software Cost Estimation With COCOMO II. Prentice Hall Edition, ISBN: 0137025769, 978013702576.
- [20] Kirk, D., & MacDonell, S. 2009. A simulation framework to support software project (re)planning, in Proceedings of the 35th Euromicro Software Engineering and Advanced Applications (SEAA) Conference. Patras, Greece, IEEE Computer Society Press, pp.285-292.

IMAGE ACQUISITION IN AN UNDERWATER VISION SYSTEM WITH NIR AND VIS ILLUMINATION

Wojciech Biegański and Andrzej Kasiński

Institute of Control and Information Engineering,
Poznań University of Technology, Poznań, Poland
wojciech.bieganski@doctorate.put.poznan.pl

ABSTRACT

The paper describes the image acquisition system able to capture images in two separated bands of light, used to underwater autonomous navigation. The channels are: the visible light spectrum and near infrared spectrum. The characteristics of natural, underwater environment were also described together with the process of the underwater image creation. The results of an experiment with comparison of selected images acquired in these channels are discussed.

KEYWORDS

underwater vision system, AUV, image enhancement, image fusion

1. INTRODUCTION

An autonomous underwater vehicle navigation could be supported by using sonar systems, dead reckoning systems and by using the computer vision [9]. The paper focuses on visual navigation, especially on improving the quality of 2D images of underwater objects. The proposed system captures images in two channels of the light spectrum. The basic assumption is that the images recorded in each band of the wavelength consist of different image features or areas. The channels are the optical spectrum (visual spectrum of light, VIS) and the near infrared band (NIR). The operational environment of designed system are inland waters (lakes or rivers), both natural and artificial, where the visibility in extremal cases, in some areas reaches no more than 20 cm (in lowland rivers). The system is designed to reduce the impact of the infavourable effects influencing the underwater image formation. The acquired images are next combined together resulting in an image consisting of more useful information than any of the two component images. The operation is called the *single-sensor image fusion*.

The hardware used for the tests presented in this paper is a trinocular vision system (TVS) designed and built for the use of inland, underwater imaging [2]. The TVS acquires images in three channels of the light spectrum: NIR, VIS and NUV (near ultraviolet). The comparison of selected images captured in the NIR and VIS channels is presented in this article.

Apart from being a sensory system for the navigation of AUVs, the exact purpose of the designed vision system depends on the kind of mission to execute. In remote mode (with the participation of the operator) the TVS could be used to support searching and rescue missions, inspection of underwater constructions and cataloguing of plants or underwater creatures.

The described vision system is a part of The Isfar Project - a hybrid of an AUV and mini-ROV class vehicle for the exploration of the inland waters [19].

2. WATER OPTICS

Natural water is an environment difficult to describe due to its various composition that is not fully identified. The nature of the underwater optical effects is strongly selective and volatile. The intensity of those phenomena has a spatial character, it depends on the location within the water body (the depth and also horizontal position). Furthermore, the intensity of the effects concerning underwater optics has a temporal character i.e. it could change during the day/night cycle and also it is seasonal. The optical water properties may also change within several years [10]. The constituents found in waters and wastewaters are divided into categories [1]:

- strongly absorbs light, particularly blue, scattering is negligible,
- total suspensoids - responsible for almost all scattering,
- mineral suspensoids - scatter light intensely but usually absorb light weakly,
- detritus - spectral absorption similar to yellow substance, also scatters light,
- phytoplankton - absorbs the light strongly with spectral selectivity and also scatters the light strongly.

The two most significant effects influencing the optical parameters of underwater environment are: light absorption and light scattering, both of them rely on the composition of the underwater environment.

2.1 Light absorption

When a photon hits a water molecule it makes that molecule oscillate, hence changes its energy level. The photon is being absorbed during the change of the energy level of the molecule. As a consequence, the radiance of the emitted light drops logarithmically as the distance from the light source grows (Lambert's Law).

The light absorption effect is described by the light absorption coefficient a . The intensity of the absorption effect strongly depends on the kind of molecules found in the optical path. The absorption coefficient a grows towards the light of lower wavelengths (IR). Absorption of light by water is minimal within the $\lambda=400$ to 500 nm (violet to green) range [8].

2.2 Light scattering

The second effect influencing the transmission of the light by water is the scattering effect. The effect occurs when the light beam changes its direction while come across the area of the non-water substance found in the optical path. The photons are being re-radiated in any or all directions with unchanged (molecular scattering) or lower (fluorescence) energy content. The last type of scattering is connected with light diffraction, refraction or reflection from suspended particles [1].

The scattering effect is the dominating effect especially in natural waters, because of the diversity and volume of constituents suspended (SOM, suspended organic matter) or dissolved (DOM, dissolved organic matter) in the environment. Moreover in natural waters the light scattering is isotropic i.e. light is scattered in every direction, even towards the light source (known as *backscattering*).

The scattering is described by the scattering coefficient b and the volume scattering function $\beta(\theta)$ (describing the intensity or radiance of light being scattered into the direction of the θ angle). The effect has significant impact on the transmission of the light by water especially for the light of shorter wavelength.

2.3 Light attenuation

The scattering and the absorption effects are inseparable in natural waters. The a and b coefficients are very difficult to measure apart. The absorption coefficient could only be measured without the influence of the scattering error in very clean water [3].

The combination of the scattering and absorption effects result in the attenuation of the light in underwater environment. The light attenuation is described by the beam attenuation coefficient $c=a+b$, it is the one of the fundamental parameters of the water quality describing its clarity. The optical parameters a , b , c and $\beta(\theta)$ are so-called *inherent optical properties* that fully specify the optical character of the water.

The most tangible and easy to measure parameter describing the optics of any light-transmitting environment is the optical transmission (or transmittance). Transmission is a ratio between the radiance of light emitted by the light source (L_0) and the radiance of the light measured at the distance r from the light source (L_r) expressed in percent:

$$T = \frac{L_r}{L_0} \cdot 100$$

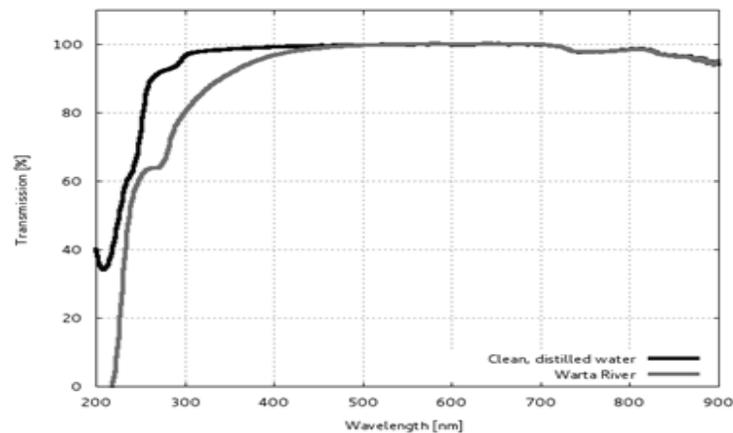


Figure 1. Optical transmission of distilled water nad water taken from Warta River near Poznań, Poland

The transmission of water samples containing the distilled water and the water taken from the Warta river, measured with the use of the spectrophotometer is presented in Fig. 1. The main differences on the graph could be observed in the 200 to 400 nm range of the light spectrum. The differences result from the attendance of the scattering effect in natural water. In the visual range (the wavelength of $\lambda=400$ to 700 nm) the transmission was invariable, still better for the distilled water. The presence of the light absorption effect occurred over the $\lambda=700$ nm (red to infrared), for both samples in an equal degree.

2.4 Underwater image formation

The optical path of photons emitted from the light source (Power-LEDs), through the object of interest immersed in the underwater environment, to the detector (a CCD camera) is shown in the diagram in Fig. 2. Other optical effects taking part in underwater imaging are: the reflection of the light rays on the surface of the glass viewfinder (two times, from both sides of the viewfinder), the refraction between air/glass and glass/water interfaces (also two times), the reflection on the surface of the detected object, letting actually *see* that object and the distortion of the lens and optical filters.

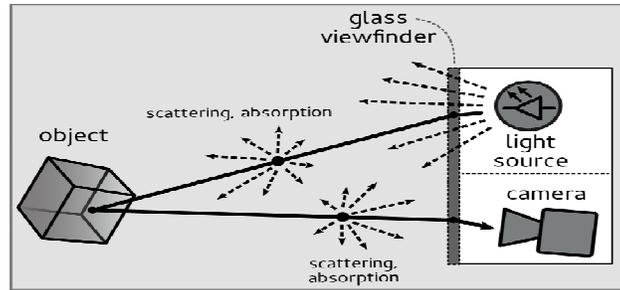


Figure 2. A process of image formation underwater

3. UNDERWATER IMAGE ENHANCEMENTS METHODS

The methods of underwater image enhancement (also called image denoising or dehazing) could be divided into two categories: methods that uses software pre-processing algorithms and hardware methods basing on the modifications of the parameters of the optical path of the light.

3.1 Software methods

The software methods of the underwater image enhancement depend mainly on image filtering: homomorphic filtering, anisotropic filtering or filtering in the wavelet domain [14],[18]. Other methods include image deconvolution, contrast equalization [14],[15] and local histogram equalization [13].

3.2 Physical methods

Hardware methods of the underwater image enhancement base on the modifications of the parameters of the optical path e.g. by using polarizers [17]. Some experiments on the the various placement of the light sources or with multi-directional fusion were also conducted [16].

A method which initial experiments were presented in this paper connects both hardware and software image enhancement methods.

4. CHANNELS OF IMAGE ACQUISITION

The image is captured in two separated channels of the light spectrum. The separation of channels is assured by using the optical filters. The light sources were selected accordingly to the desired wavebands. The system consist of one camera, three optical filters and Power-LEDs as a lighting source. The optical power of the light sources was approximately levelled, since the Power-LEDs of the same electrical power differs in optical power depending on the wavelenght of the emitted

light, the NIR LEDs have weaker optical power than white LEDs. The filters were mounted on a rotating disk driven by a servomechanism in front of the camera lens, letting to switch of the image acquisition channel. The viewfinder was made of BK-7 (borosilicate) tempered glass.

4.1 NIR channel

The near infrared spectrum is a wavelength between $\lambda=750$ nm and $\lambda=1400$ nm. There was a Schott RG-712 long pass filter used (the filter cuts all wavelengths below $\lambda=712$ nm and the Edixeon EDEI-1FA3 Power-LED (maximum optical power at $\lambda=850$ nm).

The optical parameters of natural water in NIR range differs from the parameters occurring in the VIS spectrum [12]. The light in the NIR range of the light is almost impervious to the influence of the scattering effects [11]. On the other hand the light in NIR spectrum is strongly affected by the light absorption in underwater environment [5]. The intensity of the absorption effect could depend not only on the molecular structure of the water (and dissolved/suspended substances) but also on the temperature [6], [7].

Another application of the NIR radiation was to use NIR light emitters together with the camera to observe fish. The NIR light is invisible to the them, thus the observation system does not have a notable impact on fish behaviour [4].

4.2 VIS channel

The optical (or visual) spectrum of the light were λ is situated between 380 and 780 nm. There is a pair of optical filters used in this channel: UVK-2510 UV cut-off filter and ICF-2510 IR cut off filter. The illumination comes from 3-Watt Power-LED emitting the warm white light.

5. EXPERIMENT AND RESULTS

The experiment consisted of acquiring a sequence of images in both NIR and VIS channels of the immersed object in order to compare and describe differences occurring on the images.

The experiment was conducted in laboratory environment. The water tank was the aquarium with blinded panels. The volume of the aquarium is 250 litres. The submerged object of observation is a 11x11 cm cube, where its every face is made of (or covered by) different material, that could appear in lake or river beds. Those faces are:

- a face with a marker (a chessboard) attached (used as reference),
- a metal sheet covered with rust,
- a tinplate face,
- a rubber face,
- two fabric-covered faces: with straight lines pattern and with circular blobs.

Apart from the object of interest, there were some underwater plants in the tank, black and white gravel and stones. The water comes from the water supply network. The background was a black PVC sheet.

Since the IR radiation coming from the natural light source is completely absorbed a few cm below the water surface, the authors decided to use artificial lighting only.

Two descriptions of the acquired images were proposed due to image comparison: histograms (intensity analysis) and detected edges (feature diversity analysis). Captured images are presented in Fig. 3a and Fig. 3b. On the first pair of selected images, there are three faces of the cube visible: the reference face, a tinplate face and a metal sheet face covered with rust.



Figure 3. Acquired images, a - NIR channel image, b - VIS channel image

The absorption effect causes the images to be less detailed (or darker) depending on the distance between the source (through object) and the detector. The absorption effect has stronger impact as the wavelength of the emitted light grows, thus the images acquired in the NIR had less brightness than the images in VIS channel.

On the other hand, the scattering effect lowers the contrast of captured images by brightening the water surrounding the space between the object and the detector. Since the scattering effect has stronger impact on the light of shorter wavelengths, it is noticeable on images acquired in VIS channel. Differences mentioned above are visible in the images itself, but also in the histograms presented in Fig. 4.

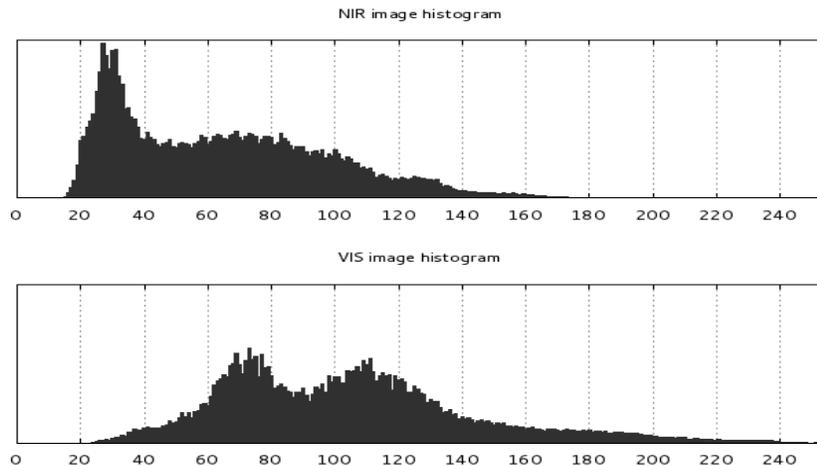


Figure 4. Histograms of 3a and 3b images

A Canny edge detector with the same threshold values for both images was used. Some edges faded on the NIR images due to the absorption effect (noticeable on the chessboard). The main differences revealed on the area where there were underwater plants. The contrast of plants was better in the NIR channel, hence more edges were detected in NIR (underwater plants reflect the

NIR light, since the energy of IR radiation is not gathered by those plants for the use in the photosynthesis process).

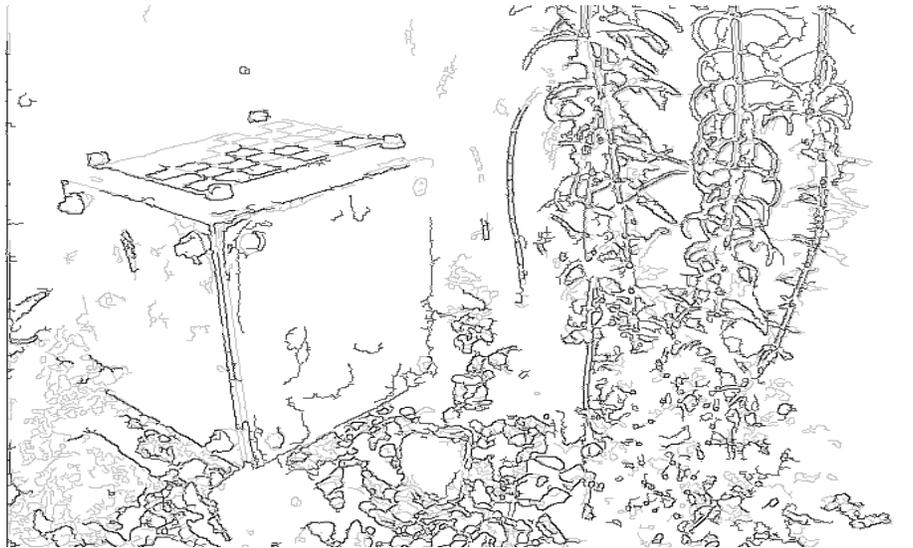


Figure 5. Detected edges, black edges - NIR channel, gray edges – VIS channel

On the second pair of images, there is a cube with three faces visible: a reference face with a chessboard, the face with a metal sheet covered with rust and a face covered with fabric with the circular pattern on it. Images and their histograms are presented in Fig. 5 and Fig. 6.



Figure 5. Acquired images, second pair, a - NIR channel image, b - VIS channel image

All tested fabrics, no matter what pattern were covered by look similarly in the NIR channel. The circles *seen* on the face of the cube in VIS channel are invisible in NIR channel (NIR light is not affected by the dye used to produce the fabric). The result of the Canny edge detector on the images is shown in Fig. 7. The black fabric, almost invisible in the VIS channel, is detectable in NIR channel, hence the NIR radiation could be used during rescue missions to detect e.g. fragments of clothing.

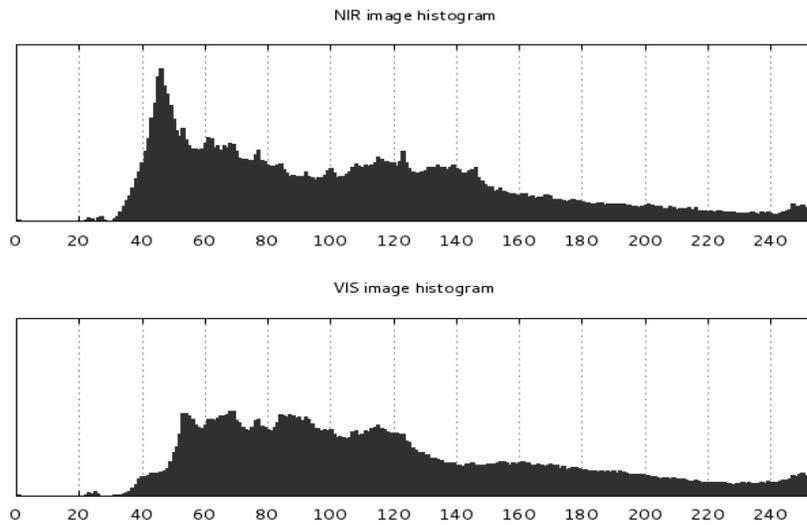


Figure 6. Histograms of 5a and 5b images

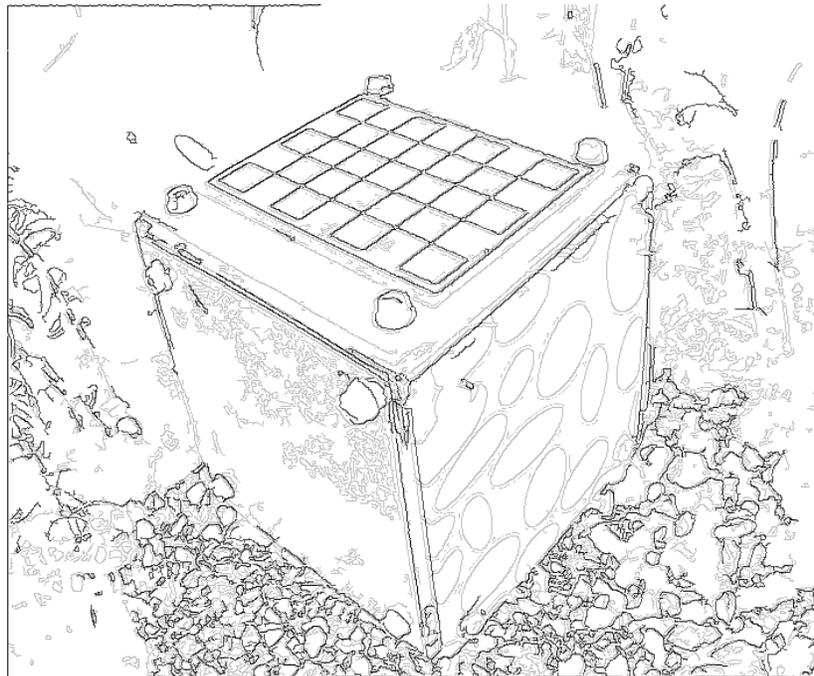


Figure 7. Detected edges, black edges - NIR channel, gray edges – VIS channel

5. CONCLUSIONS AND FUTURE PLANS

The initial tests concerning acquiring of some underwater images with the use of the two channel underwater imaging system were presented in this article. The images were next analyzed in order to find differences resulting from the acquisition channel use. Results obtained in the tests confirmed the presumptions about the principles of radiative transfer in underwater environment.

The underwater imaging with the use of the NIR radiation could find application especially in highly turbid environments such as natural, inland waters due to its resistance to the scattering effect.

Images acquired in NIR include less or equal information than images captured in VIS while imaging objects made of such materials as plastics, metals (some differences in rust-covered surfaces), rubber and, in particular, patterns on fabrics. On the other hand the underwater plants had higher contrast, thus were more distinguishable from the background, than plants captured in VIS channel. A conclusion could be drawn, that imaging with the use of NIR radiation could find be used for searching or cataloguing specific plants. Furthermore, if the presence of plants is undesirable on the images there is a possibility to use the information included in NIR image to remove the plant-filled areas from the VIS image during the image fusion process.

A weighted image fusion algorithm for the images captured in both NIR and VIS channels will be developed, where selected areas on NIR image with assigned weights could be added to (or subtracted from) the VIS image resulting with the image that would be more useful for the navigation algorithms of the underwater vehicle.

Although results are promising, some parts of the system need to be improved. In order to conduct the research concerning the development of the image fusion algorithm, the exact pixel correspondence is needed between both images. In the current system the images were not precisely matched, because the images were not captured simultaneously, since there was a delay needed for switching channels (about 150 ms). The faster servomechanism is required. Moreover, even if the images would be acquired in the same time, there is a pixel disparity resulting from the fact, that camera intrinsics and distortion coefficients are different depending on the channel (due to various refraction coefficients depending on the kind of the filter used). The light sources should need more accurate levelling of the optical power confirmed by prior tests with the use of the radiometer or pyranometer, so that the images would be captured in the same power conditions. Those inconveniences is planned to be removed in short future.

REFERENCES

- [1] R. J. Davies-Colley et al., *Colour and Clarity of Natural Waters. Science and Management of Optical Water Quality*, The Blackburn Press, Hamilton, New Zealand, 1993
- [2] W. Biegański, J. Ceranka and A. Kasiński, "Design, control and applications of the underwater robot Isfar", *Journal of Automation, Mobile Robotics and Intelligent Systems*, 02, 2011, pp.60-65
- [3] S. Tassan et al., "Light Absorption Measurements of Aquatic Particles: Status and Prospects", *Proceedings of IEEE Conference on Geoscience and Remote Sensing, IGARSS'97*, 2, 1997, pp.825 - 829
- [4] S. Chidami, "Underwater infrared video system for behavioral studies in lakes", *Limnology and Oceanography: Methods*, 5, 2007, pp. 371-378
- [5] M. Babin and D. Stramski, "Light absorption by aquatic particles in the near-infrared spectral region", *Limnology and Oceanography*, 47(3), 2002, pp. 911-915
- [6] W. S. Pegau, D. Gray and J. R. V. Zaneveld, "Absorption and attenuation of visible and near-infrared light in water: dependence on temperature and salinity", *Applied Optics*, Vol. 36, No. 24, 1997, pp. 6035-6046
- [7] V. S. Langford et al., "Temperature Dependence of the Visible-Near-Infrared Absorption Spectrum of Liquid Water", *Journal of Physical Chemistry*, 105, 2001, pp. 8916-8921
- [8] A. C. Tam and C. K. N. Patel, "Optical absorptions of light and heavy water by laser optoacoustic spectroscopy", *Applied Optics*, Vol. 18, Iss. 19, 1979, pp. 3348-3358
- [9] A. Branca, E. Stella and A. Distante, "Autonomous navigation of underwater vehicles", *Proceedings of OCEANS'98 Conference*, Vol. 1, 1998, pp. 61-65
- [10] A. D. Jassby et al., "Origins and Scale Dependence of Temporal Variability in the Transparency of Lake Tahoe, California-Nevada", *Limnology and Oceanography*, 44(2), 1999, pp. 282-294

- [11] L. H. Dawson and E.O. Hulburt, "The Scattering of Light by Water", Journal of the Optical Society of America, 27(6), 1937, pp. 199-201
- [12] K. F. Palmer and D. Williams, "Optical properties of water in the near infrared", Journal of the Optical Society of America, 64(8), 1974, pp. 1107-1110
- [13] R. Garcia et al., "On the Way to Solve Lighting Problems in Underwater Imaging", Proc. MTS/IEEE Oceans'02, Vol. 2, pp. 1018 – 1024, 2002
- [14] A. Arnold-Bos et al., "A Preprocessing Framework for automatic underwater image denoising", European Conference on Propagation and Systems, March 2005
- [15] A. Arnold-Bos et al. , "Towards a Model-Free Denoising of underwater optical images", Proc. Oceans 2005 – Europe, Vol. 1, 2005, pp. 527 – 532
- [16] T. Treibitz and Y. Y. Schechner, "Turbid Scene Enhancement Using Multi-Directional Illumination Fusion", IEEE Transactions on Image Processing, Vol. 21, No. 11, 2012
- [17] Y. Y. Schechner and Nir Karpel, "Clear Underwater Vision", Proc. Computer Vision & Pattern Recognition, Vol. I, 2004, pp.536-543
- [18] S. Bazeille et al., "Automatic Underwater Image Pre-Processing", Proc. Characterisation Du Milieu Marin, 2006
- [19] W. Biegański and A. Kasiński, "Initial tests of a trinocular vision system for the underwater exploration", Pomiar, Automatyka, Robotyka, 2, 2013

Authors

Wojciech Biegański, MSc. Eng.

Graduated from the Poznań University of Technology (2009). He is a Ph.D. student at the Institute of Control and Information Engineering of the Poznań University of Technology. His interests are the mobile robotics, especially the visual perception of robots.



Andrzej Kasiński, PhD. Eng.

Graduated from the Poznan University of Technology in 1973 and the Adam Mickiewicz University in 1974. He received the Ph. D. and D. Sc degrees from the Poznan University of Technology in 1979 and 1998, respectively. He was a visiting professor on the Delft University of Technology and the Universidad de Murcia, ENSII Cartagena. Prof. Kasiński has been the head of the Institute of Control and Information Engineering of the Poznan University of Technology since 2002. He is an author of over 150 papers and co-author of 5 patents in the fields of control theory, Pulse-Coupled Neural Network (PCNN), computer vision and biocybernetics.



VARIATION-FREE WATERMARKING TECHNIQUE BASED ON SCALE RELATIONSHIP

Jung-San Lee, Hsiao-Shan Wong, and Yi-Hua Wang

Department of Information Engineering and Computer Science,
Feng Chia University,
Taichung 407, Taiwan, ROC
leejs@fcu.edu.tw

ABSTRACT

Most watermark methods use pixel values or coefficients as the judgment condition to embed or extract a watermark image. The variation of these values may lead to the inaccurate condition such that an incorrect judgment has been laid out. To avoid this problem, we design a stable judgment mechanism, in which the outcome will not be seriously influenced by the variation. The principle of judgment depends on the scale relationship of two pixels. From the observation of common signal processing operations, we can find that the pixel value of processed image usually keeps stable unless an image has been manipulated by cropping attack or halftone transformation. This can greatly help reduce the modification strength from image processing operations. Experiment results show that the proposed method can resist various attacks and keep the image quality friendly.

KEYWORDS

Image watermarking, Discrete Cosine Transform (DCT), variation-free, coordinate system

1. INTRODUCTION

Watermarking technique is often used in anti-counterfeiting technique, and the main purpose is to solve the problem of copyright verification. It mainly marks one or more secrets and representative copyright information such as the logo of the owner in the protected digital multimedia. When this protected digital multimedia is transmitted over the insecure Internet, the secrets must be able to survive to verify the ownership[1, 9].

The digital watermarking technology can be divided into three categories: spatial domain, frequency domain, and compression domain. Spatial domain embedding technique is to modify the pixel values directly. Generally, this technique is efficient, but it is insecure once the watermark image is erased by various image processing operations. As to the frequency domain technique, it first transforms the image pixel values into coefficients via a specific conversion method such as DCT and DWT [2, 5, 6, 10]. Then the watermark bits are embedded into the coefficients. Compared with the spatial domain embedding technique, the frequency one needs more computational cost. Nevertheless, its ability to resist different image processing operations is much better. As to the compression domain watermarking [3, 4, 7, 8, 11, 12, 13], this technique is usually to compute a secret key or a codebook instead of embedding a watermark logo into the protected

image. Thus, we can obtain a lossless outcome since no pixel value is modified during the embedding procedure. By this way, we can guarantee to get a satisfactory watermarked image. But, we need extra memory to record the secret key or codebook for watermark retrieval.

Based on the above mentioned literature, it is clear that the pixel values and coefficients are the commonest component used to define the judgment condition of watermark embedding and extracting. Accordingly, once a watermarked image suffers from attacks, the modified values or coefficients must lead to the incorrect logo retrieval. To avoid this misjudgment, we aim to design a more stable estimation mechanism. Thus, we introduce the XNOR operation and voting strategy to the proposed watermarking scheme. Figure. 1 illustrates an original image and the outcomes after common signal processing operations. We can find out that the pixel value in the same position usually keeps stable unless the whole image has suffered from being seriously destroyed, such as cropping attack and halftone transformation. To enhance the stability of estimation condition, we first select two distinct pixels. Then, we apply the scale relationship of these two pixels to be the judgment condition of watermark embedding and extracting. This can make the condition more flexible even the target image has been distorted. For instance, a pair of pixel is changed from (100, 50) to (80, 60). The estimation condition will not be influenced since the relation of those two pixels still keeps the same, said $n_1 > n_2$.

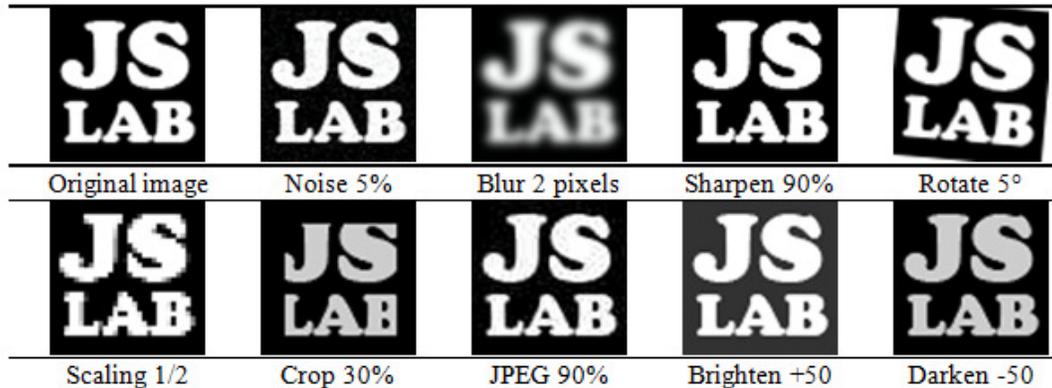


Figure.1 The common image processing operations

To prevent the estimation condition from being swayed by the modification of pixel values, the main idea of our method is to keep the target image the same. What we do to embed the watermark is to apply the XNOR operation on the watermark logo and the scale relationship of one pair of above mentioned pixels. Then we can record the outcome as a secret key. By this way, we can obtain a lossless image and get a stable estimation condition. Thus the new method can confirm the robustness and transparency.

Furthermore, how to decrease the occurrence of error extraction in a robust watermarking scheme is the challenge we are going to solve in this paper. We introduce the voting strategy to embed each watermark bit into different positions, respectively. When the watermarked image suffered from attacks, some watermarked coefficients may not be affected. Accordingly, we can determine the watermark bits through the voting strategy.

The rest of this paper is organized as follows: In Section 2, the proposed watermarking method is introduced. In Section 3, the performance is analyzed by applying various attacks to the watermarked image. Finally, the conclusion is given in Section 4.

2. THE PROPOSED WATERMARKING METHOD

Here, we introduce the detail of how to apply the scale relationship of two distinct pixels and voting strategy to perform the watermark embedding and extracting procedure without any modification on the host image.

2.1 The pixel selection rule

For the estimation condition of watermarking embedding and extracting, we use a pair of pixels as the main component. The first pixel n_1 is chosen by PRNG (Pseudo Random Number Generator), and the other one n_2 is decided according to the selection of n_1 . Actually, the difference between two neighbor pixels is usually small. That is, the scale relationship of two neighbor pixels might be the same in distinct host images. This results in the fact that we may retrieve a similar watermark logo from different host images, said a collision. Considering the uniqueness of images, two pixels at a distance might locate in distinct objects. So, the basic idea is to shift the position of n_1 for a distance s to get n_2 . As shown in Figure. 2, we have shifted node A for a distance to get node B. It is clear that these two nodes have been marked in different objects; thus representing the distinguishing characteristic of images. With the help of voting strategy, the shifting can effectively prevent the occurrence of the collision.

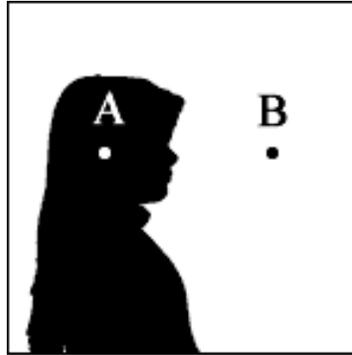


Figure. 2 The characteristic of image

We apply the coordinate system to help find the second pixel of the pair. To increase the variance and enhance the flexibility, we determine the shifting procedure according to the coordinate of the first pixel. Let (a, b) be the coordinate of n_1 and s represent the distance. The selection of n_2 is defined as Eq. (1).

$$n_2 = \begin{cases} (a, b) & \text{if } a \text{ is even and } b \text{ is even} \\ (a+s, b) & \text{if } a \text{ is odd and } b \text{ is even} \\ (a+s, b+s) & \text{if } a \text{ is odd and } b \text{ is odd} \\ (a, b+s) & \text{if } a \text{ is even and } b \text{ is odd} \end{cases} \quad (1)$$

As illustrated in Figure. 3(a), there are four possible positions of n_2 after the shifting procedure, $p_1 = n_1 = (a, b)$, $p_2 = (a+s, b)$, $p_3 = (a+s, b+s)$, and $p_4 = (a, b+s)$. For instance, assume $s = 3$ and $n_1 = (1, 1)$, we have $n_2 = (1+3, 1+3) = (4, 4)$. In case that $n_1 = (2, 1)$, we

get $n_2 = (2, 1+3) = (2, 4)$. If $n_1 = (1, 2)$, we obtain $n_2 = (1+3, 2) = (4, 2)$. Suppose $n_1 = (2, 2)$, we can infer $n_2 = n_1 = (2, 2)$. To guarantee that we can retrieve the exact watermark bit, we shall keep the case of $n_2 = n_1$.

Note that the shifting procedure is rotation-based. Once the shift distance runs over the bound, it continues from the opposite. Let us check the scenario in Figure. 3(b). If $n_1 = (3, 1)$, we have $n_2 = (3+3-5, 1+3) = (1, 4)$. As to the setting of distance s , it should be around $\left\lceil \frac{\text{length of side}}{2} \right\rceil$. The settings of a large s and a small s will result in the same situation that n_2 will be close to n_1 ; thus leading to a similar logo from two different host images.

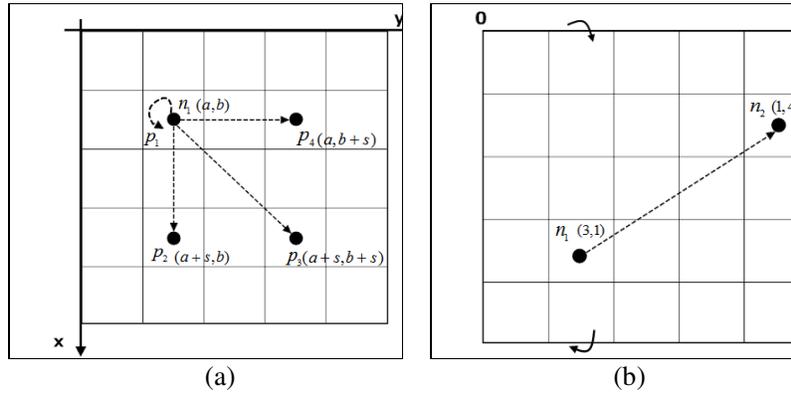


Figure. 3 The selection of pixel position

2.2 The embedding procedure

Assume that the size of host image O is $N \times N$ pixels, each pixel is denoted as $p_{i,j}$ for $0 \leq i, j < N$, and that of the watermark image W is $M \times M$ pixels. The flowchart of the embedding phase is shown in Figure. 4, and the details are given in the following. Here, we define two secret keys K_w and K_p . K_w is used to decide the order of processed watermark bit, while K_p is adopted to determine the embedding position in the original image.

- Step1. Randomly select a watermark bit w_h from the watermark image W according to K_w , for $0 \leq h \leq M \times M$. Set $v = 1$, where v is the number of vote.
- Step2. Apply K_p to PRNG to find the first pixel n_1 . Accordingly, we can obtain the second n_2 by Eq. (1).
- Step3. Get parameter f by Eq. (2).

$$f = \begin{cases} 1 & n_1 \geq n_2 \\ 0 & n_1 < n_2 \end{cases} \quad (2)$$

- Step4. Input f and w_h to XNOR operation (see Table 1) to obtain an r_m , for $m = 1, 2, \dots, h \times 3$. Record all the outcomes as a secret key.

Step5. In case $v < 3$, set $v = v + 1$ and repeat Steps 2 to 5. Repeat Steps 1 to 5 till all the watermark bits are embedded.

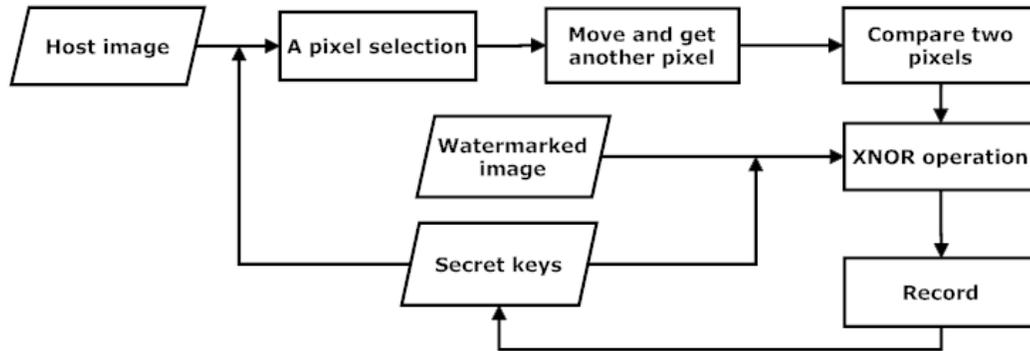


Figure. 4 The flowchart of watermark embedding

Table 1. XNOR operation

f	w_h	$r_m = \overline{f \oplus w_h}$
0	0	1
0	1	0
1	0	0
1	1	1

2.3 Watermark extracting procedure

Figure. 5 illustrates the flowchart of watermark extracting. The detail of the procedure is given as follows.

- Step1. Set $v = 1$.
- Step2. Decide the position of target watermark bit w_h according to K_w , where $0 \leq h \leq M \times M$.
Employ K_p to PRNG to find the corresponding pixel n_1 . Accordingly, we can shift n_1 to obtain the second pixel n_2 by Eq. (1).
- Step3. Compute f value according to Eq. (2).
- Step4. Extract a corresponding secret bit from r_m , where $m = 1, 2, \text{ to } h \times 3$. Apply XNOR operation to r_m and f to obtain t_x , for $x = 1, 2, 3$. Keep t_x in a temporary register.
- Step5. If $v < 3$, perform $v = v + 1$ and repeat Steps 2 to 5. Otherwise, apply the voting strategy to t_1, t_2 and t_3 to determine w_h .
- Step6. Repeat Steps 1 to 5 until all the watermark bits are retrieved.

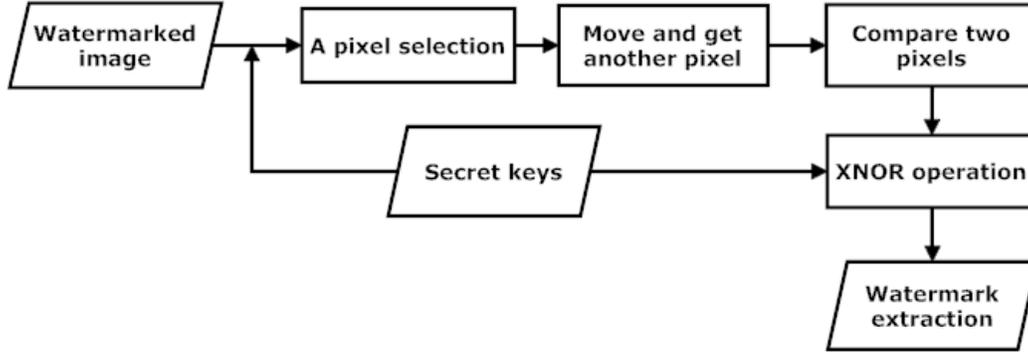


Figure. 5 The flowchart of extraction process

3. EXPERIMENTAL RESULTS

In this section, we employed JAVA to conduct all simulations to prove the practicability and robustness of the proposed scheme, including common image processing operations and attacks. Furthermore, we simulated several related works and compared the results with ours to highlight superiority. In Lin et al.'s method [8], the modulus M is set as 18 and the quality factor $f = 0.3$. In Lai's method [7], the threshold is tuned as 0.04. To obtain a better result in Run et al.'s method [12], the essential parameters are defined as $T = 10$, $\beta = 150$, $\gamma = 1.5$, and $\alpha = 0.95$.

Tools for simulating image processing operations and attacks include JAVA and Photoshop CS2. Simulation types and settings are introduced as follows.

1. Noise: Apply the Photoshop to add Gauss noise by 0.5% to 5%.
2. Blurring: Apply the Photoshop to perform Gauss blurring with the radius from one to five pixels.
3. Cropping: Use JAVA to simulate the cropping processing, including the inside cropping and the outside cropping. The inside cropping mainly concerns the object such as human faces, while the outside one focuses on removing the suburb of the image by 25%, which may destroy the reference information of watermark retrieval.
4. JPEG compression: Employ JAVA API to simulate the lossy compression according to the standard JPEG algorithm. The compression quality ranges from 30% to 90%.

In order to accurately evaluate image quality, aside from the human vision perception, we utilized the Peak Signal to Noise Ratio (PSNR) which is defined as Eq. (3).

$$PSNR(dB) = 10 \log_{10} \left(\frac{255^2 \times H \times W}{\sum_{i=1}^H \sum_{j=1}^W (x_{ij} - \hat{x}_{ij})^2} \right), \quad (3)$$

where H and W are the height and width of the image, respectively, x_{ij} is the original image pixel value at coordinate (i, j) , and \hat{x}_{ij} is the camouflaged image pixel value at coordinate (i, j) .

Moreover, the Normalized Correlation (NC) value which is defined as Eq. (4) is introduced to measure the similarity between the original watermark image and the extracted one, and $NC = [0, 1]$. The similarity between two images is higher if the value gets closer to 1.

$$NC = \frac{\sum_{i=1}^h \sum_{j=1}^w (w_{ij} \times \hat{w}_{ij})}{\sum_{i=1}^h \sum_{j=1}^w (w_{ij} \times w_{ij})}, \quad (4)$$

We also utilized Tamper Assessment Function (TAF) value to evaluate the tampered level of watermark image and the formula is defined as follow.

$$TAF = \frac{\sum_{i=1}^h \sum_{j=1}^w (w_{ij} \times \hat{w}_{ij})}{h \times w} \times 100, \quad (5)$$

where w_{ij} and \hat{w}_{ij} represent the original and extracted watermark at coordinate (i, j) , respectively.

In the following experiments, all the test images are with size of 1024×1024 pixels, and the watermark logo is a binary image with size of 64×64 pixels, as displayed in Figure. 6.

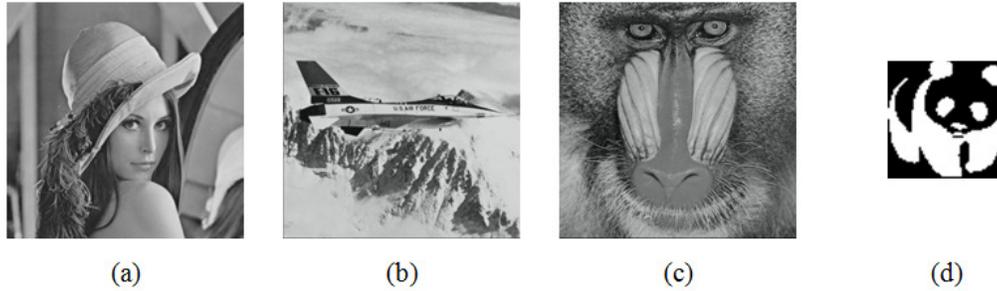
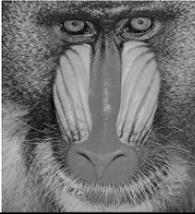
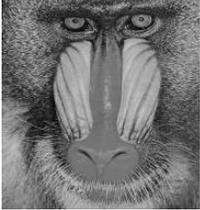
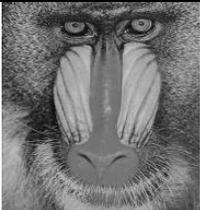
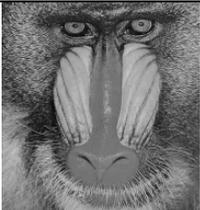
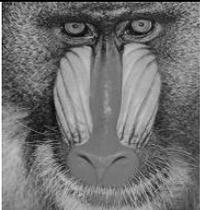


Figure. 6 Test images

Table 2 shows the simulation results of four related works and ours under three host images. Without any attacks, we can obtain the high quality watermarked images from related works. Since the host image has been modified for watermark embedding, there exist some distortions in the recovered image. Thus, we can not get lossless NC and TAF values of the extracted logo. On the contrary, the watermark logo is not really embedded in the host image in the proposed method. This results in keeping a perfect quality of watermarked image. Also, we can obtain complete NC and TAF values of the retrieved logo.

Table 2. The watermarked image and the extracted image without any attacks

		<i>Lena</i>	<i>Baboon</i>	<i>Airplane</i>
Patra et al.'s method [11]	Water-marked image			
	PSNR	43.81 dB	44.71 dB	45.54 dB
	Extracted image			
	NC	0.99	0.99	0.99
	TAF	0.54	0.56	0.46

		<i>Lena</i>	<i>Baboon</i>	<i>Airplane</i>
Lin et al.'s method [8]	Water-marked image			
	PSNR	37.55 dB	38.45 dB	38.15 dB
	Extracted image			
	NC	0.99	0.99	0.99
	TAF	0.02	0.05	0.02
Lai's method [7]	Water-marked image			
	PSNR	45.19 dB	37.3 dB	44.41 dB
	Extracted image			
	NC	0.99	0.96	0.74
	TAF	2	4.42	26
Run et al.'s method [12]	Stego image			
	PSNR	34.4 dB	35.74 dB	34.15 dB
	Extracted image			
	NC	0.85	0.86	0.84
	TAF	6.98	7.62	7.25
Proposed method	Water-marked image			
	PSNR	Infinity	Infinity	Infinity

	<i>Lena</i>	<i>Baboon</i>	<i>Airplane</i>
Extracted image			
NC	1	1	1
TAF	0	0	0

Actually, it is common that an image may suffer from malicious attacks or some signal processing operations during the network transmission or format transformation. Thus, the robustness should be taken into consideration for performance evaluation. First, we simulated the noise attack by increasing Gauss noise to the watermarked image, and the results are displayed in Table 3. It is clear that parts of pixel values may increase or decrease due to the extra noise. Considering the techniques adopted in related works, they mainly use specific transformation, including DCT, DWT, and SVD. The output of each coefficient after the transformation depends on many pixels. Once the level of noise increases, the corresponding modification and the number of affected coefficient become large. This often leads to lower down the correction ratio of extracted watermark bit. In the proposed method, we apply the scale relationship of two independent pixels for watermark embedding; thus leading to higher toleration of pixel modification in the watermarked image. With the help of voting strategy, the error rate of extracted logo can be effectively reduced, and the retrieved results can stay stable even the noise distortion becomes more serious. As shown in Table 3, the proposed method can outperform others in terms of NC value, TAF value, and human vision perception.

Table 4 illustrates the watermarked images under different levels of Gauss blurring. This operation must smooth the whole image and lower down the readability of detailed content. In particular, a small level of blurring will modify a large amount of pixels. And, this is the main reason why the frequency-based techniques can not resist the blurring attack. Nevertheless, we employ the comparison between two independent pixels to form the estimation judgment. Pixels in a pair are separated at a distance so that the difference between two corresponding pixels is usually large. Thus, the absolute difference between two pixels can stay steady under the blurring attack. This has demonstrated the robustness to blurring attack.

The cropping attack can be classified into two types: the inside cropping and the outside one. The inside cropping is mainly used to delete some objects such as face, while the outside one is often applied to cut the meaningless contour area to shorten the image. In the experiments, we used *Lena* and *Airplane* for inside cropping simulation, which are two images containing conspicuous objects. And, we cut the face of *Lena* and the body of *Airplane*. The shape of cropping could be various. For simplicity, we adopted the rectangle. The experimental results are listed in Table 5. In general, it is more difficult for a watermarking technique to withstand the outside cropping attack since the basic reference information for watermark retrieval usually locates at the suburb of the image. The main procedure of cropping is to remove some parts of the image in spatial domain. Actually, it is easy for a selected pixel to locate within the removed area in our proposed method. This must result in the inaccurate watermark extraction. However, the adoption of voting strategy has given a good solution for this weakness. As shown in the table, even the proposed method can not offer an optimal performance in this case; it still yields a recognizable watermark. This has shown that the new method has the capability of resisting the cropping attack.

To highlight the practicability of the new method, we further conducted the simulation to demonstrate its robustness to JPEG compression, which is one of the commonest compression standards

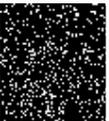
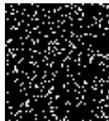
during in the field of network communications. Table 6 provides the comparison results between related works and ours under different levels of compression. The procedure of sampling and quantification is the main technique used to achieve the effective compression in JPEG standard. The adoption of quantification step can help guarantee that we can obtain a high quality compressed image after JPEG algorithm. More precisely, it only slightly varies the pixel values in the spatial domain. Thus, the new method can successfully resist this signal processing operation.

Table 3. The results under different levels of Gauss noise

Noise						
		1%	2%	3%	4%	5%
Patra et al.'s method [11]	Ex-tracted image					
	NC	0.86	0.8	0.7	0.68	0.63
	TAF	7.57	11.62	17.09	17.9	21.61
Lin et al.'s method [8]	Ex-tracted image					
	NC	0.91	0.85	0.77	0.73	0.72
	TAF	4.13	6.67	10.3	12.08	12.23
Lai's method [7]	Ex-tracted image					
	NC	0.93	0.89	0.82	0.78	0.76
	TAF	8.45	12.74	18.77	22.83	23.68

Noise						
		1%	2%	3%	4%	5%
Run et al.'s method [12]	Ex-tracted image					
	NC	0.82	0.8	0.75	0.72	0.68
	TAF	9.64	12.28	17.97	22.79	27.39
	Proposed method	Ex-tracted image				
NC	0.99	0.98	0.96	0.96	0.93	
TAF	1.44	2.42	3.86	4.47	6.08	

Table 4. The results under different levels of Gauss blurring

Blur						
		1 pixel	2 pixels	3 pixels	4 pixels	5 pixels
Patra et al.'s method [11]	Extracted image					
	NC	0.21	0.16	0.13	0.13	0.12
	TAF	39.36	43.02	44.43	44.58	44.82

		Blur				
		1 pixel	2 pixels	3 pixels	4 pixels	5 pixels
Lin et al.'s method [8]	Extracted image					
	NC	0.05	0.01	0	0	0
	TAF	43.12	44.12	44.04	44.04	44.04
Lai's method [7]	Extracted image					
	NC	0.91	0.86	0.73	0.72	0.71
	TAF	12.33	21.78	40.58	48.22	51.15
Run et al.'s method [12]	Extracted image					
	NC	0.76	0.5	0.43	0.41	0.4
	TAF	21.75	37.62	43.65	46.12	47.46
Proposed method	Extracted image					
	NC	0.99	0.98	0.97	0.96	0.96
	TAF	0.73	1.54	2.78	3.66	4.3

Table 5. The results under different sizes of cropping

		Cropping			
		Inside	Inside	Outside 25%	Outside 25%
Patra et al.'s method [11]	Extracted image				

Cropping					
		Inside	Inside	Outside 25%	Outside 25%
NC	NC	0.84	0.91	0.73	0.73
	TAF	7.18	3.86	12.3	12.7
Lin et al.'s method [8]	Extracted image				
	NC	0.8	0.92	0.9	0.67
TAF	TAF	8.72	3.69	4.52	15.06
	Lai's method [7]	Extracted image			
NC		0.75	0.99	0.99	0.97
TAF	TAF	31.57	4.2	24.24	14.6
	Run et al.'s method [12]	Extracted image			
NC		0.74	0.78	0.65	0.67
TAF	TAF	11.87	10.06	16.55	16.82
	Proposed method	Extracted image			
NC		0.92	0.96	0.88	0.87
TAF	TAF	8.23	4.22	12.38	13.72

Table 6. The results under different ratios of JPEG compression

JPEG						
		90%	70%	60%	50%	30%

						
JPEG		90%	70%	60%	50%	30%
Patra et al.'s method [11]	Extracted image					
	NC	0.93	0.89	0.82	0.79	0.69
	TAF	4.2	6.47	9.81	12.67	18.97
Lin et al.'s method [8]	Extracted image					
	NC	0.98	0.93	0.81	0.91	0.66
	TAF	0.98	3.05	8.5	4.08	14.94
Lai's method [7]	Extracted image					
	NC	0.94	0.89	0.86	0.86	0.94
	TAF	6.27	12.23	19.68	26.88	40.31
Run et al.'s method [12]	Extracted image					
	NC	0.83	0.82	0.81	0.81	0.74
	TAF	7.96	9.64	10.6	11.67	16.72
Proposed method	Extracted image					
	NC	0.99	0.99	0.99	0.99	0.99
	TAF	0.42	0.88	0.98	0.98	1.39

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we have integrated XNOR operation and voting strategy to design a watermarking scheme in the spatial domain. Based on the observation that the relativity of pixels usually keeps stable after common signal processing operations or attacks, applying the scale relationship of

two pixels to be the main component of judgment condition can effectively improve the correctness of watermark retrieval. As shown in the simulation results, the new method can resist most of signal processing operations and attacks. Specifically, it can greatly outperform related works in the cases of JPEG compression and Cropping.

REFERENCES

- [1] V. Aslantas, S. Ozer, and S. Ozturk, "Improving the Performance of DCT-based Fragile Watermarking using Intelligent Optimization Algorithms," *Optics Communications*, Vol. 282, No. 14, pp. 2806-2817, 2009.
- [2] M. Barni, F. Bartolini, and A. Piva, "Improved Wavelet-based Watermarking through Pixel-wise Masking," *IEEE Transactions on Image Processing*, Vol. 10, No. 5, pp. 783-791, 2001.
- [3] P. Bao and X. Ma, "Image Adaptive Watermarking using Wavelet Domain Singular Value Decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 1, pp. 96-102, 2005.
- [4] C.C. Chang, C.C. Lin, and C.S. Tseng, and W. L. Tai, "Reversible Hiding in DCT-based Compressed Images," *Information Sciences*, Vol. 177, No. 13, pp. 2768-2786, 2007.
- [5] J. R. Hernandez, M. Amado, and F. Perez-Gonzalez, "DCT-domain Watermarking Techniques for Still Images: Detector Performance Analysis and a New Structure," *IEEE Transactions on Image Processing*, Vol. 9, pp. 55-68, 2000.
- [6] J.R. Kim and Y.S. Moon, "A Robust Wavelet-Based Digital Watermarking Using Level-Adaptive Thresholding," *International Conference on Image Processing*, Vol. 2, pp. 226-230, 1999.
- [7] C.C. Lai, "An Improved SVD-based Watermarking Scheme using Human Visual Characteristics," *Optics Communications*, Vol. 284, pp. 938-944, 2011.
- [8] S.D. Lin, S.C. Shie, and J.Y. Guo, "Improving the Robustness of DCT-based Image Watermarking Against JPEG Compression," *Computer Standards and Interfaces*, Vol. 32, pp. 54-60, 2010.
- [9] G. C. Langelaar, I. Setyawan, and R. L. Lagendijk, "Watermarking Digital Image and Video Data," *IEEE Signal Processing Magazine*, Vol. 17, No. 5, pp. 20-46, 2000.
- [10] W. Lu, W. Sun, and H. Lu, "Robust Watermarking based on DWT and Nonnegative Matrix Factorization," *Computers and Electrical Engineering*, Vol. 35, No. 1, pp. 183-188, 2008.
- [11] J.C. Patra, J.E. Phua, C. Bornand, "A Novel DCT Domain CRT-based Watermarking Scheme for Image Authentication Surviving JPEG Compression," *Digital Signal Processing*, Vol. 20, pp. 1597-1611, 2010.
- [12] R.S. Run, S.J. Horng, W. H. Lin, T.W. Kao, P. Fan, and M. K. Khan, "An Efficient Wavelet-tree-based Watermarking Method," *Expert Systems with Applications*, Vol. 38, No. 12, pp. 14357-14366, 2011.
- [13] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia Data-embedding and Watermarking Technologies," *Proceedings of the IEEE*, Vol. 86, No. 6, pp. 1064-1087, 1998.

INTENTIONAL BLANK

REGION CLASSIFICATION BASED IMAGE DENOISING USING SHEARLET AND WAVELET TRANSFORMS

Preety D. Swami¹, Alok Jain² and Dharendra K. Swami³

¹Department of Electronics & Instrumentation Engineering,
SATI, Vidisha, India
preetydswami@yahoo.com

²Department of Electronics & Instrumentation Engineering,
SATI, Vidisha, India
alokjain6@rediffmail.com

³Department of Computer Science and Engineering, VNSIT, Bhopal, India
dhirendrakswami@gmail.com

ABSTRACT

This paper proposes a neural network based region classification technique that classifies regions in an image into two classes: textures and homogenous regions. The classification is based on training a neural network with statistical parameters belonging to the regions of interest. An application of this classification method is applied in image denoising by applying different transforms to the two different classes. Texture is denoised by shearlets while homogenous regions are denoised by wavelets. The denoised results show better performance than either of the transforms applied independently. The proposed algorithm successfully reduces the mean square error of the denoised result and provides perceptually good results.

KEYWORDS

Classification, Image denoising, Neural Network, Shearlets, Wavelets

1. INTRODUCTION

Extraction of texture regions in an image is required for interpretation of data and is a challenging job. Mostly, the methods that are employed to characterize textures are statistical in nature. Some other methods that extract texture features are those that use Gabor filtering, fractal dimensions, and wavelet transform [1]. The importance of texture detection is important with the perspective of image enhancement, image segmentation and content classification [2]. In this work a classification scheme is proposed that classifies regions in an image into homogenous regions and textures. The effectiveness of this scheme is proved by applying different transformations to the two classified areas in the image. Thus, this algorithm can be used to denoise images using a hybrid of transforms which gives better results than when denoising is done using a single transform.

The wavelet transform has proved to be a powerful tool for image denoising in the past two decades. The pioneering work of Donoho et al. [3] for image denoising paved the way for many

researchers to further exploit the multiresolution transforms for denoising purpose. The traditional wavelet transform, although, is good at denoising point singularities in signals, fails at the line singularities such as edges and at textures present in an image. This has led to the need of developing new transforms that may overcome the limitations posed by wavelets. The flowering of many multiresolution transforms, such as, brushlets [4], wedgelets [5], ridgelets [6], curvelets [7], bandelets [8], contourlets [9], waveatoms [10], shearlets [11] and ripplelets [12] has provided a handful of options for image denoising. However, the selection of a particular transform is a bit difficult, as, each of these transforms perform sparsely in specific areas of an image. A transform providing sparse representation in smooth areas may not provide sparsity at the edges or textures and vice versa.

Researchers are continuously in search of methods and transforms that can denoise the variations in an image with perfect reconstruction. For the past few years, denoising techniques, based on combination of multiple transforms have evolved. In [13], Ma et al. proposed a denoising algorithm to perform pixel fusion to result images of curvelets and wavelets approaches. The noisy image is denoised using curvelets as well as using HMT based wavelets. Then image regions are analysed with quadtree decomposition. Weighted pixel fusion method is employed to obtain the final result image.

The discrete curvelet transform can code image edges more efficiently than the wavelet transform [14-16]. On the other hand, wavelet transform, codes homogenous areas better than curvelet transform. In [14] two combinations of time invariant wavelet and curvelet transforms are used for denoising of SAR images. Both methods use the wavelet transform to denoise homogeneous areas and the curvelet transform to denoise areas with edges. The segmentation between homogeneous areas and areas with edges is done by using total variation segmentation. In [16] the areas containing edges are denoised using spatially adaptive context modelling of curvelet transform coefficients, while the remaining homogenous regions are recovered through spatially adaptive context modelling of wavelet transform coefficients. The areas containing edges and those that do not contain edges are segmented in the space domain by calculating a variance image and then thresholding it. In [17], three combinations of undecimated wavelet and nonsubsampling contourlet transforms are used for denoising of SAR images. Two methods use the wavelet transform to denoise homogeneous areas and the nonsubsampling contourlet transform to denoise areas with edges. The segmentation between homogeneous areas and areas with edges is done by using total variation segmentation. The third method is a linear averaging of the two denoising methods. A thresholding in the wavelet and contourlet domain is done by non-linear functions which are adapted for each selected subband.

Authors in [18] combine wavelet transform with both the ridgelet and the curvelet transform. The residual image gives the information about the efficiency of the method as no features are seen in it. In [19] BayesShrink wavelet is combined with BayesShrink ridgelet denoising method which performs better than each method individually. The proposed combined denoising method gains the advantage of each filter in its specific domain, i.e., wavelet for natural and ridgelet for straight regions, and produces better and smoother results, both visually and in terms of SNR.

The work in [20] utilizes features of wavelet and curvelet transform, separately and adaptively, in different regions of an image, which are identified by variance approach. The homogenous regions are denoised by wavelets and edgy information is obtained with curvelet transform. The spatially adaptive fusion technique fuses the denoised information obtained from the two transforms.

Authors in [21] proposed a multiscale and multidirectional image representation method named CBlet transform. It combines the contourlet transform with the bandeletization procedure. The contourlet transform captures image discontinuous points and links them into linear structures.

These linear structures are analysed adaptively by the bandeletization procedure and removes their correlation. The results of the fusion of denoised data from brushlet and wavelet thresholding methods are presented in [22]. Texture-based brushlet denoising is well suited for enhancement of physiological information while wavelet-based denoising is better suited for enhancement of anatomical contours. A three-dimensional multiscale edge-based data fusion algorithm is applied to combine enhanced data from these two independent denoising methods. In [23] a method is proposed for denoising in which firstly, the DTCWT is employed to obtain subbands, and then bandeletization is implemented in each subband. At last, Bayes soft-threshold shrinkage denoising in bandelet transform domain is implemented. Image with highly directional can be efficiently denoised by this method.

Wavelets and compactly supported shearlets sparsely represent point and curvilinear singularities, respectively. [24] presents an image separation method for separating images into point and curvelike parts by employing a combined dictionary consisting of wavelets and shearlets. In this, it is assumed that noise cannot be represented sparsely by either one of the two representation systems. Thus, noise can be captured in the residual.

In one of our previous works [25], an image denoising method which adaptively combines the features of wavelets, wave atoms and curvelets was proposed. It employs wavelet shrinkage to denoise the smooth regions in the image while wave atoms are employed to denoise the textures and the edges take advantage of curvelet denoising.

This paper proposes a classification technique for segmenting an image into two categories: homogenous regions and texture. The proposed classification algorithm is based on training a neural network using samples from images and some statistical measures pertaining to the two above mentioned categories present in a natural image. The neural network classifier accurately assigns the classes to the different regions. Once the classification is done the results can be used to denoise a noisy image. This work employs a combination of wavelets and shearlets for the purpose of denoising.

The organization of the paper is as follows. Section 2 deals with the proposed region classification algorithm using neural network, along with, an application of the algorithm for image denoising. Experimental results are analysed in Section 3. Finally conclusion is given in Section 4.

2. PROPOSED METHOD

In image denoising, identification of smooth, texture and other regions is a frequent requirement. Various methods have been proposed in the literature for separating smooth, texture and other regions present in the image. Texture is different from smooth areas in that they have some randomness in location, size and orientation of the texture elements [2]. Several texture descriptors are available for identifying textures. These include Gray Level Co-occurrence Matrix (GLCM), contrast, directionality, placement rules, Markov Random Field models, and filtering in the transform domain [1]. This task of classification can also be attained by machine learning in a supervised manner. In this section an attempt has been made to use neural network for texture and smooth region identification.

In this proposed classification method, training of a neural network is required which classifies an incoming data (image block) to either smooth class or texture class. For this task, the neural network is trained on some blocks of texture patches and some blocks of smooth region patches. These patches can be extracted from the image of *Barbara* as shown in Figure 1. In this figure, the region inside the red rectangles, are used to train the texture class and the areas inside the blue

squares are used to provide training of smooth regions. Each patch is divided into blocks of size 9×9 , and these 81 pixel intensity samples are provided as inputs to the neural network.

Along with the 81 pixel intensity values of an image block (texture or smooth), three texture descriptors are also used to train the neural network. These descriptors are variance, contrast and connected component count. Each descriptor is calculated for a window size of 9×9 .



Figure 1. Sample extraction for neural network training. Pixels inside the red rectangle train the texture class and pixels inside the blue square train the class corresponding to smooth regions.

The variance inside a window is calculated as the square of the standard deviation. For the calculation of connected components the pixels inside the 9×9 window are divided into two groups. For this the average grey level is computed and the pixel intensity greater than the average are assigned to the first group and the remaining pixels belong to the second group. A binary image of the 9×9 block is created in which the group one pixels are set to zero and group two pixels are set to 1. Finally connected components are computed by using the Matlab command 'bwlablel'. Contrast can be calculated by calculating the difference between the averages of the above discussed two groups of pixels. All the three descriptors are calculated for the Barbara image and are depicted in Figure 2.

The use of classification by the proposed method is applied for denoising of *Barbara* image. After classification the image regions are classified into two regions, smooth and texture. These regions will be denoised by two different transforms according to the sparse behaviour offered by them in the respective regions. Wavelets efficiently denoise the smooth regions but fail at the edges and in texture regions as they are less sparse in these regions. Shearlets are sparser than wavelets for reconstructing edges and texture but, at the same time, introduce artifacts in the smooth regions. Thus in this work wavelets denoise the smooth regions and shearlets denoise the textures. The noisy image is denoised first by wavelets [26] and then by shearlets [11]. The proposed classification scheme is applied to the wavelet denoised image for region classification and its result is used to fuse the two denoised images. The regions belonging to smooth areas are replaced by the respective wavelet denoised pixels and the regions belonging to texture are replaced by the corresponding shearlet denoised pixels. Thus both transforms work individually and effectively in different regions of the image.

3. EXPERIMENTS AND RESULTS

Feed Forward neural network is commonly used for classification task using back propagation algorithm. In our experiments, a feed forward neural network has been trained to identify a small image window for its smooth or textured nature. For this purpose experiments have been conducted on the popularly used image of *Barbara*. A set of 9×9 pixel subsets of *Barbara* have been taken and have been labelled as smooth or texture by human observation. Sufficiently large set of such training samples have been set up to achieve supervised classification using

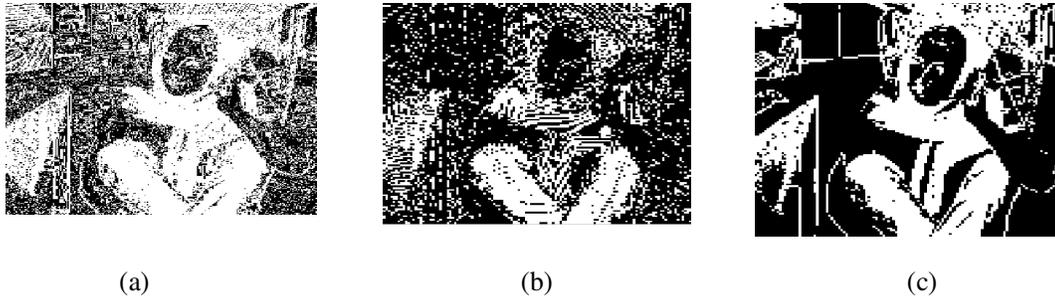


Figure 2. Texture descriptors for *Barbara* (a) Contrast, (b) Component Count and (c) Variance.

neural network. In addition to the image intensity pixel values of training data's 9×9 image window subsets, features like variance, contrast and connected component count have also been extracted for training data sample to increase distinguishing ability of different regions. Algorithms employed for extraction of texture descriptors used in this work are explained in detail in [2].

Topology of the network used is explained below:

- Intensity of the image pixel values and the texture descriptors features of the training sample (9×9 block) form the input vector to the neural network. For these training samples the target output is known in the form of '0' for *smooth* regions and as '1' for *texture* regions.
- Number of layers = 2

Number of neurons in the input layer = 84. Out of these, 81 neurons correspond to the normalized pixel intensity values of the 9×9 window. Remaining 3 neurons correspond to the three texture descriptor features: variance, contrast and connected component count.

- Output layer contains 1 node that outputs a '0' for the class *smooth regions* or a '1' for the class *texture regions*.
- Hidden layer nodes = 6.

This experiment quite satisfactorily trained the various 9×9 sub images of *Barbara*. This was tested by simulating the neural network on the training data, which gave the same desired output. Experiments were conducted in Matlab 7.0. The transfer function of all neurons was chosen as 'tansig'. The backpropagation network training function chosen was 'trainlm' and the backpropagation weight/bias learning function used was 'learngdm'. Testing was done by varying the number of hidden layer nodes from 3 to 60. Best results were obtained while taking 6 neurons in the hidden layer. In this training network, with 6 nodes in the hidden layer, the mean squared error (MSE) was of the order of 10^{-9} .



Figure 3. Denoising results of Barbara by different methods. (a) Original (b) Noisy Barbara with noise standard deviation 20 (c) Wavelets [26], PSNR= 29.53dB (d) Shearlets [11], PSNR=28.54dB (e) Wave atoms [10], PSNR=29.31dB (f) Proposed method, PSNR= 29.83dB

To compare the application of the proposed region classification method in image denoising, a 512×512 sized white Gaussian noise is added to the *Barbara* image of same size. The software for shearlets has been downloaded from [27]. The parameter employed for comparison of the denoised results is the Peak Signal to Noise Ratio (PSNR). In Figure 3, the proposed results are compared with the denoising results employing individually the wavelets, the shearlets as well as the waveatoms, which is considered for its high efficiency in denoising textures. The software for waveatoms has been downloaded from [28]. It can be observed from the figure that the proposed method yields the best PSNR. The element like artifacts present in shearlet and waveatom denoising are not present in the proposed denoising method and at the same time textures are visible with clarity.

4. CONCLUSIONS

In this paper, a classification method to separate texture from smooth regions of an image is proposed. For classification into the two above mentioned regions, a neural network is trained with sample parameters taken from smooth and texture images. The parameters selected to train the network are the pixel intensities inside a small window and variance, contrast and the connected component count of the same window. Testing of the network provides successful separation of the smooth and the textured regions.

The effectiveness of the algorithm is tested by applying a wavelet-shearlet combination for denoising of natural and texture rich *Barbara* image. The proposed denoised results are compared with the results of denoising the image individually by wavelets, shearlets and wave atoms. It is observed that using the proposed classification technique in denoising of images, improves the

PSNR significantly and results in a perceptually cleaner image as compared to employing any of these transforms in individuality.

Future research includes selection of a vivid variety of samples of each category from a large group of natural images for training of the network. The training function and training parameters of the network can also be changed and tested. The effect of a change in window size for more efficient texture descriptors can also be observed.

REFERENCES

- [1] L. Semler & L. Dettori, (2006) "Curvelet-based texture classification of tissues in computed tomography", IEEE International Conference on Image Processing (ICIP06), Atlanta, GA, USA, 8-11 Oct., pp 2165-2168.
- [2] R. Bergman, H. Nachlieli, & G. Ruckenstein, (2008) "Detection of textured areas in natural images using an indicator based on component counts", J. Electronic Imaging, Vol. 17, No. 4, pp 043003.
- [3] D. L. Donoho & I. M. Johnstone, (1994) "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 3, pp 425-455.
- [4] F. G. Meyer & R. R. Coifman, (1997) "Brushlets: A tool for directional image analysis and image compression", *Applied and Computational Harmonic Analysis*, Vol. 4, pp 147-187.
- [5] D. L. Donoho, (1999) "Wedgelets: Nearly minimax estimation of edges", *The Annals of Statistics*, Vol. 27, No. 3, pp 859-897.
- [6] E. J. Candes & D. L. Donoho, (1999) "Ridgelets: a key to higher-dimensional intermittency?", *hil. Trans. R. Soc. London. A.*, Vol. 357, No. 1760, pp 2495-2509.
- [7] J. L. Starck, E. J. Candes & D. L. Donoho, (2002) "The curvelet transform for image denoising", *IEEE Trans. Image Process.*, Vol. 11, No. 6, pp 670-684.
- [8] E. L. Pennec & S. Mallat, (2005) "Sparse geometric image representations with bandelets", *IEEE Trans. Image Process.*, Vol. 14, No. 4, pp 423-438.
- [9] M. N. Do & M. Vetterli, (2005) "The contourlet transform: an efficient directional multiresolution image representation", *IEEE Transactions on Image Process.*, Vol. 14, No. 12, pp 2091-2106.
- [10] L. Demanet & L. Ying, (2007) "Wave atoms and sparsity of oscillatory patterns", *Applied and Computational Harmonic Analysis*, Vol. 23, No. 3, pp 368-387.
- [11] K. Guo & D. Labate, (2007) "Optimally sparse multidimensional representation using shearlets", *SIAM J. Math Anal.*, Vol. 39, pp 298-318.
- [12] J. Xu, L. Yang & D. Wu, (2010) "Ripplelet: A new transform for image processing", *Journal of Vis. Commun. Image R.*, Vol. 21, pp 627-639.
- [13] L. Ma, J. Ma & Y. Shen, (2007) "Pixel fusion based curvelets and wavelets denoise algorithm", *Engineering Letters*, Vol. 14, No. 2, pp 130-134.
- [14] J. R. Sveinsson & J. A. Benediktsson, (2007) "Combined wavelet and curvelet denoising of SAR images using TV segmentation", *Proc. of IEEE International Geoscience and Remote Sensing Symposium*, pp 503-506.
- [15] Y. Li, S. Zhang & J. Hu, (2010) "Combining curvelet transform and wavelet transform for image denoising", *Proc. of 6th International Conference on Intelligent Computing*, pp 317-324.
- [16] P. D. Swami & A. Jain, (2012) "Segmentation based combined wavelet-curvelet approach for image denoising", *International Journal of Information Engineering*, Vol.2, No. 1, pp 32-37.
- [17] J. R. Sveinsson & J. A. Benediktsson, (2008) "Combined wavelet and contourlet denoising of SAR images", *Proc. of IEEE International Geoscience and Remote Sensing Symposium*, pp 1150-1153.
- [18] J. L. Starck, D. L. Donoho & E. Candes, (2001) "Very high quality image restoration by combining wavelets and curvelets", *Proc. of SPIE conference on Signal and Image Processing: Wavelets: Applications in Signal and Image Processing IX*, Vol. 4478, pp 9-19.
- [19] N. N. Kachouie & P. Fieguth, (2006) "A combined Bayesshrink wavelet-ridgelet technique for image denoising", *Proc. of IEEE International Conference on Multimedia and Expo*, Toronto, Ont., pp 1917-1920, 9-12 July.
- [20] G. G. Bhutada, R. S. Anand & S. C. Saxena, (2011) "Edge preserved image enhancement using adaptive fusion of images denoised by wavelet and curvelet transform", *Digital Signal Processing*, Vol. 21, pp 118-130.

- [21] B. Song, L. Xu & W. Sun, (2007) "Image denoising using hybrid contourlet and bandelet transforms", IEEE Proc. of Fourth International Conference on Image and Graphics, Chengdu, Sichuan, China, pp 71-74.
- [22] A. Laine, E. D. Angelini, Y. Jin, P. D. Esser & R. V. Heertum, (2004) "Fusion of brushlet and wavelet denoising methods for nuclear images", International Symposium on Biomedical Imaging (ISBI), Arlington, VA, USA, pp 1187-1191.
- [23] Z. Song & L. Yuanpeng, (2009) "A novel image denosing scheme via combining dual-tree complex wavelet transform and bandelets", Proc. of Third International Symposium on Intelligent Information Technology Application, Nanchang, China, Vol. 1, pp 509-512.
- [24] G. Kutyniok & W.Q Lim, (2012) "Image separation using wavelets and shearlets", curves and surfaces, Lecture Notes in Computer Science, Springer, Vol. 6920, pp 416-430.
- [25] P. D. Swami & A. Jain, (2012) "Image denoising by adaptive fusion of decomposed images restored using wave atom, curvelet and wavelet transform", Published online in Springer Journal of Signal, Image and Video Processing, DOI: 10.1007/s11760-012-0343-z.
- [26] A. Pizurica & W. Philips, (2006) "Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising", IEEE Trans. Image Process., Vol. 15, No. 3, pp 654-665.
- [27] www.shearlab.org
- [28] www.waveatom.org

AUTHORS

Preety D Swami received the B.E. degree in Electronics and Instrumentation from Samrat Ashok Technological Institute, Vidisha, in 1992, the M.Tech. degree in Digital Communication from Maulana Azad National Institute of Technology, Bhopal, in 2008 and the Ph. D. degree in Electronics engineering from RGPV, Bhopal in 2013. She has 15 research publications in journals and conference proceedings. She has received two best paper awards in conferences. She is a member of IET and member for life, ISTE. Her research interests include transforms, signal processing and image processing.



Alok Jain was born in Vidisha, India in 1966. He received his B.E. (Electronics & Instrumentation) degree from Samrat Ashok Technological Institute, Vidisha, M.Tech. (Computer Science and Technology) from IIT Roorkee (University of Roorkee), in 1988 and 1992, respectively. He obtained his Ph.D. degree from Thapar University (erstwhile Thapar Institute of Engineering and Technology), Patiala, India, in 2006. He is presently serving as a Professor and Head in the Department of Electronics & Instrumentation Engineering, Samrat Ashok Technological Institute, Vidisha, India. He has published more than 50 papers in journals and conference proceedings of international repute.



He authored two monographs related to filterbank and Transmultiplexer and two books on power electronics. He co-chaired the session in Int. Conf. SCI 2004 held at Orlando, USA. Dr. Jain is a life member of IE(I), IETE, ISTE, BMESI, Instrument Society of India and is the member of IET. His current research interests include digital signal processing, multirate signal processing, filterbanks, and their applications in image processing.

Dhirendra K Swami is presently Group Director, VNS Group of Institutions, Bhopal. Before joining VNS in 2008, he has served Samrat Ashok Technological Institute (SATI), Vidisha for 20 years. He obtained his Ph. D. in Computer Science and Engineering in 2007 from Rajiv Gandhi Technological University, Bhopal. He completed M. Tech in Computer Applications from IIT, Delhi in 1994 and he completed M.Sc. in Applied Mathematics in 1988. During 25 years of teaching, he has taught various courses in computer science to the students of BE, M. Tech. and MCA students. He has more than 30 research publications in journals and in conference proceedings of repute. He has edited one book and authored a book on Basic Computer Engineering. He is a member for life ISTE and senior life member of CSI. Areas of his special interest include Data Mining, DBMS, Object Oriented System and software Engineering.

SYNTHETICAL ENLARGEMENT OF MFCC BASED TRAINING SETS FOR EMOTION RECOGNITION

Inma Mohino-Herranz¹, Roberto Gil-Pita¹, Sagrario Alonso-Diaz² and
Manuel Rosa-Zurera¹

¹Department of Signal Theory and Communications, University of Alcala,
Spain

inmaculada.mohino@edu.uah.es, roberto.gil@uah.es,
manuel.rosa@uah.es

²Human Factors Unit, Technological Institute “La Marañosa” –MoD, Madrid
(Spain)

salodia@et.mde.es

ABSTRACT

Emotional state recognition through speech is being a very interesting research topic nowadays. Using subliminal information of speech, it is possible to recognize the emotional state of the person. One of the main problems in the design of automatic emotion recognition systems is the small number of available patterns. This fact makes the learning process more difficult, due to the generalization problems that arise under these conditions.

In this work we propose a solution to this problem consisting in enlarging the training set through the creation of new virtual patterns. In the case of emotional speech, most of the emotional information is included in speed and pitch variations. So, a change in the average pitch that does not modify neither the speed nor the pitch variations does not affect the expressed emotion. Thus, we use this prior information in order to create new patterns applying a pitch shift modification in the feature extraction process of the classification system. For this purpose, we propose a frequency scaling modification of the Mel Frequency Cepstral Coefficients, used to classify the emotion. This proposed process allows us to synthetically increase the number of available patterns in the training set, thus increasing the generalization capability of the system and reducing the test error.

KEYWORDS

Enlarged training set, MFCC, emotion recognition, pitch analysis

1. INTRODUCTION

Emotional state recognition (ESR) through speech is being a very interesting research topic nowadays. Using subliminal information of speech, it is possible to recognize the emotional state of the person. This information, denominated “prosody”, reflects some features of the speaker and adds information to the communication [1], [2].

The standard scheme of an ESR system consists of a feature extraction stage followed by a classification stage. Some of the most useful features used in speech-based ESR systems are the Mel-Frequency Cepstral Coefficients (MFCCs), which are one of the most powerful features used in speech information retrieval [3]. The classification stage uses artificial intelligence techniques to learn from data in order to determine the classification rule. It is important to highlight that in order to avoid loss of generalization of the results, it is also necessary to split the available data in two sets, one for training the system and other for testing it, since the data must be different in order to avoid loss of generalization of the results.

One of the main problems in the design of automatic ESR systems is the small number of available patterns. This fact makes the learning process more difficult, due to the generalization problems that arise under these conditions [4], [5].

A possible solution to this problem consists in enlarging the training set through the creation the new virtual patterns. This idea, originally proposed in [6], consists in the use of auxiliary information, denominated hints, about the target function to guide the learning process. The use of hints have been proposed several times in several applications, like, for instance, automatic target recognition [7], or face recognition [8].

In the case of emotional speech, it is important to highlight that most of the information is included in speed and pitch variations [9]. So, a change in the average pitch value that does not modify neither the speed nor the pitch variations does not affect the expressed emotion.

In this work we propose the creation of new patterns by applying a pitch shift modification in the feature extraction process of an ESR system. For this purpose, we propose a frequency scaling modification of the MFCCs. This proposed process allows us to synthetically increase the number of available patterns in the training set, thus increasing the generalization capability of the system and reducing the test error.

2. MATERIALS AND METHODS

This section explains the two main stages of an ESR system: the feature extraction stage and the classification stage, describing the configuration of the ESR system used in the experiments.

2.1. Feature extraction: Mel-Frequency Cepstral Coefficients (MFCCs)

Obtaining MFCC coefficients [10] has been regarded as one of the techniques of parameterization most important used in speech processing. They provide a compact representation of the spectral envelope, so that most of the energy is concentrated in the first coefficients. Perceptual analysis emulates human ear non-linear frequency response by creating a set of filters on non-linearly spaced frequency bands. Mel cepstral analysis uses the Mel scale and a cepstral smoothing in order to get the final smoothed spectrum. Figure 1 shows the scheme for the MFCC evaluation.

The main stages of MFCC analysis are:

- *Windowed*: In order to overcome the non-stationary of speech, it is necessary to analyze the signal in short time periods, in which it can be considered almost stationary. So, time frames or segments are obtained dividing the signal. This process is called windowed. In order to maintain continuity of information signal, it is common to perform the windowed sample with frame blocks overlap one another, so that the information is not lost in the transition between windows.
- *DFT*: Following the windowed, DFT is calculated to $x_t[n]$, the result of windowing the t -th time frame with a window of length N .

$$X_t[k] = \sum_{n=0}^{N-1} x_t[n] \cdot e^{-j2\pi nk}, 0 \leq k \leq N-1 \quad (1)$$

From this moment, phase is discarded and we work with the energy of speech signal, $|X_t[k]|^2$.

- *Filter bank*: The signal $|X_t[k]|^2$ is then multiplied by a triangular filter bank, using Equation (2).

$$E_{nt} = \sum_{k=0}^{N/2} |X_t[k]|^2 H_m[k], \quad 1 \leq m \leq F \quad (2)$$

where $H_m[k]$ are the triangular filter responses, whose area is unity. These triangles are spaced according to the MEL frequency scale. The bandwidth of the triangular filters is determined by the distribution of the central frequency $f[m]$, which is function of the sampling frequency and the number of filters. If the number of filters is increased, the bandwidth is reduced.

So, in order to determine the central frequencies of the filters $f[m]$, the behaviour of the human psychoacoustic system is approximated through $B(f)$, the frequency in MEL scale, in Equation (3).

$$B(f) = 2595 \cdot \log(1 + f/700) \quad (3)$$

where f corresponds with the frequency represented on a linear scale axis.

Therefore, the triangular filters can be expressed using Equation (4).

$$H_m[k] = \begin{cases} 0, & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, & f[m-1] \leq k < f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, & f[m] \leq k < f[m+1] \\ f[m+1] - f[m], & k \geq f[m+1] \end{cases} \quad (4)$$

where $1 \leq m \leq F$, being F the number of filter, and furthermore we have the central frequency $f[m]$ of the m -th frequency band :

$$f[m] = \frac{N}{F_s} B^{-1} \left(m \frac{B(F_s/2)}{M+1} \right) \quad (5)$$

where $B^{-1}(b) = 700(e^{b/2595} - 1)$, and F_s is the frequency sampling.

- *DCT (Discrete Cosine Transform)*: Through the DCT, expressed in Equation (6), the spectral coefficients are transformed to the frequency domain, so the spectral coefficients are converted to cepstral coefficients.

$$MFCC_{nt} = \sum_{k=1}^F \log(E_{nk}) \cos(n(k-1/2)\pi/N), \quad n = 1, \dots, F \quad (6)$$

One MFCCs are evaluated, features are determined from statistics of each MFCC. Some of the most common used statistics are the mean and the standard deviation. In is also habitual to use statistics from differential values of the MFCCs, denominated, delta MFCC, or Δ MFCCs. These Δ MFCCs are determined using Equation (7),

$$\Delta MFCC_{nt} = MFCC_{nt} - MFCC_{n(t-d)} \quad (7)$$

where d determines the differentiation shift. In this paper we use as features the mean and standard deviation of the MFCCs, and the standard deviation of Δ MFCCs with $d = 2$, since we have found that these values obtain very good results with a considerably low number of features.



Figure 1: Scheme to MFCC calculate

2.2. Classification stage: Least Square Diagonal Quadratic Classifier

The Least Square Diagonal Quadratic Classifier is a classifier that renders very good results with a very fast learning process [11] and therefore it has been selected for the experiments carried out in this paper. Let us consider a set of training patterns $x = [x_1, x_2, \dots, x_L]^T$, where each of these patterns is assigned to one of the possible classes denoted as $C_i, i = 1, \dots, k$. In a quadratic classifier, the decision rule can be obtained using a set of k combinations, as shows Equation (8)

$$y_k = w_{k0} + \sum_{n=1}^L w_{kn}x_n + \sum_{n=1}^L \sum_{m=1}^n x_m x_n v_{mnk} \quad (8)$$

where w_{kn} and v_{mnk} are the linear and quadratic values weighting respectively. Furthermore, Equation (8) can be expressed in matrix notation as shown in Equation (9).

$$\mathbf{y} = \mathbf{w}_0 + \mathbf{W}^T \mathbf{x} + \mathbf{V}^T \mathbf{x} \quad (9)$$

The particular case of a diagonal quadratic classifier is referred to the use of only the diagonal coefficients of \mathbf{V} . This leads to a simplification of Equation (8), giving Equation (10).

$$v_{mnk} = 0, \quad \forall m \neq n \quad (10)$$

With this last equation, the decision rule is obtained as shows Equation (11).

$$y_k = w_{k0} + \sum_{n=1}^L w_{kn}x_n + \sum_{n=1}^L x_n^2 v_{nnk} \quad (11)$$

The pattern matrix \mathbf{Q} , which contains the input features for classification and his quadratic value, is expressed in Equation (12).

$$\mathbf{Q} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{L1} & x_{L2} & x_{L3} & \dots & x_{LN} \\ x_{11}^2 & x_{12}^2 & x_{13}^2 & \dots & x_{1N}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{L1}^2 & x_{L2}^2 & x_{L3}^2 & \dots & x_{LN}^2 \end{bmatrix} \quad (12)$$

Being, \mathbf{V} as is expressed the Equation (13)

$$\mathbf{V} = \begin{bmatrix} w_{10} & w_{11} & \dots & w_{1L} & v_{111} & \dots & v_{1LL} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{k0} & w_{k1} & \dots & w_{kL} & v_{k11} & \dots & v_{kLL} \end{bmatrix} \quad (13)$$

So, the output of the quadratic classifier is obtained according to Expression (14).

$$\mathbf{Y} = \mathbf{V} \cdot \mathbf{Q} \quad (14)$$

Let us now define the target matrix containing the labels of each pattern as:

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & \dots & t_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{K1} & t_{K2} & t_{K3} & \dots & t_{KN} \end{bmatrix} \quad (15)$$

where N is the number of data samples, and $t_{kn}=1$ if the n -th pattern belongs to class C_k , and 0 in other case. Then, the error is the difference between the outputs of the classifier and the correct values, which are contained in the target vector:

$$\mathbf{E} = \mathbf{Y} - \mathbf{T} = \mathbf{V} \cdot \mathbf{Q} - \mathbf{T} \quad (16)$$

Consequently, the mean square error (MSE) is computed according to Equation (17).

$$MSE = \frac{1}{N} \|\mathbf{Y} - \mathbf{T}\|^2 = \frac{1}{N} \|\mathbf{V} \cdot \mathbf{Q} - \mathbf{T}\|^2 \quad (17)$$

In the least squares approach, the weights are adjusted in order to minimize the mean square error. The minimization of the MSE is obtained deriving expression (17) with respect \mathbf{V} and, using the equations of *Wiener-Hopf*, the next expression for the weight values is obtained:

$$\mathbf{V} = \mathbf{T} \cdot \mathbf{Q}^T \cdot (\mathbf{Q} \cdot \mathbf{Q}^T)^{-1} \quad (18)$$

This expression allows to determine the values of the coefficients that minimize the mean square error for a given set of features.

3. PROPOSED MFCC-BASED ENLARGEMENT OF THE TRAINING SET

As we stated in the introduction, it is important to highlight that most of the information of emotional speech is included in speed and the pitch variations [9]. So, an average change in the pitch value that does not modify neither the speed nor the pitch variations does not affect the expressed emotion.

In this paper we propose to modify the MFCC extraction in order to implement frequency scaling, allowing to create new patterns for the training set. So, the MFCCs can be easily pitch-shifted through a scale factor applied in frequency domain. This modification is applied to each pattern in the database, allowing to enlarge the training set.

Let us define the Pitch Shift Factor (P_{SF}) as a global change of the pitch, measured in semitones. Then, this shift in the pitch is equivalent to scaling the frequency with a Frequency Scale Factor (F_{SF}). So, the relationship between P_{SF} and F_{SF} can be expressed using Equation (19).

$$F_{SF} = 2^{\frac{P_{SF}}{12}} \quad (19)$$

In order to apply this frequency scaling in the MFCC process, the central frequencies $f[m]$ of the triangular filters are modified, taking into account the scaled frequency factor. So, in Equation (20) we can observe the relationship between the original and synthetic frequency.

$$f'[m] = F_{SF} \cdot f[m] \quad (20)$$

Being the new frequency scale, as shows in Equation (21)

$$f'[m] = F_{SF} \cdot \frac{N}{F_s} B^{-1} \left(m \frac{B(F_s/2)}{M+1} \right) \quad (21)$$

Figure 2 shows in linear axis, that is, the frequency is scaled the central frequency for each MFCC. In this Figure, we can observe the difference between the Normal relationships between center frequency for each coefficient, and the difference when the frequency has been scaled, that is, the center frequency is reduced when increase the number of cepstral coefficients.

As an example, the difference between the MFCCs calculate with $P_{SF}=0$ and $P_{SF}=1$ are shown in Figure 3. So, we can observe that the filter responses in logarithmic scale without frequency shift of MFCCs with a shifting in frequency of one semitone.

In order to implement the enlargement of the database using pitch shifting, two factors must be taken into account: the range of the pitch shifting (R) and the step of the pitch shifting (S).

- *Range (R)*: The range defines the maximum absolute variation in the pitch modification process in semitones. With this parameter it is possible to change the upper and lower limits of the shift variations.

- *Step (S)*: The step defines the smallest change in the pitch that is produced in the pitch shifting process in semitones.

Taking into account these two factors, it is possible to determine the enlargement factor (EF), that is, the number of times that the size of the training set is increased.

$$EF = \frac{2R}{S} + 1 \tag{22}$$

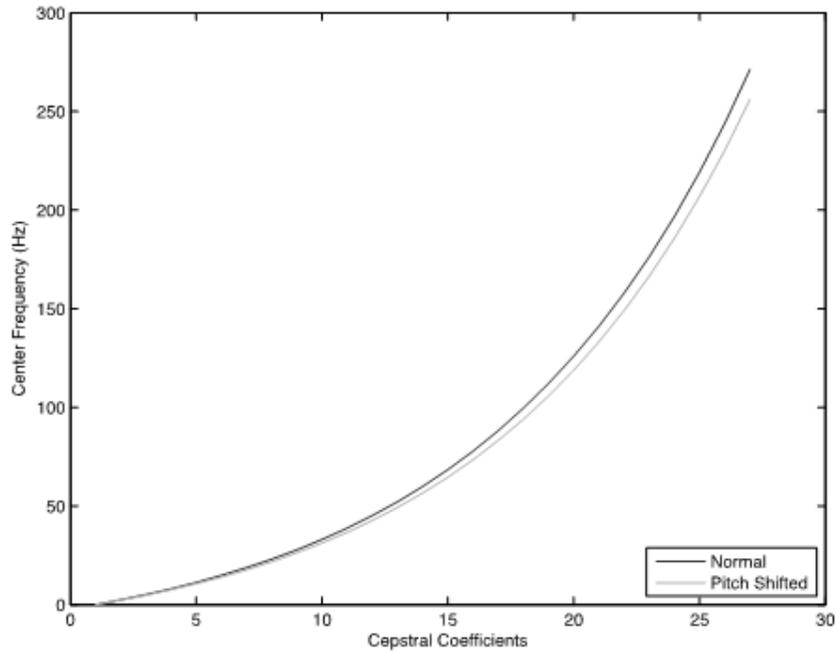


Figure 2: MFCCs for different factors. Lineal Scale.

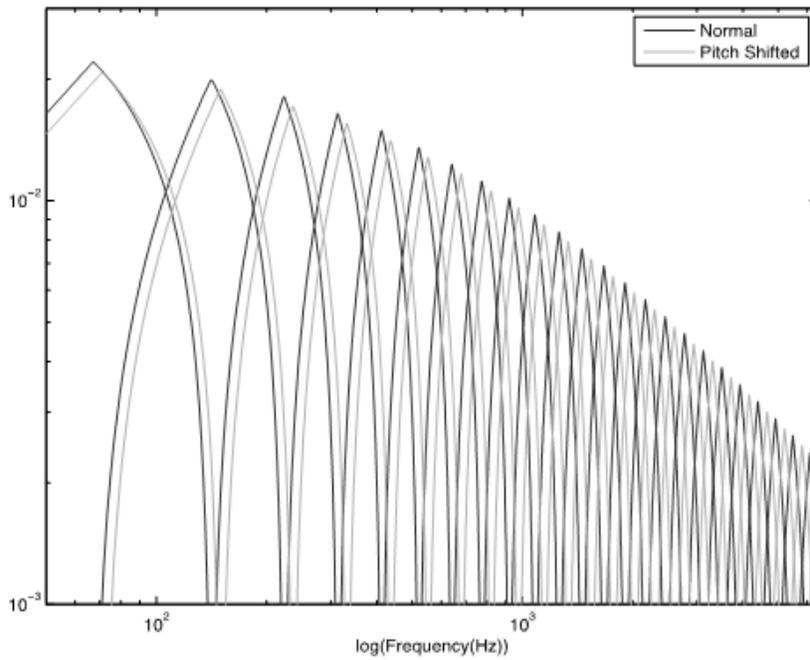


Figure 3: MFCCs for different factors. Logarithmic Scale.

4. RESULTS

4.1. Experimental setup

In this study, we have used the public database "The Berlin Database of Emotional Speech" [12]. This database consists of 800 files of 10 actors (5 males and 5 females), where each actor produces 10 German utterances (5 short and 5 longer sentences) simulating seven different emotions. These emotions are: Neutral, Anger, Fear, Happiness, Sadness, Disgust, and Boredom. The recordings were using a sampling frequency of 48 kHz and later downsampled to 16 kHz. Although this database consists of 800 files, almost 300 were eliminated, since only those utterances with a recognition rate better than 80% and naturalness better than 60% were finally chosen. So, the database consists of 535 files.

In order to evaluate the results and to ensure that they are independent of the partition between training set and test set, we have used the validation method denominated Leave One Out [13] [14]. This is a model validation technique to evaluate how the results of a statistical analysis generalize to an independent data set. This method is used in environments where the main goal is the prediction and we want to estimate how accurate is a model that will be implemented in practice.

This technique basically consists in three stages:

- First, the database is divided into complementary subsets called: training set and test set, where the test sets contains only one pattern in the database.
- Then, the parameters of the classification system are obtained using the training set.
- Finally, the performance of the classification system is obtained using the test set.

In order to increase the accuracy of the error estimation while maximizing the size of the training set, multiple iteration of this process are performed using a different partitions each time, and the test results are averaged over the different iterations.

In this paper we use an adaptation of this technique to the problem at hand, which we denominate *Leave One Couple Out*. So, we have worked with the database discussed above, which consists of 5 male and 5 female. In this case, we used 4 male and 4 female for each training set and 1 male and 1 female for each test set. This division guarantees complete independence between training and test data, keeping a balance in the gender. Therefore, our leave one couple out, is repeated 25 times, using each iteration different training and test sets. □Concerning the features, a window size of $N = 512$ has been used, which implies time frames of 32ms. We have then selected mean and standard deviation of 25 MFCCs, and standard deviation of $2-\Delta$ MFCCS, resulting in a total of 75 features, which has been used to design a quadratic classifier. □In order to complete the comprehension of the results obtained, it is necessary to analyze the error probability for training set, the error probability for test set and the enlargement factor (EF). Table 1 shows the error probability for the training set. □And in Table 3 it is possible to observe that the enlargement factor is 1 for the lowest error probability, which implies that to obtain the lowest error probability for the training set, the new patterns are not needed.

However, in the Table 2, we observe the minimum error probability for the test set is 27.36% with $S = 1/8$ and $R = 4$. Comparing these results with the one associated the Table 3, the enlargement factor in this case is 65. This implies that an $EF = 65$ is required to achieve the lowest test error rate. This error probability is much smaller than the one is obtain for Factor equal to 1.

Table 1: Error probability for the training set

Error Probability		Range										
		0	0.5	1	2	3	4	5	6	8	10	12
Step	1/16	1,74%	2,67%	4,11%	7,21%	9,62%	13,23%	14,80%	16,50%	18,61%	20,34%	22,12%
	1/8	1,74%	2,76%	4,21%	7,29%	9,69%	13,31%	14,88%	16,58%	18,70%	20,38%	22,18%
	1/4	1,74%	2,93%	4,37%	7,45%	9,81%	13,39%	14,86%	16,57%	18,69%	20,33%	22,17%
	1/2	1,74%	3,22%	4,71%	7,72%	10,03%	13,53%	15,00%	16,70%	18,87%	20,50%	22,25%
	1	1,74%	1,74%	5,07%	8,14%	10,34%	13,77%	15,13%	16,73%	18,59%	20,32%	22,07%
	2	1,74%	1,74%	1,74%	8,37%	8,37%	13,36%	13,36%	16,21%	18,02%	19,66%	21,43%

Table 2: Error probability for the test set

Error Probability		Range										
		0	0.5	1	2	3	4	5	6	8	10	12
Step	1/16	33,94%	31,77%	29,64%	27,73%	27,92%	27,47%	28,03%	28,93%	30,91%	32,71%	33,64%
	1/8	33,94%	31,77%	29,45%	27,77%	27,88%	27,36%	28,07%	28,71%	30,84%	32,85%	33,71%
	1/4	33,94%	31,73%	28,97%	27,73%	27,88%	27,62%	28,18%	29,12%	31,40%	32,97%	33,71%
	1/2	33,94%	31,51%	29,34%	27,88%	27,88%	27,55%	28,29%	29,30%	31,36%	33,00%	33,68%
	1	33,94%	33,94%	29,08%	27,92%	27,66%	28,18%	28,82%	29,60%	31,25%	33,08%	33,79%
	2	33,94%	33,94%	33,94%	28,33%	28,33%	29,27%	29,27%	29,68%	31,70%	33,53%	33,60%

Table 3: Enlargement factor

Enlargement factor		Range										
		0	0.5	1	2	3	4	5	6	8	10	12
Step	1/16	1	17	33	65	97	129	161	193	257	321	385
	1/8	1	9	17	33	49	65	81	97	129	161	193
	1/4	1	5	9	17	25	33	41	49	65	81	97
	1/2	1	3	5	9	13	17	21	25	33	41	49
	1	1	1	3	5	7	9	11	13	17	21	25
	2	1	1	1	3	3	5	5	7	9	11	13

5. CONCLUSIONS

One of the main problems in the design of ESR systems is the small number of available patterns. This fact makes the learning process more difficult, due to the generalization problems in the learning stage. In this work we propose a solution to this problem consisting in enlarging the training set through the creation the new virtual patterns. In the case of emotional speech, most of the emotional information is included in speed and pitch variations. Thus, a change in the average pitch value that does not modify neither the speed nor the pitch variations does not affect the expressed emotion. So, we use this prior information in order to create new patterns applying a pitch shift modification in the feature extraction process of the classification system. For this purpose, we propose a frequency scaling modification of the Mel Frequency Cepstral Coefficients. This proposed process allows us to synthetically increase the number of available patterns in the training set, thus increasing the generalization capability of the system and reducing the test error.

Using MFCC-based enlargement of the training set, the system has a number of patterns appropriate, and it is possible train to the system correctly. In this case, it is possible reduce the error probability in emotion recognition near 7%, which is a considerable improvement in the performance. This percentage value is very important in emotion recognition.

ACKNOWLEDGEMENTS

This work has been funded by the Spanish Ministry of Education and Science (TEC2012-38142-C04-02), by the Spanish Ministry of Defense (DN8644-ATREC) and by the University of Alcalá under project UAH2011/EXP-028.

REFERENCES

- [1] Verderis, D. & Kotropoulos, C., (2006) "Emotional speech recognition: Resources, features, and method", Elsevier Speech communication, Vol. 48, No. 9, pp1162-1181.
- [2] Schuller, B., Batliner, A., Steidl, S. & Seppi, D., (2011) "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge", Elsevier Speech Communication, Vol. 53, No. 9, pp1062-1087
- [3] Mohino, I. & Goñi, M. & Alvarez, L. & Llerena, C. & Gil-Pita, R., (2013) "Detection of emotions and stress through speech analysis", International Association of Science and Technology for Development.
- [4] Öztürk, N., (2003) "Use of genetic algorithm to design optimal neural network structure", MCB UP Ltd Engineering Computations, Vol. 20, No.8, pp979-997.
- [5] Mori, R., Suzuki, S. & Takahara, H., (2007) "Optimization of Neural Network Modeling for Human Landing Control Analysis", AIAA Infotech@ Aerospace 2007 Conference and Exhibit, pp7-10.
- [6] Abu-Mostafa, Yaser S, (1995) "Hints", MIT Press Neural Computation, Vol. 7, No. 4, pp639-671.
- [7] Gil-Pita, R & Jarabo-Amores, P & Rosa-Zurera, M & Lopez-Ferreras, F, (2002) "Improving neural classifiers for ATR using a kernel method for generating synthetic training sets", IEEE Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on, pp425-434.
- [8] Niyogi, Partha & Girosi, Federico & Poggio, Tomaso, (1998) "Incorporating prior information in machine learning by creating virtual examples", IEEE Proceedings of the IEEE, Vol. 86, No. 11, pp2196-2209.
- [9] Vroomen, J., Collier, R. & Mozziconacci, Sylvie JL, (1993) "Duration and intonation in emotional speech", Eurospeech.
- [10] Davis, S. & Mermelstein P., (1980) "Experiments in syllable-based recognition of continuous speech", IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 28, pp357-366.
- [11] Gil-Pita, R. & Alvarez-Perez, L. & Mohino, Inma, (2012) "Evolutionary diagonal quadratic discriminant for speech separation in binaural hearing aids", Advances in Computer science, Vol. 20, No. 5, pp227-232.
- [12] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. & Weiss, B., (2005) "A database of German emotional speech", Interspeech, pp15717-1520.
- [13] Chang, M.-W. and Lin, C.-J., (2005) "Leave-one-out bounds for support vector regression model selection", MIT Press Neural Computation, Vol. 17, No. 5, pp1188-1222.
- [14] Cawley, G. C. & Talbot, N. L., (2004) "Fast exact leave-one-out cross-validation of sparse least-squares support vector machines", Vol. 16, No. 10, pp1467-1475

Authors

Inma Mohino-Herranz

Year in which an academic degree was awarded: Telecommunication Engineer, Alcalá University, 2010. PhD student about Information and Communication Technologies. Area of research: Signal Processing.



Roberto Gil-Pita

Year in which an academic degree was awarded: Telecommunication Engineer, Alcalá University, 2001. Position: Associate Professor. Polytechnic School in the Department of Signal Theory and Communications. Some of his research interest include, audio, speech, image, biological signals.



Sagrario Alonso-Díaz

PhD Psychologist. Researcher in the Human Factors Unit. Technological Institute “La Marañosa” –MoD.



Manuel Rosa-Zurera,

Year in which an academic degree was awarded: Telecommunication Engineer, Polytechnic University of Madrid, 1995. Position: Full professor and dean of Polytechnic School. University of Alcalá. His areas of interest are audio, radar, speech source separation.



INTENTIONAL BLANK

A MODIFIED HISTOGRAM BASED FAST ENHANCEMENT ALGORITHM

Amany A. Kandeel¹, Alaa M. Abbas², Mohiy M. Hadhoud³, Zeiad El-Saghir¹

¹Dept. of Computer Science and Engineering,
Faculty of Electronic Engineering,
Univ. of Menoufia, 32952, Menouf, Egypt.

amany_1980@hotmail.com , zeiad40@yahoo.com

²Dept. of Electronics and Electrical Communications, Faculty of Electronic Engineering, Univ. of Menoufia32952, Menouf, Egypt.

m2alaa@hotmail.com

³Dept. of Information Technology, Faculty of Computers and Information, Univ. of Menoufia, 32511, ShebinElkom, Egypt.

mmhadhoud@yahoo.com

ABSTRACT

The contrast enhancement of medical images has an important role in diseases diagnostic, specially, cancer cases. Histogram equalization is considered as the most popular algorithm for contrast enhancement according to its effectiveness and simplicity. In this paper, we present a modified version of the Histogram Based Fast Enhancement Algorithm. This algorithm enhances the areas of interest with less complexity. It is applied only to CT head images and its idea based on treating with the soft tissues and ignoring other details in the image. The proposed modification make the algorithm is valid for most CT image types with enhanced results.

KEYWORDS

Contrast enhancement, Histogram equalization, Histogram Based Fast Enhancement Algorithm, CT image

1. INTRODUCTION

Diagnosing diseases using medical images becomes more popular using different types of imaging techniques. Computed tomography (CT) is considered as the best of them after developed in 1970's [1], especially in cancer detection [2]. Its idea depends on specializing gray level for every different organic tissue. Contrast of an image is defined as the ratio between the brightest and the darkest pixel intensities.

Histogram Equalization (HE) is considered as the most popular algorithm for contrast enhancement according to its effectiveness and simplicity. Its basic idea lies in mapping the gray

levels based on the probability distribution of the input gray levels. It flattens and stretches the dynamic range of the image's histogram, resulting in an overall contrast improvement. HE has been applied in various fields such as medical image processing and radar image processing [3, 4]. The two categories of histogram equalization are. Global histogram equalization, which is simple and fast, but its contrast-enhancement power is relatively low. Local histogram equalization, on the other hand, can effectively enhance contrast, but it requires more computations.

Global Histogram equalization is powerful in highlighting the borders and edges between different objects, but may reduce the local details within these objects [5] to overcome HE's problems. Ketcham and et al invented Local Histogram Equalization (LHE); the algorithm uses the histogram of a window of a predetermined size to determine the transformation of each pixel in the image. LHE succeeded in enhancing local details, but it depends on fixed size for windows where it may distort the boundaries between regions. It also demands high computational cost and sometimes causes over-enhancement in some portion of the image [6, 7].

There are many algorithms trying to preserve the brightness of the output image like BBHE (Brightness preserving Bi-Histogram Equalization) which separates the input image histogram into two parts based on the mean of the input image and then each part is equalized independently. There are many methods similar to BBHE like, DSIHE (Dualistic Sub-Image Histogram Equalization) where, it divides the histogram based on the median value. MDSIHE (Modified Dualistic Sub Image Histogram Equalization), A. Zadbuke made a modification on DSIHE and obtained good results [8]. MMBEBHE (Minimum Mean Brightness Error Bi-Histogram Equalization) provides maximal brightness preservation, but its results are found not good for the image with a lot of details. To overcome these drawbacks, P. Jagatheeswari and et al proposed a modification to this method. They enhanced images by passing the enhanced ones through a median filter. The median filter is an effective method for the removal of impulse based noise on the images [9]. Recursive Mean-Separate Histogram Equalization (RMSHE) is also considered as an extension to BBHE. All these methods achieve good contrast but they have some problems in gray level variation [7].

The rest of the paper is organized as follows. In section 2, the idea of A Histogram-Based Fast Enhancement Algorithm will be introduced. Then, the problems were found in this algorithm and the suggested modification is presented in section 3. Experimental results using clinical data of CT images is discussed in section 4 to demonstrate the usefulness of the proposed method. Concluding remarks are presented in section 5.

2. A HISTOGRAM BASED FAST ENHANCEMENT ALGORITHM

J. Yin and et al proposed an algorithm to enhance local interested areas in CT head images; they tried to improve the water-washed effect caused by the conventional histogram equalization algorithms as shown in Figure 1. The algorithm succeeded in removing water-washed effect. There are some important features for this algorithm like the speed and the simplicity. Its idea depends on that, most CT head images occupy the gray level 0 so they try to deal with the soft tissues by enhancing the region by using full range of all possible gray levels to enhance it in the CT head images. They analyzed these images and found that more than half of the whole range of gray levels occupies 0 level, and all CT head images have three major peaks in their histograms. The left peak is formed by background pixels, the middle peak is usually formed by soft tissues in

the CT head images, and the right peak is formed mostly by bone. For enhancement details, we need only the middle peak which formed by soft tissue [10].

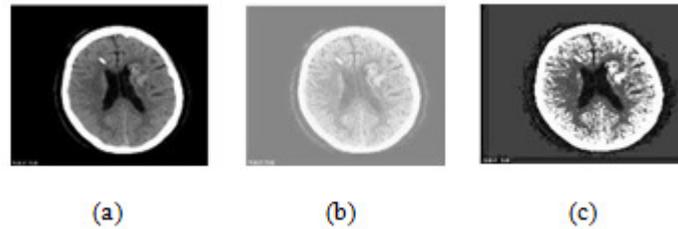


Figure 1. (a) an original CT head image (b) enhanced by conventional histogram equalization algorithm (c) Histogram-Based Fast Enhancement Algorithm.

3. A MODIFIED HISTOGRAM BASED FAST ENHANCEMENT ALGORITHM

The idea of the algorithm depends on the characteristics of CT head images. This makes the algorithm suitable for special types of images, so we tried to make a modification to this algorithm to be more appropriate for a wide range of CT images with enhanced results. The calculations of the Histogram-Based Fast Enhancement Algorithm depend on a constant value k ($0 < k < 0.4$) to evaluate how many gray levels should be ignored. This means that k remains constant for all images regardless of image characteristics, so we calculated the value of k to change with the gray levels of the picture.

First, we evaluated k as a ratio of the mean value of histogram values, which is considered an important feature of the histogram. Then we recorded these results and compared them with the Histogram-Based Fast Enhancement Algorithm; we found that there is a valuable enhancement in the results. The steps of our proposed solution remained as in the Histogram-Based Fast Enhancement Algorithm, but the change will occur in determining the k value as below.

$$k = ratio * H_{mean} \quad (1)$$

Where H_{mean} is the mean value of the histogram, which is the sum of the values divided by the number of values. Second, we performed another modification by using k as a ratio of the median value of the histogram and found that the results become better because the value depends on the characteristic of the image.

$$k = ratio * H_{median} \quad (2)$$

Where H_{median} is the median value of the histogram, it is the value which divides the values into two equal halves. At the last, we use the mode value as the most frequently occurring value in the histogram.

$$k = ratio * H_{mode} \quad (3)$$

We applied the modified algorithm to large varieties of CT images including head and lung images. To evaluate the effectiveness of the modification, we use three widely-used metrics: PSNR (Peak Signal-to-Noise Ratio), AMBE (Absolute Mean Brightness Error), and the entropy, in addition to the inspection of Visual Quality. We will show briefly how to evaluate these metrics in the next section.

3.1 Peak Signal to Noise Ratio (PSNR)

PSNR is the evaluation standard of the reconstructed image quality, and is an important measurement feature. PSNR is measured in decibels (dB). If we suppose a reference image f and a test image t , both of size $M \times N$, the PSNR between f and g is defined by.

$$PSNR(f, t) = \log_{10}(L - 1)^2 / MSE(f, t) \quad (4)$$

Where L is gray levels and MSE (Mean square error), is then defined as.

$$MSE(f, t) = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (f_{ij} - t_{ij})^2 \quad (5)$$

Note that the greater the PSNR, the better the output image quality.

3.2 Absolute Mean Brightness Error (AMBE)

It is the difference between original and enhanced image and is given as.

$$AMBE(X, Y) = |XM - YM| \quad (6)$$

Where XM is the mean of the input image $X = \{X(i, j)\}$ and YM is the mean of the output image $Y = \{Y(i, j)\}$.

We try to preserve the brightness of the image to keep the image details, so if we reduce the difference this preserve the brightness of the image.

3.3 Entropy

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. It is a useful tool to measure the Richness of the details in the output image [11].

$$Ent[P] = \sum_{i=1}^n (P_i \log_2(P_i)) \quad (7)$$

3.4 Inspection of Visual Quality

In addition to the quantitative evaluation of contrast enhancement using the PSNR and entropy values, it is also important to qualitatively assess the contrast enhancement. The major goal of the qualitative assessment is to judge if the output image is visually acceptable to human eyes and has a natural appearance [8].

4. EXPERIMENTAL RESULTS

To show the effect of the proposed modification, we apply it on different types of CT images. We use head images like the original algorithm in addition to the lung images to be validate for more image types.

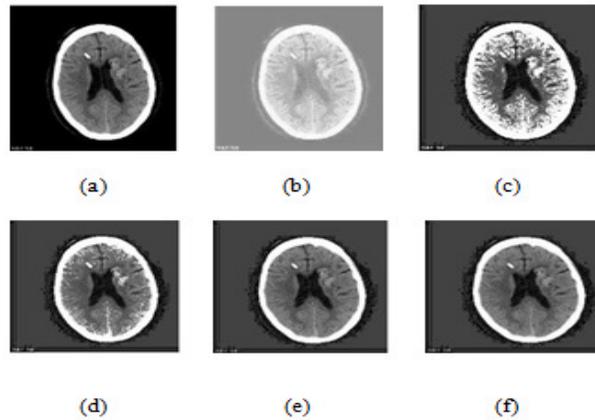


Figure 2. (a) Original CT head image (b) enhanced by conventional histogram equalization algorithm (c) enhanced by Histogram-Based Fast Enhancement Algorithm. (d) Modified Histogram-Based Fast Enhancement Algorithm using mean value (e) Modified Histogram-Based Fast Enhancement Algorithm using median value. (f) Modified Histogram-Based Fast Enhancement Algorithm using mode value

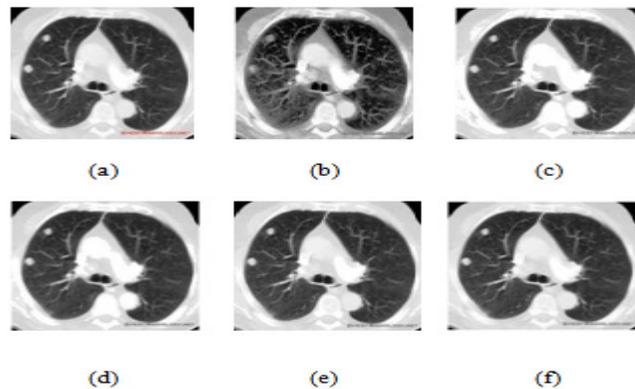


Figure 3. (a) Original CT lung image (b) enhanced by conventional histogram equalization algorithm (c) enhanced by Histogram-Based Fast Enhancement Algorithm. (d) Modified Histogram-Based Fast Enhancement Algorithm using mean value (e) Modified Histogram-Based Fast Enhancement Algorithm using median value (f) Modified Histogram-Based Fast Enhancement Algorithm using mode value

Table 1. PSNR measurement

Image	Conventional Histogram Equalization Algorithm	Histogram-Based Fast Enhancement Algorithm	Modified Histogram-Based Fast Enhancement Algorithm		
			Using Mean	Using Median	Using Mode
CThead1	6.6589	12.1687	13.9531	14.6326	14.63262
CThead2	6.7181	12.3103	14.8156	14.81723	14.81723
CThead3	4.2788	9.0203	9.8088	11.22233	11.22233
CThead4	6.7181	12.3103	14.8156	14.81723	14.81723
CTlung1	17.9699	26.9840	28.6100	32.35675	34.2496
CTlung2	19.3186	30.1420	32.3924	41.8492	43.58417
CTlung3	8.839	13.8357	13.4644	14.50487	14.5052
CTlung4	15.3099	21.5727	26.8872	31.9475	35.1296

As we mention before, the increase in the value of PSNR is considered as an enhancement in the algorithm. From Table 1 we find that there is an enhancement using the proposed modified algorithm.

Table 2. AMBE measurement.

Image	Conventional Histogram Equalization Algorithm	Histogram-Based Fast Enhancement Algorithm	Modified Histogram-Based Fast Enhancement Algorithm		
			Using Mean	Using Median	Using Mode
CThead1	111.8702	48.14349	38.0466	34.55872	34.55872
CThead2	97.365	41.37939	133.7495	30.06355	13.3852
CThead3	150.4411	78.69606	70.7215	57.66641	57.66641
CThead4	112.059	47.4835	33.5147	33.43602	33.43602
CTlung1	13.4576	4.9821	3.427556	1.969327	1.6462
CTlung2	15.1077	4.1489	3.2355	1.5521	1.369413
CTlung3	76.9961	41.85713	43.53475	35.29687	35.26233
CTlung4	13.977	11.785	5.95788	3.7327	2.9718

Our Proposed algorithm is considered one of brightness persevered algorithm so we try to reduce the difference between the brightness of input and the result image. From Table 2, we can conclude that there is an enhancement in AMBE values using the proposed algorithm.

As we will see in Table 3, there is a small increase in the Entropy values especially using the median and the mode where we have found there is a great convergence between median and mode values. As for the Inspection of Visual Quality, as we see in Figure2 and Figure3 there are some details appeared in the proposed algorithm which help in diagnostic diseases more accurate.

Table 3. Entropy measurement.

Image	Original Image	Conventional Histogram Equalization Algorithm	Histogram-Based Fast Enhancement Algorithm	Modified Histogram-Based Fast Enhancement Algorithm		
				Using Mean	Using Median	Using Mode
CThead1	0.9991	3.3235	4.608886	4.912558	5.133528	5.13352
CThead2	0.8993	4.5394	5.144703	2.077121	5.391977	5.8384
CThead3	0.9169	2.3727	2.812813	2.9972	3.377863	3.37786
CThead4	0.9997	3.3564	4.4211	5.0363	5.058119	5.05811
CTlung1	0.0022	5.8899	6.9246	7.08508	7.102342	7.0458
CTlung2	0.0695	5.9528	6.8429	6.9279	7.1620	7.20219
CTlung3	0.1575	3.3206	3.517687	3.409321	4.483833	4.50024
CTlung4	0.2065	2.5454	6.4606	6.687	6.687	6.6064

We can exclude some points from the previous results that the modified algorithm achieves greater values of PSNR, AMBE and entropy compared with Histogram-Based Fast Enhancement Algorithm. The first metric of PSNR; the propose algorithm have increased the values of PSNR; this means that less noise in the resulted image. The second metric is AMBE, it has been minimized and this means that it has preserved the brightness of the image. The third metric of entropy where it has increased; this means that more information can be extracted from the output image. We also performed statistical analysis for the results in Figure4, Figure5, and Figure6, where Figure4 shows the increment in PSNR values due to using the modification with mean, median and mode. Figure5 shows the enhancement in entropy values and Figure6 show the decrement of AMBE. There is a valuable improvement in the three parameters for the modification especially the mode where give the best results. We found that there is a range of ratio values that gives the best results for the three parameters and outside this range there aren't good results. This gives us the ability to control this ratio to obtain the best results.

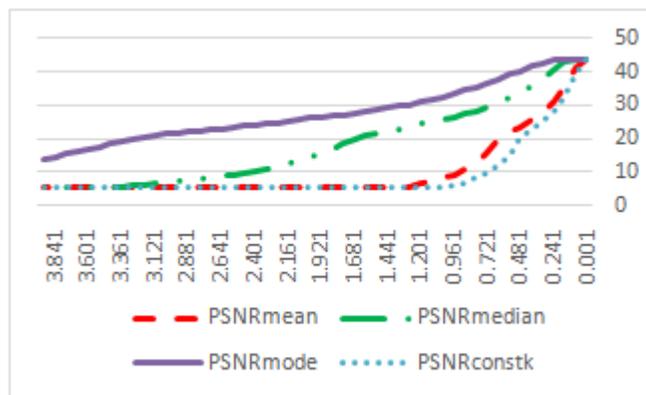


Figure4. The effect of modification on PSNR values

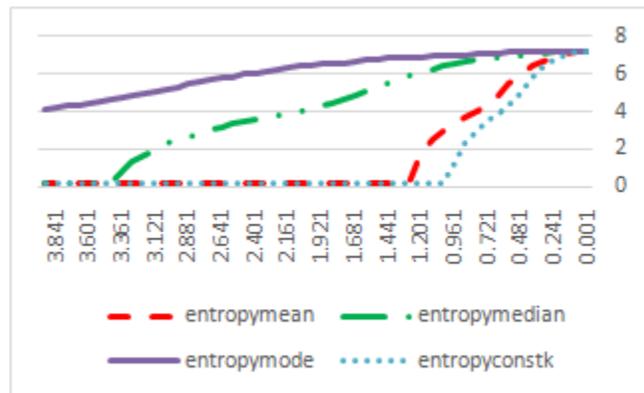


Figure5.The effect of modification on entropy values

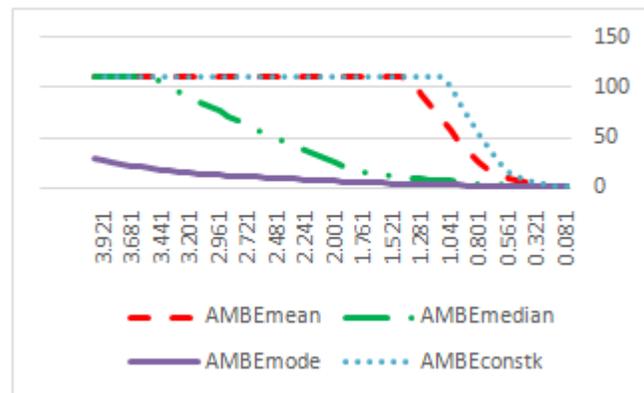


Figure6.The effect of modification on AMBE values

5. CONCLUSION

In this paper, we have presented a simple modification of Histogram Based Fast Enhancement Algorithm. First, we have showed how it succeeded in removing water-washed effect. Then discuss the proposed modification which enhances the PSNR, AMBE and entropy parameters values to be more appropriate for a wide range of CT images. In addition to the enhancements occurred to the Histogram-Based Fast Enhancement Algorithm. There are some advantages of the algorithm compared to other algorithms. It still keeps the advantage of simplicity due to less complex calculations used in the algorithm. There is another advantage of this algorithm due to its idea of using global histogram and not based on local histogram. This decreases the used time for running.

REFERENCES

- [1] S. Smith, (2002) "The Scientist and Engineer's Guide to Digital Signal Processing", Chapter 25: Special Imaging Techniques, Computed Tomography, Softcover, ISBN 0-7506-7444-X.
- [2] Dr. West, "Lung Cancer Screening Saves Lives", <http://cancergrace.org/lung/2010/11/04/lung-cancer-screening-saves-lives>.
- [3] Y. Kim, Feb.(1997) "Contrast enhancement using brightness preserving Bi-Histogram equalization", IEEE Trans. Consumer Electronics, vol. 43, no. 1, pp. 1-8.

- [4] R. Krutsch and D. Tenorio,(2011)"Histogram Equalization", Document Number: AN4318, Application Note Rev. 0.
- [5] I. Jafar and H. Ying, (2007)" Multilevel Component-Based Histogram Equalization for Enhancing the Quality of Grayscale Images", IEEE EIT, pp. 563-568.
- [6] R. Jones and T. Tjahjadi, (1993)"A study and Modification of the Local Histogram Equalization Algorithm", pattern recognition, vol2, no. 9, pp 1373-1381.
- [7] P. Shanmugavadivu, K. Balasubramania, (2010) "Image Inversion and Bi Level Histogram Equalization for Contrast Enhancement" International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 15.
- [8] A. Zadbuke,(2012)"Brightness Preserving Image Enhancement Using Modified Dualistic Sub Image Histogram Equalization", International Journal of Scientific & Engineering Research, Volume 3, Issue 2, 1 ISSN 2229-5518.
- [9] P. Jagatheeswari, S. Suresh Kumar, M. Rajaram, (2009)" Contrast Enhancement for Medical Images Based on Histogram Equalization Followed by Median Filter", the International Conference on Man-Machine Systems (ICoMMS), MALAYSIA.
- [10] J.Yin, X. Tian, Z. Tang, Y. Sun, (2006)"A Histogram-Based Fast Enhancement Algorithm for CT Head Images", Intl. Conf. on Biomedical and Pharmaceutical Engineering (ICB PE).
- [11] "Image entropy",<http://www.esrf.eu/computing/Forum/imgCIF/PAPER/entropy>.

INTENTIONAL BLANK

FINGERPRINTS IMAGE COMPRESSION BY WAVE ATOMS

Mustapha Delassi and Amina Serir.

LTIR, Faculté d'électronique et d'informatique, USTHB
BP 32 EI Alia, bab ezzouar, 16111 Alger, Algeria

ABSTRACT

The fingerprint images compression based on geometric transformed presents important research topic, these last year's many transforms have been proposed to give the best representation to a particular type of image "fingerprint image", like classics wavelets and wave atoms. In this paper we shall present a comparative study between this transforms, in order to use them in compression. The results show that for fingerprint images, the wave atom offers better performance than the current transform based compression standard. The wave atoms transformation brings a considerable contribution on the compression of fingerprints images by achieving high values of ratios compression and PSNR, with a reduced number of coefficients. In addition, the proposed method is verified with objective and subjective testing.

KEYWORDS

Image, compression, fingerprint, wavelets, wave atoms, WSQ.

1. INTRODUCTION

The fundamental goal of image compression is to obtain the best possible image quality at an allocated storage capacity. For this, data compression is one of the major challenges that are used in the majority of digital applications and specifically in the field of biometric "fingerprint", which presents the centre of interest in our work. The overall process of image compression through a series of steps: transformation, quantization and coding. The diversity at the steps led to the birth of different compression standards according to the desired application.

In image compression, it is important to get the ability of representing in a very simple way the data or the information with the minimum possible elements with allowing a loss of information, for this, the transformation has dual contribution, it decorrelates the image components and allows identifying the redundancy. Second it offers a high level of compactness of the energy in the spatial frequency domain. There have been several transforms used in data compression, Discrete Fourier Transform (DFT) and the DCT (Discrete cosines transform), DWT (Discrete wavelets transform) for images compression.

Wavelets have been widely used in image processing, such as denoising and image compression [1]. The success of wavelets with the JPEG2000 standard, and the DCT with the JPEG standard was great, but its performance is limited to a certain type of images, this is why some other standard dedicated to compression of fingerprint images are appeared, one of them is, the FBI

fingerprint image compression standard, the wavelet scalar quantization (WSQ), which is based on optimized decomposition of wavelet.

In our work we propose an algorithm for fingerprint image compression using wave atoms decomposition, compared with WSQ standard results, in order to justify the contribution that can bring wave atoms transformation for fingerprint image compression.

The remainder of this paper is divided into 4 sections. Section 2 discusses the wavelets and WSQ standard, wave atoms transform implementation details and the proposed method for fingerprint image compression is described in section 3. Experimental results are discussed in section 4 and section 5 details concluding remarks followed by acknowledgment and references.

2. WSQ STANDARD

In 1993, the FBI (Federal Bureau of Investigation) has developed a standard for compressing fingerprint images and reconstruction based on wavelet transforms, the fingerprint image is decomposed using 2D DWT, the wavelets used are the biorthogonal 9/7 discrete wavelet transform (DWT), by decomposing the image into four sub-bands of lower size in every level. The structure of the tree decomposition can be determined by applying different tests on several reference images. The tests concluded that the best tree consists of 64 sub bands [2]. These sub-bands are further decomposed in to sets which grouped the same levels of decomposition, then quantized and coded. The quantization used for each sub-bands is, the uniform scalar quantization with dead zone, and then encoded using Huffman algorithm and RLE (Run Length Encoding). The WSQ compression technique can compress fingerprint images with compression ratio ranging from 10 to 1 and 20 to 1 [2]. However, despite that the wavelet transform is effective for the detection of isotropic structures; it's not optimal for the analysis of anisotropic objects in the image (i.e., lines, contours), because they fail to follow the direction of this edge. To this effect, new multiscale geometric transforms so-called second generation have been developed, such as ridgelets [3], contourlets [4], curvelets [5], and more recently proposed, the wave atoms, which all incorporate the notion of directionality.

3. WAVE ATOMS

Wave atoms are a recent addition to the collection of mathematical transforms for harmonic computational analysis. Wave atoms are a variant of wavelet packets, they have a high frequency localization that cannot be achieved using a filter bank based on wavelet packets and curvelet Gabor atoms. Wave atoms precisely interpolate between Gabor atoms [6] (constant support) and directional wavelets [7] (wavelength \sim diameter) in the sense that the period of oscillations of each wave packet (wavelength) is related to the size of essential support by the parabolic scaling i.e. wavelength \sim (diameter)².

Different transforms based on wavelet packet, need to be represented as 'phase-space tiling', there are two different parameters α , which represent whether the multi-scale decomposition is α , and β -directional capacity if it is isotropic, Wavelets (including Multi Resolution Analysis [5], directional [8] and complex [9]) correspond to $\alpha = 1$ $\beta = 1$, for ridgelets [10] $\alpha = 1$, $\beta = 0$, Gabor transform $\alpha = 1$, $\beta = 0$ and curvelets [3] correspond to $\alpha = 1$, $\beta = 1/2$. Wave atoms are defined for $\alpha = 1$, $\beta = 1/2$. Figure 1 illustrates this classification [4].

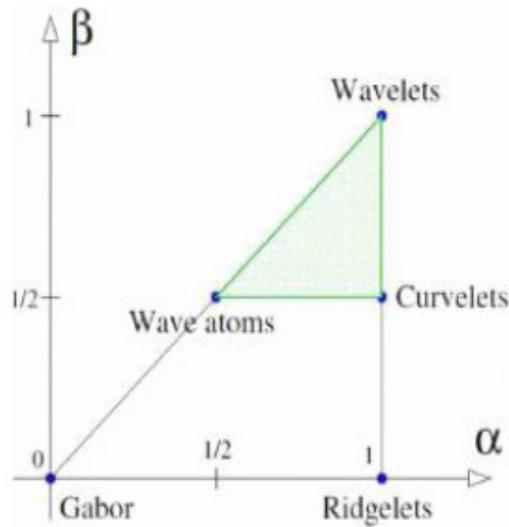


Figure 1. Diagram of (α, β) [Hadd 2009]

3.1 1D discrete wave atoms

Wave atoms are tensor products of a special type of 1D wave packets. $\psi_{m,n}^j(x)$ is a one-dimensional family of real-valued wave packets, where $j \geq 0, m \geq 0$ and $n \in \mathbb{Z}$, centered in frequency around $\pm \omega_{j,m} = 2\pi J_m$, with $c_1 2^j \leq m \leq c_2 2^j$, and centred in space around $x_{j,n} = 2^{-j}n$. One-dimensional version of the parabolic scaling states that the support of $\hat{\psi}_{m,n}^j(\omega)$ is of length $O(2^j)$ while $\omega_{j,m} = O(2^{2j})$. Filter bank-based wavelet packets is considered as a potential definition of an orthonormal basis satisfying these localization properties. The wavelet packet tree, defining the partitioning of the frequency axis in 1D, can be chosen to have depth j when the frequency is 2^{2j} . However, there is a problem associated with standard wavelet packets, that the direction in which they meet the frequency localization is rather weak [9].

Dyadic dilates and translates of $\hat{\omega}_m^0$ on the frequency axes are combined and bases function, written as:

$$\psi_{m,n}^j(x) = \psi_m^j(x - 2^{-j}n) = 2^{j/2} \psi_m^0(2^j x - n) \tag{1}$$

The coefficients $c_{j,m,n}$, for each wave number $\omega_{j,m,n}$, are obtained as a decimated convolution at scale 2^{-j} .

$$c_{j,m,n} = \int \psi_m^j(x - 2^{-j}n)u(x)dx \tag{2}$$

By Plancherel's theorem,

$$c_{j,m,n} = \int e^{i2^{-j}n\omega} \overline{\hat{\psi}_m^j(\omega)} \hat{u}(\omega) d\omega \tag{3}$$

Assuming that the function u is accurately discretized $x_k = kh, h=1/N, k=1 \dots N$, then up to some small truncation error:

$$c_{j,m,n}^D = \sum_{k=2\pi(-N/2+1:N/2)} e^{i2^{-j} \hat{u}(k)} \hat{\psi}_m^j(k) \hat{u}(k) \quad (4)$$

This equation makes sense for couples (j,m) for which the support of $\hat{\psi}_m^j(k)$ lies entirely inside the interval $[-\pi n, \pi n]$, so we may write $k \in 2\pi Z$.

3.2.2 D discrete wave atoms

A two-dimensional orthonormal basis function with 4 bumps in frequency plane is formed by individually taking products of 1D wave packets. Mathematical formulation and implementations for 1D case are detailed in the previous section. 2D wave atoms are indexed by $\mu=(j,m,n)$, where $m=(m_1,m_2)$ and $n=(n_1,n_2)$. Construction is not a simple tensor product since there is only one scale subscript j. This is similar to the non-standard or multiresolution analysis wavelet bases where the point is to enforce same scale in both directions in order to retain an isotropic aspect ratio. Eq. (1) is modified in 2D as:

$$\varphi_{\mu}^{+}(x_1, x_2) = \psi_{m_1}^j(x_1 - 2^{-j} n_1) \psi_{m_2}^j(x_2 - 2^{-j} n_2) \quad (5)$$

The Fourier transform is also separable, namely:

$$\hat{\varphi}_{\mu}^{+}(\omega_1, \omega_2) = \hat{\psi}_{m_1}^j(\omega_1) e^{-i2^j n_1 \omega_1} \hat{\psi}_{m_2}^j(\omega_2) e^{-i2^j n_2 \omega_2} \quad (6)$$

$$\varphi_{\mu}^{-}(x_1, x_2) = H \psi_{m_1}^j(x_1 - 2^{-j} n_1) H \psi_{m_2}^j(x_2 - 2^{-j} n_2) \quad (7)$$

The recombination between the relations (6) and (7), by dual orthonormal basis give:

$$\varphi_{\mu}^{(1)} = \frac{\varphi_{\mu}^{+} + \varphi_{\mu}^{-}}{2}, \quad \varphi_{\mu}^{(2)} = \frac{\varphi_{\mu}^{+} - \varphi_{\mu}^{-}}{2} \quad (8)$$

provides basis functions with two bumps in the frequency plane, symmetric with respect to the origin, hence directional wave packets (oscillating in one single direction). Together, $\varphi_{\mu}^{(1)}$ and $\varphi_{\mu}^{(2)}$, form the wave atom frame and may be denoted jointly as φ_{μ} . The price to pay in considering both $\varphi_{\mu}^{(1)}$ and $\varphi_{\mu}^{(2)}$, is an increase of a factor 2 in the redundancy.

Wave atom algorithm is based on the apparent generalization of the 1D wrapping strategy to two dimensions and its complexity is $O(N^2 \log N)$.

Figure 2 represents the wave atom tiling of the special frequency plane. The size of the squares doubles when the scale j increases by 1. At a given scale j, squares are indexed by m1, m2 starting from zero near the axes.

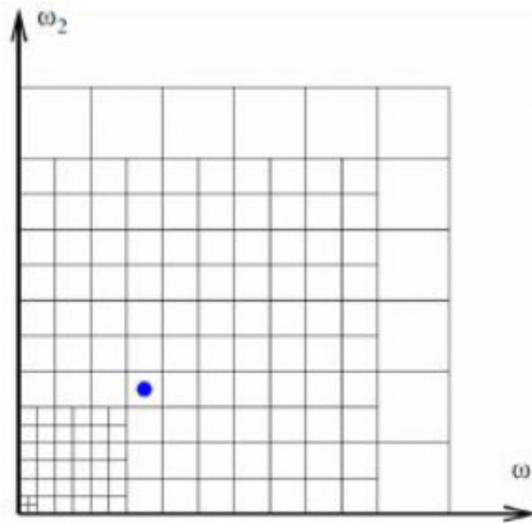


Figure 2. wave atom tiling of the frequency plane

4. RESULTS AND DISCUSSION

4.1 Transforms studies

Several fingerprint images compression standards was used in order to reduce their size, quoting JPEG with CDT transforms, WSQ, and JPEG 2000 with DWT transforms, wavelet transform has brought a significant contribution, but for the feature of the fingerprint images, these transformations have not given a good compression results, for this, other transforms was used to better characterize the fingerprints images.

Measures such as MSE (mean squared error) or PSNR (peak signal to noise ratio), correspond to the numerical analysis of pixel values before and after compression, these values are very general and do not always reflect the quality of the reconstructed image. For an image I of size $n \times m$ pixels, we define:

$$MSE = \frac{1}{n \times m} \sum_{i=1}^n \sum_{j=1}^m (I - \hat{I})^2 \quad (9)$$

$$PSNR = 10 \log_{10} \frac{(256-1)^2}{MSE} \quad en \ dB \quad (10)$$

\hat{I} represents the compressed image.

The purpose of each transformation is to concentrate the information on few coefficients, for this we are going to precede tests to reconstruct the image with a portion of the coefficients.

In this paper we propose algorithms for fingerprint images compression based on wavelets and wave Atoms transform. First, we apply the different transforms on the original image in order to obtain the energy maximization on a reduced number of coefficients; following the low coefficients will be set to zero by a threshold. The image is reconstructed in this case only by the

most significant coefficients, and calculates the PSNR for determining the quality of the reconstructed image.

Tables 1 and 2 present a comparative study between different transformations, based on wavelets transforms and wave atoms, and their effects on reducing the number of coefficients required for image reconstruction.

Table 1. PSNR's results with wavelets transforms

Coefficients selected (%)	10	20	30	50	80	90	100
wavelets	33.95	34.05	34.25	34.33	34.39	34.40	34.44
WP	34.45	34.58	34.66	34.80	35.20	35.32	35.40
WSQ	34.08	34.15	34.36	34.42	34.73	34.78	34.85

From the Tables 1 and 2 we note that the PSNR increases when the percentage of coefficients selected for image reconstruction increase. This applies to all transformations, by varying the threshold value. However, this increase differs from decomposition to another, and in the same context, we see that it is not a large scale for wavelet-based transformations. While with the transformation in wave atom, the PSNR is growing strongly and the values are higher compared to other processing while the number of coefficients used for reconstruction is much less important.

Table 2. PSNR's results with wave atoms transforms

Coefficients selected (%)	1	2	3	5	6	9	10
wave atoms	34.70	36.24	37.74	40.61	41.94	45.53	47.04

We conclude that the transformation wave Atoms can better focus the information useful for the image reconstruction, on few coefficients, this give more importance to use the wave atoms transform for the compression of fingerprints images.

4.2 Compression algorithms

In this section, we realized the several compression schemes based on transformations wavelets and wave atoms.

All these methods are based on the same compression scheme: we start by the transformation, after the quantization and finally coding. So the difference between these schemes lies in the choice of the transformation and the strategy of quantification. For the quantization, we adopt the uniform scalar quantization with dead zone for the WSQ algorithm compression, while for the

compression algorithm based on wave atoms transform we chose non-uniform scalar quantization, these quantization are followed by Huffman and RLE (Run length coding) coding.

Table 3. Comparative results between algorithms

Algorithms	RATE	PSNR (dB)	Visual assessment
wavelets	6,17	31,20	Poor
WSQ	5,75	31,36	Poor
Wave atoms	18.00	35. 04	Very good

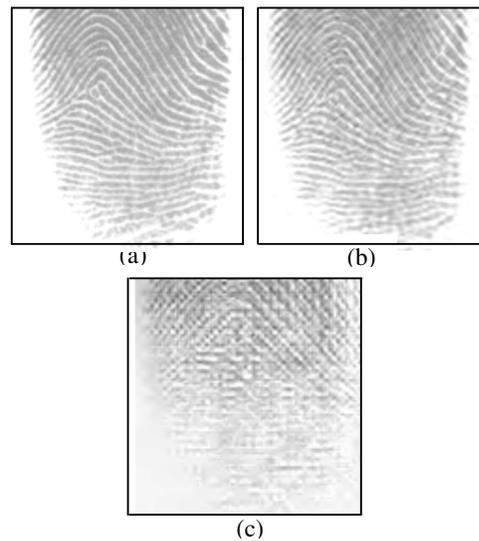


Figure 3. (a) Original image (b) Compressed image with wave atoms algorithm, (c) Compressed image with WSQ algorithm

As shown in the table 3 we can say that the wave atoms algorithm gives the best results PSNR with high compression rate.

Figure 3 represents a sample fingerprint image compressed and reconstructed by wave atoms and WSQ algorithms; we can see that wave atoms algorithm provides excellent results for the image reconstruction.

To appreciate the compression results, we could apply minutiae detection. Figure 4 illustrates the detected minutiae for original and decompressed images, we can notice that the wave atoms prevue the local structures.

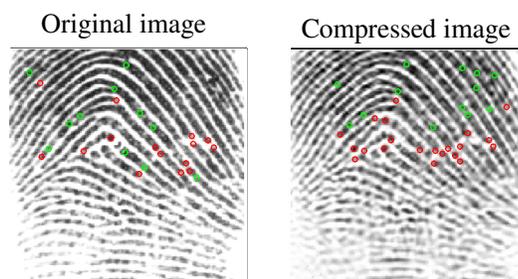


Figure 4. Minutiae detection on original and compressed image by wave atoms

The obtained results show that wave atoms transform is more appropriate to fingerprints images compression than wavelet transform

5. CONCLUSION

Compression of fingerprint images based on wave atoms provides better results, it can improved the PSNR with high RATE compared to WSQ fingerprint standard compression, we have also shown that wavelets are not all time appropriate to different type of images and on of them the images with curvatures and lines.

ACKNOWLEDGEMENTS

Thank's everyone

REFERENCES

- [1] S. G. Mallat. "A Wavelet Tour of Signal Processing". 2nd Edition, San Diego : Academic Press, 1999.
- [2] Zehira Haddad, Azeddine Beghdadi, Amina Serir, Anissa Mokraoui, "A new fingerprint image compression based on wave atoms transform", IEEE Xplore, 2009.
- [3] M.N. Do and M. Vetterli, "The finite ridgelet transform for image representation", IEEE Trans Image Processing, vol. 12, no 1, pp. 16-28, 2003.
- [4] M.N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation", IEEE Trans. Image Processing, vol. 14, no 12, pp. 2091-2106, 2005.
- [5] E.J. Candès, L. Demanet, D.L. Donoho and L. Ying, "Fast discrete curvelet transforms", Multiscale Model. Simul., pp. 861-899, 2005.
- [6] S. Mallat, "A wavelet Tour of Signal Processing", Second Edition, Academic Press, Orlando-SanDiego, 1999.
- [7] J.P. Antoine and R. Murenzi, "Two-dimensional directional wavelets and the scale-angle representation", Sig. Process, vol. 52, pp. 259-281, 1996.
- [8] L. Villemoes, "Wavelet packets with uniform time-frequency localization", in Comptes Rendus Math, vol. 335, no 10, pp. 793-796, 2002.
- [9] L. Demanet and L. Ying, "Wave Atoms and Sparsity of Oscillatory Patterns", Appl. Comput. Harmon. Anal., vol. 23, no 3, pp. 368-387, 2007.
- [10] L. Demanet and L. Ying, "Scattering in Flatland: Efficient Representations via Wave Atoms", in Found. of Comput. Math, 2008.

VISUAL TRACKING USING PARTICLE SWARM OPTIMIZATION

J.R.Siddiqui and S.Khatibi

Department of Computing, Blekinge Institute of Technology, Sweden.

ABSTRACT

The problem of robust extraction of visual odometry from a sequence of images obtained by an eye in hand camera configuration is addressed. A novel approach toward solving planar template based tracking is proposed which performs a non-linear image alignment for successful retrieval of camera transformations. In order to obtain global optimum a bio-metaheuristic is used for optimization of similarity among the planar regions. The proposed method is validated on image sequences with real as well as synthetic transformations and found to be resilient to intensity variations. A comparative analysis of the various similarity measures as well as various state-of-art methods reveal that the algorithm succeeds in tracking the planar regions robustly and has good potential to be used in real applications.

KEYWORDS

camera tracking, visual odometry, planar template based tracking, particle swarm optimization

1. INTRODUCTION

Accurate relative position estimation by keeping track of salient regions of a scene can be considered to be the core functionality of a navigating body such as a mobile robot. These salient regions are often referred to as “Landmarks” and the process of position estimation and registration of landmarks on a local representation space (i.e. a Map) is called SLAM (Simultaneous Localization and Mapping). The choice of landmarks and their representation depends on the environment as well as the configuration of a robot. In the case of vision based navigation, feature oriented land-marking is often employed, where features can be represented in many ways (e.g. by points, lines, ellipses and moments)[1]. Such techniques either do not exploit rigidity of the scene[2]-[4] or geometrical constraints are loosely coupled by keeping them out of the optimization process [5][7]. These techniques can therefore have inaccurate motion estimation due to small residual errors incurred in each iteration which make motion estimations inaccurate as these errors get accumulated. In order to mitigate this, an additional correction step is often added which either exploits a robot’s motion model to predict the future state using an array of extended Kalman-Filters [8] or minimizes the integrated error calculated over a sequence of motion [9].

Generally, feature-oriented ego-motion estimation approaches [10][11]follow three main steps; feature extraction, correspondence calculation and motion estimation. The extracted features are mostly sparse and the process of extraction is decoupled from motion estimation. Sparsification and decoupling makes a technique less computationally expensive and also allows it to handle large displacements in subsequent images, however accuracy suffers when the job is to localize a

robot and map the environment for a longer period of time. Since finding correspondences is itself an error-prone task, a large portion of the error is introduced in a very early phase of motion estimation.

There is another range of methods that utilize all pixels of an image region when calculating camera displacement by aligning image regions and hence enjoy higher accuracy due to exploitation of all possible visual information present in the segments of a scene [12]. These methods are termed “direct image alignment” based approaches for motion estimation because they do not have feature extraction and correspondence calculation steps and work directly on image patches. Direct methods are often avoided due to their computational expense which overpowers the benefits of accuracy they might provide, however an intelligent selection of the important parts of the scene that are rich in visual information can provide a useful way of dealing with the issue [13]. In addition to being direct in their approach, such methods can also better exploit the geometrical structure of the environment by including rigidity constraints early in the optimization process. The use of all visual information in a region of an image and keeping track of gain or loss in subsequent snapshots of a scene is also relevant, since it is the way some biological species navigate. For example, there are evidences that desert ants use the amount of visual information which is common between a current image and a snapshot of the ant pit to determine their way to the pit [14].

An important step in a direct image alignment based motion estimation approach is the optimization of similarity among image patches. The major optimization technique that is extensively used for image alignment is gradient descent although a range of algorithms (e.g. Gauss-Newton, Newton-Raphson and Levenberg-Marquardt [9][15]) are used for calculation of a gradient descent step. Newton’s method provides a high convergence rate because it is based on second order Taylor series approximation, however, Hessian calculation is a computationally expensive task. Moreover, a Hessian can also be indefinite, resulting in convergence failure. These methods perform a linearization of the non-linear problem which can then be solved by linear-least square methods. Since these methods are based on gradient descent, and use local descent to determine the direction of an optimum in the search space, they have a tendency to get stuck in the local optimum if the objective function has multiple optima. There are, however, some bio-inspired metaheuristics that mimic the behavior of natural organisms (e.g. Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO) [16]-[18]) or the physical laws of nature to cater this problem [19]. These methods have two common functionalities: exploration and exploitation. During an exploration phase, like any organism explores its environment, the search space is visited extensively and is gradually reduced over a period of iterations. The exploitation phase comes in the later part of a search process, when the algorithm converges quickly to a local optimum and the local optimum is accepted as the global best solution. This two-fold strategy provides a solid framework for finding the global optimum and avoiding the local best solution at the same time. In this case, PSO is interesting as it mimics the navigation behavior of swarms, especially colony movement of honeybees if an individual bee is represented as a particle which has an orientation and is moving with a constant velocity. Arbitrary motion in the initial stage of the optimization process ensures better exploration of the search space and a consensus among the particles reflects better convergence.

In this paper, the aim is to solve the problem of camera motion estimation by directly tracking planar regions in images. In order to learn an accurate estimate of motion and to embed the rigidity constraint of the scene in the optimization process, a PSO based camera tracking is performed which uses a non-linear image alignment based approach for finding the displacement of camera within subsequent images. The major contributions of the paper are: a) a novel approach to planar template based camera tracking technique which employs a bio-metaheuristic

for solving optimization problem b) Evaluation of the proposed method using multiple similarity measures and a comparative performance analysis of the proposed method.

The rest of the paper is organized as follows: In section 2 the most relevant studies are listed, in section 3 the details of the method are described, section 4 explains the experimental setup and discussion of the results, and section 5 presents the conclusion and potential future work.

2. RELEVANT WORK

There are many studies that focus on feature oriented camera motion estimation by tracking a template in the images. However, here we focus on the direct methods that track a planar template by optimizing the similarity between a reference and a current image. A classic example of such a direct approach toward camera motion estimation is the use of a brightness constancy assumption during motion and is linked to optical flow measurement [12]. Direct methods based on optical flow were later divided into two major pathways: Inverse Compositional (IC) and Forward Compositional (FC) approaches [20]-[22]. The FC approaches solve the problem by estimating the iterative displacement of warp parameters and then updating the parameters by incrementing the displacement. IC approaches, on the other hand, solve the problem by updating the warp with an inverted incremental warp. These methods linearize the optimization problem by Taylor-series approximation and then solve it by least-square methods. In [23] a multi-plane estimation method along with tracking is proposed in which region-based planes are firstly detected and then the camera is tracked by minimizing the SSD (Sum of Squared Differences) between respective planar regions in 2D images. Another example of direct template tracking is [23] which improves the tracking method by replacing the Jacobian approximation in [21] with a Hyper-plane Approximation. The method in [23] is similar to our method because it embeds constraints in a non-linear optimization process (i.e. Levenberg-Marquardt [9]) although it differs from the method proposed here since the latter employs a bio-inspired metaheuristic based optimization process which maximizes the mutual information in-between images and also the proposed method does not use constraints among the planes.

3. METHODOLOGY

The problem that is being addressed deals with estimation of a robot's state at a given time step that satisfies the planarity constraint. Let $\mathbf{x}(x^t, x^r) \in \mathbb{R}^6$ be the state of the robot with $x^t \in \mathbb{R}^3, x^r \in \mathbb{R}^3$ being the position and orientation of the robot in Euclidean space. Let's also consider I, I_r to be the current and reference image, respectively. If the current image rotates $\mathbf{R} \in \mathbb{SO}(3)$ and translates $t \in \mathbb{R}^3$ from the reference image in a given time step then the motion in terms of homogeneous representation $\mathbf{T} \in \mathbb{SE}(3)$ can be given as:

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \hat{\mathbf{s}}(x^r) & x^t \\ \mathbf{0}^T & 1 \end{bmatrix} \quad (1)$$

where $\hat{\mathbf{s}}$ is the skew symmetric matrix. It is indeed this transformation that we ought to recover given the current state of the robot and reference template image.

3.1 Plane Induced Motion

It is often the case that the robot's surrounding is composed of planar components, especially in the case of indoor navigation where most salient landmarks are likely to be planar in nature. In such cases the pixels in an image can be related to the pixels in the reference image by a projective homography \mathbf{H} that represents the transformation between the two images [24] If

$\mathbf{p} = [u, v, 1]^T$ be the homogeneous coordinates of the pixel in an image and $\mathbf{p}^r = [u^r, v^r, 1]^T$ be the homogenous coordinates of the reference image then the relationship between the two set of pixels can be written as given in equation 2.

$$\mathbf{p} \propto \mathbf{H}\mathbf{p}^r \quad (2)$$

Let's now consider that the plane that is to be tracked or the plane which holds a given landmark has a normal $\mathbf{n}_r \in \mathbb{R}^3$, which has its projection in the reference image I_r . In case of a calibrated camera, the intrinsic parameters, which are known, can be represented in terms of a matrix $\mathbf{K} \in \mathbb{R}^{3 \times 3}$. If the 3D transformation between the frames is \mathbf{T} , then the Euclidean homography with a non-zero scale factor can be calculated as:

$$\mathbf{H}(\mathbf{T}, \mathbf{n}_r) \propto \mathbf{K}(\mathbf{R} + t\mathbf{n}_r^T)\mathbf{K}^{-1} \quad (3)$$

3.2 Model-Based Image Alignment

The next step after modeling the planarity of the scene is to relate plane transformations in the 3D scene to their projected transformations in the images. For that reason a general mapping function that transforms a 2D image given a projective homography can be represented by a warping operator w and is defined as follows:

$$\mathbf{w}(\mathbf{H}, \mathbf{p}^r) = \left[\frac{h_{11}u^r + h_{12}v^r + h_{13}}{h_{31}u^r + h_{32}v^r + h_{33}}, \frac{h_{21}u^r + h_{22}v^r + h_{23}}{h_{31}u^r + h_{32}v^r + h_{33}} \right]^T \quad (4)$$

If the normal of the tracked plane is known then the problem to be addressed is that of metric model based alignment or simply model based non-linear image alignment. It is the transformation $\mathbf{T} \in \mathbb{SE}(3)$ that is to be learned by warping the reference image and measuring the similarity between the warped and the current image. Since the intensity of a pixel $\mathbf{I}(\mathbf{p})$ is a non-linear function, we need a non-linear optimization procedure. More formally, the task is to learn an optimum transformation $\hat{\mathbf{T}} = \mathbf{T}(\mathbf{x})$ that maximizes the following:

$$\max_{\mathbf{x} \in \mathbb{R}^6} \psi \left(I_r \left(\mathbf{w}(\mathbf{H}(\hat{\mathbf{T}}, \mathbf{n}_r), \mathbf{p}^r) \right), \mathbf{I}(\mathbf{p}) \right) \quad (5)$$

where ψ is a similarity function and $\hat{\mathbf{T}}$ is updated as $\hat{\mathbf{T}} \leftarrow \mathbf{T}(\mathbf{x})\hat{\mathbf{T}}$ for every new image in the sequence.

3.3 Similarity Measure

In order for any optimization method to work effectively and efficiently, the search space needs to be modeled in such a way that it captures the multiple optima of a function but at the same time suppresses local optima by enhancing the global optimum. It is also important that such modeling of similarity must provide enough convergence space so that the probability of missing the global optimum is minimized. This job is performed by a selection of similarity measure that is best suited for a given problem context. An often used measure is SSD (Sum of Squared Differences) that can be given as:

$$\psi_{SSD} = \sum_i^N \left(I_r(\mathbf{w}(\mathbf{H}, \mathbf{p}^r)) - \mathbf{I}(\mathbf{p}) \right)^2 \quad (6)$$

where ‘ N ’ is the total number of pixels in a tracked region of the image.

Similarly, another relevant similarity measure is the cross correlation coefficient of the given two data streams. Often a normalized version is used to restrict the comparison space to the range [0, 1]. The normalized cross correlation between a current image patch I and a reference image patch I_r , with μ, μ_r being their respective means, can be written as:

$$\psi_{NCC} = \sum_{i,j} \frac{(I_r(i,j) - \mu_r)(I(i,j) - \mu_I)}{\sqrt{(I_r^2(i,j) - \mu_r)(I^2(i,j) - \mu_I)}} \quad (7)$$

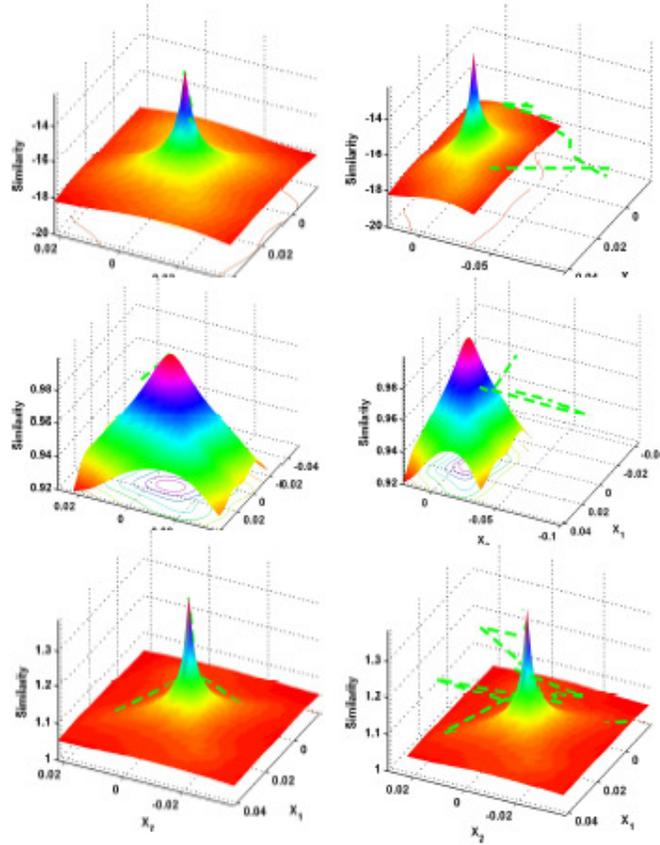


Fig. 1: Convergence surface of various similarity functions along with motion of a PSO particle on its way towards convergence depicted by green path. First Row: Sum of squared difference ($\log(\psi_{SSD})^{-1}$), Second Row: Normalized Cross Correlation (ψ_{NCC}), Third Row: Mutual Information (ψ_{MI}).

The similarity measures presented in equation 6 and 7 have the ability to represent the amount of information that is shared by the two data streams; however, as can be seen in figure 1, the convergence space and the emphasis on the global optimum need improvement. A more intuitive approach for measuring similarity among the data is Mutual Information (MI), taken from information theory, that measures the amount of data that is shared between the two input data streams [25]. The application of MI in image alignment tasks and its ability to capture the shared information have also proven to be successful [26],[27]. The reason for avoidance of MI in robotics tasks has been its relatively higher computational expense, since it involves histogram

computation. However, the gains are more than the losses, so we choose to use MI as our main similarity measure. Formally, the MI between two input images can be computed as:

$$\begin{aligned}\psi_{MI} &= E(\mathbf{I}) + E(\mathbf{I}_r) - E(\mathbf{I}_r, \mathbf{I}) \\ E(\mathbf{I}) &= -\sum_{i=0}^{N_I} \rho_I(i) \log(\rho_I(i)) \\ E(\mathbf{I}_r, \mathbf{I}) &= -\sum_{i=0}^{N_I} \sum_{j=0}^{N_I} \rho_{\Pi}(i, j) \log(\rho_{\Pi}(i, j))\end{aligned}\quad (8)$$

where $E(\mathbf{I}), E(\mathbf{I}_r, \mathbf{I}), N_I$ are the entropy, joint entropy and maximum allowable intensity value respectively.

Entropy according to Shannon [23] is the measure of variability in a random variable \mathbf{I} , whereas 'i' is the possible value of \mathbf{I} and $\rho_I(i) = \Pr(i == \mathbf{I}(\mathbf{p}))$ is the probability density function. The log basis only scales the unit of entropy and does not influence the optimization process. In other words, since $-\log(\rho_I(i))$ represents the measure of uncertainty of an event 'i' and $E(\mathbf{I})$ is the weighted mean of the uncertainties, the latter represents the variability of random variable \mathbf{I} . In the case of images we have pixel intensities as possible values of a random variable, so the probability density function of a pixel value can be estimated by a normalized histogram of the input image. The entropy for the input image is therefore a dispersion measure of the normalized histogram. This is the same in the case of joint entropy since there are two variables involved, so the joint probability density function is $\rho_{\Pi}(i, j) = \Pr(i == \mathbf{I}_r(\mathbf{p}^r) \cap j == \mathbf{I}(\mathbf{p}))$ where i, j are possible values of image \mathbf{I}_r, \mathbf{I} respectively. The joint entropy measures the dispersion in the joint histogram. This dispersion provides a similarity measure because when the dispersion in the joint histogram is small then the correlation among the images is strong, giving a large value of MI and suggesting that two images are aligned, while in case of large dispersions, MI would have a small value and images would be unaligned.

3.4 Optimization Procedure

The problem of robust retrieval of Visual Odometry (VO) in subsequent images is challenging due to the non-linear and continuous nature of the huge search space. The non-linearity is commonly tackled using linearization of the problem function; however, this approximation is not entirely general due to challenges in exact modeling of image intensity. Another route to solve the problem is to use non-linear optimization such as Newton Optimization which gives fairly good convergence due to the fact that it is based on Taylor series approximation of the similarity function. However, it requires computation of the Hessian which is computationally expensive and also it must be positive definitive for a convergence to take place.

The proposed method seeks the solution to the optimization problem presented in equation 5. In order to find absolute global extrema and not get stuck in local extrema we choose a bio-inspired metaheuristic optimization approach (i.e. PSO). Particle Swarm Optimization (PSO) is an evolutionary algorithm which is directly inspired by the grouping behavior of social animals, notably in the shape of bird flocking, fish schooling and bee swarming. The primary reason for interest in learning and modeling the science behind such activities has been the remarkable ability possessed by natural organisms to solve complex problems (e.g scattering, regrouping, maintaining course, etc.) in a seamlessly and robust fashion. The generalized encapsulation of such behaviors opens up horizons for potential applications in nearly any field. The range of problems that can be solved range from resource management tasks (e.g intelligent planning and scheduling) to real mimicked behaviors by robots. The particles in a swarm move collectively by

keeping a safe distance from other members in order to avoid obstacles while moving in a consensus direction to avoid predators while maintaining a constant velocity. This results in behavior in which a flock/swarm moves towards a common goal (e.g. a Hive, food source) while intra-group motion seems random. This makes it difficult for predators to focus on one prey while it also helps swarms to maintain their course, especially in case of long journeys that are common, e.g., for migratory birds. The exact location of the goal is usually unknown as it is in the case of any optimization problem where the optimum solution is unknown. A pictorial diction of the robot's states represented as particles in an optimization process can be seen in figure 2.

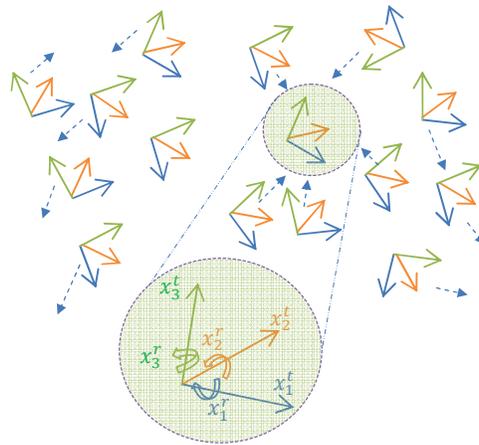


Fig. 2: A depiction of PSO particles (i.e. robot states) taking part in an optimization process. Blue arrows show the velocity of a particle and the local best solution is highlighted by an enclosing circle.

PSO is implemented in many ways with varying levels of bioinspiration reflected in terms of the neighborhood topology that is used for convergence [28]. Each particle maintains its current best position p_{best} and global best g_{best} position. The current best position is available to every particle in the swarm. A particle updates its position based on its velocity, which is periodically updated with a random weight. The particle that has the best position in the swarm at a given iteration attracts all other particles towards itself. The selection of attracted neighborhood as well as the force to which the particles are attracted depends on the topology being used. Generally a PSO consists of two broad functions: one for exploration and one for exploitation. The degree and extend of time that each function is performed depends again on the topology being used. A common model of PSO allows more exploration to be performed in the initial iterations while it is gradually decreased and a more localized search is performed in the later iterations of the optimization process. There can be multiple topologies, some of which are presented in figure 3. A global topology considers all the particles in the swarm and thus converges faster but potentially to a local optimum. The local best topology provides relatively more freedom and only a set of close neighbors are allowed to be attracted to the best particle in the swarm. This allows the algorithm to converge more slowly, but chances of finding the global optimum are increased. Another subtle variation to this neighborhood selection could be that it is performed dynamically, e.g. being increased over the number of iterations, making the system more explorative in the start and more exploitative in later stages of convergence so that a global optimum is achieved.

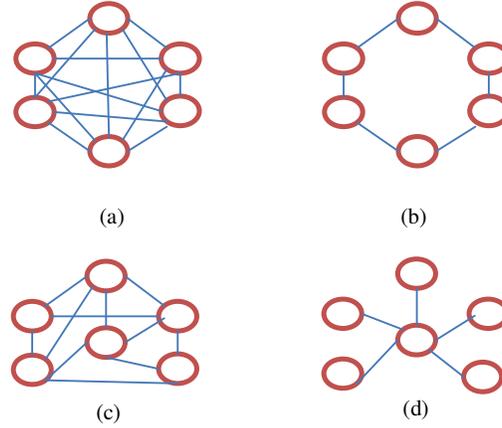


Fig. 3: Various topologies used in PSO based optimization processes. (a) Global topology uses all neighbors, (b) Circle Topology consider 2-neighbours, (c) Local Topology uses k-neighbors, and (d) Wheel Topology considers only immediate neighbors.

The process of PSO optimization starts with initialization of the particles. Each particle is initialized with a random initial velocity \mathbf{v}_i and random current position \mathbf{x}_i represented by a k dimensional vector where ' k ' is the number of degrees of freedom of the solution. The search space is discretized and limited with a boundary constraint $|\mathbf{x}_i| \leq \mathbf{b}_i$, $\mathbf{b}_i \in [b_l, b_u]$ where b_l, b_u are lower and upper bounds of motion in each dimension. This discretization and application of boundary constraints helps reduce the search space assuming that the motion in between subsequent frames is not too large. After initialization, particles are moved arbitrarily in the search space to find the solution that maximizes the similarity value as given in equation 8. Each particle updates its position based on its own velocity and the position of the best particle in the neighborhood. The position and velocity update is given in equation 9:

$$\begin{aligned} \mathbf{x}_i(t+1) &= \mathbf{x}_i(t) + \mathbf{v}_i(t+1) \\ \mathbf{v}_i(t+1) &= \omega \mathbf{v}_i(t) + \alpha_r \sigma_r \mathbf{c}_i + \alpha_s \sigma_s \mathbf{s}_i \end{aligned} \quad (9)$$

where ω is the inertial weight and is used to control the momentum of the particles. When a large value of inertial weight is used, particles are influenced more by their last velocity and collisions might happen with very large values. The cognitive (or self-awareness) component of the velocity update is represented by $\mathbf{c}_i = \mathbf{x}_i^b - \mathbf{x}_i(t)$ where \mathbf{x}_i^b is the personal best solution of the particle. Similarly, the social component is represented as $\mathbf{s}_i = \mathbf{x}_i^g - \mathbf{x}_i(t)$ where \mathbf{x}_i^g is the best solution in the particle's neighborhood. Randomness is achieved by $\sigma_c, \sigma_s \in [0,1]$ for cognitive and social components respectively. The constant weights α_c, α_s control the influence of each component in the update process. The final step in the PSO algorithm is evaluating against the termination criteria. There could be a range of strategies for terminating the algorithm: a) using a fixed number of iterations until improvement in the solution is observed, b) reaching a maximum number of consecutive iterations with no change in the solution, c) exceeding a threshold on the maximum similarity value. The first strategy would be fast but may fail to return a true optimum due to being deceived by a local best solution. The second strategy would be better than the first one in returning the global best solution; however it might also not guarantee the best solution. Third strategy is good in the sense that it can find the best solution, but it could make the algorithm continue for an infinite number of iterations. There is no general strategy for termination; rather, each problem situation needs to be adapted. If faster computation is the ultimate concern then the first option would be a good choice; similarly the third option would be

good in situations where the best solution is to be found regardless of the time taken. A weighted combination of termination criteria proves better suited for the problem at hand.

3.5 Tracking Method

The proposed plane tracking method consists of three main steps: initialization, tracking and updating. These steps are given as follows:

- 1) The planar area in the image that is to be tracked is initialized in the first frame and an initial normal of the plane is provided. If the plane normal is not already known then a rough estimate of the plane in the camera coordinate frame is given. The search space of the problem is discretized and constrained within an interval. PSO is initialized with a random solution and a suitable similarity function is provided.
- 2) The marked region in the template image is aligned with a region in the current image and an optimum solution of the 6-dof transformation is obtained. The optimization process continues until it meets one of the following conditions: (i) max number of iterations is reached, (ii) the solution has not improved in a number of consecutive iterations, or (iii) a threshold for solution improvement is reached.
- 3) The global camera transformation is updated and process repeats.

4. EXPERIMENTAL RESULTS

In order for the system to be evaluated the experiment setup must consider the basic assumptions namely: planarity nature of the scene and small subsequent motion. The planarity nature of the scene means that there should be a dominant plane in front of the camera whose normal is either estimated by using another technique or using an approximated unit normal without scale, however the rate of convergence and efficiency is affected in the latter case. The second important assumption of the system is that the amount of motion in subsequent frames is small as large motions increase the search space significantly. In addition to this, the planar region that is to be tracked must be textured so as to provide good variance of similarity while being transformed. Keeping these assumptions in mind, the algorithm is evaluated for both simulated and real robotic transformations and results are recorded.

4.1 Synthetic sequence

The experimentation process using a benchmark tracking sequence is performed. The sequence consists of a real image with a textured plane facing the camera and its 100 transformed variations whereas the motion within the subsequent transformation is kept small. The tracking region is marked in the template image in order to select the plane and the optimization algorithm is initialized. The tracking method succeeds in capturing the motion as it can be seen in figure 4. In order to test behavior of the similarity measures, the method is repeated with all three similarity functions and error surface is analyzed which can be seen in the figure 1 which also show the path of a particle in the swarm on its way toward convergence. It was found that MI provides better convergence surface than other two participating similarity measures and hence it is used for later stages of the evaluation process.

As it could be interesting to determine whether the algorithm could cater variation in the degree of freedom, the sequence is run multiple times with different dimensions of the solution that is to be learned. The increase in the number of parameters to be learned affects the convergence rate, however the algorithm successfully converges all the variations as can be seen in the figure 5. However, with an increase in degree of freedom, the search space expands exponentially making it harder to converge in the same number of iterations as needed for lower degrees of freedom.

This can be catered by multiple ways; a) increasing the overall number of iterations needed by the algorithm to converge b) increasing the number of iterations dedicated for exploration and c) putting more emphasis on the exploration by setting the appropriate inertial and social weights in equation 9. The learning of all 6 parameters of a robotic motion is a challenging task and is often reduced to 4 DOF by supplementing the two DOF from the IMU however using the optimization based approach such as presented in this paper could be used for such challenging tasks and could successfully solve the problem. It is important to note here that the proposed method only uses single plane and hence only one constraint or more formally it is an unconstrained plane segment tracking while introducing more planes and introduction of their inter-plane constraints could lead to significant increase in stability and accuracy.

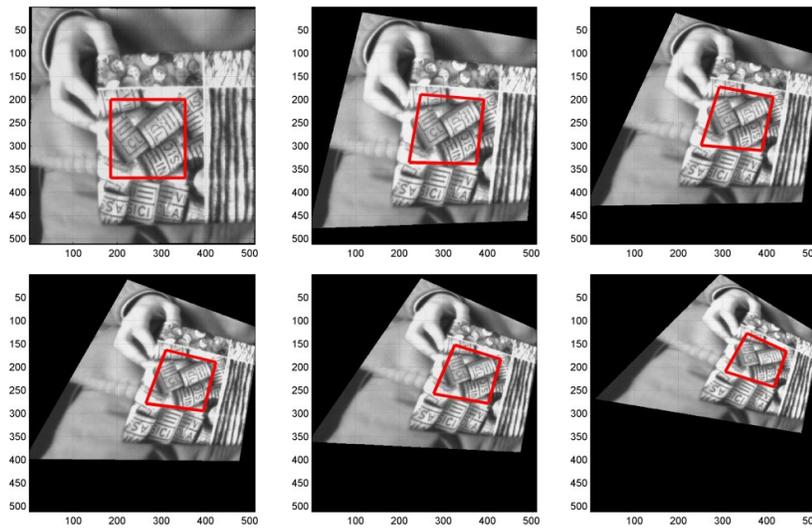


Figure 4: The result of the tracking when applied on a benchmarking image sequence with synthetic transformations.

4.2 Real Sequence

The proposed tracking method is also tested on a real image benchmarking sequence that is recorded by a downward looking camera mounted on a quadrocopter [30]. The sequence consists of 633 images with the resolution 752x480 recorded by flying the quadrotor in a loop. The important variable that was unavailable in case of real images is the absolute normal of the tracked plane. There could be two ways to solve this problem: using an external plane detection method to estimate the normal or using a rough estimate of the plane and leave rest to optimization process. The former approach is more preferable and could lead to better convergence rate however, to show the insensitivity of the proposed method to absolute plane normal and depth estimates, we use the latter approach for evaluation. The rest of the parameterization and initialization process is similar to simulated sequence based evaluation process as described earlier.

It could be visualized in figure 6 that even though initial transformation of the marked region was not correct and absolute normal was unknown; the tracking method learns the correct transformation over a period of time and successfully tracks the planar region.

A through error analysis is provided in the figure 7 which shows that the proposed tracking method has good tracking ability with minimal translation and rotational error when the motion is kept within the bounds of search space. A good way to keep the motion small is attained by using high frame-rate cameras.

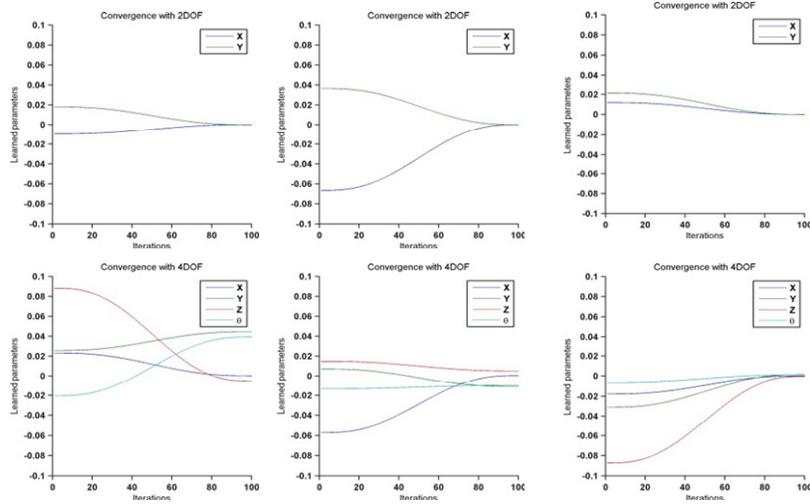


Figure 5: Convergence with variation in DOF and similarity measures. The columns represent similarity measures SSD, NCC and MI respectively and rows represent DOF 2 and 4 respectively.

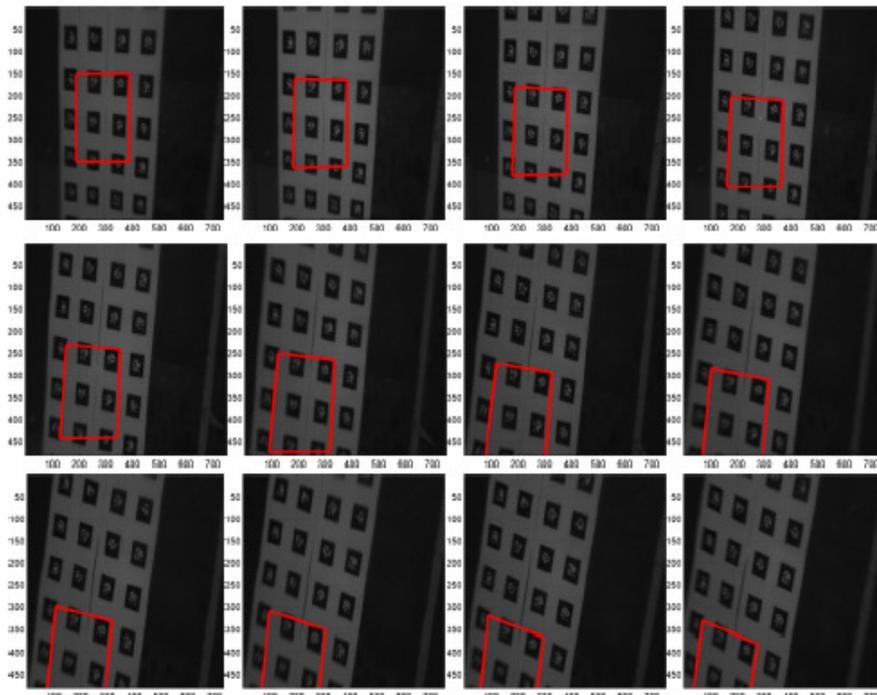


Figure 6: The result of the tracking when applied on a benchmarking image sequence with real transformations.

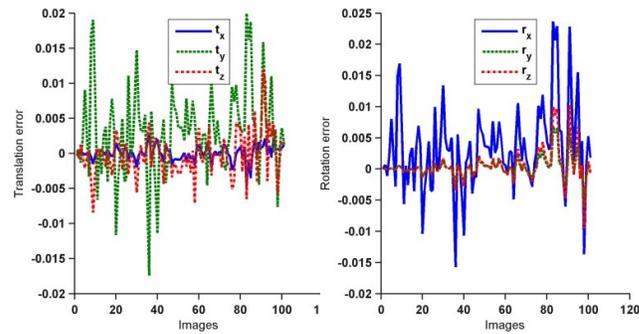


Figure 7: Transformation error for image sequence.

4.3 Comparative analysis

The proposed method performs an optimization based tracking so comparative analyses with variations of PSO and also with other state of the art methods help us determine its significance in real applications. In figure 8 a comparison of the multiple variations of PSO is presented. The Trelea PSO is good at converging to optimum similarity in all cases although its convergence rate is not fastest due to being explorative in nature. PSO common on the other hand finds its way quickly toward solution although it may not find global optimum due to being more exploitative in nature. A group of three state of the art plane tracking methods (IC, FC and HA) are applied on the same image sequence and a normalized root mean squared error is measured for the image sequence and the number of iterations. As it can be visualized from the figure 9 that the algorithm successfully beats IC and HA while it nears the performance of the forward compositional approach.

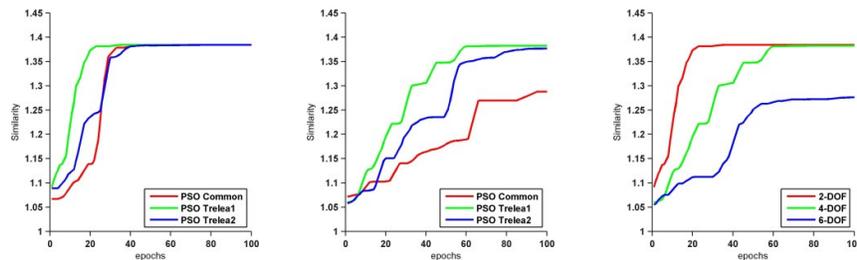


Figure 8: comparison of the convergence rate of variations of PSO along with convergence behavior of best performing version of PSO with different degree of freedom.

It can be noted that the IC and HA misses the track of the plane after 40th iteration, most probably due to intensity variation that is introduced in the sequence for which Taylor series approximation failed to capture the intensity function. As a comparison if we check the performance of the methods with different degree of freedom (see figure 10), we can see that the proposed PSO-Track method performs decently.

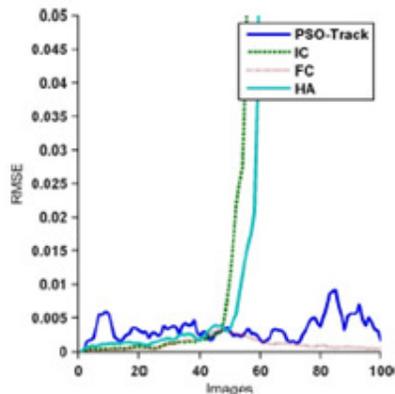


Figure 9: Performance comparison of various tracking methods.

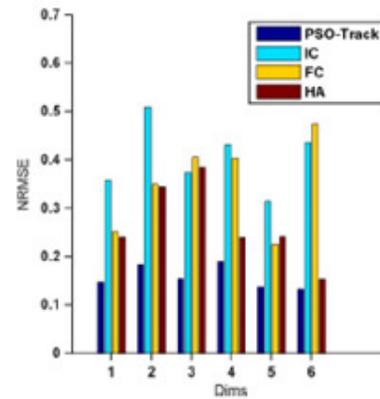


Figure 10: Performance comparison of various tracking methods with variation in DOF.

5. CONCLUSIONS

In this paper, we presented a novel approach toward solving camera tracking by tracking a projection of the plane. A non-linear image alignment is adopted and correct parameters of the transformation are recovered by optimizing the similarity between the planar regions. A thorough comparative analysis of the method over simulated and real sequence of images reveal that the proposed method has ability to track planar surfaces in when the motion within the frames is kept small. The insensitivity of the method toward intensity variations as well as to unavailability of true plane normal is also tested and algorithm has been found resilient to such environmental changes.

REFERENCES

- [1] P. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," *Vis. Algorithms Theory Pr.*, pp. 278–294, 2000.
- [2] E. Eade and T. Drummond, "Scalable monocular SLAM," in *Computer Vision and Pattern Recognition*, 2006 IEEE Computer Society Conference on, 2006, vol. 1, pp. 469–476.
- [3] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Computer Vision*, 2003. Proceedings. Ninth IEEE International Conference on, 2003, pp. 1403–1410.
- [4] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point ransac," in *Robotics and Automation*, 2009. ICRA'09. IEEE International Conference on, 2009, pp. 4293–4299.
- [5] G. Klein and D. Murray, "Parallel tracking and mapping on a camera phone," in *Mixed and Augmented Reality*, 2009. ISMAR 2009. 8th IEEE International Symposium on, 2009, pp. 83–86.
- [6] C. Pirschheim and G. Reitmayr, "Homography-based planar mapping and tracking for mobile phones," 2011, pp. 27–36.
- [7] D. Wagner, D. Schmalstieg, and H. Bischof, "Multiple target detection and tracking with guaranteed framerates on mobile phones," in *Mixed and Augmented Reality*, 2009. ISMAR 2009. 8th IEEE International Symposium on, 2009, pp. 57–64.
- [8] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proceedings of the National conference on Artificial Intelligence*, 2002, pp. 593–598.
- [9] J. More, "The Levenberg-Marquardt algorithm: implementation and theory," *Numer. Anal.*, pp. 105–116, 1978.

- [10] H. Zhou, P. R. Green, and A. M. Wallace, "Estimation of epipolar geometry by linear mixed-effect modelling," *Neurocomputing*, vol. 72, no. 16–18, pp. 3881–3890, Oct. 2009.
- [11] H. Zhou, A. M. Wallace, and P. R. Green, "Efficient tracking and ego-motion recovery using gait analysis," *Signal Process.*, vol. 89, no. 12, pp. 2367–2384, Dec. 2009.
- [12] M. Irani and P. Anandan, "About direct methods," *Vis. Algorithms Theory Pr.*, pp. 267–277, 2000.
- [13] G. Silveira, E. Malis, and P. Rives, "An efficient direct approach to visual SLAM," *Robot. IEEE Trans.*, vol. 24, no. 5, pp. 969–979, 2008.
- [14] A. Philippides, B. Baddeley, P. Husbands, and P. Graham, "How Can Embodiment Simplify the Problem of View-Based Navigation?," *Biomim. Biohybrid Syst.*, pp. 216–227, 2012.
- [15] A. Bjorck, *Numerical methods for least squares problems*. Society for Industrial Mathematics, 1996.
- [16] D. E. Goldberg, "Genetic algorithms in search, optimization, and machine learning," 1989.
- [17] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Neural Networks, 1995. Proceedings., IEEE International Conference on, 1995*, vol. 4, pp. 1942–1948.
- [18] Y. K. Baik, J. Kwon, H. S. Lee, and K. M. Lee, "Geometric Particle Swarm Optimization for Robust Visual Ego-Motion Estimation via Particle Filtering," *Image Vis. Comput.*, 2013.
- [19] E. Aarts and J. Korst, "Simulated annealing and Boltzmann machines," 1988.
- [20] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.
- [21] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.
- [22] F. Jurie and M. Dhome, "Hyperplane approximation for template matching," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 24, no. 7, pp. 996–1000, 2002.
- [23] D. Cobzas and P. Sturm, "3d ssd tracking with estimated 3d planes," in *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on, 2005*, pp. 129–134.
- [24] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, vol. 2. Cambridge Univ Press, 2000.
- [25] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
- [26] N. Dowson and R. Bowden, "A unifying framework for mutual information methods for use in non-linear optimisation," *Comput. Vision–ECCV 2006*, pp. 365–378, 2006.
- [27] N. Dowson and R. Bowden, "Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 30, no. 1, pp. 180–185, 2008.
- [28] M. Günther and V. Nissen, "A comparison of neighbourhood topologies for staff scheduling with particle swarm optimisation," *KI 2009 Adv. Artif. Intell.*, pp. 185–192, 2009.
- [29] S. Benhimane and E. Malis, "Homography-based 2d visual tracking and servoing," *Int. J. Robot. Res.*, vol. 26, no. 7, pp. 661–676, 2007.
- [30] G. H. Lee, M. Achtelik, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "A benchmarking tool for MAV visual pose estimation," in *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on, 2010*, pp. 1541–1546.
- [31] M. Warren, D. McKinnon, H. He, and B. Upcroft, "Unaided stereo vision based pose estimation," 2010.
- [32] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Inf. Process. Lett.*, vol. 85, no. 6, pp. 317–325, 2003.

FUZZY INFERENCE SYSTEM FOR VOLT/VAR CONTROL IN DISTRIBUTION SUBSTATIONS IN ISOLATED POWER SYSTEMS

Vega-Fuentes E¹, León-del Rosario S², Cerezo-Sánchez J M³, Vega-Martínez A⁴

Dept. of Electronics Engineering and Automatics Institute for Applied
Microelectronics University of Las Palmas de Gran Canaria, Spain

¹evega@diea.ulpgc.es, ²sleon@diea.ulpgc.es,
³jcerezo@diea.ulpgc.es, ⁴avega@diea.ulpgc.es

ABSTRACT

This paper presents a fuzzy inference system for voltage/reactive power control in distribution substations. The purpose is go forward to automation distribution and its implementation in isolated power systems where control capabilities are limited and it is common using the same applications as in continental power systems. This means that lot of functionalities do not apply and computational burden generates high response times. A fuzzy controller, with logic guidelines embedded based upon heuristic rules resulting from operators at dispatch control center past experience, has been designed. Working as an on-line tool, it has been tested under real conditions and it has managed the operation during a whole day in a distribution substation. Within the limits of control capabilities of the system, the controller maintained successfully an acceptable voltage profile, power factor values over 0,98 and it has ostensibly improved the performance given by an optimal power flow based automation system.

KEYWORDS

Distribution Automation, Fuzzy Inference Systems, Isolated Power Systems, Reactive Power Control, Voltage Control.

1. INTRODUCTION

In March 2007, EU leaders set the "20-20-20" targets, an European commitment with climate and energy objectives for 2020. Subsequently enacted through the climate and energy package in 2009, it implies: A 20% reduction in EU greenhouse gas emissions from 1990 levels; Raising the share of EU energy consumption produced from renewable resources to 20%; And a 20% improvement in the EU's energy efficiency.

Automation and control of medium voltage (MV) networks has been defined as one of the pillars underpinning smart grids [1], the way to achieve the "20-20-20" targets.

Automation of MV network scope may include: voltage regulation, minimize subtransmission losses restraining the reactive power flow through the transformers, deal with distributed generation and storage batteries, self healing (location and isolation of faults) and reconfiguration.

Optimal power flow (OPF) based applications are usually found to conduct the reactive power/voltage control in the transmission and subtransmission system [2], [3], [4], [5]. Control variables such as on-load tap changer (OLTC) in main transformers, shunt capacitor banks located on the feeders and at the secondary bus of the substation and even tie-switches located on the feeders to reconfigure the network are optimized in order to minimize bus voltage variations and subtransmission losses [6].

Contrary to great quantity of works on transmission systems and distribution feeders, the papers on voltage/reactive power control of a distribution substation are rather limited [7], [8]. Reports on transmission systems [9], [10] ignore the singularities of isolated and small power systems where there is not subtransmission, voltage levels greater than 36 kV are treated as transmission and in deregulated market its operation is not distributor responsibility. Shunt capacitor banks are not installed to work stepwise due to breakers costs. There are no shunt capacitors on the feeders, and large consumers have their own capacitors although the distributor has no way to manage their connection.

In fact limited size of these network has led distribution network utilities to consider any specific development negligible, so the same applications as in continental power systems are used. This means that lot of functionalities do not apply and that the computational burden generates high response times.

In practice, operators at distribution and dispatch control center, with heuristic rules based on their past experience decide better strategies for voltage/reactive power regulation than these unsuitable programs. However as described in previous paragraphs, power system networks, strongly based on SCADA systems, are expected to evolve for Intelligent SCADA, and distribution automation as essential piece of smart distribution, is requested.

In this paper, a soft computing system based in fuzzy logic is proposed to deal with voltage/reactive power regulation in the Canary Islands (Spain) distribution substations. Its guiding principle, "exploit tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low costs solution" fits exactly with the purpose of the pursued system.

The controller has been implemented in the control centre because of operative reasons for the pilot test, however the objective is to embed it in the substation control units at distribution substations.

This paper's composition is as next, section 2 describes the system under study, the control devices and their operation policy. Section 3 introduces optimal power flow applications for voltage and reactive power control problem and their inefficiencies and limitations when applied to isolated and small power systems and with the substation automation issue. Section 4 proposes a fuzzy control system to deal with the regulation task. The report of a pilot test carried out in a substation working under real conditions is presented.

2. DESCRIPTION OF THE SYSTEM

The analyzed system is part of a 66/20 kV distribution substation as shown in Fig. 1, it is composed by a 66 kV primary bus, a 20 kV secondary bus and a 50 MVA power transformer

equipped with OLTC to regulate the secondary bus voltage and keep it nearby its specified value under changing load conditions. This is accomplished with 6 taps under and 15 taps over nominal settings, with voltage steps of 1.46 %.

Located at the secondary bus of the substation there are shunt capacitor banks capable of providing 4.2 MVar but connected through one only breaker, limiting the reactive power compensation to an all or nothing configuration. The operation policy is to limit the reactive power flow through the transformer but avoiding recirculation, setting a maximum value of leading power factor beyond which the capacitive way of working of the transformer would produce undesirable effects on generation, specially when demand falls at off-peak hours.

Reactive power management policies remain insufficiently developed, mainly due to the local nature of the problem and the complexity of developing a competitive market for the provision of reactive power [11]. Thus, voltage and reactive power control is normally included in the category referred to as ancillary services, which are necessary for the efficient and economical provision of active power.

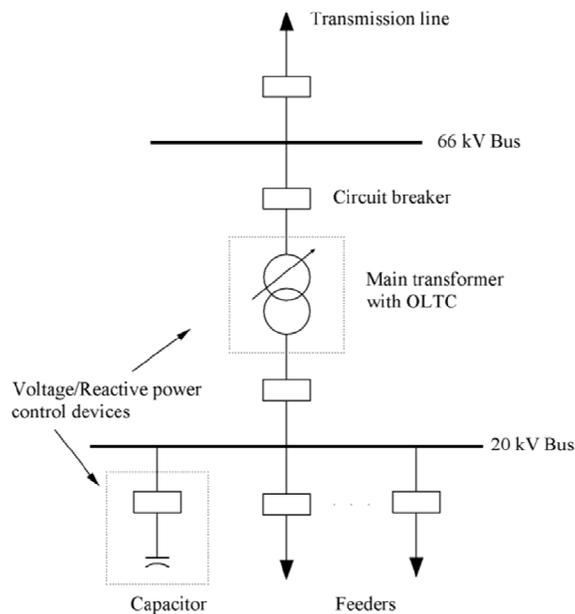


Figure 1. Part of a 66/20 kV distribution substation in the Canary Islands

The feeders connected to secondary bus include industrial and domestic customers. Constant impedance, constant current and constant power load models should be considered in order to describe the relationship between load demands and bus voltage. This distribution network has a slightly meshed topology but it is radially operated.

The SCADA system manages six stand alone electrical power networks, almost one per island, over 52 HV/MV substations and about 900 remotely controlled MV distribution centers. There are two frontends coexisting, the main one communicates with remote terminal units (RTUs) using IEC 60870-5-104:2006 protocol. The other one use WISP+ extended protocol, installations which still communicate with this frontend are gradually migrating to the main one.

The RTUs are solved with substation control units which carry out the communications and data management of substation protection, control, and metering intelligent electronic devices (IEDs) using IEC 60870-5-101:2003 protocol. They also provide a local human-machine interface.

Despite the classes of precision of the metering transformers at substations are 0.2 – 0.5 for the variables involved in voltage/reactive power regulation and at IEDs metering input cards the accuracy is 0.1%, the measurement resolution at SCADA is only: 100 V, 10 kW, 10 kVAr and 1 tap position. This is an important matter because as the refresh time of the value is just 4 seconds it is common the observation of oscillations of hundreds of volts and one of the only two control outputs, the tap changer, has voltage steps of 1.46% that is about 320 V. So, the effects of data uncertainty should be taken into account.

3. OPTIMAL POWER FLOW FOR REACTIVE POWER/VOLTAGE CONTROL PROBLEM

OPF applications improve objective functions, usually minimizing subtransmission losses or flattening the voltage profile with the aim of attaining a better quality of service. To deal with this, OPF compares the current status with that achievable by acting on control variables, in this case connecting shunts or changing transformer taps. If the calculated new state improves the objective function then outputs are acted.

The computational burden is high, due to power flow calculations for every secondary bus in all HV/MV substations. The response time is high too, so the solutions to under or overvoltage are delayed except when limits are exceeded, in those cases warnings are activated and operators act. Furthermore, as oscillations of the metered values are not taken into account and response time is high, the current status at the moment of the result control action may be quite different than that used for calculations. It has been pointed that required data for power flow calculations exceed the quality available.

Other aims such as limitation of switching number of capacitor or OLTC in a day, daily scheduling, and load characterization are dismissed.

Power flow calculations need data from multiple locations so local implementations at substations are discarded because of the huge amount of data exchange required.

Fig. 2, shows the voltage profile at secondary bus of the substation under study during a random day. The power factor at the boundary between transmission and distribution is shown in Fig. 3.

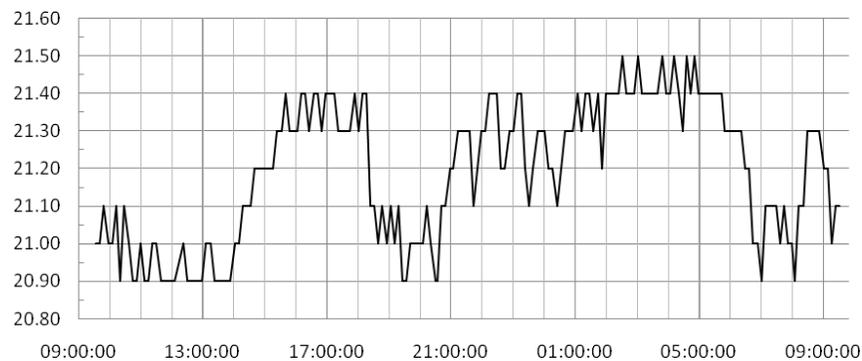


Figure 2. Voltage profile (y axis, kV) at secondary bus working with OPF

The desired value for voltage at secondary bus is 21.0 kV and the operation limits are defined at 20.3 and 21.6 kV. At those points the system alerts operators with low or high voltage alarms. It may be observed that although the profile shows the voltage inside operation zone, there are peaks during the afternoon and during the night. A better OLTC performance should be desirable. It acted 8 times keeping the tap position 2 from 23:23:38h even with 21,5 kV till 07:05:25h in the morning when voltage dropped to 20.9 kV.

Nowadays there is not a specific limit for power factor although in order to reduce losses in transmission network and to improve power transformer performance, values over 0.98 are desired. The profile shows leading power factor values close to 0.93 at 03:00 am and at 04:00 am. From 23:00 pm to 08:00 am the current is leading the voltage making the power transformer work in a capacitive way with its adverse associated effects. Despite of this, capacitor banks remained connected all day long. This illustrates the bad behavior of the running OPF based automatism.

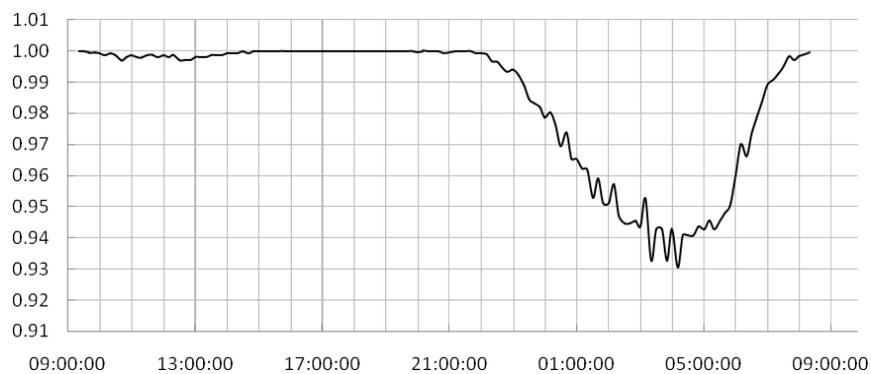


Figure 3. Power factor profile (y axis) at HV/MV boundary working with OPF

In paradigm of deregulated markets, the System Operator manages reactive power through three aspects applicable to all agents connected to the transmission network which are generators, transmission service providers and consumers:

- Reactive power constraints.
- Reactive power actually demanded.
- Payment for the service (penalties and incentives).

Distribution networks utilities are considered to be large consumers, this means that connection conditions in form of power factor constraint bands will be defined at every boundary transformer so reactive power profiles like shown would be penalized.

4. FUZZY LOGIC FOR REACTIVE POWER/VOLTAGE CONTROL PROBLEM

In point of fact systems such as this under study has few control options for reactive power/voltage control, so any trained operator would solve the task at least as fine as the OPF, with a lower response time and surely in a better way because furthermore from the regulation aim, other issues would be regarded such as time of day and expected increase of load, which leads to keep the voltage at the secondary bus a little bit over nominal values, the switching number of shunts and OLTC, issues to take into account in order to extend their life time, as well

as imprecision, uncertainties and oscillations given by sensors, transmitters, acquisition systems and its digitalization process that show a partial truth of reality.

The way proposed to transmit operators knowledge and their heuristic rules to deal with these aims to an automatic system is through fuzzy logic, a tool for embedding structured human reasoning into workable algorithms.

Many fuzzy inference systems has been proposed for voltage/reactive power control [12], [13], [14], [15] and even adaptive neuro-fuzzy inference systems [16] which provide a method for the fuzzy modeling procedure to learn information about the data set, in order to compute the membership function parameters that best allow the associated fuzzy inference system to track the given input/output data. However, the fuzzy logic in these cases has been oriented to:

- Estimation of sensitivities of load profiles.
- Find a feasible solution set to achieve buses voltage improvement and then identify the particular solution which most effectively reduces the power loss.
- Manage reactive power resources in a transmission network.
- Optimize an objective function which includes fuzzified variables based on a forecast of real and reactive power demands.

But always in a loop with a power flow routine that evaluates the progressive effect of control action until a criterion is met. These algorithms probably improve the performance given by the OPF based control system but preserve all disadvantages that disadvice it for isolated and small power systems. On the other hand, as commented in previous section, power flow calculations and its associated data traffic do not fit well with substation automation task.

A fuzzy inference system (FIS) has been designed as an on-line, real time tool, to give the proper dispatching strategy for capacitor switchings and tap movements such that satisfactory secondary bus voltage profile and main transformer power factor are reached.

The fuzzy controller acts directly from the result of its inference system. No further calculations are executed in order to estimate the future status achievable and to compare it to the current one. As previously mentioned, it is assumed that operators at distribution and dispatch control center, may decide good strategies for voltage/reactive power regulation within the limits of the control capabilities of the system. So the work done consisted in the transmission of this knowledge from operators to the controller through heuristic rules based on their past experience, allowing it working in automatic way.

The input variables defined were: voltage at secondary bus, reactive power flow through the transformer measured at HV winding, tap position and shunt capacitor status switched on or off. Fig. 4 shows the fuzzy model membership function for the input Voltage.

Where the fuzzy linguistic sets for the input Voltage were: very low (VL), low (L), low on-peak time period (LP), good (G), high (H), high on-peak time period (HP) and very high (VH).

The ouput variables defined were orders to move up or down the OLTC and switch on or off the shunt capacitor banks. Fig. 5 shows the fuzzy model membership function for the output Taps.

The inference system was designed using Matlab, it was Mamdani type with 14 rules. The logic guidelines embedded were like:

- If (Reactive_power is High) and (Tap is Normal) and (Shunt_Off is Disconnected) then (Tap is -2)(Capacitor is Connect).
- If (Voltage is H) and (Reactive_power is Good) and (Tap is not Tap1) then (Tap is -1).

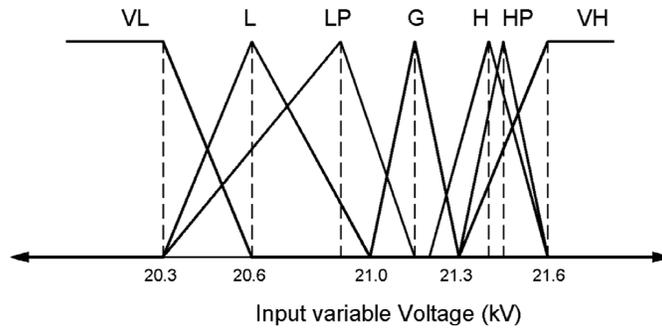


Figure 4. Sample of the fuzzy model membership functions. Input Voltage

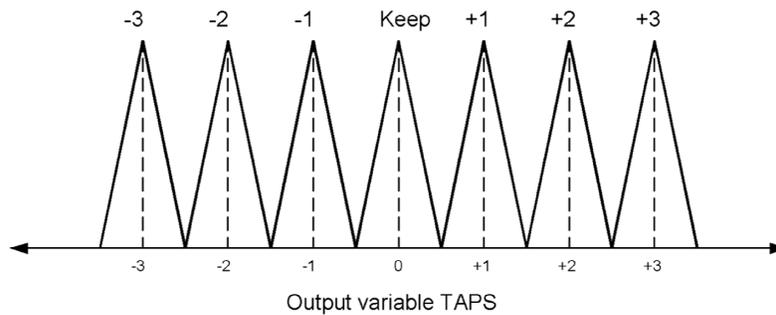


Figure 5. Sample of the fuzzy model membership functions. Output Taps

Fig. 6 shows the voltage profile at secondary bus of the substation and Fig. 7 shows the power factor at the boundary between transmission and distribution registered one week later than those showed working with OPF, this time working with the designed FIS.

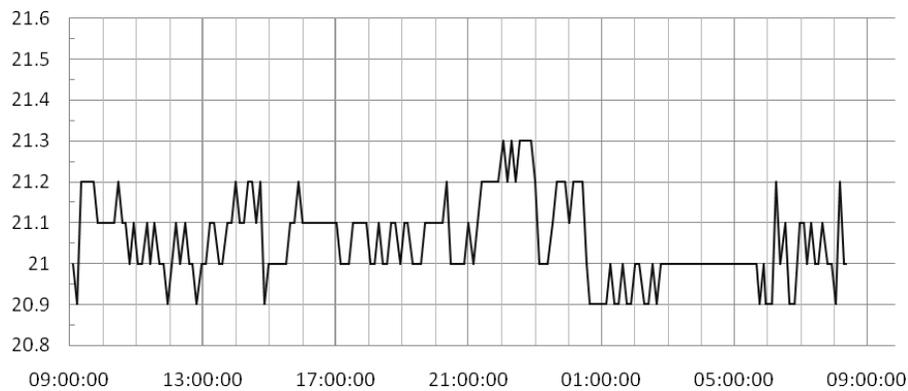


Figure 6. Voltage profile (y axis, kV) at secondary bus working with FIS

The voltage profile has been flattened and the power factor has been kept 23 hours of the day over 0.99 and always over 0.98. The OLTC just acted 11 times, a value well below the maximum defined (the maximum switching number of shunts and OLTC were defined as 30 and 6 respectively in order to extend their life time). The capacitor banks were disconnected from the secondary bus at 23:55:30h and were reconnected again at 08:13:19h.

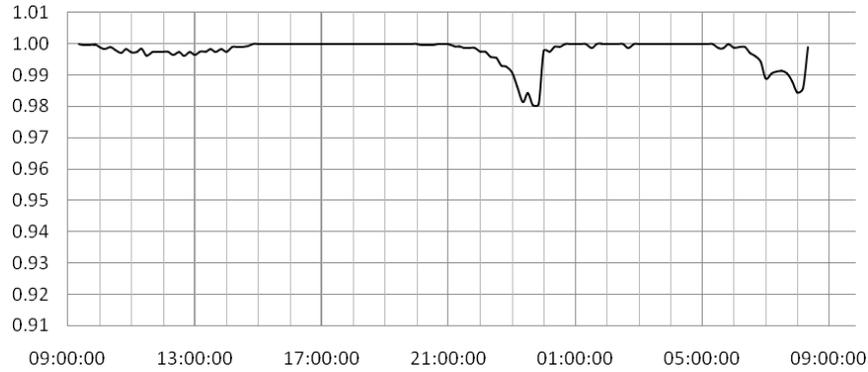


Figure 7. Power factor profile (y axis) at HV/MV boundary working with FIS

Fig. 8 shows the voltage profile at secondary bus comparison between two random days (april 16 and 22) working with an OPF based controller and one day (april 23) working with a FIS based controller.

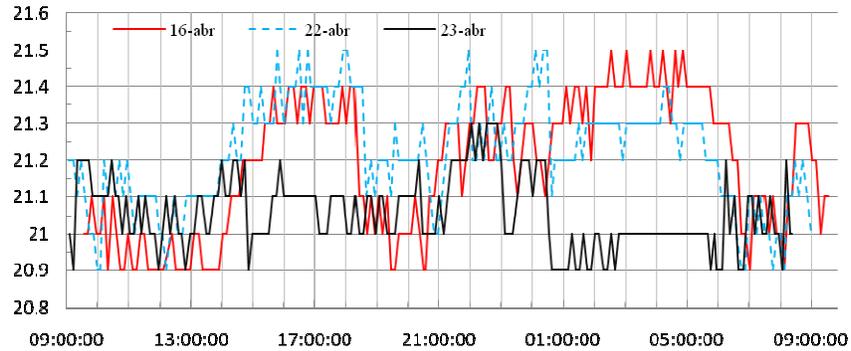


Figure 8. Voltage profile (y axis, kV) at secondary bus comparison

Table 1 shows statistical summary related to these voltage profiles. Average (\bar{U}), maximum deviation (DM) and mean deviation (Dm) both referred to objective value calculated using expressions (1) and (2).

$$D_M = \text{Max} |U_i - 21| \quad (1)$$

$$D_m = \frac{\sum_{i=1}^n |U_i - 21|}{n} \quad (2)$$

where U_i is the voltage value at the measure i , and n is the number of registered measurements.

Table 1. Statistical summary of voltage profiles results.

	OPF (Apr-16)	OPF (Apr-22)	FIS (Apr- 23)
\bar{U}	21.1914 kV	21.2178 kV	21.0537 kV
D_M	0.5000 kV	0.5000 kV	0.3000 kV
D_m	0.2192 kV	0.2251 kV	0.0792 kV

The average voltage has been kept exactly in the reference value considering its measurement resolution. The maximum deviation has been reduced 200 V and the mean deviation had a 64 % drop off.

In order to evaluate the benefits of reactive power compensation, losses at transmission network with the system working with OPF based automatic control and with the FIS based one are compared. For this purpose only Joule effect losses are taken into account as they are the most important ones. The relationship is evaluated by expression (3).

$$\phi = \frac{Losses_{FIS}}{Losses_{OPF}} = \left(\frac{\cos \varphi_{OPF}}{\cos \varphi_{FIS}} \right)^2 \quad (3)$$

As power factor does not remain constant in either case, neither does the losses relationship. Its profile is shown in Fig. 9. Although there are values over 1, mostly at instants after connection and before disconnection of the capacitor banks due to the limit reactive power recirculation policy embedded in the FIS, values of 0.8660 are reached. It implies 13.40 % losses reduction.

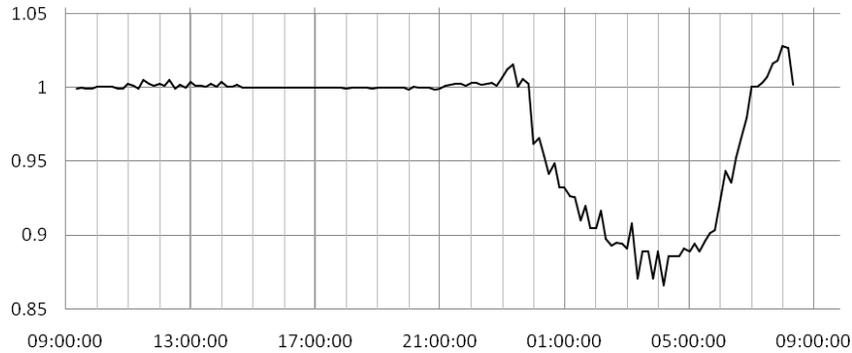


Figure 9. Losses ratio profile. FIS/OPF

The average losses ratio has been calculated using (4).

$$\bar{\phi} = \frac{\sum_{i=1}^n \left(\frac{\cos \varphi_i_{OPF}}{\cos \varphi_i_{FIS}} \right)^2}{n} \quad (4)$$

where $\cos \varphi_i$ is the power factor value at the measure i , and n is the number of registered measurements.

Table 2 shows average losses ratio results for one day period comparison and during the interval during which FIS based system decided to disconnect capacitor banks and OPF based did not.

Table 2. Average losses ratio between OPF and FIS based controllers.

	$\bar{\phi}$	Loss Reduction
24h	0.9750	2.50 %
23:55:30h to 08:13:19h	0.9312	6.88 %

The average loss reduction for all day comparison was 2.50 % and in the interval during which controllers had different behavior the average loss reduction was 6.88 %. In order to achieve acceptable and valid results, days with the same load profile were compared. For this purpose it was selected the same day of the week before the day the FIS based controller was tested.

A better performance could be affordable with the capacitor banks working with several switches with staggered configuration however the results obtained ostensibly improve those achieved with the OPF based automatic control.

5. CONCLUSIONS

A fuzzy inference system has been designed as an on-line, real time tool, to give the proper dispatching strategy for capacitor switchings and tap movements for control of reactive power/voltage in a distribution substation.

The system passed all the security matters around power systems and was tested in a real case under real conditions. It managed the operation at dispatch control center during a whole day for one distribution substation in the Canary Islands.

Within the limits of the control capabilities of the system, the fuzzy controller maintained successfully an acceptable and flattened voltage profile, reducing 36.13% (from 219.2V to 79.2V) the mean deviation referred to objective value. The reduction of computational burden has been significative and has led to lower time response when submitted to changing operation conditions. The average voltage has been kept exactly in the reference value considering its measurement resolution.

The power factor values are kept over 0.98 and has ostensibly improved the performance given by an optimal power flow based automation system. The ratio of losses improvement during a day period has been estimated in 2.50 % reaching 13.40 % loss reduction at certain times of the day .

It has been demonstrated that for distribution substations in stand alone electric power systems with limited control actions, fuzzy inference systems solve properly the voltage/reactive power control task.

The logic proposed may be embedded in the substation control unit. This would release the control centre of these tasks, although retaining the ability to monitor, supervise and even configure the inference system. It should be taken into account in the challenge of substation automation in isolated and small power systems.'

SYMBOLS

- \bar{U} : Average voltage.
 D_M : Maximum voltage deviation.
 D_m : Mean voltage deviation.
 ϕ : Losses ratio.
 $\bar{\phi}$: Average losses ratio.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Endesa Distribución Eléctrica, S. L. and its staff at dispatch control center in Canary Islands for providing the valuable system data and dispatch experience.

REFERENCES

- [1] Buchholz B. M. and Styczynski Z. B., (2011) "The three pillars of Smart Distribution realized by IEC 61850 Communications", presented at the 16th Int. Conf. on Intelligent System Applications to Power Systems, Crete, Greece.
- [2] Sun D., Ashley B., Brewer B., Hughes A. and Tinney W.F., (1984) "Optimal power flow by Newton approach", IEEE Trans. Power Apparatus and Systems, vol. 103, no. 10, pp. 2864-2880.
- [3] Sun D. I., Hu T. I., Lin G. S., Lin C. J. and Chen C. M., (1988) "Experiences with implementing optimal power flow for reactive scheduling in the Taiwan power system", IEEE Trans. Power Systems, vol. 3, no. 3, pp. 1193-1200.
- [4] Bjelogrić M., Calović M. S. and Babić B. S., (1990) "Application of Newton's optimal power flow in voltage/reactive power control", IEEE Trans. Power Systems, vol. 5, no. 4, pp. 1447-1454.
- [5] Huneault M. and Galianu F. D., (1991) "A survey of the optimal power flow literature", IEEE Trans Power Systems, vol. 6, no. 2, pp. 762-770.
- [6] Augugliaro A., Dusonchet L., Favuzza S. and Riva E., (2004) "Voltage regulation and power losses minimization in automated distribution networks by an evolutionary multiobjective approach", IEEE Trans. Power Systems, vol. 19, no. 3, pp. 1516-1527.
- [7] Baran M. E. and Hsu M-Y., (1999) "Volt/Var control at distribution substations", IEEE Transactions on Power Systems, vol. 14, no. 1, pp. 312-318.
- [8] Lu F. C. and Hsu Y-Y., (1997) "Fuzzy dynamic programming approach to reactive power/voltage control in a distribution substation", IEEE Trans. Power Systems, vol. 12, no. 2, pp. 681-688.
- [9] Khiat M., Chaker A., Gómez A. and Martínez J.L., (2003) "Reactive power optimization and voltage control in the Western Algerian transmission system: a hybrid approach", Electric Power Systems Research, vol. 64, no. 1, pp. 3-10.
- [10] Trigo A. L., Martínez J. L., Riquelme J. and Romero E., (2011) "A Heuristic Technique to Determine Corrective Control actions for Reactive Power Flows", Electric Power Systems Research, vol. 81, no. 1, pp. 90-98.
- [11] Gross G., Tao S., Bompard E. and Chicco G., (2002) "Unbundled reactive support service: key characteristics and dominant cost component", IEEE Trans. on Power Systems, vol. 17, no. 2, pp. 283-289.
- [12] Starrett S. K., Anis W. R., Rust B. P. and Turner A. L., (1999) "An on-line fuzzy logic system for voltage/var control and alarm processing", conference paper, IEEE Power Engineering Society Winter Meeting, vol. 1, pp 766-771.
- [13] Su C. T. and Lin C.T., (2001) "Fuzzy-based voltage/reactive power scheduling for voltage security improvement and loss reduction", IEEE Trans. Power Delivery, vol. 16, no. 2, pp. 319-323.
- [14] Rahideh A., Gitizadeh M. and Rahideh A., (2006) "Fuzzy logic in real time voltage/reactive power control in FARS regional electric network", Electric Power Systems Research, vol. 76, no. 1, pp. 996-1002.
- [15] Miranda V., Moreira A. and Pereira J., (2007) "An improved fuzzy inference system for voltage/var control", IEEE Trans. Power Systems, vol. 22, no. 4, pp. 2013-2020.

- [16] Ramakrishna G. and Rao N. D., (1999) "Adaptive neuro-fuzzy inference system for volt/var control in distribution systems", Electric Power Systems Research, vol. 49, no. 2, pp. 87-97.

AUTHORS

Eduardo Vega-Fuentes received his M.Sc. degree in electrical engineering from the University of Las Palmas de Gran Canaria (ULPGC), Spain, in 1998. He is currently pursuing the Ph.D. degree in the same university. Since 2004 he has been with the Department of Electronics Engineering and Automatics in ULPGC as an Associate Professor in Systems Engineering and Automatics. His main research interests include distribution efficiency, optimal power system operation and distribution automation.



Sonia León-del Rosario received her M.Sc. degree in Electrical Engineering from the University of Las Palmas de Gran Canaria (ULPGC), Spain, in 1998. She is currently working towards the Ph.D. degree in Electrical Engineering in the ULPGC. She is in the Department of Electronics Engineering and Automatics at the same university working as Associate Professor. Her research interests include intelligent techniques applied to Forecasting and Power Systems.



Juan Manuel Cerezo-Sánchez received his M.Sc. degree in Telecommunications Engineering from the University of Las Palmas de Gran Canaria (ULPGC), Spain, in 1993. He is currently working towards the Ph.D. degree in Telecommunications Engineering in the ULPGC. He is in the Department of Electronics Engineering and Automatics working as Professor at the same University. His research interests are SCADA systems and Industrial Communications



Aurelio Vega-Martínez received his M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Las Palmas de Gran Canaria (ULPGC), Spain, in 1987 and 1992 respectively. He is in the Department of Electronics Engineering and Automatics of the ULPGC. His main research interests include distribution automation, power system operation and control systems design.



TOWARDS UNIVERSAL RATING OF ONLINE MULTIMEDIA CONTENT

Lawrence Nderu¹, Nicolas Jouandeau², and Herman Akdag²

¹Jomo Kenyatta University of Agriculture and Technology, Kenya

¹nderu@jkuat.ac.ke

²University of Paris 8 –LIASD, France

{n, akdag}@ai.univ-paris8.fr

ABSTRACT

Most website classification systems have dealt with the question of classifying websites based on their content, design, usability, layout and such, few have considered website classification based on users' experience. The growth of online marketing and advertisement has led to fierce competition that has resulted in some websites using disguise ways so as to attract users. This may result in cases where a user visits a website and does not get the promised results. The results are a waste of time, energy and sometimes even money for users. In this context, we design an experiment that uses fuzzy linguistic model and data mining techniques to capture users' experiences, we then use the k-means clustering algorithm to cluster websites based on a set of feature vectors from the users' perspective. The content unity is defined as the distance between the real content and its keywords. We demonstrate the use of bisecting k-means algorithm for this task and demonstrate that the method can incrementally learn from user's profile on their experience with these websites.

KEYWORDS

Website Classification, Fuzzy Linguistic Modeling, K-Means Clustering, Web Mining.

1. INTRODUCTION

The Internet has become a major source of information. Individuals and organizations depend on the Internet in one way or another. It can be asserted that the web is the largest available repository of data with the largest number of users [1]. Therefore, the web can be viewed as a meeting point of providers of information and services and their consumers.

Since its early days the Internet has seen remarkable growth. This growth is fueled by millions who provide high quality, trustworthy content. However, in this favorable landscape a good number of providers may seek to profit by promising users resources which eventually they cannot or do not provide. This leads to cases whereby a user might end up wasting time, energy and even sometimes money. This situation may also create a disillusioned user. Our goal is to develop a website clustering system that takes into consideration the previous users' experiences.

The paper is set out as follows: First, we introduce some literature on web mining and fuzzy linguistic modeling. Secondly, an experiment is proposed to compute the agreement between what a web site claims it provides and what in-fact it provides based on users browsing patterns. Third, a discussion is provided to demonstrate the feasibility and effectiveness of the proposed model.

Finally, some conclusions are presented at the end of this paper.

2. WEB MINING AND FUZZY LINGUISTIC MODELING

2.1 Web Mining

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This includes the automatic search of information resources available on-line, i.e. Web content, data mining and discovery of users access patterns from Web servers [2].

As information technology grows, users can aggregate and store mass data more easily and efficiently [3]. Data mining can help users analyze and retrieve valuable information from mass data sources available online [4]. Due to the existence of a variety of websites that claim to provide the information that the users are looking for, users find themselves dealing with the question of identification of sources that deliver what they claim. Web mining can be used to analyze user's browsing behavior and provide suitable information for future users of the same websites. Important data can be obtained from Web servers or proxy servers such as log files, user profiles, registration data and user sessions or transactions [5]. From these data we can discover valuable information about the websites visited.

2.2 Fuzzy Linguistic Modeling

Fuzzy linguistic modelling [6] is a very useful kind of fuzzy linguistic approach used for modeling the computing with words process as well as linguistic aspects [7]. When collecting details from the user's on-line activities some aspects are qualitative while others are quantitative. Fuzzy linguistic modelling has been widely used and has provided very good results, it deals with qualitative aspects that are presented in qualitative terms by means of linguistic variables [8], [9], [10], [1]. The 2-tuple Fuzzy Linguistic representation model provides the following advantages over classical models [1].

- 1) The linguistic domain can be treated as continuous, while in the classical models it is treated as discrete.
- 2) The linguistic computational model based on linguistic 2-tuples carries out processes of computing with words easily and without loss of information.
- 3) The results of the process of computing with words are always expressed in the initial linguistic domain.

The model used in this paper to evaluate the conformance of what the web site says and what it actually delivers uses a set of quality criteria related to the Web sites and a computation instrument of quality assessments. We assume that the quality of a web site is measured through users perceptions on the services offered through its Web site [11] and that users have an objective to achieve when they decide to visit a certain website.

Users are invited to fill in a survey built on a set of quality criteria. To measure quality, conventional measurement tools used by the customers are devised on cardinal or ordinal scales. However, the scores do not necessarily represent user preferences. This is because respondents have to internally convert preference to scores and the conversion may introduce distortion of the preference [3]. For this reason, we use fuzzy linguistic modeling to represent the user's perceptions and tools of computing with words to compute the quality assessments [7]. The subjectivity and vagueness in the assessment process is dealt with using the fuzzy logic [12]. Multiple raters are often preferred rather than a single rater to avoid the bias and to minimize the partiality in the decision process. Figure 1 shows an example of the way the Fuzzy computation with words was

used in the experiment in one of the question to the users. Does the website provide accurate information? The choices were None-N, Very Little- VL, Little- L, Medium- M, High- H, Very High- VH and Perfect- P.

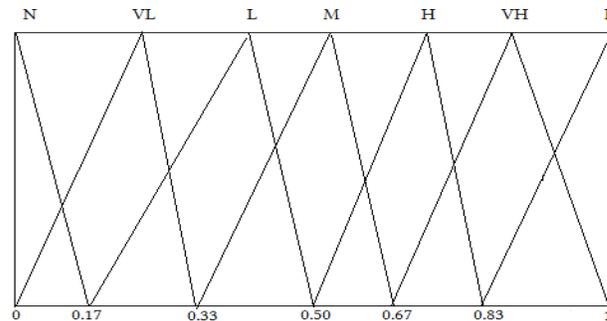


Figure 1. Linguistic variable "Accurate".

3. CLUSTERING FUZZY WEB ACCESS PATTERNS BASED ON K-MEANS

By making use of the log files details and analysis of the results obtained from the online questioners, it is possible to adapt the paradigm of clustering. The key effectiveness of the clusters is an intuitive distance function [13]. Since the content of each single page $p \in W$ can be represented by a feature vector of term frequencies, the whole details measurable properties as seen from the users is represented by set of feature vectors. One assumption made by this approach is that sites that fulfills a user needs as seen from the users pattern mining of related content but varying size will become very similar with respect to the Sum of Minimum Distances (SMD) [14]. Let U_1, U_2 be two users and let $f: P \rightarrow N^d$ be a feature transformation that returns the feature vector of a users' profile $p \in P$ where P is the set of all profiles. The concept of SMD is discussed by Hans et al in [15]. Using these details we create $S_1, S_2 \dots S_3$ as centroids that represent the various log features. Figure 2 shows the modeling of the clusters based on the fact that websites are different, based on the fact that users have different expectations when visiting the websites.

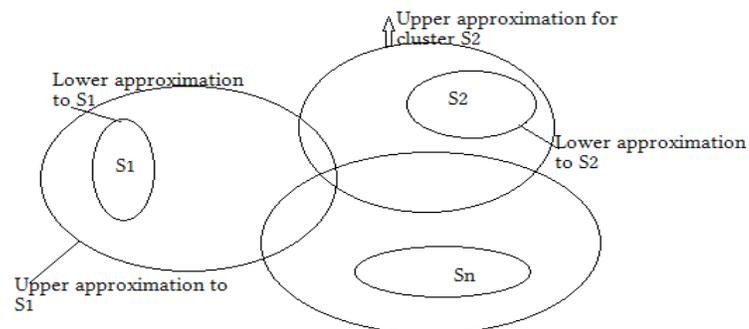


Figure 2. The Clustering Model.

4. EXPERIMENTAL SETUP

An experiment was performed with the aim of collecting data from two sources, log analysis and online questionnaires. Students were used as experimental subjects. It is a known fact that this could lead to results that are artificial, particularly when students are asked to perform tasks which they have little or no experience. After identifying this as a potential problem, a number of websites were selected. The selection criterion was that a specific objective needed to be achieved by the website users. The time users took on this websites was then read from the freely provided Google API. A total of 199 randomly selected samples were collected, this included the time taken in a website and the time it would take to carry out the specific objective on the website.

Several methodologies have been suggested for rating websites [17], [2], since our objective was to cluster websites based on what it promises, a new criteria needed to be developed. The criteria developed identified the following as the important features with respect to our objective: accuracy, believable, relevant, details, value, revisit, and deliverability. These factors were obtained from the literature. Table 1, shows the online questionnaire with results from one rater of the websites. In this example on Table 1, the user is rating a website that he has already visited. The questionnaires were administered for ten visited URLs for which already the time taken by users while visiting this websites had already been noted. The fact that we had 199 log details for users and only selected 10 URLs to evaluate was dictated by a number of factors, the major one being the evaluation that we had carried out about the nature of the websites that we were to use in the study, but to introduce randomness this information was not available to the users. Table 2 shows the analyzed results for the ten URLs. The arithmetic mean for the 2-tuple linguistic aggregation operators was used for calculation.

Table 1. Website Online Questionnaire Example.

		None	Very Little	Little	Medium	High	Very High	Perfect
1.	Accurate information							X
2.	Believable information						X	
3.	Relevant information							X
4.	The right detail of information							X
5.	Value for your time			X				
6.	Would you visit it again				X			
7.	Does it deliver							X

Table 2. Analyzed data from the questionnaires.

URL	Accuracy	Believable	Relevant	Details	Value	Revisit	Deliverability
1	8.72	8.33	8.88	7.89	8.33	8.55	8.49
2	2.39	2.78	3.11	1.45	1.40	1.62	2.17
3	8.00	8.44	8.22	8.44	8.22	7.88	8.33
4	2.72	2.78	3.17	3.67	3.50	4.00	2.33
5	8.27	7.44	7.22	7.22	7.22	8.11	8.16
6	3.22	2.95	4.56	2.89	2.33	3.33	3.83
7	8.55	8.49	6.72	8.16	7.50	9.32	7.33
8	6.78	8.11	8.33	7.22	8.11	7.77	7.22
9	3.50	3.67	2.67	4.33	2.83	2.83	3.67
10	2.89	4.06	2.72	3.50	3.56	3.17	4.17

5. RESULTS AND ANALYSIS

The Bisecting K-means algorithm was used. Table 3, shows the results obtained from the clustering program and Figure 3, shows the clusters for the ten URLs. The websites used for this experiment URL1, URL2..., URL10 were well analyzed with respect to what the websites promised to deliver. Any new website that promises to deliver what these websites were delivering can now be clustered with respect to the created clusters.

Table 3. Clustering Using k-Means.

Cluster	Data Unit	Sum	Time(Seconds)
1	0	59.2	5.3
	2	57.5	5.2
	4	54.1	5.5
	6	56.1	6.0
	7	53.5	5.6
2	1	14.9	2.7
	3	22.2	3.0
	5	23.1	3.1
	8	23.5	2.9
	9	24.1	3.6
Clustered Data: K=2			

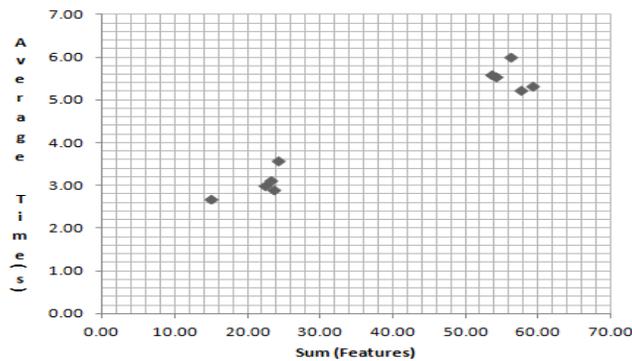


Figure 3. Clusters for the ten URLs.

6. CONCLUSION

In this paper, we proposed a new solution to automatic classification of websites based on the level of satisfaction of the previous users'. The experiment shows that clustering can be used as a way of telling how far a website is from the ideal users' website. The results so far point to an interesting direction in the sense that previous users' experience can be used to rank website and provide a metric for classification of websites. Combinations of these results are key point to developing future websites classification system.

REFERENCES

- [1] E. Herrera-viedma, A. G. Lopez-herrera, and C. Porcel, "Evaluating the Information Quality of Web Sites : A Methodology Based on Fuzzy Computing With Words," vol. 57, no. 4, pp. 538–549, 2006.
- [2] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining : Information and Pattern Discovery on the World Wide Web * 1 Introduction," pp. 558–567, 1997.
- [3] C. Chen, W. Tai, and C. Chu, "A preference perception system on website by combining fuzzy set with data mining technology," *Int. J. Inf. ...*, vol. 19, no. 1, pp. 93–105, 2008.
- [4] U. M. Fayyad, "Data Mining and Knowledge Applications in Astronomy Discovery in Databases : Science and Planetary," pp. 1590–1592, 1996.
- [5] R. Kosala, B.- Heverlee, and H. Blockeel, "Web Mining Research : A Survey," vol. 2, no. 1, 2000.
- [6] F. Herrera, E. Herrera-Viedma, and J. Verdegay, "Direct approach processes in group decision making using linguistic OWA operators," *Fuzzy Sets Syst.*, no. ii, 1996.
- [7] Y.-J. Xu and Z.-J. Cai, "Method Based on Fuzzy Linguistic Judgement Matrix and Trapezoidal Fuzzy Induced Ordered Weighted Geometric (TFIOWG) Operator for Multi-Attribute Decision-Making Problems," 2007 *Int. Conf. Wirel. Commun. Netw. Mob. Comput.*, no. 1, pp. 5752–5755, Sep. 2007.
- [8] L. Feng and T. S. Dillon, "to Provide Explanatory Semantics for Data Warehouses," vol. 15, no. 1, pp. 86–102, 2003.
- [9] M. Decision-making, F. Herrera, and L. Martínez, "A Model Based on Linguistic 2-Tuples for Dealing with Multigranular Hierarchical Linguistic Contexts," vol. 31, no. 2, pp. 227–234, 2001.
- [10] G. Bordogna and G. Pasi, "A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval : A Model and Its Evaluation," vol. 44, no. 2, pp. 70–82, 1993.
- [11] L. Hidalgo and F. J. C. J. L. G. E. Herrera-viedma, "Applying Fuzzy Linguistic Tools to Evaluate the Quality of Airline Web Sites," 2007.
- [12] L. a. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Inf. Sci. (Ny)*, vol. 8, no. 3, pp. 199–249, Jan. 1975.
- [13] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets Syst.*, vol. 158, no. 19, pp. 2095–2117, Oct. 2007.
- [14] Y. Sharon, J. Wright, and Y. Ma, "Minimum sum of distances estimator: robustness and stability," *Am. Control Conf. 2009. ...*, pp. 524–530, 2009.
- [15] H. Kriegel and M. Schubert, "Classification of Websites as Sets of Feature Vectors.," *Databases Appl.*, pp. 127–132, 2004.
- [16] W. Hung and R. J. Mcqueen, "Developing an Evaluation Instrument for e-Commerce Web Sites from the First-Time Buyer ' s Viewpoint," pp. 31–42, 2003.
- [17] S. J. Barnes and R. T. Vidgen, "AN INTEGRATIVE APPROACH TO THE ASSESSMENT OF E-COMMERCE QUALITY," no. August 1998, pp. 114–127, 2000.
- [18] F. Herrera, "A 2-tuple fuzzy linguistic representation model for computing with words - Fuzzy Systems, *IEEE Transactions on*," vol. 8, no. 6, pp. 746–752, 2000.

QUERY PROOF STRUCTURE CACHING FOR INCREMENTAL EVALUATION OF TABLED PROLOG PROGRAMS

Taher Ali¹, Ziad Najem², and Mohd Sapiyan¹

¹Department of Computer Science, Gulf University for Science and
Technology, Kuwait

ali.t@gust.edu.kw, sapiyan.m@gust.edu.kw

²Department of Computer Science, Kuwait University, Kuwait

najem@cs.ku.edu.kw

ABSTRACT

The incremental evaluation of logic programs maintains the tabled answers in a complete and consistent form in response to the changes in the database of facts and rules. The critical challenges for the incremental evaluation are how to detect which table entries need to change, how to compute the changes and how to avoid the re-computation. In this paper we present an approach of maintaining one consolidate system to cache the query answers under the non-monotonic logic. We use the justification-based truth-maintenance system to support the incremental evaluation of tabled Prolog Programs. The approach used in this paper suits the logic based systems that depend on dynamic facts and rules to benefit in their performance from the idea of incremental evaluation of tabled Prolog programs. More precisely, our approach favors the dynamic rules based logic systems.

KEYWORDS

Incremental evaluation of tabled Prolog, Incremental tabulation for Prolog queries, Justification based truth maintenance systems, Tabulation, Memoing.

1. INTRODUCTION

Tabled resolution for logic programs [1] mitigates some of the well-known problems of Prolog, including the tendency to fall into infinite loops, repeating subcomputations, and the unsatisfactory semantics of negation. The implementations of tabling [2, 3, 4,5] have become stable and efficient. The incremental evaluation of logic programs [6] maintains the tabled answers complete and consistent in response to the changes in the database of facts and rules. The basic idea behind incremental tabulation is that when some facts or rules change in a program, the system recomputes only the results affected by the change, instead of re-evaluating and tabling the query answers from scratch. The critical challenges for the incremental evaluation are how to detect which table entries need to change, and how to compute the changes. One of the efficient approaches to achieve these challenges is to use the symbolic support graph [7]. The symbolic

support graph caches the dependencies between the tabled answers to propagate the changes to the tables when the related facts/rules are added/deleted. This approach requires to cache the answers of the query in a table along with the support graph to maintain the completeness and correctness of tabled answers.

```

: -table connected/2
edge(a,b) %F1
edge(a,c) %F2
edge(b,d) %F3
edge(c,d) %F4
edge(d,e) %F5
edge(f,g) %F6
connected(X,Y) : -edge(X,Y). %R1
connected(X,Y) : -edge(X,Z),connected(Z,Y). %R2

```

Figure 1: Translative closure program of the directed edge relationship

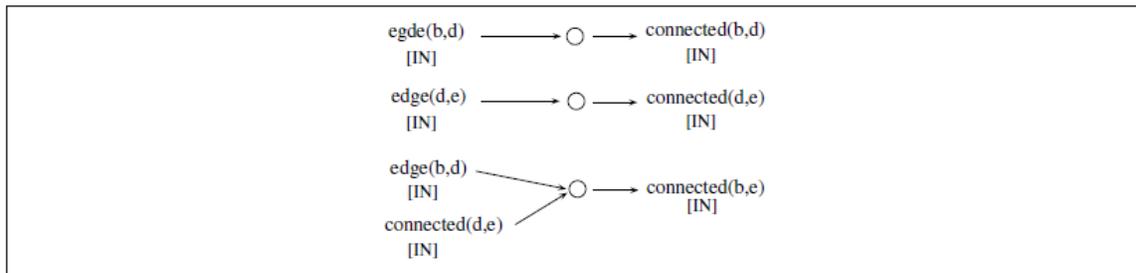


Figure 2: JTMS network installed by THE SYSTEM after proving query ? - *connected* (b, Y) for the first time.

The other challenge for the incremental evaluation is to avoid the re-computation which is required to update the tabled answers due to the changes in the related database of facts and rules. The current technique [8] uses extra data structures (dynamic dependency graph) to interleave the propagation of deletion and insertion operations caused by the updates of facts and rules. The technique tries to minimize the challenging problem of re-computation which is caused by the updates. This paper presents an alternative approach to incremental tabulation that is capable of working in non-monotonic situations. The main idea is to cache the proof generated by the deductive inference engine rather than the end results. In order to be able to efficiently maintain the proof to be updated, the proof structure is converted into a justification-based truth-maintenance (JTMS) network [9, 10].

2. CACHING THE QUERY PROOF AS A JTMS NETWORK

The main idea of our approach is to cache the proof generated by the deductive inference engine rather than caching the end results. The proof structure is converted into a justification-based truth-maintenance (JTMS) network. JTMS saves the dependency between deduced facts and the facts used to make the deduction in order to be able to efficiently cache the proof structure. The system translates every successful branch of a query into a JTMS network that links the facts and

the rule to the answer generated by that branch. Consider the evaluation of the query: $? - \text{connected}(b, Y)$ with respect to the PROLOG program of Figure 1. Figure 2 shows the justifications installed by the system when it proves the query $? - \text{connected}(b, Y)$ with respect to the PROLOG program of Figure 1. These justifications represent the proof structure of the query $? - \text{connected}(b, Y)$. A justification is installed for each complete branch of the SLD-tree. When a query is reevaluated, the system returns the answers of the query by collecting the IN consequences of each query's JTMS justification.

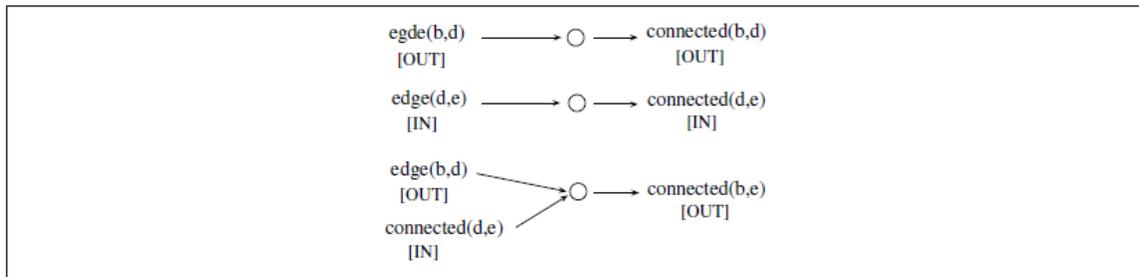


Figure 3: JTMS network of Figure 2 after retracting the fact $\text{edge}(b, d)$ to the database of Figure 1.

When changes in database take place, we have to ensure that the proof structure is both sound and complete. Consider the following changes to the base facts of Figure 1 after caching the proof structure of the query $? - \text{connected}(b, Y)$ for the first time:

1. Retracting the fact $\text{edge}(b, d)$

The system has to ensure that whenever base facts participating as antecedents in any justification are asserted/retracted, the effect of this assertion/retraction should be propagated through the JTMS justifications in order to keep the proof structure sound. Achieving this is not difficult since changing the state of any antecedent that is asserted/retracted to/from the database requires marking the label from IN/OUT or vice versa, and after that, propagating the effect of this change through the whole network. Figure 3 shows the effect of retracting the fact $\text{edge}(b, d)$ from the database of the Figure 1. The first effect of this retraction is on the first justification since $\text{edge}(b, d)$ is in the antecedent list of that justification. This results in marking $\text{connected}(b, d)$ from IN to OUT. Since $\text{edge}(b, d)$ is in the antecedent list of the 3rd justification, the result of outness propagation marks $\text{connected}(b, e)$ from IN to OUT. This method of propagating inness/outness ensures that whenever the query is re-evaluated, the returned results by the system are valid answers regardless whether or not the database has been changed. Note that ensuring the soundness of the proof structure does not require any PROLOG inference work.

2. Asserting the fact $\text{edge}(b, f)$

Here the situation is more complicated. the system has to take care about the effect of asserting new data that was not available when a query was evaluated for the first time. This is important since asserting new data to the database may add to the set of results that are already available for the query or even remove some of them. The system handles this problem by monitoring the nodes that may contribute to some new results of the query.

Whenever a new fact that is related to a monitored node is asserted, query resumption takes place to update the query's cached proof structure. Referring back to the example of Figure 1, when the system proves the query $?-connected(b, Y)$ for the first time, it marks the nodes that will participate in resuming this query when new data is asserted. Those nodes come from the right hand side of program rules, i.e. $edge(X, Y)$ and $connected(Z, Y)$. Whenever new data that is related to the marked nodes is asserted, the query $?-connected(b, Y)$ resumes its work to update the proof structure of the query. Figure 4 shows the effect of asserting the fact $edge(b, f)$ to the database of Figure 1 on the JTMS network of Figure 3. Three new justifications have been installed upon resuming the query after the assertion of $edge(b, f)$. An important point that should be mentioned here is that, in order to keep the proof structure complete, the system has to use the help of the PROLOG inference engine.

3. Asserting the fact $edge(b, d)$

The retracted fact $edge(b, d)$ is asserted back to the base facts of Figure 1. The system is going to change the label of the TMS node attached to this fact from OUT to IN and then propagates the effect of this change in label throughout the JTMS network. Figure 5 shows the effect of asserting back the fact $edge(b, d)$ to the database of Figure 1 on the JTMS network of Figure 4.

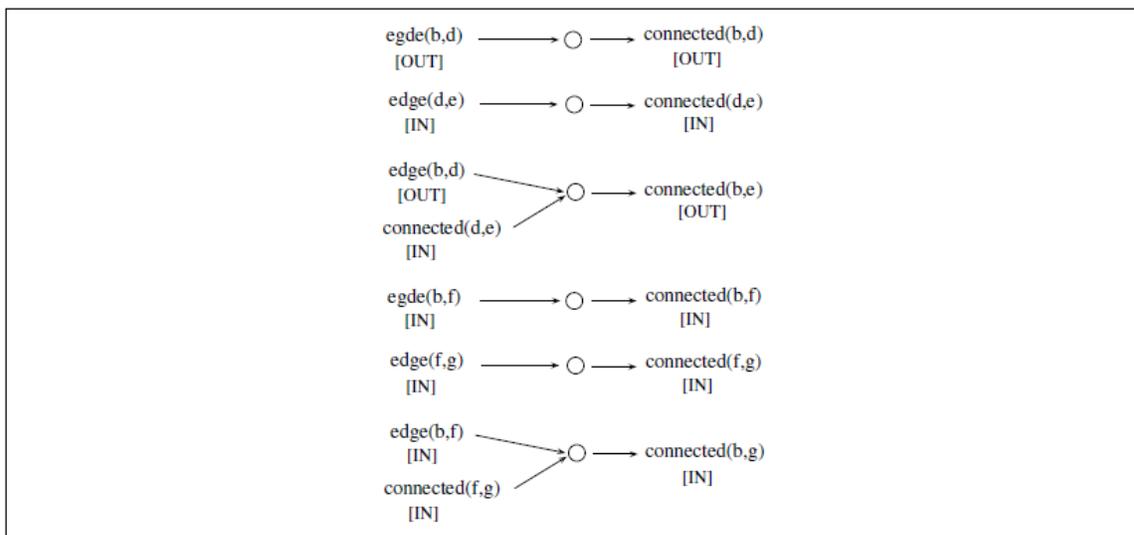


Figure 4: JTMS network of Figure 3 after asserting the fact $edge(b, f)$ to the database of Figure 1.

3. IMPLEMENTATION

The main objective of this research is to provide a PROLOG system which supports incremental tabulation by using the justification-based truth-maintenance system, and this is what we achieved. The system evaluates the query only once with maintaining enough information to ensure both consistency and completeness of the collected solutions as the dynamic state changes. When the query is re-evaluated, the system returns the cached answers which are always up to

date. There are two approaches to integrate tabling support into existing PROLOG systems. The first approach is to modify and extend the low-level engine. The advantage of this approach is the run-time efficiency, however, the drawback is that it is not efficiently portable to other Prolog systems because the engine level modifications are slightly more complex and time consuming. This approach is used by the XSB [2] system. XSB is the only PROLOG implementation so far that supports incremental tabulation. The second approach to incorporate tabled evaluation into existing PROLOG systems is to apply the source level transformations to a tabled program, and then use external tabling primitives to provide direct control over the search strategy. This idea was first explored by Fan and Dietrich [11] and later used by Rocha, Silva and Lopes [12] to implement tabled PROLOG systems. The main advantage of this approach is the portability of applying it on different PROLOG systems. The drawback is of course the efficiency, since the implementation is not at a low level. Our implementation approach is based on applying the source level transformations to a tabled program. We named our approach as JLOG (Justification-based Logic), the idea of this name came from the word PROLOG (Programming in logic).

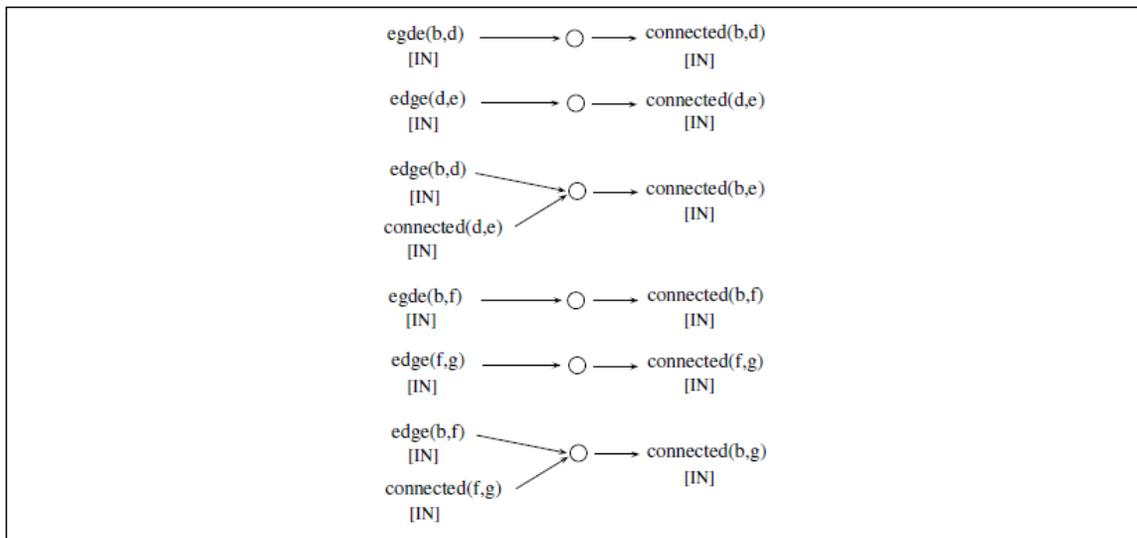


Figure 5: JTMS network of Figure 4 after asserting back the fact *edge(b,d)* to the database of Figure 1.

4. RESULTS AND DISCUSSION

To have a look at the performance of the system, JLOG is to compare it with:

1. Normal PROLOG (NP) implementations [13] that do not support tabulation.
2. Tabled PROLOG (TP) implementations [4, 2] that support monotonic (static facts and rules) logic systems.
3. Incremental tabled PROLOG (ITP) implementations [2] that supports non-monotonic (dynamic facts and rules) logic systems. This is considered to be the main assessment factor since the main objective of this research is to support incremental tabulation. Our

benchmark is the XSB system since it is the only PROLOG implementation that so far supports the incremental tabulation.

The main assessment factors for testing the performance of our approach are categorized into the following:

1. Evaluating the query for the first time

We execute PROLOG queries on normal, tabled, incremental tabled PROLOG and JLOG. The execution time of these queries is analyzed and compared among the four systems.

2. Re-evaluation of a query

Once a query is evaluated for the first time, the same query is re-evaluated again on normal, tabled, incremental tabled PROLOG and JLOG. The execution time of re-evaluating this query is analyzed and compared among the four systems.

```

edge(Sem, S1, S2) :-    reg(Sem, Course, S1, Section),
                       reg(Sem, Course, S2, Section),
                       S1 < S2.

connected(Sem, X, Y) :- edge(Sem, X, Y).
connected(Sem, X, Y) :- edge(Sem, X, M), connected(Sem, M, Y).

```

Figure 6: Translative closure PROLOG program to find the connected students in a certain semester.

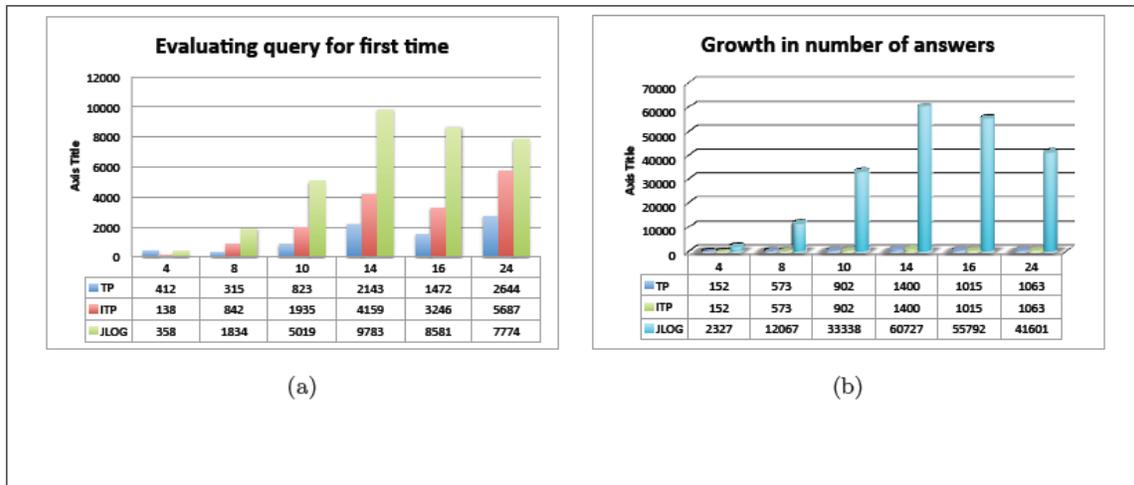


Figure 7: Statistics of evaluating the query $connected(Sem, 946, Y)$.

3. Evaluating a subquery related to a previously proven query

We compare the time it takes to evaluate subqueries related to some previously proven queries on normal, tabled, incremental tabled PROLOG and JLOG.

4. The cost of maintaining the cached proof structure up-to-date for a previously proven query

After the query is proven for the first time, we assert/retract PROLOG facts or rules to/from the PROLOG program that would change the state of the cached answers of the query.

We tested the performance of JLOG by implementing a small business intelligence [14], or a reporting tool for a mid-size University. We have chosen this approach rather than the standard benchmark dataset to test the system on real data. The objective is to observe if the system is able to work under real applications. Graph reachability is a classic problem with many applications in the real-world. The graph reachability problem has been used as a benchmark in any PROLOG tabled implementation. We mapped the graph reachability to the student information database using the following scenarios:

- Picking a certain student in an academic semester, we would like to know the set of students that can be reached from, connected to, this particular student. We used the assumption that all students registered in the same class are connected to each other, i.e. we add an edge between each couple of students registered in the same class (There are so many scenarios in the student information system that can be mapped to the reachability problem, we just picked one example). Then we apply the transitive closure; if student X is connected to Y, Y is connected to Z; we conclude that X is connected to Z.
- In graph theory, a connected component of an undirected graph is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the super graph. In a student information database, we would like to know how many connected components of students exist in a certain semester. Each student registered in the current semester is represented as a vertex in the graph. Whenever two students are registered in the same class, we add an edge between these two students (vertices) in the graph

Figure 6 shows the transitive closure PROLOG program to find the connected students in a certain semester. The first rule in the program connects each couple of students registered in the same class. The second rule uses the transitive relation to connect students indirectly. Given the enrollment data up to a certain academic year, we would like to list all the students connected to a particular student. For example, the query *connected(Sem,946,Y)* finds the list of students connected to the student number 946 in all the semesters that exist in the database of facts. Figure 7 presents the statistics of evaluating the query *connected(Sem,946,Y)* for the first time. The graph that is going to be constructed from the relation *edge/3* contains cycles which yields that this query suffers from infinite loop in Np while it terminates successfully in all tabled (Tp, ITp, JLOG) runs. The query generates a lot of redundant answers which are neglected by Tp and ITp. These answers are not neglected by JLOG, hence it is suffering from overhead when the query is proved for the first time. To test the correctness (soundness) and completeness of the cached

answers, we picked samples of the data such that the number of edges (facts), coming from the *reg/4* predicate, is between 2 to 4kb. The sample takes snapshot of the data before the first day of classes, i.e. start of add/drop period in the university. First we evaluate the general query related to this program which is *connected(Sem,X,Y)*. We pass the semester values for which we are looking the list of connected students. Then, we use the following scenarios to test the soundness and completeness of the cached proof structure:

1. We track all the changes that take place on the predicate *reg/4* starting from the 1st day of add drop period until the end of semester. When a student drops (soundness) a class, the related PROLOG fact is retracted. When a student adds a new class, then the fact is asserted. This can be a new fact (completeness) if the student is adding the class for the first time, or it can be an old fact (soundness) because the student dropped the class after registering it for the first time and then decided to reenroll back in the class. The current version of JLOG updates the JTMS network, attached to the cached query, whenever the assert/retract command is executed. This means that the query proof structure is always updated and returns the correct answers. Figure 8 shows the statistics of maintaining the soundness and completeness of the query *connected(Sem,X,Y)* based on the changes in the predicate *reg/4*. For the same add/drop events ITp (XSB) is faster than JLOG. The reason behind this difference in the performance is coming from the fact that JLOG is updating the JTMS network after each assert or retract command, while ITp is handling the situation through batch processing since it is implemented at low level. When the tables were updated after each assert or retract command in ITp, the performance of the system was degraded. For example, for the semester 1101, ITp takes 47608 milliseconds to update the query answers through batch processing, see Figure 8. This time jumps to 12,972,825 milliseconds when we tried to update the tables after each assert/retract command. JLOG updates the JTMS network after each assert/retract in 120,842 milliseconds which is significantly lower than the time taken by ITp to handle the events one by one.

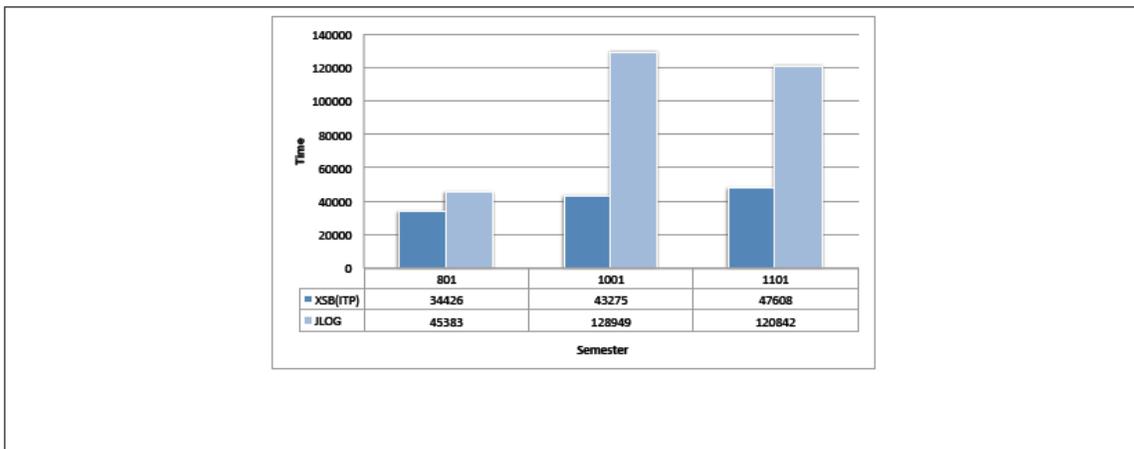


Figure 8: Statistics of maintaining the soundness and completeness of the query *connected(Sem,X,Y)* based on the changes in the predicate *reg/4*.

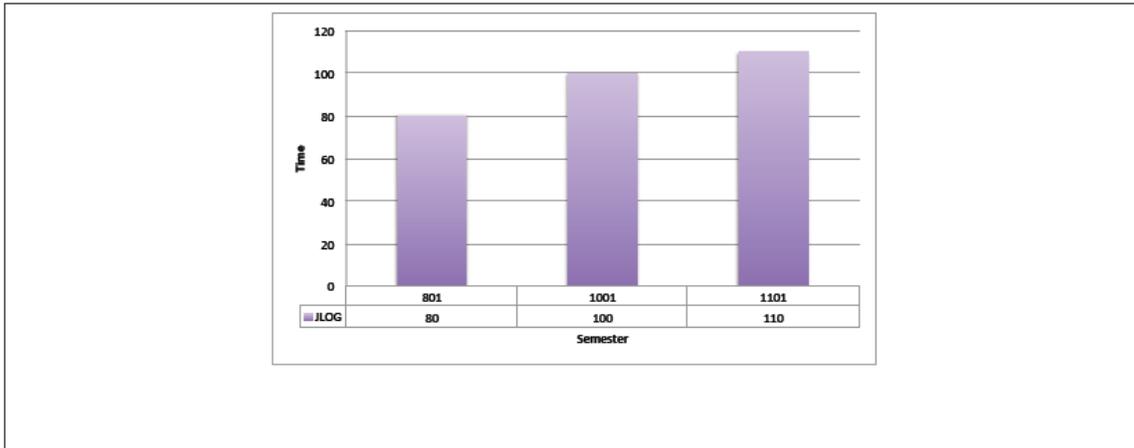


Figure 9: Maintaining the soundness of the query $connected(Sem, X, Y)$ after retracting of the rule $connected(X, Y) : -edge(X, M), connected(M, Y)$ from the PROLOG program of Figure 6.

2. The second scenario is used to test soundness of the query which is related to assertion/retraction of rules in the program of Figure 6. Consider the case where we would like to know the list of students who are connected directly and we want to exclude the tuple of students who are connected indirectly. This can be achieved by retracting the third rule in the program of Figure 6 which connects the students indirectly. In order to be able to retract the rule, the predicate $connected/2$ must be defined as incrementally dynamic. Once $connected/2$ is defined as dynamic predicate, the query $connected(Sem, X, Y)$ does not terminate in ITp. ITp fails to handle the query due to an infinite loop. JLOG handles the situation smoothly. Figure 9 shows the time taken by JLOG for maintaining the soundness of the query $connected(Sem, X, Y)$ after retracting of the rule $connected(X, Y) : -edge(X, M), connected(M, Y)$ from the PROLOG program of Figure 6. JLOG handles the retraction easily because it is a single event used to update the JTMS network and does not require any inference work from the PROLOG side.

5. CONCLUSION

This paper proposed a general framework of a subsystem integrated with PROLOG inference engine (SWIPROLOG, YAP-PROLOG, XSB, ..., etc) that uses justification-based truth maintenance system to support incremental tabulation that can work under nonmonotonic logic. Our system evaluates a query only once, maintaining enough information to ensure both consistency and completeness of the collected solution as the dynamic state changes. The main idea of the system is to cache the proof generated by the PROLOG inference engine as a JTMS network rather than saving the end results as it is the case for most tabling systems. The approach presented in this paper is suitable for a query that depends on dynamic information to be evaluated repeatedly as the dynamic state changes. There are few advantages of our approach. The first advantage comes from caching the query answers in one consolidated subsystem (JTMS Network). Then the evaluation of subqueries requires no inference work. Another system advantage is related to handling the assertion/retraction of rules. On the other hand, the approach is suffering from few limitations. The first limitation of the system is the inability to handle queries with infinite answers. The other limitation of current approach occurred when the query is evaluated for the first time. JLOG is paying sufficient overhead since it caches the proof structure

of the query rather than the end results. JLOG is not a good choice for the queries generating a large number of answers. The large number of answers for a query requires large JTMS network to be installed for the query in order to cache the proof structure. We need to study carefully the memory usage of JLOG and see how this issue can be resolved by controlling or compacting the memory management for the JTMS network.

REFERENCES

- [1] Weidong Chen and David S. Warren. Tabled evaluation with delaying for general logic programs. *J. ACM*, 43(1):20–74, January 1996.
- [2] Terrance Swift and David Scott Warren. Xsb: Extending prolog with tabled logic programming. *TPLP*, 12(1-2):157–187, 2012.
- [3] Ricardo Rocha, Fernando Silva, Ricardo Rocha Fern, and Vítor Santos Costa. Yapstab: A tabling engine designed to support parallelism. 2000.
- [4] Vítor Santos Costa, Ricardo Rocha, and Luís Damas. The yap prolog system. *TPLP*, 12(1-2):5–34, 2012.
- [5] Neng-Fa Zhou, Isao Nagasawa, Masanobu Umeda, Keiichi Katamine, and Toyohiko Hirota. Bprolog: A high performance prolog compiler. In Takushi Tanaka, Setsuo Ohsuga, and Moonis Ali, editors, *IEA/AIE*, page 790. Gordon and Breach Science Publishers, 1996.
- [6] Diptikalyan Saha and C. R. Ramakrishnan. Incremental evaluation of tabled logic programs. In *ICLP*, pages 392–406, 2003.
- [7] Diptikalyan Saha and C. Ramakrishnan. Symbolic support graph: A space efficient data structure for incremental tabled evaluation. In Maurizio Gabbrielli and Gopal Gupta, editors, *Logic Programming*, volume 3668 of *Lecture Notes in Computer Science*, pages 235–249. Springer Berlin / Heidelberg, 2005.
- [8] Diptikalyan Saha and C. Ramakrishnan. A local algorithm for incremental evaluation of tabled logic programs. In Sandro Etalle and Miroslaw Truszczyński, editors, *Logic Programming*, volume 4079 of *Lecture Notes in Computer Science*, pages 56–71. Springer Berlin / Heidelberg, 2006.
- [9] Truong Quoc Dung. A revision of dependency-directed backtracking for jtms. In Günther Görz and Steffen Hölldobler, editors, *KI*, volume 1137 of *Lecture Notes in Computer Science*, pages 57–60. Springer, 1996.
- [10] Gerhard Brewka, David Makinson, and Karl Schlechta. Jtms and logic programming. In *LPNMR*, pages 199–210, 1991.
- [11] Changquan Fan and Suzanne Wagner Dietrich. Extension table built-ins for prolog. *Softw. Pract. Exper.*, 22(7):573–597, July 1992.
- [12] R. Rocha, C. Silva, and R. Lopes. Implementation of Suspension-Based Tabling in Prolog using External Primitives. In J. Neves, M. Santos, and J. Machado, editors, *Local Proceedings of the 13th Portuguese Conference on Artificial Intelligence, EPIA'2007*, pages 11–22, Guimarães, Portugal, December 2007.
- [13] Jan Wielemaker, Tom Schrijvers, Markus Triska, and Torbjörn Lager. SWI-Prolog. *Theory and Practice of Logic Programming*, 12(1-2):67–96, 2012.
- [14] Oksana Grabova, Jerome Darmont, Jean-Hugues Chauchat, and Iryna Zolotaryova. Business intelligence for small and middle-sized enterprises. *SIGMOD Record*, 39(2):39–50, June 2010.

AUTHORS

Tahir M. Ali received his BSc and Ms from Kuwait University and PhD from University of Malaya. He is currently an Assistant Professor of Computer Science in Gulf University for Science and Technology, and also serving as the IT director. His main research interest is in field of Artificial Intelligence (AI), in particular, logic programming and scheduling algorithms.



Ziad H. Najem received his BSc from Kuwait University and Ms and PhD from University of Illinois at Urbana-Champaign. Prior to joining the Department of Computer Science at Kuwait University in 1999, Dr. Najem worked as a Scientific Researcher at Kuwait Institute for Scientific Research.



Mohd Sapiyan Baba is currently a Professor of Computer Science in Gulf University for Science and Technology, Kuwait. He was a lecturer in University of Malaya for more than 30 years, teaching Mathematics and Computer Science courses, and supervised numerous students for their research projects at undergraduate and postgraduate levels. His main research interest is in field of Artificial Intelligence (AI), in particular, the application of AI in Education



INTENTIONAL BLANK

DESIGN, IMPLEMENT AND SIMULATE AN AGENT MOTION PLANNING ALGORITHM IN 2D AND 3D ENVIRONMENTS

Haissam El-Aawar¹ and Hussein Bakri²

¹Associate Professor, Computer Science/Information Technology Departments,
LIU, Bekaa-Lebanon

e-mail: haisso@yahoo.com., haissam.aawar@liu.edu.lb.

²Instructor, Computer Science/Information Technology Departments, LIU,
Bekaa-Lebanon,

e-mail: hussein.bakri@liu.edu.lb.

ABSTRACT

This article presents a computer simulated artificial intelligence (AI) agent that is able to move and interact in 2D and 3D environments. The agent has two operating modes: Manual Mode and Map or Autopilot mode. In the Manual mode the user has full control over the agent and can move it in all possible directions depending on the environment. In addition to that, the designed agent avoids hitting any obstacle by sensing them from a certain distance. The second and most important mode is the Map mode, in which the user can create a custom map, assign a starting and target location, and add predefined and sudden obstacles. The agent will then move to the target location by finding the shortest path avoiding any collision with any obstacle during the agent's journey.

The article suggests as a solution, an algorithm that can help the agent to find the shortest path to a predefined target location in a complex 3D environment, such as cities and mountains, avoiding all predefined and sudden obstacles. It also avoids these obstacles during manual control and moves the agent to a safe location automatically.

KEYWORDS

Motion Planning Algorithm, Artificial Intelligence, Real Time Motion, Automatic Control, Collision Avoidance.

1. INTRODUCTION

An agent motion planning algorithm plays an important role in many applications of AI, robotics and virtual reality especially in navigating agents in 2D and 3D environments. Designing an artificially intelligent agent is a complex task especially in a non-observable or partially observable environment where the level of uncertainty is very high. The best way to describe this case is through the following scenario: imagine yourself in a dark room, where you can't see anything and you need to find your way out of the room, following the shortest path to the door and avoiding all objects in the room. Imagine all that and add to it an extra 3rd dimension, like the case of flying agents (planes) where the environment becomes more complex, obstacles are more unpredictable, and agent's motion becomes more difficult.

Complex and precise data is required from sensors of the plane in order for it to respond in almost real time manner to sudden obstacles or contingencies faced during flight. Although this work is completely simulated on a computer, we suggest that the algorithm of Map Mode and obstacle avoidance can be used in any real environments and on any real agents. It should be noted that the algorithm still needs to be tested on a real airplane drone in a real world environment.

This work describes the design, implementation and simulation of an algorithm that can help an agent to find its way to a target location automatically and without human intervention in a complex, continuous and partially observable 2D and 3D environments. The agent in these environments may face many static and/or moving obstacles, for example a plane may face buildings, birds or other planes. The algorithm should avoid these obstacles without abandoning its pre- mentioned goals. The algorithm is implemented using C# object oriented programming language and using multiple data structures like arrays, queues and trees. Finally, the algorithm is simulated on a computer in 2D and 3D environments (precisely in 3D Virtual City) which contains virtual real-time obstacles.

2. RELATED RESEARCH

In this section we provide a brief overview of some of prior works related to path planning in AI.

2.1 Motion Planning

Motion planning is a term used in robotics. Motion planning algorithms are used in many fields, including bioinformatics, character animation, video game AI, computer-aided design and computer-aided manufacturing (CAD/CAM), architectural design, industrial automation, robotic surgery, and single and multiple robot navigation in both two and three dimensions. A classical version of motion planning is sometimes referred to as the Piano Mover's Problem [1].

Motion planning is a fundamental problem in robotics. It may be stated as finding a path for a robot or agent, such that the robot or agent may move along this path from its starting position to a desired goal position without colliding with any static obstacles or other robots or agents in the environment. This problem interacts with other important problems, such as real-time motion control, sensing and task planning.

The basic motion planning problem is stated as: Given a start pose of the robot, a desired goal pose, a geometric description of the robot and a geometric description of the world, the objective is to find a path that moves the robot gradually from start to goal while never touching any obstacle [1, 2].

2.2 Probabilistic Roadmap Methods

One of the most important classes of motion planning methods addressed in the literature [3, 4, 5, 6, 7, 8] is the probabilistic roadmap methods (PRMs).

A roadmap, RM, is a union of one-dimensional curves such that for all start and goal points in C_{free} that can be connected by a path. In order to compute collision-free paths for robots of virtually any type moving among stationary obstacles (static workspaces) the Probabilistic Roadmap Method/Planner (PRM) is used.

PRM uses randomization extensively to construct roadmaps in C spaces. Heuristics functions are used without calculations for sampling C obstacles and C Spaces.

PRM planner can be applied to robots with many degrees of freedom (dof). The method consists of two phases: a learning (construction) phase and a query phase.

In the learning phase, a roadmap is a graph where the nodes are built of collision-free configurations and where the edges are built of collision-free paths by repeating the two following steps:

- Choose a certain random configuration of the robot, check the collision rate and repeat this step until the random configuration is free of collisions.
- Try to connect the former configuration to the roadmap using a fast local planner.

To find a path in the query phase, the idea is to connect initial and goal configurations to the roadmap and to search the roadmap for a sequence of local paths linking these nodes. The path is thus obtained by a Dijkstra's shortest path query [2, 3, 9]

2.3 Robot's (Agent's) Workspace (Environment)

A robot is defined in motion planning problems as an object or as versatile mechanical device which can move, rotate and translate. It can be polymorphic i.e. taking several forms like rigid object or a manipulator arm, a wheeled or legged vehicle, a free-flying platform (a plane) or a combination of these or a more complex form like or a humanoid form – equipped with actuators and sensors under the control of the computing system [4]. Furthermore, a robot is a reprogrammable, multi-functional manipulator designed to perform a variety of tasks through variable programmed motions, such as moving material, parts, tools, or specialized devices [10].

In motion planning algorithms, robots can move in a myriad of environments consisting from 2D, 3D to even N-dimensional. As an abstracted version that aims to solve the problem, a robot can be represented as a point in space taking translational coordinates (x, y, z) . Normally in 3D workspace, it is a recurring theme to use six parameters: (x, y, z) for locating the position of the robot and (α, β, γ) for its rotation at every point [2].

2.4 AI Planning

Planning is essential to intelligent and rational agents, in the sense of achieving their autonomy and flexibility through the construction of sequences of actions to achieve their goals. Planning as a sub discipline in the artificial intelligence field has been an area of research for over three decades. Throughout the history of research in this domain, the distinction between planning and problem solving has been indefinable [1, 2].

In AI, the term planning takes a more discrete flavor than a continuous one. Instead of moving a piano through a continuous space, problems solved tend to have a discrete nature like solving a Rubik's cube puzzle or sliding a tile puzzle, or building a stack of blocks. These types of discrete models can still be modeled using continuous spaces but it seems more convenient to define them as finite set of actions applied to a discrete set of states. Many decision-theoretic ideas have recently been incorporated into the AI planning problem, to model uncertainties, adversarial scenarios, and optimization [1, 2, 11, 12].

2.5 Computational Environments

The article's implementation and simulation are executed on Microsoft Visual Studio 2010 (using C# language). The 3D simulation is implemented on Windows XNA Game Studio 4.0, which

uses the XNA framework. Microsoft XNA Game Studio is used to build interactive games on windows based computers, X-box 360 and windows mobile [22].

The **XNA framework** helps making games faster. Typically XNA framework is written in C#, and uses DirectX and Direct3D which is designed to virtualize 3D hardware interfaces for windows platforms, so instead of worrying about how to get the computer to render a 3D model, user may take more focus view on the problem and gets the 3D data onto screen. An XNA tutorial [23] is used to create a flight simulator and produce a flying aircraft in a true 3D city which is used to apply the algorithm.

Direct X is recommended in order to have the optimal performance in the 3D Simulator [24].

C sharp (C#) was developed by Microsoft (.Net) as Common Language Infrastructure (CLI) to be a platform-independent language in the tradition of Java [13, 14, 25]. The C# language is intended to be a simple, modern, general-purpose, object-oriented programming language. It offers a strong typing, functional, class-based programming, which makes programming much easier and rigid. C# helps programmers to initiate parallel threads, called asynchronies methods, and this brews for us a full control of the dataflow to govern the AI agent.

3. MOTION PLANNING ALGORITHM

As mentioned before, the algorithm has two main modes: **Manual and Map mode**. In manual mode the user has full control over the agent and can move it in any possible direction. In this mode there is a special feature called **Contingency Mode**. This mode allows the agent to sense the obstacle from a certain distance and avoid it by moving to a safe position away from it. The Second mode is the map mode. In this mode the user assigns a starting point, target point and a number of obstacles. The agent then must apply the algorithm in order to find the shortest path to the target and move to it avoiding all kinds of obstacles [15, 16, 17, 18].

The development of algorithm passed through three main phases:

1. The Design of the algorithm phase was the first phase where the algorithm was designed to meet as much as possible the following requirements: optimality, completeness, and acceptable time and space complexities.
 - a. Optimality means that the algorithm must always find the lowest cost path or the shortest path during the search process.
 - b. Completeness means that the algorithm must always find a solution when a solution exists so it must return the path if it exists or null if it does not.
 - c. Acceptable time and space complexities are very important, because creating an optimal and complete algorithm is useless if it has slow running time and consumes considerable amount of memory.
2. The implementation of the algorithm using a programming language was the second phase. The preferred algorithm implementation language chosen was an OOP language (even it is preferable to be an open source or fully supported language like Java, VB, or C#). This phase is important for the next phase which is the simulation, because it will use the implementation of the algorithm and apply it on the agent.
3. The simulation of the algorithm on a computer using 3D graphics engine was the third phase were a virtual city was created representing the environment and a virtual plane representing the agent. The manual mode with obstacle avoidance and the map mode were implemented and simulated using the algorithm that we will discuss in this paper.

4. IMPLEMENTATION AND TESTING

This section covers the Map mode algorithm's design and its implementation. It also covers the implementation of the environment (i.e. Simulator) on which the algorithm was tested and simulated using virtual agents in 2D and 3D environments.

4.1 Description of the Environment

In AI, the environment is classified based on many properties (single vs multi-agent, stochastic vs deterministic, discrete vs continuous...). In this case the agent is moving in a partially observable, single-agent, stochastic and continuous environment which is considered one of the hardest environments in AI.

- **Partially Observable Environment:** the agent's sensors have no access to all states of the environment. In this project, the agent can't determine which location has an obstacle and which does not unless the obstacles are in near proximity (in sensors range).
- **Single Agent Environment:** there is only one agent in the environment, and there are no other objects that can decrease or increase the agent's performance measure
- **Stochastic Environment:** the next state of the environment is not determined by the current state and/or the actions executed by the agent [11].
- **Dynamic Environment:** is when the environment is changing continuously while the agent is thinking or acting. In this project, obstacles can appear randomly and suddenly.
- **Continuous Environment:** the number of clearly defined percepts and actions is unknown. The speed and location of the agent sweeps through a range of continuous values and do so smoothly over time [11].

4.2 The Map Mode Algorithm's Design in 2D Environment

In map mode, the agent must move automatically and autonomously from its current position to a predefined target location on a map following the shortest path possible and avoiding predefined or sudden obstacles on the map.

4.2.1 Map Generation

The search process starts from the target location and ends when the starting location is found. Each location will have a specific cost which is equal to the distance to the target location. Therefore, the cost of the target (Goal) location is equal to 0.

Since the agent can move in four directions in 2D environment (up, down, left and right) then each neighbor of the target is checked:

- If it is not an obstacle.
- If it is not a starting location or target location.
- If it is inside the boundary of the environment.
- If it is not visited.

If all these conditions are satisfied, then the neighbor's cost will be equal to 1 since they are only one step away from the target. After that, all these neighbors are visited and all their possible neighbors' cost will be equal to 2. The map generation continues in this manner where the cost of every neighbor is equal to the cost of the location that is currently visited plus 1.

The Map Generation pseudo code is shown in Figure 1.

```

Set search node to target node.
Set the TempCost to 1.
L1:
  Check an unchecked adjacent node If (not the start node And inside environment boundary
  And not Obstacle)
  {
    Mark the node as checked.
    Set its weight to TempCost.
  }

  If all adjacent nodes are checked then
  {
    Increment TempCost.
    Set search node to a sub node.
    If all sub nodes are checked then return
  }

repeat L1.
}

```

Figure 1. Map Generation Pseudo Code

After the Map Generation algorithm is executed the map is shown in Figure 2.

14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
18	17	16	15	14	13	12	11	10	9	8	7	6	5	4
19	18	17	16	15	14	13	12	11	10	9	8	7	6	5
20	19	18	17	16	15	14	13	12	11	10	9	8	7	6
21	20	19	18	17	16	15	14	13	12	11	10	9	8	7
22	21	20	19	18	17	16	15	14	13	12	11	10	9	8
23	22	21	20	19	18	17	16	15	14	13	12	11	10	9
24	23	22	21	20	19	18	17	16	15	14	13	12	11	10
25	24	23	22	21	20	19	18	17	16	15	14	13	12	11
26	25	24	23	22	21	20	19	18	17	16	15	14	13	12
27	26	25	24	23	22	21	20	19	18	17	16	15	14	13
-2	27	26	25	24	23	22	21	20	19	18	17	16	15	14

Figure 2. Map Generation

Where the yellow location is the starting location and the green location is the target location.

4.2.2 Finding the Shortest Path Algorithm in Map Mode:

After the map is generated, finding the shortest path becomes simple. Therefore, a simple Hill Climbing algorithm is used to find the shortest path.

The Hill Climbing algorithm is used because:

- It is implemented as a loop that continually moves in the direction of the “best” neighbour with increasing value that is, uphill [1], it can also be implemented in the opposite way i.e. in the direction of the neighbour with decreasing value that is downhill
- It terminates when it reaches a “peak” where no neighbour has a higher value (Climbing to the top) or where no neighbour has a lower value (descending the Hill to the bottom).
- The algorithm does not maintain a search tree, so only the current state and the value of its objective function are stored saving by that a lot of space.

Hill Climbing “does not look ahead beyond the immediate neighbours of the current state” this is why it is a greedy local search algorithm [11].

4.2.3 Applying Hill Climbing

In this situation the search will start from the starting location until finding the target. Since the target has cost = 0 then it will be needed to get down from a required location on the hill to find the global minimum, which is the lowest point in the hill. The algorithm should always choose the “best neighbor”, because the cost represents the distance to the goal. Note: all obstacles have cost = 1000, Unvisited node = 800 and starting node cost = -2. Finding the shortest path algorithm pseudo code is shown in Figure 3.

```

Set search node to starting node.
L1:
Check all adjacent node If (not the start node and inside environment boundary){
If adjacent node is target then return target is found.
else
If there is no possible route exists then return no route exist.
else
Set search node to a sub node that has the lowest cost.
Mark the visited search node.
Repeat L1

```

Figure 3. Shortest Path Algorithm

Normally, Hill Climbing Algorithm is not complete and not optimal because of several limitations explained in [11, 12] and may other textbooks. The enhancement done here that make it complete in this case is the way the map is generated initially that does not allow the appearances of local minima or plateaus.

Giving the constraints that we set for the movements allowed of the agent in 2D, the algorithm finds the best path with lowest cost and effort. The altered algorithm is also complete since it always gives a solution whether a route exists or not. A route does not exist if the target or starting locations are surrounded with obstacles. This algorithm has also a low memory (space) complexity of ($O(n)$), since there are no routes or nodes that are stored, and there is only one loop that is iterating till the target is found.

4.3 The Algorithm's Implementation in 2D Environment

The 2D Environment is considered easier to deal with since there is no height (z-axis). The coordinates of the locations on the map are defined in terms of x-axis and y-axis coordinates. For that, we are using a 2 dimensional matrix to define the coordinates of each location. The 2D environment consists of 15 x 15 locations, so the array of locations is defined as: a [15, 15].

In the map based mode, the user sets a starting point, target point and obstacles on the map, and the agent must move automatically from its location (starting point) to the target point avoiding obstacles (predefined and sudden) applying the shortest path algorithm.

4.3.1 Map Generation Implementation

Based on the algorithm design, the cost of every location must be stored on the map and since we are using an array to store these locations, then the value of each member of this array will be equal to the cost; if the target is at x=1 and y=1 then a[1][1]=0. In order to visit every position and all its neighbors, a FIFO queue is needed to store the visited locations. This idea resembles the method used in Breadth first search in trees where all nodes are visited level by level, where each level contains nodes of equal costs in order to find the goal node which is in this case the starting node. The implementation of the algorithm is shown in Figure 4.

```

Initial state: all elements in the array = 800
After assigning the target and starting point a[tx,ty]=0 and a[stx,sty]=-2
Put the target point in FIFO queue Q
While the queue is not empty {
String xy = Q.get() //Ex: xy = 2,8
Int x = xy.x
Int y = xy.y
Int c = a[x,y]
If (left , right , up and down neighbors =800 & neighbor inside boundary){
a[neighborX,neighborY] = c + 1
Q.Put(neighbor)
}}

```

Figure 4. Used code for Algorithm

where, tx and ty are coordinates of target location.

stx and sty are coordinates of starting location. Obstacles locations have cost = 1000.

4.3.2 Finding Shortest Path Implementation

As mentioned before, a simple Hill Climbing algorithm can be used to find the shortest path and simulating the movement of the agent. In this case, we want to reach a global minimum which is zero (target point). So, we need to choose the neighbor with the lowest cost. The algorithm implementation pseudo code is shown in Figure 5.

```

int x = stx int y=sty
While(x!=stx && y!=sty)
{
Find Min Cost Neighbor ( Check left, right , up and down neighbor)
If(Min Cost = 800 or 1000) { return no route }
Else
{
x=MinNieghbor.X // MinNieghbor is neighbor with lowest Min Cost
y=MinNieghbor.Y
Mark MinNieghbor }}

```

Figure 5. Algorithm Pseudo code

As shown in the code in figure 5, the best location that is selected next is the one with the lowest cost. There are two ways to detect that there is no route that exists: the first way is when the best neighbor is undiscovered (cost = 800) which means that the target or agent is surrounded with obstacles. The second way is when the best neighbor has cost = 1000 which means that the starting location is surrounded directly with obstacles (see Figure 6).

14	13	12	11	10	9	8	7	6	5	4	3	4	5	6
13	12	11	10	9	8	7	6	5	4	3	2	3	4	5
12	11	10	9	8	7	6	5	4	3	2	1	2	3	4
11	10	9	8	7	6	5	4	3	2	1	0	1	2	3
12	11	10	9	8	7	6	5	4	3	2	1	2	3	4
13	12	█	█	█	█	█	█	█	█	█	2	3	4	5
14	13	14	15	16	█	800	800	800	800	█	3	4	5	6
15	14	15	16	17	█	-2	800	800	800	█	4	5	6	7
16	15	16	17	18	█	800	800	800	800	█	5	6	7	8
17	16	17	18	17	█	800	800	800	800	█	6	7	8	9
18	17	18	17	16	█	█	█	█	█	█	7	8	9	10
19	18	17	16	15	14	13	12	11	10	9	8	9	10	11

Figure 6. Blocked path example

It should be mentioned that 1000 and 800 and all other numbers used here can be increased or changed depending on how much the environment is big.

In order to simulate the Map Mode and the Direct Drive Mode, a form containing a Map of clickable 15 x 15 buttons is used. Each button on the map represents a location from [0, 0] till [14, 14] ([Row, Column]).

The user first assigns starting location which will color the pressed button with yellow. Then, the user presses another button which will be the target location, and will be colored green. After that, the user can control the agent freely, and move it up, down, left, and right using the W, S, A, D keyboard keys respectively. The user is also able to add obstacles on the map by pressing a location which will be then marked with black. Whenever the user wants, he can press the generate map and the find path buttons. The first will change the text of each button to a number that represents the current cost of location represented by this button. The find path button will trace the best or shortest path in red color. Figure 7 shows the starting location (yellow), target location (green), sudden obstacles (black), shortest path (red). In this image the agent was moving already toward the target and sudden obstacles (in black) were put on the map. The algorithm recalculates the shortest path and chooses adequately the new path.

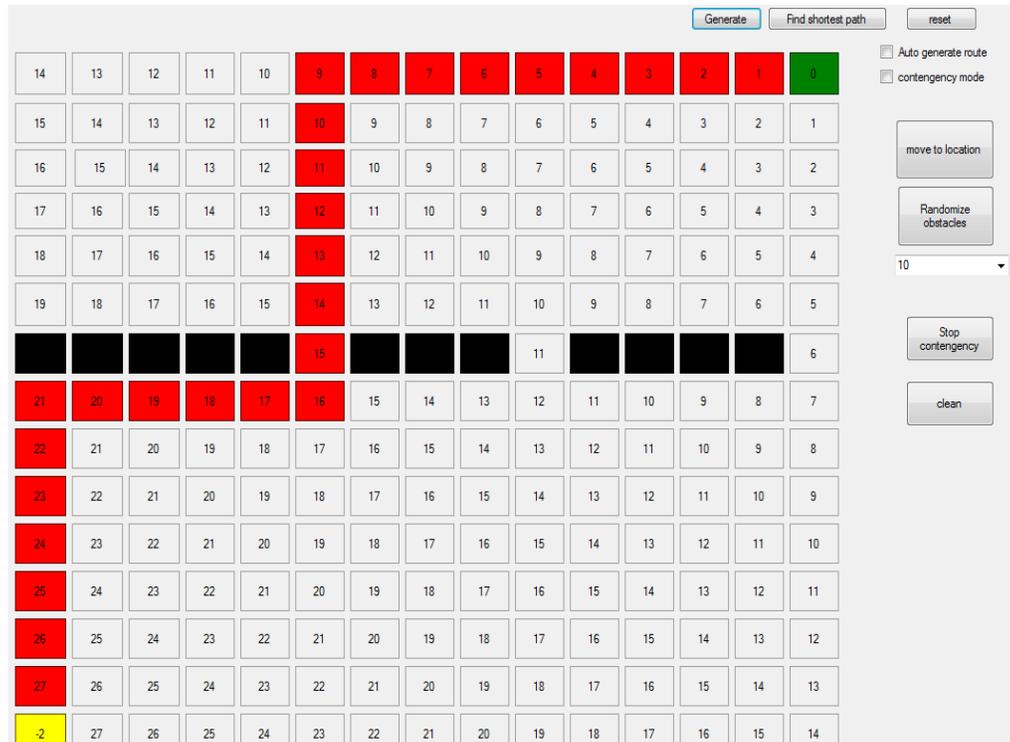


Figure 7. Path Finding

4.4 The Map Mode Algorithm in 3D Environment

In the 3D environment there is an extra dimension (Z-axis or height) which makes the movement of the agent more complicated. Consider that the agent is flying in a city which contains buildings of different heights and shapes. This agent needs to find the shortest path to a destination. It is logical that if the agent changes its height while moving, it will probably reach the target location faster than moving on the path of the roads of the city. But since our environment is partially observable and could change at any moment, therefore, changing the height is not always an optimal solution (like if the obstacle was a tall mountain or building). So first, the algorithm tries to find a shortest path on the same height like in 2D mode, but the difference in that if no route was found (obstacles are surrounding completely the agent or the target) the application will not return that no route exists. It will return that no route exists at the current height of the agent and offers the user the option to choose to move the agent following the shortest path considering that there are no obstacles on that height (see Figure 8).

```

Set search node to starting node.
L1:
Check all adjacent node If (not the start node And inside environment boundary){
If adjacent node is target then return target is found.
Else
If there is no possible route exists then
remove all obstacles on the map
Call Generate Map and find shortest path
else
Set search node to a sub node that has lowest cost.}
Mark the visited search node.
Repeat L1

```

Figure 8. Algorithm Development

The question now is how the agent will avoid the obstacles?

The answer is simple: as long as the agent is sensing that an obstacle exists in front of it at a certain distance, it goes up. Consider the case where the obstacle is a building, when the agent senses that an obstacle exists, it will increase its height till the agent is higher than the maximum height of the building.

What are the changes in the three main methods (Generate, find and move) in the map mode?

- The generate map method remains the same.
- Finding the shortest path has some changes when no route exists on the same height, since this method detects whether a route is found or not.
- Moving the agent as in 2D depends on marked route made by the find shortest path algorithm. When there is no route on the same height. If the agent faces an obstacle while moving on the shortest path it will avoid it by flying over it.

4.4.1 The 3D Simulator

In order to test the algorithm, we needed a virtual but realistic environment, obstacles and a virtual agent to represent a plane. Since the algorithm was implemented using C#, we had to use a 3D engine which is based on this language. The best choice was XNA Game Studio 4.0 in Microsoft Visual Studio 2010. A 3D City was built and a plane can now move in it (see Figure 9).



Figure 9. 3D screenshot

Figure 9 shows a complex environment where buildings (obstacles) of different sizes can be created.

In order to apply the map mode algorithm, to add obstacles and to retrieve the agent's location, a windows form which contains a map that work on the same concept as the 2D simulator (buttons, colors and labels) was created. The map form, loads first and through it, the user can start the 3D simulator. The user can assign a starting location and a target location and add obstacles wherever he wants on the map. The obstacles will appears suddenly as buildings in the simulator.

During manual control the user can view the agent's position on the map (the location will be colored yellow). There are also two labels which represent the x-axis and y-axis coordinates of the plane. We added an extra label in order to count the steps that the agent has taken since the start of the simulator. The random map button adds a random building in random locations. The user can start map mode by pressing its button after assigning start and target location. After that, the plane will start moving according to the red path that can be viewed on the map plus the current location of the plane.

4.4.2 Map Generation Optimization

Since we have experimented that the map generation algorithm has the highest running time and highest space complexity, we tried to apply some improvements.

We measured empirically the running time using a diagnostic tool provided by visual studio called Stop Watch (actually it is a famous method for empirical analysis of algorithms that is based on the same concepts in different languages). This tool is able to measure the running time of a method and return the time in milliseconds [19, 20, 21].

Upon the results of using this analysis tool we made the following improvements:

- Resetting the queue after every move.
- Stop searching when the starting location was found instead of discovering every location on the map.

The second solution was very effective and had a great impact on the efficiency and execution time of the algorithm according to the distance between the agent's location and target point, where the improvement was bigger when the distance is closer, and it starts to decrease as the distance increases. The improvement in time was between 0 and 95% according to the distance. Many tests were also performed to measure the performance and stability of the algorithm and the simulator, especially on threads. These tests had to make sure that each thread does not interfere with the other, and it's terminated when its job finishes or the application closes.

Other tests covered the manual control of the agent, where turning and moving were tested in order to represent the most accurate and realistic navigation that resembles that of a real plane.

5. REAL WORLD SCENARIO

This project's purpose is to be implemented on a real flying agent in a real environment. For a successful implementation of the algorithms presented in this paper, the agent needs to be equipped with powerful and accurate sensors and other hardware tools like speedometers or accelerometers, gyroscope and any useful hardware for an agent in flight.

The speed of the agent should be measured at each instant. Configurations of the environment with all sudden obstacles should be continuously generated. Precise location of obstacles given to the map generation algorithm is crucial for any success navigation in a 3D environment.

6. CONCLUSION

The problem of path planning in intelligent autonomous vehicles is still topical research field although it has been studied by the research community for many decades. Any intelligent autonomous vehicles have to include path planning into their deliberation mechanisms [16].

The article's aim (in addition to manual control) is to find as much as possible a complete and logical solution for automatic agent navigation in a dynamic two and three dimensional environment, where the definitions of the obstacles and restricted areas along with the agent's position are altered after each search.

In manual mode, the objective was to simulate the movement control of the agent, and automatic obstacle avoidance.

Through The Microsoft XNA and Visual Studio, we were able to create an environment that resembles a real city that can be easily manipulated by the user, and we chose the airplane to

represent the AI agent in this 3D environment. The movement of the plane was smooth and mimics every possible direction that a real plane can perform. In addition, the agent was able to sense all obstacles from a certain distance and avoid them by moving to another direction.

In the map mode which was the main focus of the project, the agent is required to move on a predefined map from one location to another avoiding all kinds of obstacles on the shortest path possible inside the boundary of the environment. The solution consisted of three main algorithms: map generation, finding the shortest path and agent's navigation algorithm. First, Map Generation was used in order to change the partially observable environment to a fully observable environment at some time 'T' (since the environment is still dynamic), by assigning costs to every location in the map. The implementation of this algorithm was based on the iterative implementation of breadth first search with some modifications. Second, finding the shortest path is the search algorithm which finds the lowest cost path to the target. The implementation of this algorithm was based on the Hill Climbing algorithm (descent version), i.e. in this case the agent is on the top of the hill and needs to find the shortest path to the bottom (Global minimum). The Map generation part avoids a lot of limitations of Hill Climbing. Third, moving the agent was a simple algorithm which consists of moving and directing the agent according to the shortest path previously generated, and must be generated with every step that the agent makes in such continuous and dynamic environment.

7. FUTURE WORK SUGGESTIONS

The following are some of future work suggestions that could be applied to the algorithm:

- More optimization can be achieved in the map generation algorithm. This part of the algorithm has the highest complexity in the project, and it must be executed at every location in order to check if the path is still the optimal one or not. In addition to that improving the way that the agent can sense obstacles is very important. Through sensor's data in a real world environment, the agent can build a knowledge base about the properties of the obstacles (sizes, shapes, motions), this allow the agent to act accordingly.
- We also hope in the future to try the Map Mode algorithm on a real plane in a real environment. The main aim of any future project will be the construction of a small plane with the ability to sense obstacles through sensors like sonars per example. This plane can be controlled by a smartphone or a computer via wireless radio connection or via the internet. The plane should have the ability to control itself in case it encounters obstacles, and should able to move on the map automatically (according to the algorithm presented in this paper). Other information might be necessary to be taken into consideration like the position, altitude, distance and signal strength of the plane and many others. More empirical and mathematical analysis on the algorithms in this paper will expanded in subsequent work.

ACKNOWLEDGEMENTS

We would like to express their special thanks of gratitude to the president of Lebanese International University HE Abdel Rahim Mourad for his continuous encouragement of research activities in the university. Secondly, we would like also to thank the LIU Bekaa campus administration for their full support during the project and last but not least special thanks go to the students Youssef Al Mouallem and Hassan Al Kerdy for their contribution and insights of the project.

REFERENCES

- [1] Steven M. LaValle (2006) "Planning Algorithms", Cambridge University Press, Cambridge University Press.
- [2] Antonio Benitez, Ignacio Huitzil, Daniel Vallejo, Jorge de la Calleja and Ma. Auxilio Medina (2010) "Key Elements for Motion Planning Algorithms", Universidad de las Américas – Puebla, México.
- [3] Kavraki L. E., Svestka P., Latombe J.-C., Overmars M. H. (1996) "Probabilistic roadmaps for path planning in high-dimensional configuration spaces", IEEE Transactions on Robotics and Automation, vol. 12, No 4, pp566–580, doi:10.1109/70.508439.
- [4] Juan-Manuel Ahuactzin and Kamal Gupta (1997) "A motion planning based approach for inverse kinematics of redundant robots: The kinematic roadmap", IEEE International Conference on Robotics and Automation, pp3609-3614, Albuquerque.
- [5] Amato N. M. & Wu Y (1996) "A randomized roadmap method for path and manipulation planning", IEEE Int. Conf. Robot. and Autom., pp113–120.
- [6] Amato N., Bayazit B., Dale L., Jones C. & Vallejo D (1998) "Choosing good distance metrics and local planner for probabilistic roadmap methods", Procc. IEEE Int. Conf. Robot. Autom. (ICRA), pp630–637.
- [7] M. LaValle and J. J. Kuffner (1999) "Randomized kinodynamic planning", IEEE Int. Conf. Robot. and Autom. (ICRA), pp473–479.
- [8] M. LaValle, J.H. Jakey, and L.E. Kavraki (1999) "A probabilistic roadmap approach for systems with closed kinematic chains", IEEE Int. Conf. Robot. and Autom.
- [9] Geraerts, R.; Overmars, M. H. (2002) "A comparative study of probabilistic roadmap planners", Proc. Workshop on the Algorithmic Foundations of Robotics (WAFR'02), pp43–57.
- [10] C. Ray Asfahl (1985) "Robotos and Manufacturing automation", John Wiley & Sons, Inc.
- [11] Russell and Norvig (1 April 2010) "Artificial Intelligence: A Modern Approach", 3rd edition by - Pearson. ISBN-10: 0132071487, ISBN-13: 978-0132071482.
- [12] George F. Luger (2009) "Artificial Intelligence: Structures and Strategies for Complex Problem Solving", 6th edition, Pearson College Division.
- [13] Paul Deitel and Harvey Deitel (2011) "C# 2010 for programmers", Prentice Hall, 4th edition.
- [14] John Sharp (2011) "Microsoft Visual C# 2010 Step by Step", Microsoft Press.
- [15] Haissam El-Aawar and Mohammad Asoud Falah (April-2009) "An Integration of a High Performance Smartphone with Robotics using the Bluetooth Technology", Communications of SIWN (ISSN 1757-4439).
- [16] David Sislak, Premysl Volf & Michal Pechoucek (2009) "Flight Trajectory Path Planning", Proceedings of ICAPS 2009 Scheduling and Planning Applications woRKshop (SPARK), pp76-83.
- [17] David Sislak, Premysl Volf, Stepan Kopriva & Michal Pechoucek (2012) "AgentFly: Scalable, High-Fidelity Framework for Simulation, Planning and Collision Avoidance of Multiple UAVs. In Sense and Avoid in UAS: Research and Applications", Wiley: John Wiley&Sons, Inc., pp235-264.
- [18] Benitez A. & Mugarte A. (2009) "GEMPA:Graphic Environment for Motion Planning Algorithm", In Research in Computer Science, Advances in Computer Science and Engineering, Vol. 42.
- [19] Sedgewick, Robert (2002) "Algorithms in Java", Parts 1-4. Vol. 1. Addison-Wesley Professional.
- [20] Levitin, Anany (2008) "Introduction To Design And Analysis Of Algorithms", 2/E. Pearson Education India.
- [21] Goodrich, Michael T., & Roberto (2008) "Tamassia. Data structures and algorithms in Java", John Wiley & Sons.
- [22] <http://www.microsoft.com/en-us/download/details.aspx?id=23714>
- [23] <http://www.riemers.net> visited at 1/5/2013
- [24] <http://www.techopedia.com/definition/27489/activity-diagram> visited at 15/5/2013
- [25] <http://msdn.microsoft.com> visited at 7/5/2013.

AUTHORS

Haissam El-Aawar is an Associate Professor in the Department of Computer Science and Information Technology at the Lebanese International University where he has been a faculty member since 2009.



Haissam completed his Ph.D. and M.Sc. degrees at the State University "Lviv Polytechnic" in Ukraine. His research interests lie in the area of Artificial Intelligence, theory of complexity, microprocessors evaluation, CISC- and RISC-architectures, robotics control, mobility control and wireless communication.

Hussein Bakri is an instructor in the department of Computer Science and Information Technology at Lebanese International University. He has completed his M.Sc. degree at the University of St Andrews in United Kingdom and he is now continuing his Ph.D. His research interests lie in the area of Artificial Intelligence, software engineering, virtual worlds and virtual reality and cloud computing.



INTENTIONAL BLANK

A ROS IMPLEMENTATION OF THE MONO-SLAM ALGORITHM

Ludovico Russo, Stefano Rosa, Basilio Bona¹ and Matteo Matteucci²

¹Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy
{ludovico.russo, stefano.rosa, basilio.bona}@polito.it

²Dipartimento di Elettronica, Informatica e Bioingegneria,
Politecnico di Milano, Milano, Italy
matteo.matteucci@polimi.it

ABSTRACT

Computer vision approaches are increasingly used in mobile robotic systems, since they allow to obtain a very good representation of the environment by using low-power and cheap sensors. In particular it has been shown that they can compete with standard solutions based on laser range scanners when dealing with the problem of simultaneous localization and mapping (SLAM), where the robot has to explore an unknown environment while building a map of it and localizing in the same map. We present a package for simultaneous localization and mapping in ROS (Robot Operating System) using a monocular camera sensor only. Experimental results in real scenarios as well as on standard datasets show that the algorithm is able to track the trajectory of the robot and build a consistent map of small environments, while running in near real-time on a standard PC.

KEYWORDS

SLAM, Mono-SLAM, Mapping, Mobile robotics

1. INTRODUCTION

In several application scenarios mobile robots are deployed in an unknown environment and they are required to build a model (map) of the surroundings, as well as localizing therein.

Simultaneous localization and mapping (SLAM) applications now exist in a variety of domains including indoor, outdoor, aerial and underwater and using different types of sensors such as laser range finders, sonars and cameras [4]. Although, the majority of those approaches still rely on classical laser range finders, the use of vision sensors provides several unique advantages: they are usually inexpensive, low-power, compact and are able to capture higher level information compared to classical distance sensors. Moreover, human-like visual sensing and the potential availability of higher level semantics in an image make them well suited for augmented reality applications.

Visual SLAM approaches are usually divided in two main branches: smoothing approaches based on bundle adjustment, and filtering approaches based on probabilistic filters. The latter are divided in three main classes: dense, sparse and semantic approaches. Dense approaches ([17], [14], [22]) are able to build dense maps of the environment, which make the algorithms more robust but at the same time heavy in terms of computational requirements; indeed, most of these

approaches are able to work in real-time only when dedicated hardware (e.g., a GPU or FPGA) is used. Sparse approaches ([15], [7], [12]) address the problem of computational requirements by sparsifying the map; obviously this choice impacts on the robustness of the solution. These algorithms require less computational efforts because they try to allocate in memory only the most significant key points representing the map; for this reason, they are natural candidates for a real time Visual SLAM implementation. Semantic approaches ([10],[20]), extract higher level semantic information from the environment in order to build a more robust and compact map.

Parallel Tracking and Mapping (PTAM) was proposed in [15] as a sparse approach based on a monocular camera targeted to augmented reality applications. The main idea of PTAM is to divide tracking and map updating phases. Camera pose tracking is performed at each time step, by comparing the new frame with the current map using feature matching techniques; FAST features [18] are used for matching. Map updating is performed only on a set of key frames and when the current camera position estimation is precise enough. The algorithm requires an initialization phase in which the same features are viewed from different points of view. The most important limitation of the algorithm is the impossibility to handle occlusions.

A promising solution is the Mono-SLAM algorithm, originally proposed by Davison et al. in [12]. In this approach, the map and the camera pose are stored as stochastic variables and the system evolution is estimated by an incremental Extended Kalman Filter (EKF). Inverse depth parametrization can be used for the representation of point features, which permits efficient and accurate representation of uncertainties [16]. By using an approach known as active search paradigm [11], the algorithm is able to speed-up feature matching, since interesting points in the each new frame are looked for only in the most probable regions. In addition, the algorithm does not need an initialization phase and its probabilistic nature makes it more robust to occlusions. The most evident limitation of the algorithm is the fact that, when the map becomes too large, the EKF processing phase becomes too computationally heavy to be computed in real-time [21]. However, some solutions have been proposed to solve that problem too [13].

In this work we present an implementation of the Mono-SLAM algorithm using the ROS [3] framework. The developed ROS node takes as input the images captured from a monocular camera and outputs the trajectory of the camera, as well as a point map representing the environment around the robot. The algorithm is able to run in near real-time on a standard PC with no dedicated hardware for small and medium length trajectories. The rest of the paper is organized as it follows: in Section 2 we briefly recall the formulation of the Mono-SLAM problem; in Section 3 we describe our implementation of the algorithm; in Section 4 we show experimental results that validate the effectiveness of the approach; finally in Section 5 we draw some conclusions and discuss about future extensions of our implementation.

2. PROBLEM FORMULATION

As in classical SLAM approaches based on laser scanners, the robot pose is described as a stochastic variable with Gaussian distribution, and the map of the environment is sparse. The environment is described by a limited set of features $\{f_i\}$, i.e., measurable geometrical entities (points in the case of Mono-SLAM). Features are described as gaussian variables, as well.

The system state, identified by the robot pose and the map, is represented at any time $t = k\Delta t$, where Δt is the time elapsed since the previous step, as a stochastic variable with Gaussian distribution

$$\hat{\boldsymbol{\mu}}_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (1)$$

having $\boldsymbol{\mu}_k$ mean and covariance $\boldsymbol{\Sigma}_k$.

The state vector encapsulates the information on both camera pose and world features

$$\boldsymbol{\mu}_k = (\mathbf{x}_k^T \quad \mathbf{f}_1^T \quad \cdots \quad \mathbf{f}_m^T)^T \quad (2)$$

Information concerning the camera motion at any instant is encoded in the vector built as

$$\mathbf{x}_k = (\mathbf{r}_k^T \quad \mathbf{q}_k^T \quad \mathbf{v}_k^T \quad \boldsymbol{\omega}_k^T)^T \quad (3)$$

where the vector \mathbf{r} and the quaternion \mathbf{q} represent the pose of the camera reference frame C while vectors \mathbf{v} and $\boldsymbol{\omega}$ are the linear and angular velocities of C with respect to the world reference frame W.

The function used to predict the evolution of the camera state is given by

$$\mathbf{x}_{k+1|k} = \mathbf{g}_v(\mathbf{x}_k) = \begin{pmatrix} \mathbf{r}_k + (\mathbf{v}_k + \mathbf{V}_k) \Delta t \\ \mathbf{q}_k \times \text{quat}((\boldsymbol{\omega}_k + \boldsymbol{\Omega}_k) \Delta t) \\ \mathbf{v}_k + \mathbf{V}_k \\ \boldsymbol{\omega}_k + \boldsymbol{\Omega}_k \end{pmatrix}, \quad (4)$$

Where $\text{quat}((\boldsymbol{\omega}_k + \boldsymbol{\Omega}_k) \Delta t)$ is the quaternion corresponding to the rotation $(\boldsymbol{\omega}_k + \boldsymbol{\Omega}_k) \Delta t$ obtained from the axis-angle representation while \mathbf{V}_k and $\boldsymbol{\Omega}_k$ are the noise vectors affecting respectively linear and angular velocities. Features are considered static so they do not need a prediction model.

The measurements function necessary to perform the update step of the EKF filter is given by

$$\mathbf{h}(\boldsymbol{\mu}) = \begin{pmatrix} h(\mathbf{f}_1, \mathbf{x}_k) \\ \vdots \\ h(\mathbf{f}_m, \mathbf{x}_k) \end{pmatrix} \quad (5)$$

where the projection function $h()$ is a function able to project a 3D features in the image plane using the predicted camera pose.

3. IMPLEMENTATION

The software has been developed using the Robot Operating System (ROS) [3] in C++ under Linux. The OpenCV library [1] was used for image processing and the Eigen3 library [2] was used for matrix operations.

The architecture of the implemented solution is described in Figure 1. Each new frame f_k is acquired from the camera sensor topic by the capture and preprocess block, which performs some preprocessing and feeds the frame to the Mono-SLAM super-block, which is the main block in charge of processing the frame in order to reconstruct the camera motion and the map. The output of the Mono-SLAM main block at each step is the estimated state of the system, as in equation (2).

The feature matching block is in charge of matching, in each new frame, the predicted features contained in the set $\mathcal{P}'_{k|k-1}$. In order to improve matching $k|k-1$ performances, we implemented an active search technique [11]: candidates for new corresponding features are searched only in the most probable areas (which are modeled using ellipsoids), where there is the 99% probability of finding them. This block uses information from the predicted measurements and its related covariance $S_{k|k-1}$ in order to compute the measurement vector y_k .

When y_k has been computed, the algorithm performs the standard EKF update and prediction steps. The innovation vector $e_k = y_k - \hat{y}_{k|k-1}$ is computed; then the filter performs update and prediction steps in order to estimate, respectively, the updated state $\mu_{k|k}$ with its covariance $\Sigma_{k|k}$ and the predicted state $\mu_{k+1|k}$ with its covariance $\Sigma_{k+1|k}$. In the proposed implementation, the update step includes the 1-Point RANSAC algorithm [9] for outliers rejection, in order to improve robustness.

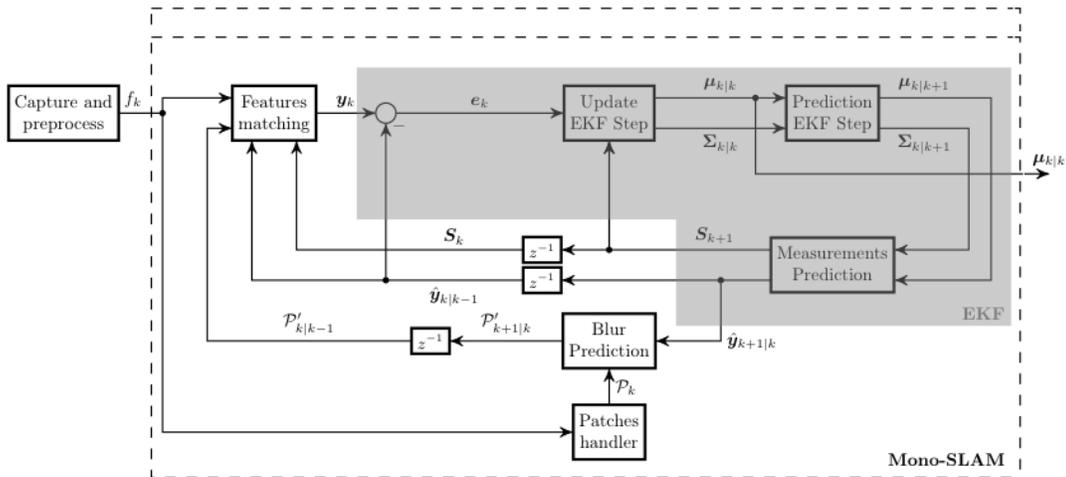


Figure 1. Architecture of the proposed solution.

The patches handler block is in charge of managing the \mathcal{P}_k set. In particular, it is used to find the best features to track at the very beginning (i.e., when the first frame is being acquired) and to delete old features that are no more useful (e.g., when they exit from the frame bounds) and must be replaced. More details are reported in Section 3.1.

Blur prediction block is used to predict the motion blur affecting patches when the camera undergoes quick motion. Details are reported in Section 3.2.

Note that the z^{-1} block represents a delay equal to the inverse of the camera frame rate, which is used to store predicted information until a new frame has been acquired.

3.1. Feature matching

The task of features matching consists in finding correspondences for a set of features when a new frame is acquired. Each feature is associated with a descriptor, which will be discussed later in this subsection.

Once a new image is acquired, new meaningful features are found and their descriptors are computed. Then, correspondences are found between the new features and the features from the previous frame. Matching techniques which are fast enough to be used in real-time are usually not robust enough for finding correct correspondences. Luckily, using a probabilistic approach gives an advantage, since the regions where we look for correspondences can be reduced to the ones where the features are more probable to be found. In other words, the search for matches is performed in a region around the prediction of the measurements and the size of the region is related to the covariance matrix of that features; intuitively, the larger the uncertainty on the position of the features, the wider the searching region should be.

In Mono-SLAM, that region is given by projecting the ellipsoid related to the prediction of the feature f_i in the image plane, where s is a number specifying the size of the region, usually $s = 2$ or $s = 3$, i.e., 95% or 99% of probability of finding the feature. Please note that the size of this region is given by $s^2 S_i$, where S_i is the 2×2 submatrix of S_s related to the measurements estimation of. Hence, the matching is performed on the ellipsoid S_i having center in p_i and size $s^2 S_i$ mathematically described as follows:

$$S_i = \left\{ p \in \mathcal{I} \mid (p - p_i)^T S_i^{-1} (p - p_i) \leq s^2 \right\}.$$

If the feature can be matched inside this search region, then it is considered as successfully matched. The image coordinates z_i , denoting the position of the feature in the new frame, are appended to the measurement vector z_k and the feature will contribute towards the correction of the estimates mean μ_k and covariance Σ_k . Otherwise the predicted measurement for f_i will be removed from h_k , since z_k there is no corresponding measurement in. Subsequently, the corresponding rows of are deleted from the Jacobian H_k of vector h_k . This method, known as active search, makes the algorithm to be much more robust and it allows also to use very simple and fast features descriptors to perform matching. For this reason, in this work a simple patch matching techniques was used, as in the original work in [12].

A patch is defined as a sub image of given dimensions (usually small) extracted from an image around a specific pixel. In our work the center of each patch is found by an interest point detector. Patch matching requires to find the position of the center of the patch inside the original image, i.e., the interest point. The normalized cross correlation operator is the default solution to perform patch matching. Given two patches P and P' of the same dimensions, the cross correlation score between the two patches is defined as

$$C_{NCC}(P, P') = \frac{1}{n} \sum_u \sum_v \frac{(P(u,v) - \bar{P})(P'(u,v) - \bar{P}')}{\sqrt{\sigma_P^2 \sigma_{P'}^2}}, \quad (6)$$

where n is the number of pixels contained in each patch and \bar{P} and σ_P^2 are respectively mean and covariance of the intensity of the pixels of P , defined as

$$\bar{P} = \frac{1}{n} \sum_u \sum_v P(u, v),$$

$$\sigma_P^2 = \frac{1}{n} \sum_u \sum_v (P(u, v) - \bar{P})^2 = \frac{1}{n} \sum_u \sum_v P^2(u, v) - \bar{P}^2.$$

Patch matching is performed computing C_{NCC} between the patch to find and each patch having center in the search region S_i , defined above. The center of the patch which maximizes C_{NCC} is chosen as best match.

3.2. Blur correction

When the camera motion is very pronounced, patch matching could fail due to the amount of motion blur affecting the acquired camera frames. To handle motion blur, each patch is pre-blurred using the speed information contained in the predicted state, and the blurred patches are used to perform matching. Unlike alternative solutions that restore the whole image in order to handle blur, this solution uses the Mono-SLAM paradigm in order to reduce computational efforts, by applying blur prediction only to few patches instead of the whole image. More details on the implementation of this approach are reported in a previous work [19].

3.3.1-Point RANSAC

1-Point RANSAC algorithm for EKF filters has been originally proposed in [9]. The algorithm is composed by two parts: the first one is in charge of selecting low-innovation inliers, while the second one selects high-innovation inliers. Low-innovation inliers are selected by executing RANSAC using a single feature (point) to generate hypotheses and select the best consensus set. Unlike classical RANSAC algorithm, the hypotheses are generated using also information given by the EKF filter: new hypotheses are generated by performing EKF update on the selected points only. Then, a consensus set is created by collecting all measured points which lay inside a certain probability ellipsoid (given by a threshold) centered in the predicted measurements obtained from the computed hypothesis.

Low-innovation inliers are elements of the consensus set of the best hypothesis after RANSAC execution. They are assumed to be generated by the true model since they are at a small distance from the most supported hypothesis. The remaining points could be both inliers and outliers, even if they are far from the supported hypothesis. This is due to the fact that the point chosen to generate the best hypothesis could not contain all the information needed to correctly update the state. For instance, it has been explained that distant points are useful for estimating camera rotation, while close points are needed to estimate translation. In the RANSAC hypotheses generation step, a distant feature would generate a highly accurate 1-point hypothesis for rotation, while translation would remain inaccurately estimated. Other distant points would, in this case, have low innovation and would vote for this hypothesis. But as translation is still inaccurately estimated, nearby points would presumably exhibit high innovation even if they are inliers.

High-innovation inliers are selected after a partial EKF update step involving only the low-innovation inliers selected by RANSAC. After this partial update, most of the correlated error in the EKF prediction is corrected and the covariance is greatly reduced. This high reduction will be exploited for the recovery of high-innovation inliers: as correlations have weakened, consensus for the set will not be necessary to compute and individual compatibility will suffice to discard inliers from outliers. For each point discarded by the RANSAC algorithm, we check if it is individually compatible by verifying whether it lies in a fixed-sized (multiple of its covariance)

probability region or not. If the check is passed, the point is added to the high-innovation inliers set. After that all points are checked, a second update involving the high-innovation inliers is performed.

3.4. Inverse depth coding

In order to tackle the problem of features initialization and features at infinity, in an alternative representation was proposed called inverse depth, which represents each feature with six parameters

$$\boldsymbol{\psi} = (x^c \ y^c \ z^c \ \theta \ \phi \ \rho)^T, \quad (7)$$

where $\mathbf{r}^c = (x^c \ y^c \ z^c)^T$ is the optical center of the camera the first time the feature is observed, θ and ϕ are respectively azimuth and elevation of the feature with respect to the image coordinate system and $\rho = 1/d$ is the so-called inverse depth of \mathbf{f} , where d is the distance of the feature from the optical center the first time it is observed. The 3D position of the features in \mathbf{W} can be computed as

$$\mathbf{y} = \mathbf{r}^c + \frac{1}{\rho} \boldsymbol{\eta}(\theta, \phi) \quad (8)$$

$$\boldsymbol{\eta}(\theta, \phi) = (\sin \theta \cos \phi \ -\sin \phi \ \cos \theta \cos \phi)^T \quad (9)$$

The advantage of using inverse depth encoding is that it allows to compute a normalized vector parallel to \mathbf{y} so defined

$$\hat{\mathbf{u}} = \frac{1}{d} \mathbf{y} = \rho \mathbf{r}^c + \boldsymbol{\eta}(\theta, \phi), \quad (10)$$

which is computable also in cases of features at infinity, i.e., with $\rho \rightarrow 0$. Note that inverse depth features can not represent point with zero depth, because this implies $\rho \rightarrow \infty$. This is not a problem because a features with zero or very small depth implies a real point coincident or very near to the camera optical center, i.e., inside the optics of the camera.

4. ROS IMPLEMENTATION DETAILS

The algorithm has been fully implemented in ROS as a node called Mono- SLAM. The node subscribes to an image topic containing the video stream coming from a video camera (in our work the topic is published by camera1394 or gscam ROS nodes). The full source code for our implementation will soon be available online on the repository of our laboratory¹.

The Mono-SLAM node publishes a set of topics which are the camera pose, the camera trajectory, the reconstructed point-map. Other topics which are useful for debugging purposes are the image frames showing the tracked features as well as their state (new, normal, discarded by RANSAC), and the covariances of all features. The topics can be visualized using the rviz ROS node. Figure 2 shows the different subscribed and published topics.

¹<https://github.com/rrg-polito>

Some examples of the node in action are shown in Figure 3 and 4. Figure 3 shows some frames elaborated by the algorithm. Several information are given:

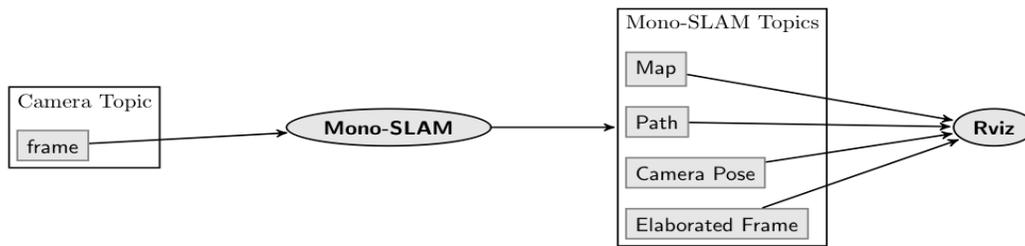


Figure 2. ROS graph of the developed solution.

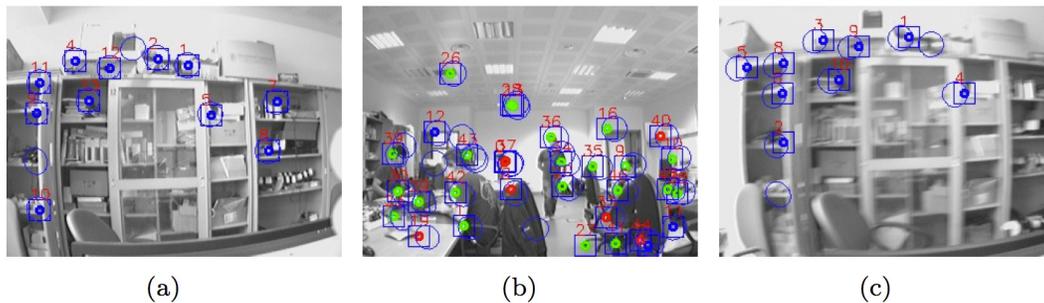
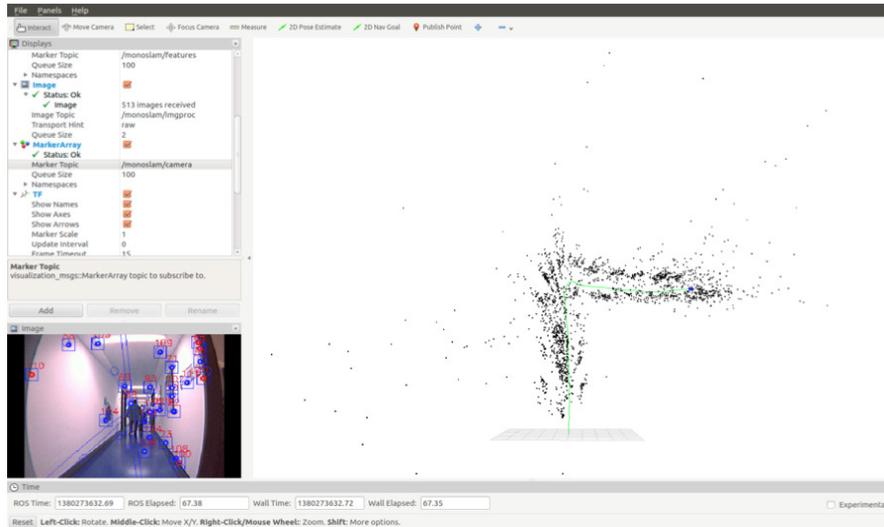
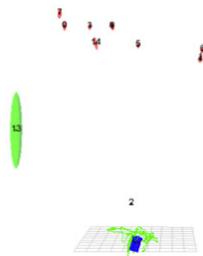


Figure 3. Examples of the algorithm in action. Positions of the patches are shown in each frame, as well as the ellipsoids that represent their predicted positions in the next frame. (a) shows a standard frame. (b) shows the 1-Point RANSAC algorithm working: blue patches are low-innovation inliers, green patches are high-innovation inliers and red patches are rejected measurements. (c) shows enhances prediction.

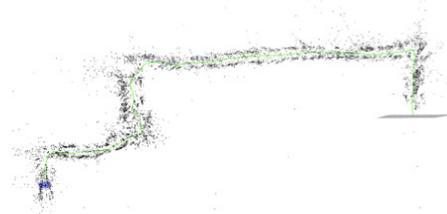
correctly matched patches are shown by blue rectangles, features selected as low-innovation inliers are shown as blue points, features selected as high-innovation inliers are shown as green points, features which are discarded by the 1-point RANSAC algorithm are shown as red points and finally the high probability region in which each feature should lie in the next frame is marked by an ellipse. Moreover, Figure 4 shows some examples of the visual output provided by rviz ROS node. The camera trajectory and the reconstructed map are shown. Each 3D feature is shown together with its covariance ellipsoid. The green covariance refers to features in inverse depth coding while the red ones refer to features in Euclidian representation. Please note that, due to limits of rviz, the covariance of the inverse depth features is represented by projecting the features in the Euclidian space, i.e., always as an ellipsoid, instead of correctly representing that covariance in the inverse depth space, i.e., with a conic shape.



(a)



(b)



(c)

Figure 4. Output from the algorithm can be visualized using the rviz ROS node. (a) shows a screenshot of the complete rviz environment: camera image with detected features is shown at the bottom-left, while in the main area the resulting 3D map is shown as well as the camera pose and its trajectory (in green). (b) shows the visualization of the ellipsoids representing the covariance on the pose of the features in 3D. Green ellipsoids are associated to inverse depth features while red ellipsoids to Euclidian features. (c) shows a larger map reconstructed by the algorithm.

5. EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithm, experimental tests were carried out using a standard monocular camera. Moreover the algorithm has been tested on a standard benchmarking dataset for robotic systems. All the experiments have been carried out on a standard PC equipped with an Intel i7 3.4 GHz CPU and 4 GB of RAM.

5.1. Hand held camera

We first tested the algorithm using a simple handheld FireWire camera. We tested robustness to occlusions by introducing an object (hand) in front of the camera while the algorithm was running. Figure 5 shows that the approach is able to reject occluded points and to match them correctly again when the occluding object is removed from the scene.

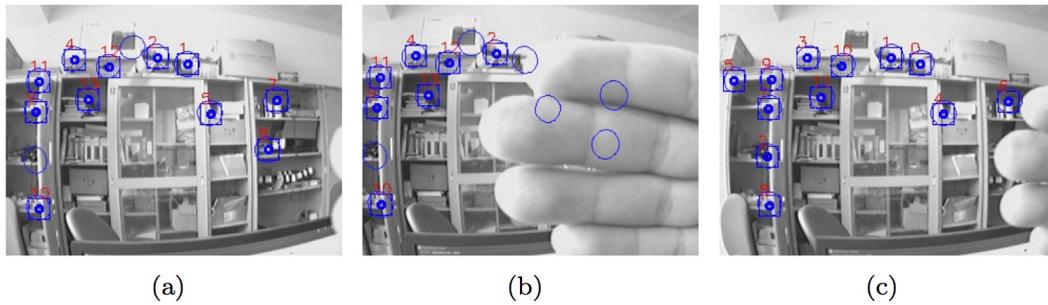


Figure 5. Testing robustness to occlusions. (a) The algorithm is working normally. (b) Patches occluded by the hand are discarded. (c) The hand is removed from the scene patches are matched again.

In another experiment a moving object (person) is present in the scene. In Figure 6 a person is moving in front of the camera. RANSAC is able to reject moving features, which are not matched. Occluded features are also discarded.

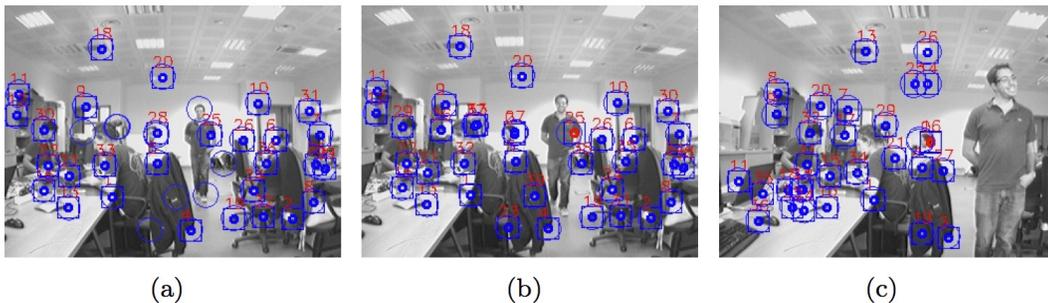


Figure 6: Testing robustness to moving objects. In (a) the person starts moving. In (b) feature 25 is rejected by RANSAC. In (c) occluded features are removed from the state.

5.2. Rawseeds dataset

We also tested our algorithm on the datasets freely available from the Rawseeds Project [5]. These are high-quality multi-sensor datasets, with associated ground truth, of rovers moving in large environments, both indoor and outdoor. For each sensor, calibration data is provided. For the experiments we used the video stream coming from the front camera of the rover in both indoor and outdoor scenarios.

The robot trajectory estimated by our approach was compared with the ground truth available from the datasets. The ground truth is composed by the trajectory obtained from a multi-camera system and visual tags mounted on the robot, which is not available for the whole length of the trajectory, and the estimated trajectory coming from a standard SLAM algorithm based on laser scanner sensors; for the outdoor dataset GPS data is used for the ground truth.

For indoor experiments, we used the Biccoca_2009-02-26a dataset², in which the robot is moving in an indoor dynamic environment. Some results obtained from this dataset are shown in Figure 7. It is possible to note that the algorithm was able to perform with good accuracy on small and medium scales (10 or plus meters), while on large scales the trajectory is far from the real one.

²<http://www.rawseeds.org/rs/datasets/view/6>

This is due to the fact that the Mono-SLAM algorithm does not explicitly perform loop closing. Loop closing only happens when new patches are correctly matched to previous patches, but this is difficult after large displacements of the camera. Another issue is that the algorithm is not able to correctly measure rotations in some cases. This is due to the lack of significant features in some areas (e.g., when facing a wall) or in presence of too many moving objects in front of the robot.

For outdoor experiments Bovisa_2008-10-04 dataset³ has been used. Also this second experiment (see Figure 8) shows that the algorithm works very well on small and medium scales, while on large scales the estimated trajectory drifts from the real one, as in the previous experiment.

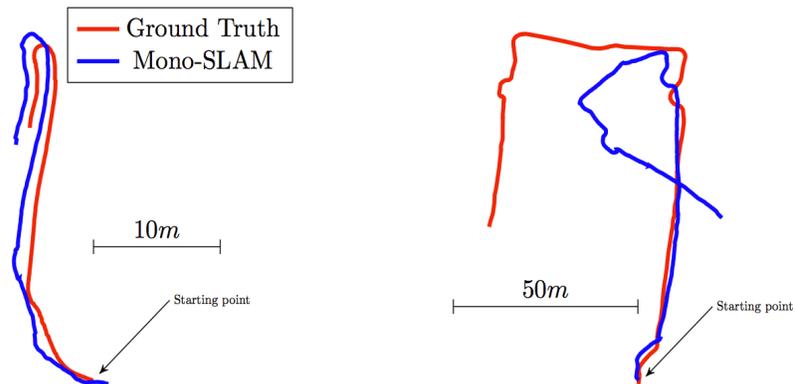


Figure 7. Results on the Bicocca 2009-02-26a dataset. Note that the small errors on estimated angles as well as scale drift both lead to a big difference in the final trajectories.

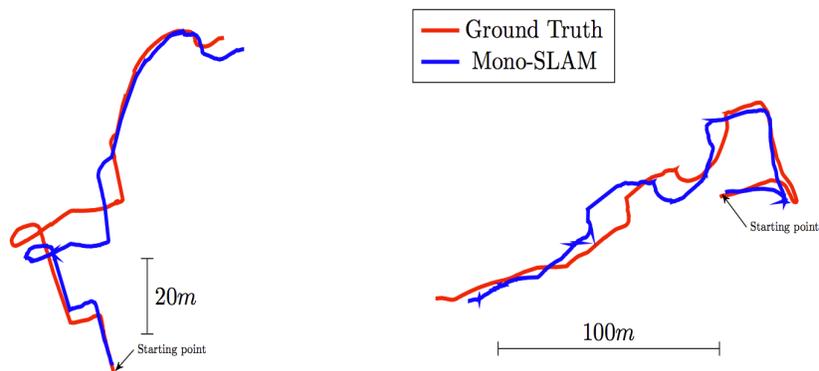


Figure 8: Results on the Bovisa 2008-10-04 dataset.

³<http://www.rawseeds.org/rs/datasets/view/7>

6. CONCLUSIONS

We presented a ROS implementation of the Mono-SLAM algorithm. The node is able to run in near real-time and is robust enough on small and medium scales to be used for retrieving the trajectory of the camera as well as re-constructing a 3D point-map of the environment. It should be noted that the performances of the algorithm are heavily influenced by the choice of parameters, in particular noise covariances. Moreover, some robustness issues still remain in the developed algorithm in the case of highly dynamic environments. Some solutions have been implemented in order to improve robustness, and actually the algorithm is able to work in dynamic environments and to correctly manage occlusions. Finally, when the camera trajectory is large, the computational time increases too much for meeting real-time constraints.

Future work will be devoted to improve the robustness of the approach on larger scales by implementing a way to detect and manage loop closings. Moreover, other sensors will be included when available, such as accelerometers, magnetometers, gyroscopes, and wheel odometry. Finally, more complex applications of the Mono-SLAM algorithm are under consideration. These applications concern a multi-robot extension of the algorithm; integration of high-level semantic informations in the algorithm; and an extension of the solution with multi-camera systems.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

- [1] OpenCV library. Website. <http://opencv.org>.
- [2] Radish: The robotics data set repository. Website. <http://eigen.tuxfamily.org/>.
- [3] Ros (robot operating system). Website. <http://www.ros.org>.
- [4] Josep Aulinas, Yvan Petillot, Joaquim Salvi, and Xavier Llado'. The slam problem: a survey. In Proceedings of the 2008 conference on Artificial Intelligence Research and Development: Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence, pages 363–371, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.
- [5] Andrea Bonarini, Wolfram Burgard, Giulio Fontana, Matteo Matteucci, Domenico Giorgio Sorrenti, and Juan Domingo Tardos. Rawseeds: Robotics advancement through web-publishing of sensorial and elaborated extensive data sets. In In proceedings of IROS, volume 6, 2006.
- [6] Simone Ceriani, Giulio Fontana, Alessandro Giusti, Daniele Marzocchi, Matteo Matteucci, Davide Migliore, Davide Rizzi, Domenico G Sorrenti, and Pierluigi Taddei. Rawseeds ground truth collection systems for indoor self-localization and mapping. *Autonomous Robots*, 27(4):353–371, 2009.
- [7] Javier Civera, Andrew Davison, and JMMontiel. Dimensionless monocular slam. *Pattern Recognition and Image Analysis*, pages 412–419, 2007.
- [8] Javier Civera, Andrew J Davison, and J Montiel. Inverse depth parametrization for monocular slam. *Robotics, IEEE Transactions on*, 24(5):932–945, 2008.
- [9] Javier Civera, Oscar G Grasa, Andrew J Davison, and JMM Montiel. 1-point ransac for extended kalman filtering: Application to real-time structure from motion and visual odometry. *Journal of Field Robotics*, 27(5):609–631, 2010.
- [10] Javier Civera, Dorian Glvez-Lpez, Luis Riazuelo, Juan D. Tardos, and J. M. M. Montiel. Towards semantic slam using a monocular camera. In IROS, pages 1277–1284. IEEE, 2011.
- [11] Andrew J Davison. Active search for real-time vision. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 66–73. IEEE, 2005.
- [12] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1052–1067, 2007.

- [13] Ankur Handa, Margarita Chli, Hauke Strasdat, and Andrew J Davison. Scalable active matching. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1546–1553. IEEE, 2010.
- [14] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011.
- [15] Georg Klein and David Murray. Parallel tracking and mapping on a camera phone. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 83–86. IEEE, 2009.
- [16] JMM Montiel, Javier Civera, and Andrew J Davison. Unified inverse depth parametrization for monocular slam. *analysis*, 9:1, 2006.
- [17] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011.
- [18] Edward Rosten and Tom Drummond. Fusing points and lines for high performance tracking. In *International Conference on Computer Vision*, pages 1508–1515. Springer, 2005.
- [19] L. O. Russo, G. Airo` Farulla, M. Indaco, Rolfo D. Rosa, S., and B. Bona. Blurring prediction in monocular slam. *International Design and Test Symposium, IEEE Conference*, 2013.
- [20] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, June 2013.
- [21] Hauke Strasdat, J. M. M. Montiel, and Andrew J. Davison. Editors choice article: Visual slam: Why filter? *Image Vision Comput.*, 30(2):65–77, February 2012.
- [22] Thomas Whelan, Michael Kaess, Maurice Fallon, Hordur Johannsson, John Leonard, and John McDonald. Kintinuuous: Spatially extended kinectfusion. 2012.

INTENTIONAL BLANK

PRE-RANKING DOCUMENTS VALORIZATION IN THE INFORMATION RETRIEVAL PROCESS

Chkiwa Mounira¹, Jedidi Anis¹ and Faiez Gargouri¹

¹Multimedia, InfoRmation systems and Advanced Computing Laboratory
Sfax University, Tunisia
m.chkiwa@gmail.com, jedidianis@gmail.com, faiez.gargouri@isimsf.rnu.tn

ABSTRACT

In this short paper we present three methods to valorise score relevance of some documents basing on their characteristics in order to enhance their ranking. Our framework is an information retrieval system dedicated to children. The valorisation methods aim to increase the relevance score of some documents by an additional value which is proportional to the number of multimedia objects included, the number of objects linked to the user particulars and the included topics. All of the three valorization methods use fuzzy rules to identify the valorization value.

KEYWORDS

Information Retrieval, Pre-ranking valorization, Fuzzy Logic, Fuzzy Rules

1. INTRODUCTION

In the information retrieval process, younger users have particularities concerning what they really need in results. In this paper we study those particularities in order to have maximum coverage of relevant documents. To do it, we present the Pre-ranking Documents Valorization which aims to increase some documents relevance score by an additional value in order to enhance their ranking. In order to make the additional value be proportional to the document characteristics we use fuzzy logic principles. The pre-ranking document valorization takes place after running the querying process which finds the relevant documents. Our framework materializes collaboration between two axes: the Semantic Web [1 and 2] and the Fuzzy Logic [3, 4 and 5]. We use semantic web technologies as RDF [6 and 7] to annotate semantically our collection of web documents and SPARQL [8] for querying the annotation. Also, we use Fuzzy rules to find the exact value added to a score relevance in order to enhance the ranking of the correspondent document. The rest of the paper is organized as follows: in section 2 we introduce some preliminaries about our framework which is an information retrieval system dedicated for children. Section 3 we present the pre-ranking document valorization by explaining its three types. Finally section 4 concludes the paper.

2. FRAMEWORK

To ensure a high quality of use, available information retrieval systems dedicated for children have common highlighted facts:

- Security: Since the web covers a large amount of uncontrollable data, security represents the first factor taken into account to create a safe information retrieval system dedicated for children.
- Design: it represents the main visual factor to call users attention.
- Profile: it represents the common way used to personalize a search process taking into account the user's personal interests.
- Querying: this factor makes the difference between information search engines even if it follows outwardly the same demarche: the annotation, the query/document matching, the ranking and the result display.

In addition of considering the cited facts, our particularity resides in the “pre-ranking document valorisation” which is localized in the figure below.

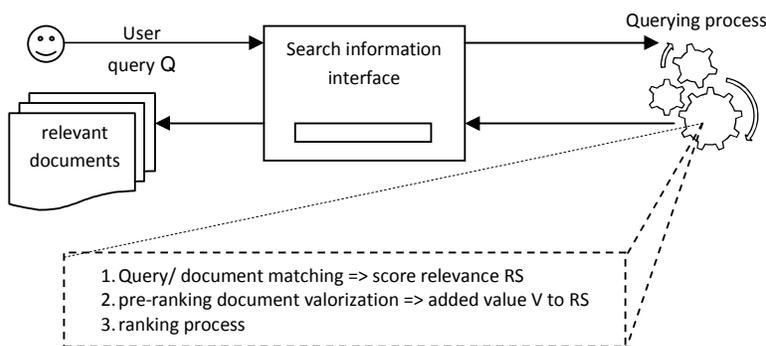


Figure 1. Our querying particularity

In order to explore young web users searching interests, we do a survey covering 723 children between 8 and 15 years. The results show that more than 75% of them prefer results rich by multimedia objects (especially pictures and video sequences). Likewise, more than 70% of surveyed children want to see in returned results information dealing with their own particulars (own country, avenue, friends ...) especially with the success of social networks use. In the next section we present two flexible methods that give the priority to relevant results rich by multimedia objects and/or dealing with the current user particulars. In the other side, we consider the fact that a returned document may be exhaustive and deal with different topics; this fact could misplace a user attention especially if it is a child. We propose therefore a method that avoid “multi-topic” documents and make priority to documents focusing only on query main topic. The three pre-ranking valorization methods are submitted to fuzzy rules which mainly use variables representing the user ages; Figure 2 shows the membership function representing web young user's age as input set which ranges from 3 to 14

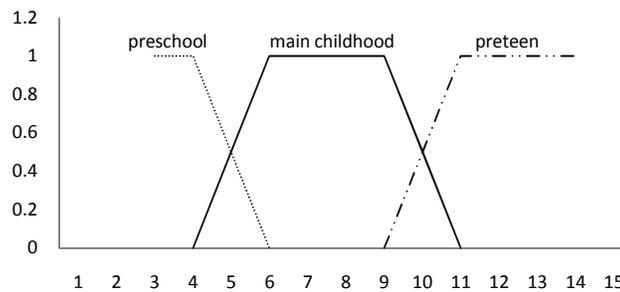


Figure 2. Fuzzy sets representing the different childhood periods of young web users.

3. PRE-RANKING DOCUMENT VALORIZATION

After running a classical query/document process, we get the score relevance of each document to the query. The pre-ranking document valorization aims to increase score relevance of relevant documents depending on their particularities. It's much easier and time-saver for the system to handle only relevant documents and not all the collection documents. For that, we choose to not run the document valorization while the querying process but after it. We define three types of document valorization: the “multimedia richer” document valorization, the “Nearest personal information” document valorization and “same-topic” document valorization. All of the three types of the pre-ranking document valorization are submitted to the application of fuzzy rules.

3.1. “Multimedia Richer” Document Valorization

Given that younger user are interested more in multimedia objects (images, video sequences ...) while the information retrieval process. The “multimedia richer” document valorization aims to increase the relevance score of relevant document including more multimedia objects. This fact is submitted to fuzzy rules aiming to decide the value added V to score relevance proportionally to the user age UA and the Number of Multimedia Objects in the Document $NMOD$.

- If UA is in preschool period and $NMOD$ is high then V is high.
- If UA is in main childhood period and $NMOD$ is high then V is medium.
- If UA is in preteen period and $NMOD$ is high or medium then V is low.

The Number of Multimedia Objects in the Document $NMOD$ is an input set represented by triangular membership functions which ranges from Min to Max (see figure 3). Min and Max are variables representing respectively the minimum number of multimedia objects included in a document in the collection and the maximum number of multimedia objects included in a document in the collection.

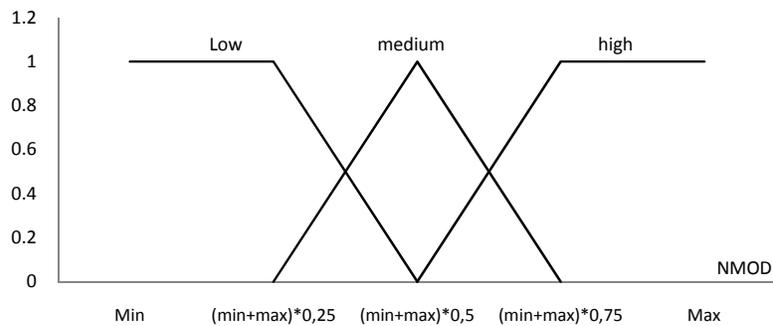


Figure 3. Fuzzy sets representing the variation range of NMOD.

3.2. “Nearest Personal Information” Document Valorization

As we mention before, more than 70% of surveyed children want to see in returned results information dealing with their own particulars (own country, own avenue, own school, friends ...). The idea is to familiarize the information searching concept to children through the maximum coverage of their own particular in the information searching process. Referring to this observation, we suppose that results will be considered relevant if they include some personal data. The nearest personal information document valorization exploits the user age UA and the Number of Personal information Items found in a Document $NPID$ as numerical inputs of the fuzzification operation submitted to the fuzzy rules. A personal information item found in a document may be a piece of text, a picture, or any multimedia object dealing with the current user particularities. The fuzzy rules listed below make decision about the value V added to a document relevance score in order to valorize documents dealing with user personal information:

- If UA is in main childhood period and $NPID$ is high then V is medium.
- If UA in preteen period and $NPID$ is high then V is high.
- If UA in preteen period and $NPID$ is medium or low then V is low.

As the $NMOD$ variable, The $NPID$ is an input set represented by triangular membership functions which ranges from 0 to Max (see figure 4). Max is a variable representing the maximum number of personal information items concerning a user found in a document of the collection and 0 means that a document didn't include any information items concerning the current user.

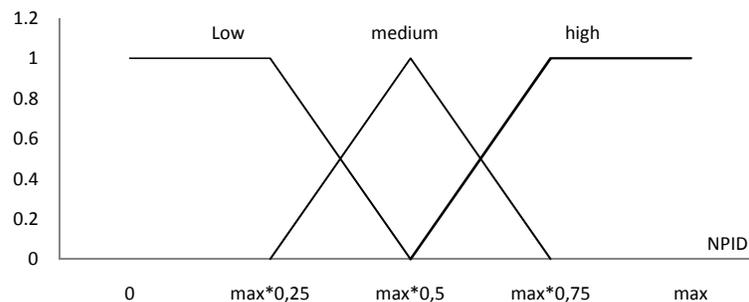


Figure 4. Fuzzy sets representing the variation range of NPID.

3.3. “Same-Topic” Document Valorization

A “Same-topic” document valorization aims to increase relevance score of document referring to the topics mentioned therein. The Figure 5 gives a structural view of this type of document valorization. The meta-document is introduced in [9] and it is able to annotate multimedia objects as well as web documents in a way that ensures its reusability. The querying process matches the user query with the meta-documents in order to identify the score relevance of the document to the query. We define the “topic cloud” as groups of weighted terms concerning a given topic. Simply, we collect potential terms representing a given topic to construct a topic cloud. The terms’ weights express the ability of each term to represent the topic. After running a usual querying process matching the query and the meta-documents, we get the relevance score for each annotated resource or document. At this point, the topic clouds are used to enhance ranking results in the benefit of relevant documents focusing mainly on query interests.

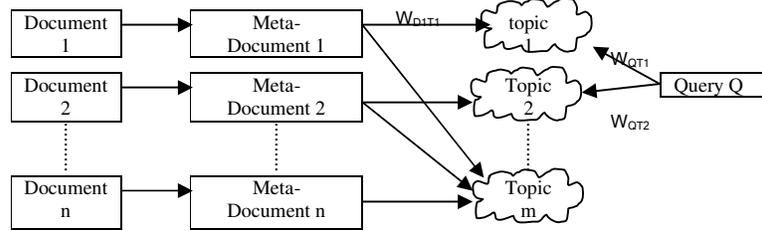


Figure 5. The “Same-topic” document valorization.

To run the “Same-topic” document valorization, first, we establish the meta-document/topic weighted links W_{DT} . W_{DT} expresses the potential topics mentioned by the document. To assign a weight W_{DT} to a meta-document/topic link, we simply sum the weights of topic terms existing in the meta-document. Then we establish query/topics weighted links which express the ability of each topic to represent the query. To assign a weight W_{QT} to a Query/Topic link, we use the classic similarity measure between two weighted terms vectors:

$$W_{QT} = \text{sim}(Q, T_i) = \frac{\sum_{j=1}^t w_{qj} * w_{t_{ij}}}{\sqrt{\sum_{j=1}^t (w_{qj})^2 * \sum_{j=1}^t (w_{t_{ij}})^2}} \quad (1)$$

The next step of the “Same topic” document valorization is to calculate for each document his topic similarity relative to the query in order to increase or decrease its relevance score in terms of the value of the topic similarity. The topic similarity TS is calculated as follow:

$$TS(Q, D) = \sum_{i=1}^k |W_{QT_i} - W_{DT_i}| \quad (2)$$

The main goal of a “Same topic” document valorization is to increase relevance of documents focusing on the same query topics and valorize “mono-topic” documents to users at an early age in order to facilitate its comprehension. The $TS(Q, D)$ value is optimal when its value is minimal; this means that the query and the document are focusing on the same topics with approached values. Contrariwise, if the TS is high, this means that the document deals with other topics in addition to the query topics. Finally, the increase value V affected to a document Relevance Score RS is based on the following fuzzy rules:

- If UA is in (pre-school or main childhood period) and TS is low then V is high
- If UA is in preteen period and TS is low or medium then V is medium

Remains to mention that this valorization method is inspired from our previous work [10], except that we exclude the rule that decrease relevance score of documents having a high TS value. This exclusion allows to not penalizing some document because of the diversity of topics included therein. Also, we include the UA variable in this valorization method instead of score relevance RS. The *TS* variable is an input set represented by triangular membership functions which ranges from 0 to Max (see figure 6). Maximum *TS* means that the current document deals with several topics in addition to the query topics and a null *TS* means that the document and the query are dealing with the same topics.

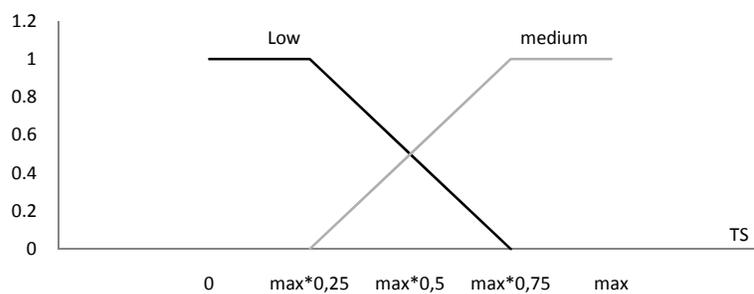


Figure 6. Fuzzy sets representing the variation range of TS.

After the application of valorization methods separately to a document D_i we sum the eventually deduced values to get the value V_i which is added to the document score relevance in order to get the new score relevance. After applying the valorization process to the relevant documents set found in the querying process, we pass finally to the ranking procedure. Table 1 summarizes the document valorization process.

Table 1. Recapitulation of the pre-ranking document valorization process.

Valorization method	Inputs variables		output
Multimedia Richer	NMOD	UA	v_1
Nearest Personal Information	NPID	UA	v_2
Same-Topic	TS	UA	v_3
Valorization process	Document D_i		$V_i = \sum_{k=1}^3 v_k$

4. CONCLUSION

In this paper, we present three methods to valorize score relevance of some documents depending on their characteristics concerning the multimedia included objects, the user particulars and the topics mentioned therein. In this work, we use fuzzy rules to define in a flexible way the value added to the score relevance of valorized documents. Our framework is under development and it represents an information retrieval system dedicated for kids. We are working in the short-term on the identification of the relevant range and shape of the membership function representing the added value V usable on the three valorization methods.

REFERENCES

- [1] Tim Berners-Lee, James Hendler and Ora Lassila the Semantic Web. Scientific American: Feature Article: The Semantic Web: May 2001
- [2] W3C. Semantic Web <http://www.w3.org/standards/semanticweb/>

- [3] Zadeh, L.A.: Fuzzy sets. *Information and Control* (1965) 338–353
<http://www-bisc.cs.berkeley.edu/Zadeh-1965.pdf>
- [4] Lotfi A. Zadeh: Knowledge Representation in Fuzzy Logic. *IEEE Trans. Knowl. Data Eng.* 1(1): 89-100 (1989)
- [5] Lotfi A. Zadeh: A Summary and Update of "Fuzzy Logic". *GrC* 2010: 42-44
- [6] Manola, F. Miller, E., Beckett, D. and Herman, I. 2007. *RDF Primer*
<http://www.w3.org/2007/02/turtle/primer/>
- [7] RDF Working Group. *Resource Description Framework (RDF)* <http://www.w3.org/RDF/>
- [8] Eric Prud'hommeaux, W3C. *SPARQL Query Language for RDF* W3C Recommendation 15 January 2008 <http://www.w3.org/TR/rdf-sparql-query/>
- [9] Jedidi A. (July 2005). « modélisation générique de documents multimédia par des métadonnées : mécanismes d'annotation et d'interrogation » Thesis of « Université TOULOUSE III Paul Sabatier », France.
- [10] Chkiwa, M., Jedidi, A., Gargouri, F. (October 2013). *Handling Uncertainty in Semantic Information Retrieval Process*. *URSW 2013*: 29-33, Sydney Australia

INTENTIONAL BLANK

AUTO LANDING PROCESS FOR AUTONOMOUS FLYING ROBOT BY USING IMAGE PROCESSING BASED ON EDGE DETECTION

Bahram Lavi Sefidgari¹ and Sahand Pourhassan Shamchi²

¹Department of Computer Engineering, EMU, Famagusta, Cyprus
Bahram.lavi@emu.edu.tr

²Department of Mechanical Engineering, EMU, Famagusta, Cyprus
Sahand.pourhassan@hotmail.com

ABSTRACT

In today's technological life, everyone is quite familiar with the importance of security measures in our lives. So in this regard, many attempts have been made by researchers and one of them is flying robots technology. One well-known usage of flying robot, perhaps, is its capability in security and care measurements which made this device extremely practical, not only for its unmanned movement, but also for the unique manoeuvre during flight over the arbitrary areas. In this research, the automatic landing of a flying robot is discussed. The system is based on the frequent interruptions that is sent from main microcontroller to camera module in order to take images; these images have been distinguished by image processing system based on edge detection, after analysing the image the system can tell whether or not to land on the ground. This method shows better performance in terms of precision as well as experimentally.

KEYWORDS

Quadcopter, Flying Robot, Image Processing, Edge Detection, Auto Landing

1. INTRODUCTION

Quadcopter, also named as quadrotor helicopter, is a vehicle that moves with electric motors. Basically there are four upward rotors which help quadcopter for any kind of manoeuvres within its flying region. Since the quadcopter is classified as unmanned aerial vehicle (UAV), it is believed that by the increasing demand for autonomous UAVs, quadcopters are going to be developed in autonomous control system. In the last decade, due to the military and security reasons many attempts had been conducted related to this issue [1,2,3]. These days, the quadcopter is taken into consideration by many of the robotic researchers regarding its complicated structure. Such tendencies provide a platform for us to find answers to the landing challenges faced by this technology. So by the help of a specified marks or (Signs), device can recognize its landing place. This can be done by aids of image processing to detect and extract the predefined image.

A common limitation to the quadcopters is the battery life during flight [4], and our homemade quadcopter usually fly about 20 minutes with its on-board battery. With this constraint, the best method for automatic landing is to hire the afore-mentioned technique. Using one extra micro controller and small camera can be helpful to save the battery power.

The aim of this research is to develop a real-time system for detecting and tracking the specific place for landing the quadcopter and in this regard, the edge detection have been applied. The system is implemented experimentally. The main purpose of this work is to employ the quadcopter as a safety and security robot in wide range areas.

2. BACKGROUND

Quadcopter itself is not a novel technology; instead, controlling it is considered as a new challenge [5-11]. Basically, the automatic landing in quadcopters is a challenging issue for researchers. This paper will provide a guide by using scientific methods to landing off a quadcopter.

2.1. Quadcopter Dynamic Model

Quadcopter robot is controlled by various angular speeds produced from each motor. It has four rotors arranged in cross shape. The front and back rotors are rotating counter of clockwise direction and the left and right rotors are rotating on clockwise side [12]. This structure is shown in figure 1.

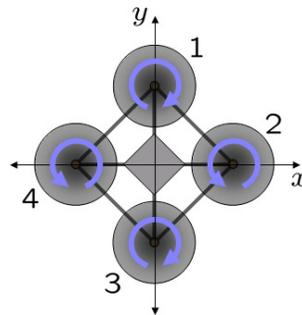


Figure 1. Whole structure of rotations in quadcopter

The quadcopter, which is considered in this work, is classified as a small size of flying robot with limited weight, less than 1 kg. Integration of a homemade flying robot with small camera is required, as well as having an adequate embedded image processing platform. The prototype of implementation with detail is shown in figure. 2.



Figure 2. Prototype of quadcopter with detail

The control of quadcopter which was provided by PID controller and coefficients was improved by genetic algorithm [13, 14] which was implemented on AVR microcontroller. The execution of the overall system was evaluated in real-time experiments. A small CMOS camera module with photographic array 320*240 was embedded as shown in figure 3.

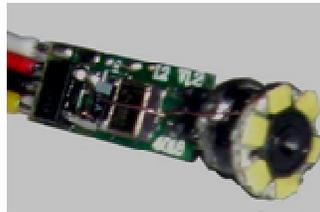


Figure 3. CMOS camera module

2.2. Image Processing in Quadcopters

Landing place detection consists of detecting and tracing special shapes which was pre-defined for landing the quadcopter. There are many algorithms that are developed for image processing, such as HOG and LBP [15], Haar wavelets and EOH [16], region covariance matrix [17-18], partial least squares [19], edge detection[20], stereo vision, monocular vision, laser and vision [21-25], sonar and vision, and thermal vision [26-29]. To the best of our knowledge, most of the image processing algorithms can be used in our paper; however, based on our previous experience in edge detection, this algorithm will be employed for detecting special shapes.

3. LANDING PLACE DETECTION METHODS ON QUADCOPTER

3.1. System Overview

Landing Detection and tracing system extract the exact place which is defined for landing the vehicle. It recognizes an object from the altitude (approximately up to 10meter) and employs geometric and edge detection algorithms to check whether the object has circular shapes or not. The extraction of circular shapes, which defines our place for landing, could be used instead of high quality image. In this regard, by using interruption emitted from main microcontroller to

take frequent images and identify landing scope. Figure 4 presents the diagram of proposed system and figure 5 presents the algorithm which is implemented in this paper.

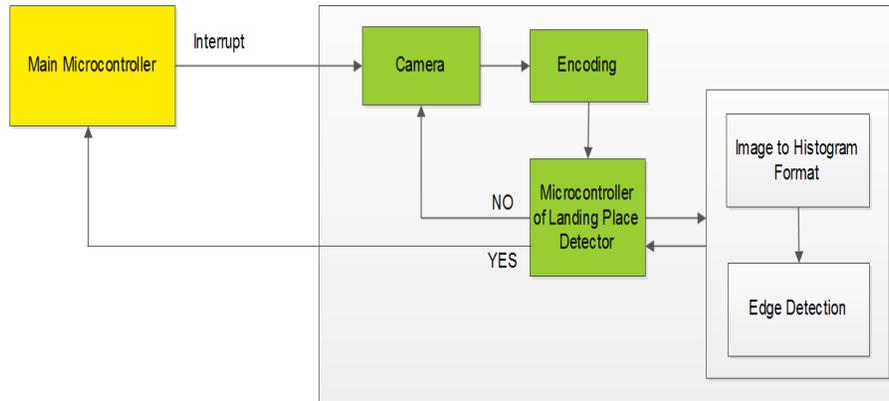


Figure 4. Block diagram of landing place detection in quadcopter

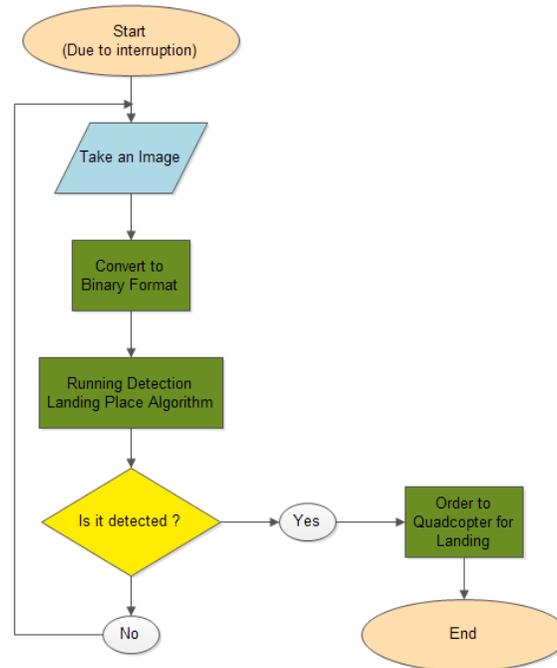


Figure 5. Implemented algorithm for detecting landing place of quadcopter

After interruption, which is sent out from main micro controller, the camera takes some frequent images and at the same time sends these to the encoding module for conversion into a binary format; Converted images sent into the landing place detector and the trace of shape will begin. Table 1 illustrates the list of modules and their methods which are employed in this paper. Those details will be presented in section below.

Table 1. List of modules

Module Name	Input	Output	Method
Main Microcontroller	Final Decision	Interrupt Order to Landing off	Dynamic Programming Dynamic Programming
Camera	Take an image	Send to Encoding	-
Encoding	Image	Binary format of Image	Dynamic Programming
Histogram	Binary Image	Geometric Histograms	Pattern Machine
Edge Detection	Geometric Histograms	Final Decision	Morphology Operation

First of all, for checking the accuracy of detection and tracing, we considered our case in MATLAB software, after utilization and getting the satisfactory results, next was implementation. A suitable algorithm which was written in C++ programming language was used for robotic programming and it creates understandable file that is programmed for SMD AVR micro controller. Figure 6 present our result in MATLAB software.

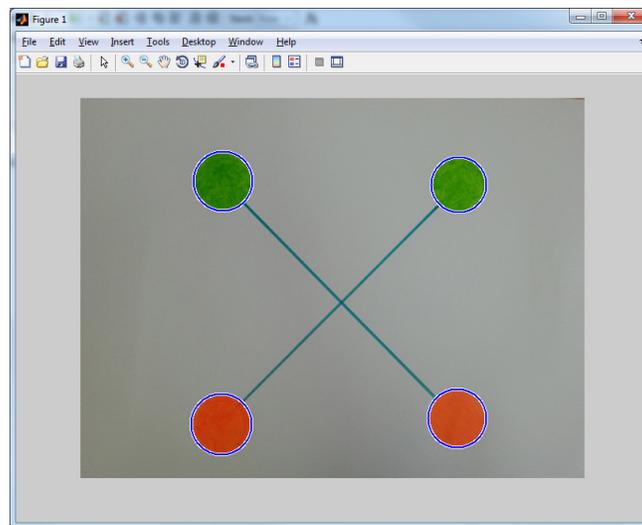


Figure 6. Result of detection circles which are defined for landing off quadcopter in MATLAB shapes.

3.2. Main Microcontroller

The micro controller which was used in this paper is in the same category of AVR family. Not only microcontroller has variety of duties on quadcopter, but also it keeps sending the interruption to camera module in order to catch our predefined image. The interruption which is determined for this case is Timer1 of micro controller. By using it, micro controller will be able to work in parallel mode.

3.3. Camera

The camera module is essential for detection; obviously, it takes the image for detecting and tracing landing place and sends each of them to the next module, encoding module, the detecting process will begin.

3.4. Encoding

To employ image processing algorithm, the image should convert to a binary format. An image in binary format has only two possibilities for each pixel. Numerically, those values are 0 for black and 1 until 255 for white. This converter is so important for understanding the image processing algorithm which is used in this paper.

3.5. Landing Place Detector

The Histogram that is based on geometric histogram is valued in the horizontal and vertical directions; similarly it can be done by Euclidean distance. This sequence executes for each pixel in binary format to achieve the degree of vectors. For each histogram, degree of match between model feature M_j and image feature I is calculated by:

$$D_j = \sum_{i=1}^n \sqrt{I_i \cdot M_{ji}} \quad (1)$$

Due to the geometric histogram of the binary format, another calculation would be the edge detection. The specific algorithm is employed to Euclidean distance in order to transform the image. First of all, for 2D images, the algorithm distinguishes the image resolution in X and Y directions and the edge pixels should be defined. Secondly, for each edge pixel there is an exact metric distance, so each pixel of the metric is correlated with a matching area of the binary image and they are assigned a distance value via Euclidean distance to get nearest pixels in binary format image. For two pixels point $P(x_1, y_1)$ and $Q(x_2, y_2)$ in 2D image the classical Euclidian distance is defined by:

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

In order to extract the background distance via Euclidian distance formulation, the image resolution is utilized in X and Y dimensions, therefore:

$$\sqrt{(x_2 - x_1)^2 \times x_{resolution}^2 + (y_2 - y_1)^2 \times y_{resolution}^2} \quad (3)$$

These formulas are applied for two areas in the image. Background and circular shapes in image which is taken from camera.

4. EXPERIMENTS AND RESULTS

Regarding to our proposed implementation methods in MATLAB software, to detect and trace circular shapes which are defined for landing place of quadcopter, we got good results about different images. So it is assumed that the quadcopter could carry out by this way. We have worked hard to ensure our proposition in experimental practice. See the experimental video clip on (<https://www.youtube.com/watch?v=zfPzSbz93rE>).

5. CONCLUSIONS

In this research paper, image processing technology based on edge detection was utilized to detect the landing spot of a quadcopter. Basically, the purpose of this study is to detect the landing place of homemade quadcopter, which has a great sensitivity in bright environments.

The hierarchical classifier that is implemented in this issue uses two distinct detectors. First one is the timer of main micro controller used to send interruptions to the camera. By frequent images that were taken via the camera, quadcopter can identify its predefined landing place. Meanwhile second detector has to apply the detection algorithm on the image to detect and trace the landing place. This method is examined by MATLAB software and satisfactory outcomes are achieved, and also this approach is very accurate and performs to detect landing place. However, its landing on exact place and recognition of the spot in foggy places are challenging issues and needs further improvements. It is believed that by this technique another step is taken toward autonomous flying robots.

REFERENCES

- [1] Teppo Luukkonen, "Modelling and control of quadcopter," Independent research project in applied mathematics, Espoo, August 22, 2011.
- [2] H. Huang, G. M. Hoffmann, S. L. Waslander, and C. J. Tomlin, "Aerodynamics and control of autonomous quadrotor helicopters in aggressive maneuvering," IEEE International Conference on Robotics and Automation, pp. 3277–3282, May 2009.
- [3] B. Lavi, "HUMAN BODY DETECTION AND SAFETY CARE SYSTEM FOR A FLYING ROBOT", Second International Conference on Advanced Information Technologies and Applications (ICAITA), pp. 329–337, November 2013.
- [4] Achtelik, M. C., Stumpf, J., Gurdan, D., & Doth, K. M. (2011, September). Design of a flexible high performance quadcopter platform breaking the mav endurance record with laser power beaming. In Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on (pp. 5166-5172). IEEE.
- [5] Kunzmann, J., & Berry, C. A. (2011). Investigation in the Control of a Four-Rotor Aerial Robot.
- [6] S. Bouabdallah, P. Murrieri, and R. Siegwart, Design and Control of an Indoor Micro Quadrotor, Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA), 5(26), New Orleans, LA, April 26 - May 1, 2004, 4393-4398.
- [7] B. Erginer and E. Altug, Modeling and PD Control of a Quadrotor VTOL Vehicle", Proceedings of the 2007 IEEE Intelligent Vehicles Symposium (IV'07), Istanbul, Turkey, June 13-15, 2007, 894-899.
- [8] S. Bouabdallah and R. Siegwart, Towards Intelligent Miniature Flying Robots, (Berlin, GE: Springer, 2006).
- [9] P. Pounds, R. Mahony, P. Hynes, and J. Roberts, Design of a Four-Rotor Aerial Robot, Proceedings of the 2002 Australasian Conference on Robotics and Automation, Auckland, Australia, November 2 - 29, 2002.
- [10] P. Pounds, R. Mahony, Corke, P. Corke, and J. Roberts, Towards Dynamically-Favourable Quadrotor Aerial Robots, Proceedings of the 2004 Australasian Conference on Robotics & Automation (ARAA), Canberra Australia, December 6 - 8, 2004.
- [11] A. Tayebi and S. McGilvray, Attitude Stabilization of a VTOL Quadrotor Aircraft, IEEE Transactions on Control Systems Technology, 14, May 2006, 562-571.
- [12] González, I., Salazar, S., Torres, J., Lozano, R., & Romero, H. (2013). Real-Time Attitude Stabilization of a Mini-UAV Quad-rotor Using Motor Speed Feedback. Journal of Intelligent & Robotic Systems, 70(1-4), 93-106.
- [13] A. Tayebi and S. McGilvray, "Attitude stabilization of a four-rotor aerial robot," 43rd IEEE Conference on Decision and Control, vol. 2, pp. 1216–1221, 2004.
- [14] Fatan, M., Lavi, B., Vatankhah, A., "An Adaptive Neuro PID for Controlling the Altitude of Quadcopter robot", 18th international conference on methods and models in automation and robotics, August 2013
- [15] X. Wang, T. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 32–39.

- [16] D. Geronimo, A. Sappa, A. Lopez and D. Ponsa, "Adaptive image sampling and windows classification for on-board pedestrian detection," in Proc. 5th Int. Conf. Comput. Vis. Syst., 2007.
- [17] Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 7, pp. 989–993, Jul. 2008.
- [18] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 10, pp. 1713–1727, Oct. 2008.
- [19] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in Proc. IEEE Int. Conf. Image Process., 2002, vol. 1, pp. 900–903.
- [20] Lee, K. D., Nam, M. Y., Chung, K. Y., Lee, Y. H., & Kang, U. G. (2013), "Context and profile based cascade classifier for efficient people detection and safety care system", Multimedia Tools and Applications, 1-18.
- [21] Satake, J., Miura, J.: Robust Stereo-based Person Detecting and Tracking for a Person Following Robot. In: Proc. ICRA 2009 workshop on person detection and tracking. Kobe, Japan (2009)
- [22] Wilhelm, T., Böhme, H.-J., Gross, H.-M.: A multimodal system for tracking and analyzing faces on a mobile robot. Robot. Auton. Syst. 48(1), 31–40, (2004), European Conference on Mobile Robots (ECMR '03)
- [23] Medionia, G., R.J. François A., Siddiquia, M., Kima, K., Yoonb, H.: Robust real-time vision for a personal service robot. Comput. Vis. Image Understanding 108(1–2), 196–203, Special Issue on Vision for Human-Computer Interaction, October–November 2007
- [24] Li, L., Koh, Y.T., Ge, S.S., Huang, W.: Stereo-based human detection for mobile service robots. Control, Automation, Robotics and Vision Conference, 2004, ICARCV, vol. 1, pp. 74–79 (2004)
- [25] Bellotto, N., Hu, H.: Multisensor-based human detection and tracking for mobile service robots systems, man, and cybernetics, Part B: cybernetics. IEEE Trans. 39(1), 167–181 (2009)
- [26] Böhme, H.J., Wilhelma, T., Keya, J., Schauera, C., Schrötera, C., Großa, H.-M., Hempelb, T.: An approach to multi-modal human-machine interaction for intelligent service robots. Robot. Auton. Syst. 44(1), 83–96 (2003)
- [27] Meis, U., Oberlander, M., Ritter, W.: Reinforcing the reliability of pedestrian detection in far-infrared sensing. 2004 IEEE Intelligent Vehicles Symposium, pp. 779–783, 14–17 June 2004
- [28] Mudaly, S.S.: Novel computer-based infrared pedestrian data-acquisition system. Electron. Lett. 15(13), 371–372 (1979)
- [29] Nanda, H., Davis, L.: Probabilistic template based pedestrian detection in infrared videos. IEEE Intell. Veh. Symposium 1, 15–20 (2002)
- [30] Bertozzi, M., Broggi, A., Fascioli, A., Graf, T., Meinecke, M.-M.: Pedestrian detection for driver assistance using multiresolution infrared vision. IEEE Trans. Veh. Technol. 53(6), 1666–1678 (2004)

AUTHORS

Bahram Lavi Sefidgari received a B.S degree from Islamic Azad University of Tabriz, Iran. Currently, He is a post graduate student in Computer Engineering in Eastern Mediterranean University (EMU) and work in computer center of university. His research interest is Artificial Intelligent in robotics.



Sahand Pourhassan Shamchi received a B.S degree from Islamic Azad University of Tabriz, Iran. Currently, He is a post graduate student in Mechanical Engineering in Eastern Mediterranean University (EMU). His research interest is manufacturing.



INTELLIGENT MULTI-AGENT FUZZY CONTROL SYSTEM UNDER UNCERTAINTY

Ben Khayut¹, Lina Fabri² and Maya Abukhana³

¹Department of R&D, IDTS at Intelligence Decisions Technologies Systems,
Ashdod, Israel

ben_hi@hotmail.com

²Department of R&D, IDTS at Intelligence Decisions Technologies Systems,
Ashdod, Israel

lina.fabri@gmail.com

³Department of R&D, IDTS at Intelligence Decisions Technologies Systems,
Ashdod, Israel

maya.kh111@gmail.com

ABSTRACT

The traditional control systems are a set of hardware and software infrastructure domain and qualified personnel to facilitate the functions of analysis, planning, decision-making, management and coordination of business processes. Human interaction with the components of these systems is done using a specified in advance script dialogue "menu", mainly based on human intellect and unproductive use of navigation. This approach doesn't lead to making qualitative decision and effective control, where the situations and processes cannot be structured in advance. Any dynamic changes in the controlled business process make it necessary to modify the script dialogue. This circumstance leads to a redesign of the components of the entire control system. In the autonomous Fuzzy Control System, where the situations are unknown in advance, fuzzy structured and artificial intelligence is crucial, the redesign described above is impossible. To solve this problem, we propose the data, information and knowledge based technology of creation Situational, Intelligent Multi-agent Control System, which interacts with users and/ or agent systems in natural and other languages, utilizing the principles of Situational Control and Fuzzy Logic theories, Artificial Intelligence, Linguistics, Knowledge Base technologies and others. The proposed technology is defined by a) methods of situational fuzzy control of data, information and knowledge, b) modelling of fuzzy logic inference, c) generalization and explanation of knowledge, d) fuzzy dialogue control, e) machine translation, f) fuzzy decision-making, g) planning and h) fuzzy control of organizational unit in real-time under uncertainty, fuzzy conditions, heterogeneous domains, multi-lingual communication in Fuzzy Environment.

KEYWORDS

Intelligent fuzzy control

1. INTRODUCTION

1.1. Analysis

Data, information and knowledge are one of the main *components* in the world development. Their correct use leads to the adoption of *relevant* decisions and *effective* control in *Intelligent Multi-agent Fuzzy Control System (IMAFCS)*.

Traditional Control Systems are using a "Menu" principle to interact with their users (analysts, experts, managers). Each professional within its subject area, spends an *unproductive* time to navigate the menu dialog instead of solving their day to day tasks. They can be more productive in interacting with the applications in a more natural and less predefined way, where the system itself should find the required and relevant data, information or knowledge to support their tasks. Changes in business process of organizational unit entail changes in the script menu dialog. This leads to a permanent and continuous *redesign* of the Control system. In these circumstances, since the technology doesn't support the business process need, it is impossible to make relevant real-time decisions and control. To avoid such a discrepancy it becomes necessary to use the principle of *Situational Control* [6] in the design of the *Intelligent Control System (ICS)*. Given the uncertainty of the environment, as well as increased complexity of business processes and technologies, the use of *Fuzzy Logic* [1] and Artificial Intelligence becomes necessary. A multi-lingual *Linguistic Processor (LP)* should be used to support human-machine interaction to incorporate a business process into the control system. For real time decision making and *fuzzy control* in a *fuzzy environment* we suggest to use of Fuzzy Logic Inference [13], Generalization and Explanation of knowledge [15], Dialog Control [11] and other methods. Obviously, the considered *IMAFCS* should be able to support a variety of *subject areas*. The solution of these tasks in this article is focused on the *systematic* approach of *modeling, planning and controlling* of *linguistic* and *subject area* data, information, knowledge, *fuzzy logic inference* and others, by mapping the *objectives* and *constraints* in *fuzzy environment*.

The *novelty* of the technology which designs *IMAFCS* consists of:

- *Modelling* and *situational fuzzy control* of data, information and knowledge for implementing an *automatic* fuzzy inference and finding a correct, accurate, timely and adequate *decision*, taking into account a current *situation* and impact of *fuzzy environment*.
- Using of resulting *decision*, criteria and purpose for providing of *modelling, planning and control* of the business process in the fuzzy environment.
- Converting and deriving images, concepts, meanings from *natural languages* in various subject areas and serializing them into the *bases* of data, information and knowledge.
- Use of these *bases* for *multi-lingual* human - machine *communication* using methods of dialog control, generalization and explanation of knowledge in the *intelligent fuzzy control system*.
- Use of *properties* of a) *atomicity* of data, b) *relationality* of information, c) *figurativeness* of knowledge for their *integration* and *aggregation*.
- Using *methods* of *wisdom, intuition* and *behavior* and others to obtain decision of high quality and precision.

The analysis of the state of scientific research in the field of design *intelligent control systems* showed that the directions and methods of implementation are related mainly to their *functionality*. In this context, we will hold a brief of comparative analysis of the functionality of the *IMAFCS*, offered by us and other authors.

In [12] is given an interesting overview of the approach to the problem of *Fuzzy Control*. Our approach *differs* from the mentioned, the fact that in addition to proposed methods of formalization we taking into account the principles of *situational control*, *artificial intelligence* and others. This allows realizing fuzzy control in situation, which *unknown in advance*. At the same time, we have developed *modelling* techniques [16] based on the *managed* data, information and knowledge [10] that allows finding relevant solution with the *desired accuracy* in the *circumstances*. This accuracy is implemented using *intelligent agents* of analysis, decision-making, planning and others by using the values of fuzzy membership function.

In [17] is represented linguistic approach for solving decision problems under linguistic information using Multi-criteria decision making, linguistic modeling, aggregation and linguistic choice functions methods on base of rank ordering among of the alternatives for choosing the best of them.

The main difference between our systematic approach and the proposal is a:

- *Generalized notion of linguistic variable* of Fuzzy Logic, by which we evaluate and take into account not only the *morphological*, *syntactic* and *semantic*, but also, *behavioural*, *psychological* and other aspects of the terms (atomic units) of Natural Language (NL).
- *Situational Fuzzy Control in Fuzzy environment*, by which we control not only information, but also data, knowledge, decisions, agents and others.
- *Decision-making* process is based not only on using the *rank* for estimation of the alternatives, but also on *automatic Fuzzy Logic Inference*, *Planning*, *Control* of alternatives, situations and other units.
- *Multi-lingual interaction*, *generalization*, *explanation*, *serialization*, *storage* and *actualization of knowledge* in fuzzy conditions, *heterogeneous subject areas*, where the situations are *unknown in advance*, fuzzy structured and not clearly regulated.

In [19] are considered adjustable autonomous agents that possess partial knowledge about the environment. In a complex environment and unpredictable situations these agents are asked the help of human on base of the model, called HHP-MDP (Human Help Provider MDP) and requests, which *are set* in advance.

The comparative analysis of these and other works, associated with our work, showed, that there is no *integrated*, *systematic* and *linguistic* approach to the problem of *situational fuzzy control* in a *fuzzy environment*, including the techniques of situational control of fuzzy data, information and knowledge, modeling, planning, decision-making, dialog control and situational fuzzy control of the *organizational unit*, based on the achievements of *Fuzzy Logic*, *Situational Control*, *Artificial Intelligence*, *Linguistics* and others.

In this article, we present the results of our studies and the approach to the design of *IMAFCS* using our developed methods and tools.

1.1. Terminology

Data is organized in the memory and are perceived by the person or machine as facts, numbers, words, symbols, lines and other items of information. They are not related to each other and are found intexts, pictures and othermaps of reality.

Information is a group of *related data*, organized in the memory that respond to the questions of "who", "what", "where", "when" and others.

Knowledge is the *image* or domain *model*, extracted from *information* and organized in memory, which in *itself* are interpreted, structured, linked, associated, transformed, compared, upgraded, activated, analyzed, deduced, built, serialized and so on, in real time. The mentioned *image* or *domain* should respond to questions "why" and "how", consider the impact of environment and specificity of subject area, satisfying the criteria and purpose of existence.

Wisdom is a method of perceiving reality and achieving a unique solution (answer) on the basis of intelligence, archival knowledge (experience), principles and *inference* in a *certain situation*.

Intuition is a method of perceiving reality and a achieving a unique solution (answer) on the basis of intelligence, archival knowledge (experience) principles and *unique inference* in an *extreme situation*.

Modelling decisions is defined as construction of a new conceptual situation and a state of controlled units (fuzzy data, information, knowledge, inference and others), which meets the criteria and purposes of the information system in fuzzy environment. The purposes are functions of the information system.

Planning decisions is defined as a use of modelling results to create a sequence of alternative decisions that will match to the situation and the state of information system in the subsequent stages of management of the organizational unit.

Decision-making is defined as a process of modelling *fuzzy logic inference* [13] for selection the relevant decision from limited number of alternative decisions, obtained during the *planning decisions*.

Fuzzy Control is the process of using the modelling results of planning and decision-making in fuzzy environment, in order to implement a control action on the units (data, information, knowledge, decisions, organizational unit and others) to shift them and their control system to a new state, that matches a specified criterion.

Under the *fuzzy logic inference* [13] we mean *procedure* for determining the vector of internal and external *output* fuzzy variables $b_k^i \in V_k^m$ using a *new* vector of the values of the *input* fuzzy variables $a_k^i \in U_k^m$, which *transforms* the *IMAFCS* in its *new state*. This *procedure* is implemented on the extensional, intentional and reformative levels of modelling knowledge [16].

Under the *Dialog Control* we means the process of presenting partners of common commands(questions) to each other and providing by them targeted actions (issuing replies) relevant to the subject of the dialogue and the situations in which it occurs.

Intelligent fuzzy control system is understood as a *knowledge based* system, which is reliably electronic *autonomous* system, and which a) operates at a high-level operating system b) connected to the Internet, d) executes a native or cloud-based applications, e) analyzes the collected data, information and knowledge, and e) realizes the human-machine functions for solving problems in fuzzy environment.

Traditional Control System is an Information System, which is working on base of "menu" scenarios and is not autonomous.

According Wikipedia the *Organizational Unit* O^U (Figure 2, Figure 3) represents a single organization with multiple units (departments) within that organization.

The *business process* is an activity or set of activities in *organizational unit* O^U that will realizes a specific organizational goal.

Subject area understood by us as branch of knowledge and technologies, where the organizational units are functioning.

Environment is the surrounding reality, consisting of organizational units, information systems, robots, agents, agent systems and so on, which interact with each other under the influence of the environment.

A *multi-agent system (MAS)* [18] is a computational system where *agents* cooperate or compete with others to achieve some individual or collective task.

Agent is a *real-world* or *artificial* entity, which is a *person* (in the first case) and an *object* (in the second case), and which are capable of performing some *action* or *service* or otherwise, *interacting* with other entities.

Figure 1 depicts nesting of the above-defined concepts.

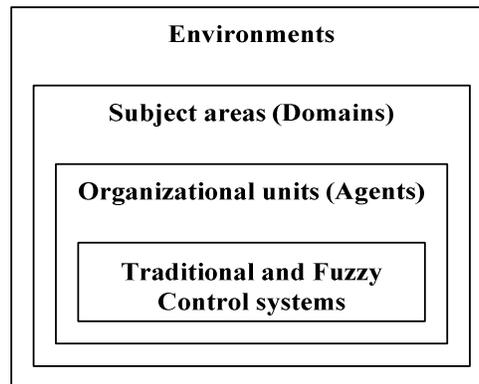


Figure 1. Control systems and environments

Thus, the above defined and implemented in the computer concepts are *agents*.

Combining, nesting and integrating the agents into the groups according to their *objectives* and *functional* features turn them into an *IntelligentMulti-agent System* in the paper.

Fuzzy Control provides a formal methodology for representing, manipulating, and implementing a human's heuristic knowledge about how to control a system [12].

According Wikipedia the *Fuzzy Control System* is a control system based on fuzzy logic – a mathematical system that analyzes analog input values in terms of logical variables that take on conditions values between 0 and 1, in contrast to classical or digital logic, which operates on discrete values of either 1 or 0 (true or false, respectively).

2. CONTENT

2.1. Conception

The main purpose of the *IMAFCS* function is to *facilitate* the analysis, planning, decision-making, fuzzy control and coordination of organizational unit on the basis of the intended *target* and the current *situation*. Under the *facilitating* in *IMAFCS* is understood the *full use of intelligence* of the system to perform these functions by interacting with it in natural language. In this connection the operation of the system should be based on the perception, processing and synthesizing knowledge in real time.

Given the fact that these processes occur in unexpected situations and *Fuzzy Environment*(people, robots, nature, space, hardware, software, other information and agent systems, and so on) becomes necessary to use the theories *Situational Control*[6], *Fuzzy Logic* [1] and other methods.

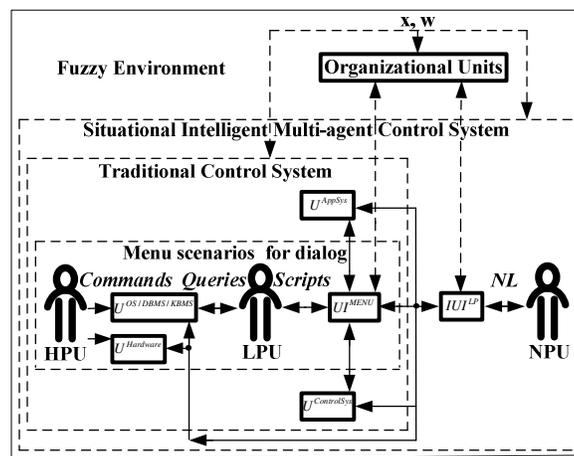


Figure 2. Human Interaction in Situational, Intelligent Multi-agent and Traditional Control Systems.

The paper proposes the *creation technology* of the *IMAFCS* that *integrates* the achievements of Situational Control, Fuzzy Logic, Linguistics, Artificial Intelligence and others for *realization* of the following *functionalities*:

- *To use of models* of representation of *linguistic* and *subject area* data, information and knowledge in *Fuzzy Environment*.
- *Situational fuzzy* Modelling, Decision-making, Control and Planning in conditions of the absence, incompleteness, vagueness and ambiguity of knowledge.
- *Fuzzy Logic Inference*.
- *Generalization* and *Explanation* of knowledge.
- *Fuzzy Dialog Control*.
- *Linguistic Processor* and *Multi-lingual* interaction in *Fuzzy Environment*.

The main features of the proposed technology of design *IMAFCS* lies in the use of methods:

- Situational control of data, information and knowledge and implementation by it of automatic processes of inference, making the right decisions, generalization and explanation knowledge, dialog control, planning and management of organizational unit.

- Creation *LP*, which is a part of the Intelligent Interface and enabling people to interact with him in Natural Language *without* "menu" dialog.

The *group of highly professional users* (HPU) (Figure 2) includes users, which were trained in the use and maintenance of hardware, software, networks, data, information and knowledge bases.

A *group of less professional users* (LPU) in (the mentioned above resources) are developers: system analysts, application programmers, testers, operators and others.

The *group of not professional users* (NPU) is composed of experts in their field, who use the *functionality* of information system to solving their *functional* tasks. This group includes the *decision makers, analysts, experts, consultants, managers* and other experts in their subject area.

UI^{MENU} is the *user interface* of dialogue menu in Traditional Information System, with whom interacts user LPU using *script dialogue language*. This user also interacts with interfaces $U^{OS/DBMS/KBMS}$, using the query language.

$U^{OS/DBMS/KBMS}$ are *group of interfaces*, used by users HPU for system support the Operation Systems (OS), Data Base Management Systems (DBMS) and Knowledge Base Management Systems using *commands languages*

U^{AppSys} , $U^{ControlSys}$ are respectively, *Application* and *Control* systems, which realize the *functionalities* of the considered SIIS and other systems in it.

UI^{LP} is *intellectual interface* with built in LP and with whom interacts the user NPU using *natural language*. This *interface* is connected with interfaces $U^{OS/DBMS/KBMS}$, UI^{MENU} and with systems U^{AppSys} , $U^{ControlSys}$ for the use of the existing functionalities of other information systems. Thus, *IMAFCS* integrates the *functionality* of existing *traditional control systems* to solve *simple* problems and its *intellectual capabilities* to solve *complex* problems in *fuzzy environment*.

The $U^{Hardware}$ is *hardware resources*, supported by user HPU.

The interface IUI^{LP} interacts with the organizational entities, performing all functionality of ICS by returning to user NPU the results in the required form. Similar work is done by the user LPU, which interacts with the *organizational units* through the interface UI^{MENU} .

Intellectualization of labor of the LPU and HPU users *is not* considered in this paper and will be the subject of further study. These studies involve the introduction of artificial intelligence in hardware, software and networks, and the inclusion of these groups in the NPU group.

2.2. Methods

2.2.1. The models and methods of representation of linguistic and subject data, information and knowledge in Fuzzy Environment.

In [13], [10] were extended the concept of a *linguistic variable*, formal and *semiotic models* with using the principle and method of *situational control* [6] by taking into account of the accepted

methods of representation, organization, integration, processing and synthesis of data, information and knowledge [3-5], [7,8], [14]. Through the use of *fuzzy sets* theory and *situational control* model were defined *linguistic* and *thematic* units, attributes, corteges and dictionary entries in linguistic and thematic relational bases of data, information and knowledge.

These *models* define a conceptual means of presenting and structuring of data, information, knowledge in fuzzy environment, and are also used for modeling, planning, decision-making and control in ICS.

Thus, the *intellectuality* of the *Data, Information and Knowledge Control System* consists in providing of interaction of decision-maker with consultants and experts (last among themselves) in order to organize *dialogue* between them in a *natural language*.

2.2.3. Situational Fuzzy Modelling, Decision-making, Control and Planning in conditions of the absence, incompleteness, vagueness and ambiguity of knowledge.

In order to control an organizational unit it is required to know its structure, the purpose of its existence and its control criteria [6].

The task becomes more complicated when there is a need to control organizational units in real time, in situations unexpected in advance, using variety of natural languages and subject areas. In these circumstances, arises a problem of decision making in fuzzy environment [2] based on the data, information and knowledge.

The solution to this problem implemented by a) methods of modelling, planning and controlling of linguistic and subject area data, information, knowledge, fuzzy inference and others, b) mapping the objectives and constraints in fuzzy environment [16,13].

2.2.4. Fuzzy Inference

Given the complex character of functioning of the *ICS*, its design is impossible without the use of theories situational control [6], fuzzy sets [1] and the proposed above of models of representation, synthesis, modelling, planning and management of data, information and knowledge.

Therefore, the modeling method of fuzzy inference can be applied to data control system, satisfying the following principles [13], [10]:

- All information about the data, information and knowledge (about the organizational unit) may be communicated to the control system as a set of phrases of Natural Language.
- Control model is fundamentally should be open and never ends the creation of the final formal model.
- Description of the data management (information, knowledge) process is possible in the form of natural phrases and \ or another language.

In these circumstances, the proposed modeling method of fuzzy inference is implemented by a system of situational data control and displays *alinguistic approach* to the problem. The method allows realizing the inductive and deductive inference in natural language in integrated subject areas, based on incoming fuzzy fragments (parcels) of the language.

To do this, we used the heuristic *algorithms, methods* of wisdom, intuition, behaviour and other algorithms and methods that invoke the modules of modelling of data, information and knowledge. The algorithms and methods uses generalized linguistic variables, fuzzy sets, rules

and facts (situations), previous decisions and their subsets (segments), extracted from bases of data, information and knowledge to obtain relevant decision.

The detailed presentation of the fuzzy logic inference is given in [13].

2.2.5. Generalization and Explanation of knowledge

The task of *generalization* of knowledge is reduced to finding the *target (unique)* situation Q_i of data, information and knowledge by using of their *current* situation Q_j and process of *control* them on base of *model* [15].

The *decision*, correspond to the found situation of the data, information and knowledge, shifts them from the current situation Q_j into a new Q_i .

This *decision (action)* determines the *impact rules* I on data (information and knowledge), which must be met in the overall situation S_i , so that they and control system would correspond to the new (changed) situation Q_i .

The *target function* in the model of generalization and explanation of knowledge defines the purpose of control of data, information and knowledge. The purpose may be the stirring up of processes of modelling, decision-making, planning, control, generalization and explanation knowledge and other.

2.2.6. Dialog Control

In this paper, we propose an approach to the control of the dialogue [11] in ICS using the modelling and control data, information and knowledge [10], fuzzy inference [13], generalization and explanation knowledge [15] and others.

Together with the second control systems, Dialogue is controlled by *planning* system, which uses the model, created by interpreter of dialogue U^D (Figure 3) and sub systems of fuzzy inference U^I , generation and explanation of knowledge U^G . Those sub systems provide processing and forming the input and output messages of natural language in knowledge management system using *bases* of *subject area* and *linguistic* data, information and knowledge.

The IUI^L is a *Intelligent User Interface (IUI)* with the inclusion of LP. Together with the Manager the IUI^L , U^D , U^I , U^G are representing the Decision-making system. The U^A (analyst, reviewer) and U^E (expert, approver) are intelligent subsystems, which are support the decision-making process.

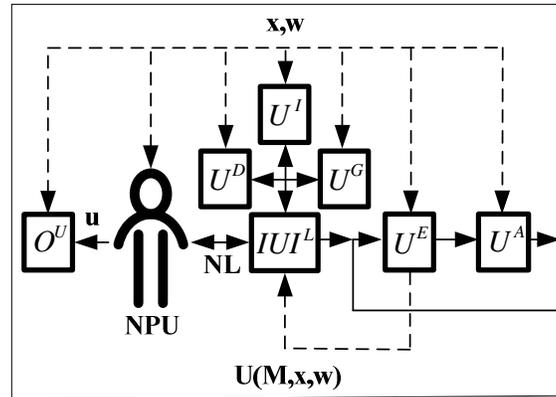


Figure 3. Dialog Control in human interaction.

2.2.7. Linguistic Processor and Multi-lingual interaction in Fuzzy Environment

The users interact with IUI using natural language, which is processed by LP. It is represented by the Interpreter (D) and Synthesizer (R) (Figure 2).

Its implementation involves the *functionality* [3, 5] of:

- Providing meaningful *machine translation* for identification concepts.
- The *adequacy* of the mapping meanings of the concepts expressed by the termin in a particular language in the context of a particular subject area.
- Opportunity to endow a specific term by attributes of *grammar, logic, semantics, pragmatics, psychology* and others in accordance with its meaning in a particular subject area and the context of use.
- Implementation of the "*understanding*" by means the algorithms of synthesis output expressions of natural language based on *logical-semantic* characteristics of intra-linguistic representations of meaning.
- Resolution *disambiguation* expressions used languages by sampling lexical and semantic characteristics and vocabulary of languages and conduct on the basis of their lexical and semantic analysis to determine these matching of input and output language equivalents.
- The possibility of implementing a system of automatic machine dictionaries and thesauri in linguistic databases with conceptual connection to the subject databases and knowledge.

Implementation of Machine Translation terminological phrases includes the blocks of:

- Logical-semantic, grammatical analysis of input combinations and the identification of concepts of its output equivalents.
- Prior authorization of lexical ambiguity on input phrases using conceptual codes.
- Extraction of grammatical information.
- Grammatical ambiguity resolution based on the stems of input combinations.
- Extracting logical information.
- Final resolution of lexical ambiguity.
- Resolution of lexical and grammatical ambiguity of input and output combinations of stems.
- Logical and semantic disambiguation of input and output terminological expressions.
- Identification of concepts and the formation of natural language expressions of output terminological phrases.

The LP realizes transformation target's actions, which are expressed in NL and other languages. The method of realization of the LP is represented in [3].

3. CONCLUSIONS

The proposed methods and technology are oriented for design and development of *autonomous Smart/ Intelligent Multi-agent Fuzzy Control System*, which will be operated in fuzzy environment, interacting with people and other systems and agents in different languages and dissimilar subject areas, where the situations and factors of influence on the control unit cannot be determined and structured in advance.

A distinctive feature of this approach is to target the modelling and control of fuzzy linguistic and subject data, information, knowledge, alternative solutions, objectives and constraints in order to find accurate, relevant and right decisions, which are suitable to the situation, with regard to external and internal influences of the fuzzy environments on the system.

The results of this work focused on the creation of *autonomous, situational, intelligent, multi-agent information control* systems of robots, unmanned production and of Apparatus, functioning in the fuzzy environment and unforeseen situations in advance.

REFERENCES

- [1] L. A. Zadeh, (1970) "Fuzzy Sets", Information and Control, 8, New York, pp. 338-359.
- [2] R. E. Belman & L. A. Zadeh, (1970) "Decision-making in fuzzy environment", Management Sciences, California, vol. 14, pp. 141-164.
- [3] B. Z. Khayut, G. I. Nikolaev, A. N. Popeskul, V. A. Chigakovsky, (1983) "Realization Linguistic Data Base for Machine translation foreign languages texts using DBMS INES", Collection of papers, International Machine Translation Symposium, Moscow, pp. 236-237.
- [4] B. Z. Khayut & G. I. Nikolaev, (1984) "Data organization for interactive systems "understanding" of natural language and multilingual interaction", Workshop, Interactive systems and their practical application, Chisinau, pp. 165-166.
- [5] G. I. Nikolaev & B. Z. Khayut, (1985) "Logical-semantic method of identifying noun phrases of natural language", Workshop, Logic and combinatorial methods in artificial intelligence and pattern recognition, Chisinau, pp. 79-80.
- [6] D. A. Pospelov, (1986) "Situational Control", Science, Moscow, 288 pages.
- [7] B. Z. Khayut, (1986) "An approach to the integration of heterogeneous databases in design of intelligent access", Workshop, Interactive systems and their practical application, Russia, Kalinin, pp. 164-165.
- [8] B. Z. Khayut, (1986) "The method of constructing an integrated system of intelligent access and data control", Workshop, Methods for building integrated software systems of data control, Russia, Kalinin, pp. 179-180.
- [9] G. I. Nikolaev, Y. N. Pechersky, B. Z. Khayut, (1986) "The conceptual approach to the presentation and synthesis of knowledge about natural language and subject area for the "man-intelligent robots", Workshop, Automation and robotizing of manufacture with the use of microprocessor-based products, Chisinau, pp. 67-68.
- [10] B. Z. Khayut & Y. N. Pechersky, (1987) "Situational Data Control", VINITI, Deposited manuscript, Moscow, 29 pages.
- [11] Y. N. Pechersky & B. Z. Khayut, (1988) "Approach to Dialog Control in the situational decision-making system", Workshop, Interactive tools in automated control systems, Chisinau, pp. 169-170.
- [12] K. M. Passino & S. Yurkovich, (1988) "Fuzzy Control", Ohio State University, Addison-Wesley, Menlo Park, California, pages 502.
- [13] B. Z. Khayut, (1989) "Modelling of fuzzy logic inference in the decision making system", Science, Issue 110, Academy of Sciences of Moldova, Chisinau, pp. 134-144.

- [14] B. Z. Khayut & Y. N. Pechersky, (1989) "Knowledge modelling in decision support system" VINITI, Deposited manuscript, Moscow, 12 pages.
- [15] Y. N. Pechersky & B. Z. Khayut, (1989) "Generalization of knowledge in decision support system", Workshop, Questions of development computer facilities, Chisinau, pp. 107–108.
- [16] B. Z. Khayut, (2000) "The method of modelling and decision-making using fuzzy data", fourth collection of scientific papers, Creative searches of scientific Israel today, Ashkelon, pp. 130–133.
- [17] F. Herrera & E. Herrera-Viedma, (2000) "Linguistic decision analysis: steps for solving decision problems under linguistic information", Elsevier Science, Fuzzy sets and systems, vol. 115, Issue 1, pp. 67-82.
- [18] J. W. Krupansky, (2005) "Advancing the Science of Software Agent Technology", *Agivity*.
- [19] N. Cote, A. Canu, M. Boused, A. Mouaddib, (2012) "Humans-Robots Sliding Collaboration Control in Complex Environments with Adjustable Autonomy", Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conferences, vol. 2, pp. 146-153.

AUTHORS

Ben Khayut graduated MSc in *Mathematics* at the University of Chisinau and PhD in *Mathematical Cybernetics* (completed studies) at the Institute of Mathematics of the Academy of Sciences of the Moldova in the field of *applied mathematics, computer science, artificial intelligence, fuzzy logic*, where he defended the pre-thesis "*Data Control in decision-making on natural language*". Has experience (over 30 years) in *design and development of Information Control systems* in industry using *linguistic* and *subject area* data, information and knowledge bases, and *algorithms* of machine translation as researcher, designer, programmer, project manager, using modern scientific methods, software and hardware. Have 25 scientific papers. In 2007 he founded the *Intelligence Decisions Technologies Systems* (IDTS).



Lina Fabri graduated *Bachelor, Economics and Management* at the College of Business Management. Have 14 years of successfully leading and executing *Business Intelligence and Technology* strategies and solutions to drive business value. Proven ability to *define* and *drive technology* solutions and services with solid track record of high performing standards and excellence. Has good *scientific* ability and *experience* in *R&D* of Information Control Knowledge based Systems, Planning and Decision-making using business strategy methods in various subject areas using methods of *Situational Control, Linguistics* and *Artificial Intelligence*.



Maya Abukhana graduated *Information systems and Management* at the College of Business Management. Have 9 Years of successfully leading and executing Business Projects of Financial and Trading systems to drive business value. Proven ability to *define* and *drive technology* solutions and services with solid track record of high performing standards and excellence. Has a good scientific ability and experience in R&D in the design of knowledge bases, architecture and algorithms for complex financial control systems, as well methods of human interaction using Linguistics and Database technologies.



INTELLIGENT AND PERVASIVE ARCHIVING FRAMEWORK TO ENHANCE THE USABILITY OF THE ZERO-CLIENT- BASED CLOUD STORAGE SYSTEM

Keedong Yoo

Department of Management Information Systems, Dankook University,
Cheonan, Republic of Korea
kdyoo@dankook.ac.kr

ABSTRACT

The cloud storage-based zero client technology gains companies' interest because of its capabilities in secured and economic management of information resources. As the use of personal smart devices such as smart phones and pads in business increases, to cope with insufficient workability caused by limited size and computing capacity most of smart devices have, the necessity to apply the zero-client technology is being highlighted. However, from the viewpoint of usability, users point out a very serious problem in using cloud storage-based zero client system: difficulty in deciding and locating a proper directory to store documents. This paper proposes a method to enhance the usability of a zero-client-based cloud storage system by intelligently and pervasively archiving working documents according to automatically identified topic. Without user's direct definition of directory to store the document, the proposed ideas enable the documents to be automatically archived into the predefined directories. Based on the proposed ideas, more effective and efficient management of electronic documents can be achieved.

KEYWORDS

Intelligent archiving, Cloud storage, Zero-client, Automatic document summarization

1. INTRODUCTION

The zero-client technology, or the empty can-like PC technology, is an emerging ECM(enterprise content management) technology by integrating the VDI(virtual desktop infrastructure) into the cloud storage environment to securely manage and utilize companies' intellectual resources in a more efficient and pervasive manner. Comparing to the service through conventional thin-client technology, the zero-client technology can not only more securely manage documents by minimizing the amount of working documents stored in operator's personal workstation, but also more economically maintain computing systems by directly downloading required software patches and updates in a real time basis. As the needs to apply personal smart devices such as smart phones and pads widely used nowadays increase, the zero-client technology can play a very essential role in coping with insufficient workability caused by limited size and computing capacity most of smart devices have.

The cloud storage, a model of networked corporate storage where data is stored in virtualized pools of storage, provides the Internet-based data storage as a service. One of the biggest merits of cloud storage is that users can access data in a cloud anytime and anywhere, using any types of network-enabled user devices [1]. Amazon Web Services S3 (<http://aws.amazon.com/s3>), Mosso (<http://www.rackspacecloud.com>), Wuala (<http://www.wuala.com>), Google Drive (<http://drive.google.com>), Dropbox (<http://www.dropbox.com>), uCloud (<http://www.ucoud.com>), and nDrive (<http://ndrive.naver.com>) are typical examples of corporate and personal cloud storage services. All of these services offer users transparent and simplified storage interfaces, hiding the details of the actual location and management of resources [2]. Once a document is stored in the cloud storage, a user can access and download the document anytime and anywhere under the condition that designated access right has been granted. Because of the advantages in storing and extracting information resources, more companies are implementing the online storage under the cloud storage environment.

While the cloud storage can deliver users various benefits, it also has technical limits in network security as well as in privacy [3]. In addition, from the viewpoint of usability, many users also point out a very serious problem in using cloud storage-based zero client environment, which is the difficulty in storing and retrieving documents. Since the directories in the cloud storage has been defined and structured by companies' decision, most of users are not accustomed to them. To store a working document in the cloud storage, a user has to decide a proper directory that exactly coincides with the contents of the document. Since the directories are naturally varied and the overall structure of directories is complicated, deciding a proper directory is not an easy work: sometimes a user can go astray by the confusion in deciding and locating target directory. Also, when a user tries to retrieve a document, he/she may spend not a little time to locate the file because too many directories exist. Since the directories are not defined and provided by him/herself, relatively much time to make a user be accustomed. Therefore, any automated assistance in concluding the target directory is indispensably needed by analysing contents of the given document with respect to directories in the cloud storage. Since any keywords or topics extracted from the document stand for the possible title of the directory under which the document must be stored, a user can easily complete his/her job to store and retrieve documents. In retrieving a document from the storage, more accurate and fast searching can be made because each document has been archived into the topic-based directory.

This research tries to enhance the usability of a zero-client-based cloud storage system by intelligently and pervasively archiving working documents according to automatically identified topic. To do so, this research suggests not only a framework to automatically extract the predefined directory-specific topic of a working document by applying an automatic document summarization technique, but also required sample codes to pervasively archive documents under the automatically determined directory.

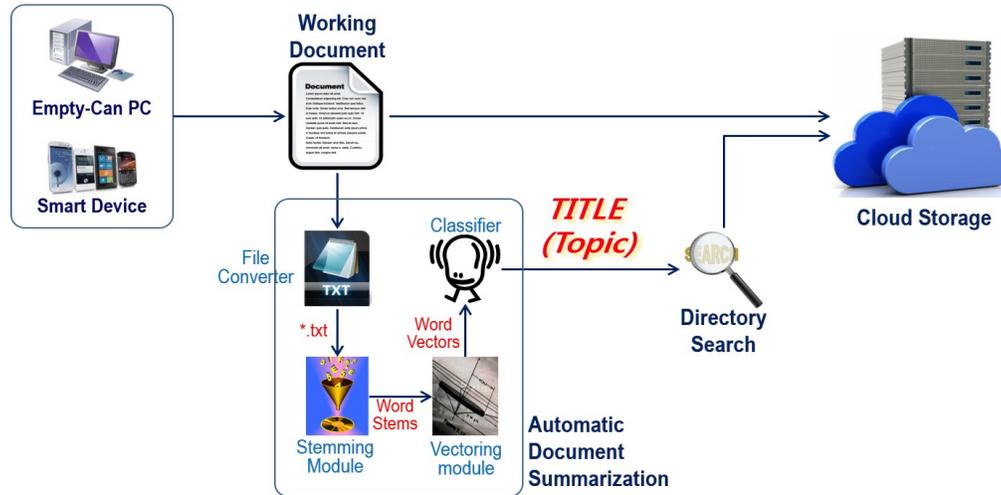


Figure 1. Framework for intelligent and pervasive archiving

2. FRAMEWORK FOR INTELLIGENT AND PERVASIVE ARCHIVING

Figure 1 shows the suggested framework for intelligent and pervasive archiving. Since the cloud storage plays the role of VDI-enabled database, pervasiveness in document storing and retrieving can be guaranteed. Intelligent archiving can be attained by two functionalities: one is automatic document summarization to automatically extract a title of the given document, and the other is automatic directory search to locate given document onto predefined directory according to the extracted title.

Once a working document is created by users using their empty-can PC or smart devices, it must be archived in the cloud storage-based corporate repository because users' terminals are not equipped with internal storages. To archive the working document intelligently, the title of the document must be automatically determined by analysing the words included in the document, and corresponding directory in the cloud storage must be automatically concluded also according to the topic or title of the document. Therefore, as a user finishes creating the document and tries saving it, the module for automatic document summarization initiates its function to extract the title of the working document according to the procedures as follows;

2.1. File format converting

The file format of the working document can be varied with the types of software used in creating the document. To guarantee the efficiency of analysis to extract the title of a given document, the file format must be normalized (or standardized) into analysable one in advance to the rest of procedures involve [4]. In this research, the file formats are designed to be converted into the '.txt' format to promote the readability of following modules.

2.2. Stemming

Once the document formats have been normalized, words in the document must be also normalized so that only stems of each word can be considered by separating inflectional and derivational morphemes from the root, the basic form of a word. For example, the root of the English verb form 'preprocessing' is 'process-'; the stem is 'pre-process-ing', which includes the

derivational affixes ‘pre-’ and ‘-ing’, but not the present progressive suffix ‘-ing’. After stemming each word, non-necessary stems must be eliminated to promote the efficiency of analysis by setting lists of stop words which need to be filtered out ahead to further analysis.

2.3. Vectorizing

Based on the word stems from the phase of stemming, each stem must be vectorized to extract a document vector. In many cases, the TF/IDF (Term Frequency/Inverse Term Frequency) is usually used, and this study also apply it. TF/IDF is a statistical technique to assess relative importance of a word in a document. A high weight in TF/IDF is obtained by a high term frequency and a low document frequency of the term in the whole collection of documents; the weight therefore tends to filter out common terms. The word with the highest TF/IDF is deemed as the topic of a document.

2.4. Classifying

Resultant topic of a document can be identified by plotting the document vector onto a given vector spaces prepared by predefined category-based sample data. Therefore, to promote the accuracy of classification, the quality of sample data is very crucial, and therefore a corpus which is a collection of predefined categories with sufficient number of example documents must be formally examined. In conventional text mining area, a classifier is based on various algorithms such as SVM (Support Vector Machine), Naïve Bayes, and k-NN(Nearest Neighbors), etc. In this research, a SVM-based classifier is implemented as an example because SVM was reported to outperform other algorithms [5, 6]. The accuracy of SVM-based classification was also verified as satisfactory as up to 90% if the prediction model was sufficiently trained using a formal corpus like Reuter-21578 [7].

The identified topic of a document, then, must be migrated to the directory searching module to conclude the possible directory under which the document archives. The title of the document needs to be formulated by combining the topic, the document creator’s ID, and the time of archiving so that the document can be uniquely identified.

3. TOPIC IDENTIFICATION BASED ON AUTOMATIC DOCUMENT SUMMARIZATION

Automatic summarization is the process for making reduced version including the most important points of a given document using the functionality of computer programs. Making summaries automatically is an indispensable work as the amounts of information and documents increase. The Summly, an iPhone-based automatic summarization application developed by Nick D’Aloisio (<http://summly.com/>) and acquired by Yahoo.com is a typical example proves the importance of automatic summarization techniques nowadays. There exist two approaches to automatic summarization: extraction and abstraction. Extractive methods select a subset of existing words, phrases, or sentences in the original text to make a summary. Abstractive methods build an internal semantic representation and then use natural language generation techniques to make a summary that is closer to what a human might generate. Abstractive methods can give a liberal translation and therefore perform more comprehensive and realistic summarization. However, because of burdens in implementing and training a prediction model used in concluding keywords or keyphrases by projecting word vectors onto the n-dimensional corpus-based space, extractive methods are more widely used rather than abstractive methods.

Conventional approaches of extractive methods usually require training the prediction algorithm, or a classifier, using predefined category-based sample data, and this type of learning procedure is called as the supervised learning. In supervised learning, each set of sample data is composed of a pair of a document (or a word) and its associated category in the form of a vector. By reading sample data, a prediction algorithm can form a vector space constructed by given categories, and therefore can put the vector of a given document (or word) onto corresponding location within the vector space. While the supervised methods can produce reliable outputs based on pre-validated data, they have limitations in application caused by the large amount of training data as well as by the quality of data sets. Usually a number of documents with identified keywords or keyphrases are required to train a classifier, and therefore burdens in time and computing capacity are indispensably exhibited. Moreover, wrong results can be outputted in case of data with biased subject being inputted. Therefore, to meet with these limitations, unsupervised methods, such as TextRank [8] and LexRank [9], which eliminate the process of training using sample data are gaining much interest. The TextRank algorithm exploits the structure of the text itself to determine keyphrases that appear 'central' to the text in the same way that PageRank [10] selects important Web pages. Because the TextRank enables the application of graph-based ranking algorithms to natural language texts, it produces results independent to the training data and language types.

4. PROTOTYPE DESIGN

The prototype system is designed to initiate the function of topic extraction simultaneously with the user's trial to save the working document. Indexing the document by tagging the identified topic with user's ID and time, the prototype transmits and stores the document into the cloud storage. A dialogue between a user and the prototype is also needed to check whether the resultant topic is proper or not. If the user confirms that topic has no problem, the prototype transmits the file to the cloud storage with tagging required information about the user's ID and the time of archiving: Automatic archiving can be completed. Figure 2 shows the sequence of functions the prototype has.

The Stemming and Vectorizing module are implemented by using 'Word stemming tool' and 'Vector creating tool' of 'Yale', an open source environment for KDD(Knowledge Discovery and Data mining) and machine learning [11], respectively. As announced previously, this research deploys the SVM algorithm as the classification method. Therefore, using the LibSVM [12], a classifier is implemented. Following codes show the procedures to convert file format and to make the classifier read the file.

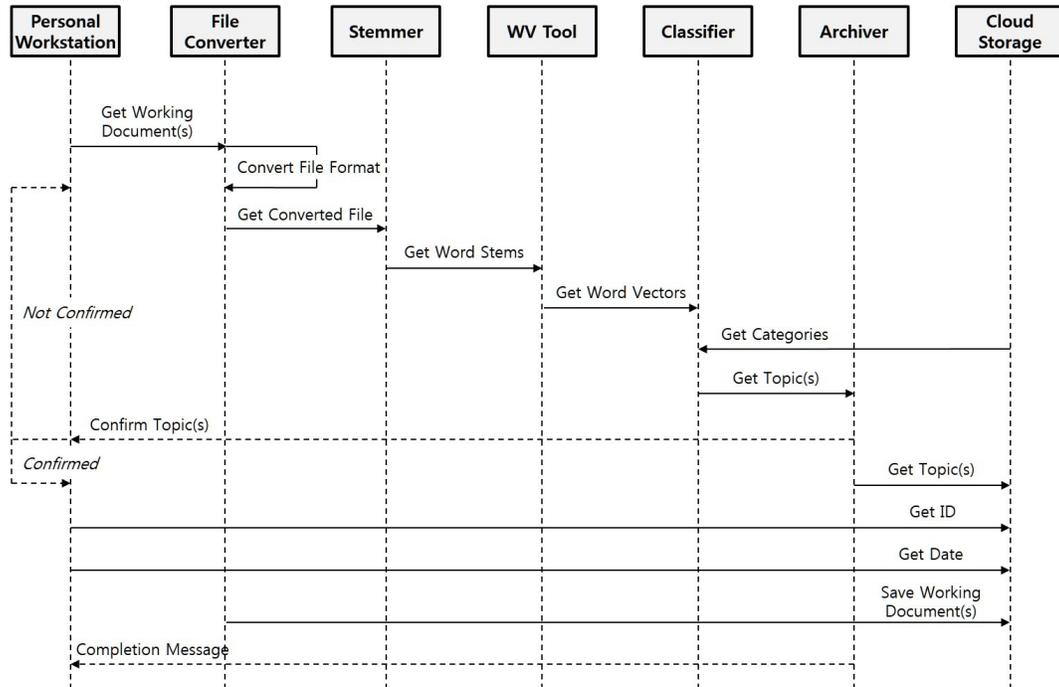


Figure 2. Execution sequence of the prototype system

```

if(predict_probability == 1) {
    if(svm_type == svm_parameter.EPSILON_SVR || svm_type ==
svm_parameter.NU_SVR) {
        System.out.print("Prob. model for test data: target value = predicted
value + z,\nz:      Laplace distribution e^(-
|z|/sigma)/(2sigma), sigma="+svm.svm_get_svr_probability(model)+"\n");
    }
    else {svm.svm_get_labels(model, labels);
prob_estimates = new double[nr_class];
output.writeBytes("labels");
for(int j=0;j<nr_class;j++)
output.writeBytes(" "+labels[j]);
output.writeBytes("\n");
}
}
while(true) {
    String line = input.readLine();
    if(line == null) break;
    StringTokenizer st = new StringTokenizer(line, " \t\n\r\f:");
    double target = atof(st.nextToken());
    int m = st.countTokens()/2;
    svm_node[] x = new svm_node[m];
    for(int j=0;j<m;j++) {
        x[j] = new svm_node();
        x[j].index = atoi(st.nextToken());
        x[j].value = atof(st.nextToken());
    }
}
  
```

Identified topic needs to be combined with creator's (user's) ID and the time to archive as the following codes show.

```
SimpleDateFormat dateFormat = new SimpleDateFormat("yyyyMMdd", Locale.KOREA);
String recordedDate = dateFormat.format(new Date());
String tagName = topicName + "-" + userID + "-" + recordedDate;
System.out.printf("Indexing tag is %s\n", tagName);
JOptionPane.showMessageDialog(null, tagName);
```

Under the condition that the cloud storage has i directories ($i=1,2,\dots,n$), the document must be archived in one of existing directories. To search proper directory coincide with the identified topic, the 'Hash' function can be used, and corresponding programming codes can be as follows;

```
HashMap<String, Integer> categoryHash = new HashMap<String, Integer>();
categoryHash.put("Topic1", 0);
categoryHash.put("Topic2", 1);
categoryHash.put("Topic3", 2);
.
.
.
categoryHash.put("Topici", i-1);

int indexOfCategory = categoryHash.get(topicName);
System.out.printf("Searching result is %d index\n", indexOfCategory);
JOptionPane.showMessageDialog(null, indexOfCategory);
```

If the searching result is correct and the user confirm it, then a message of processing archiving needs to be sent to the user, as the following codes show;

```
try {
    BufferedWriter tagFile = new BufferedWriter(new FileWriter(filePath));
    tagFile.write(tagName);
    tagFile.close();
} catch (IOException e) {
    System.err.println(e);
    System.exit(1);
}
JOptionPane.showMessageDialog(null, "The document is to be saved as '" +
filePath + "'");
```

Finally the document is to be archived in the concluded directory with the title of 'topic-ID-date', and a message informing the completion of archiving is to be notified to the user with displaying the title and location of archiving, as following codes show;

```
String msg = "The topic of working document is '" + topicName + "' ?";
int ret = JOptionPane.showOptionDialog(null, msg, "Message Window",
JOptionPane.YES_NO_OPTION, JOptionPane.PLAIN_MESSAGE, null, null, null);
switch (ret) {
case JOptionPane.YES_OPTION:
    JOptionPane.showMessageDialog(null, "The file '" + tagName + "' has been
archived.");
    break;

case JOptionPane.NO_OPTION:
    JOptionPane.showMessageDialog(null, "user canceled");
    break;
}
```

5. CONCLUSIONS

Zero-client-based cloud storage is gaining much interest as a tool for centralized management of organizational documents. Besides the well-known cloud storage's defects such as security and

privacy protection, users of the zero-client-based cloud storage point out the difficulty in browsing and selecting the storage directory because of its diversity and complexity. To resolve this problem, this study proposes a method of intelligent document archiving by applying an automatic summarization-based topic identification technique. Since the cloud storage plays the role of VDI-enabled database, pervasive document storing and retrieving can be naturally enabled. Although not a few researches also tried to enhance the functionality of corporate archiving systems, no research has suggested the intelligent archiving by automatically attaching the title of documents to leverage the usability of zero-client-based cloud storage, which is the main contribution of this study.

Issues in this paper remain points to discuss concerning technical limitations and future works. Especially, discussions around the algorithms for automatic document summarization need to be addressed, because the application efficiency of SVM is doubted because of the burden in training the prediction model. Training the prediction model via server-side computing might be a solution for this problem, however the computing load a server must endure can also keep increasing as the use of smart devices increase. Therefore, approaches of unsupervised methods can yield very effective solutions to meet this problem. However, more formal and statistical validation on the performance of the unsupervised methods is required to acquire the reputation supervised methods have gained without the smallest strain. Meanwhile, a formal corpus must be developed to guarantee the performance of conventional text mining techniques, because most of conventional algorithms in the area of text mining are much dependent upon the quality of corpus. More formal, general and universal corpus must be developed so that the results from applying the corpus can be unbiased and objective. Since the corpus can be applied in setting the directories of cloud storage, this supplement can also make up for the applicability of intelligent document archiving suggested by this study.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A5A2A01017530).

REFERENCES

- [1] Liu, Q., Wang, G., & Wu, J., (2012) "Secure and privacy preserving keyword searching for cloud storage services", *Journal of Network and Computer Applications*, Vol.35, No.3, 927-933.
- [2] Pamies-Juarez, L., García-López, P., Sánchez-Artigas, M., & Herrera, B., (2011) "Towards the design of optimal data redundancy schemes for heterogeneous cloud storage infrastructures", *Computer Networks*, Vol.55, 1100-1113.
- [3] Svantesson, D. & Clarke, R., (2010) "Privacy and consumer risks in cloud computing", *Computer Law & Security Review*, Vol.26, 391-397.
- [4] Kim, S., Suh, E., & Yoo, K., (2007) "A study of context inference for Web-based information systems", *Electronic Commerce Research and Applications*, Vol.6, 146-158.
- [5] Basu, A., Watters, C., & Shepherd, M., (2003) "Support Vector Machines for Text Categorization", *Proceedings of the 36th Hawaii International Conference on System Sciences*, Vol.4.
- [6] Meyer, D., Leisch, F., & Hornik, K., (2003) "The support vector machine under test", *Neurocomputing*, Vol.55, 169-186.
- [7] Hsu, C.W., Chang, C.C., & Lin, C.J., (2001) "A Practical Guide to Support Vector Classification: LibSVM Tutorial". In <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- [8] Mihalcea, R. & Tarau, P., (2004) "TextRank: Bringing order into texts", *Proceedings of EMNLP*, Vol.4, No.4.
- [9] Erkan, G. & Radev, D.R., (2004) "LexRank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research*, Vol.22, No.1, 457-479.
- [10] Brin, S. & Page, L., (1998) "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, Vol.30, 1-7.

- [11] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T., (2006) "YALE: Rapid Prototyping for Complex Data Mining Tasks", Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06).
- [12] Chang, C. & Lin, C., (2011) "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, Vol.2, No.3, 1-27.

AUTHOR

Keedong Yoo is an associate professor in the Department of MIS at Dankook University, South Korea (kdyoo@dankook.ac.kr). He has B.S. and M.S. in Industrial Engineering from the POSTECH (Pohang University of Science and Technology), South Korea; and a Ph.D. in Management and Industrial Engineering from the POSTECH. His research interests include knowledge management and service; intelligent and autonomous systems; context-aware and pervasive computing-based knowledge systems.



INTENTIONAL BLANK

EFFICIENTLY PROCESSING OF TOP-K TYPICALITY QUERY FOR STRUCTURED DATA

Jaehui Park¹ and Sang-goo Lee²

¹Electronics and Telecommunications Research Institute, Daejeon, Korea

jaehui@etri.re.kr

²School of Computer Science and Engineering, Seoul National University

sglee@snu.ac.kr

ABSTRACT

This work presents a novel ranking scheme for structured data. We show how to apply the notion of typicality analysis from cognitive science and how to use this notion to formulate the problem of ranking data with categorical attributes. First, we formalize the typicality query model for relational databases. We adopt Pearson correlation coefficient to quantify the extent of the typicality of an object. The correlation coefficient estimates the extent of statistical relationships between two variables based on the patterns of occurrences and absences of their values. Second, we develop a top-k query processing method for efficient computation. TPFilter prunes unpromising objects based on tight upper bounds and selectively joins tuples of highest typicality score. Our methods efficiently prune unpromising objects based on upper bounds. Experimental results show our approach is promising for real data.

KEYWORDS

Typicality, Top-k query processing, Correlation, Lazy join, Upper bound

1. INTRODUCTION

Analyzing typical characteristics of objects is an effective method to understand the semantics of the objects in real-world data sets. Traditional studies in cognitive science [1, 2] have noted that a measure of typicality generally improves people’s judgment, whether some objects to be “better examples” for a given concept (or a category). For example, consider a user who wants to learn a concept, mammals, using a zoology data set. Based on typicality analysis, lions may be more useful example than whales because lions have typical attributes of mammals, such as quadruped (four legs). Finding typical instance is a useful application for reflecting semantics of whole data set by only using a limited set of objects. Therefore, lions and bears are better examples than whales and platypuses when we introduce a conceptual knowledge of mammals to children. Following general understandings in cognitive science, we adopt intuitions from typicality analysis to information retrieval tasks, especially, rankings. In this paper, we focus on a ranking model for objects with categorical attributes in a large database using the concept of typicality. Moreover, several processing techniques are proposed to improve the efficiency of retrieval in large scale data sets.

More precisely, we first investigate the problem of applying the notion of typicality analysis into ranking of database query results. Motivated by [3], we propose a novel model, typicality query model, for relational databases. From the definition [3], a typical object shares many attribute values with other objects of the same category, and few attribute values with objects of other categories. Given a query, which determines a specific category, computing common attribute values of objects is crucial for typicality query. In this paper, statistical relationships based on correlation analysis [4, 5] are adopted to specify the amount of the common attribute values for queries. Furthermore, the correlation analysis naturally provides for quantification of common attribute values of objects in not only a set of a single category but also multiple categories. However, constructing comprehensive dependency model for every correlation yields unreasonably high computational costs. Therefore, we develop the typicality query model by introducing limited independence assumption on attribute values for efficient computation. Previous studies [6, 7] have proved that the assumption reduces a significant amount of computations without deteriorating the quality of rankings over structured data.

Secondly, we propose a method to find top-k typical objects efficiently. Despite the significance of the topic that users are more interested in the most important, that is, top-k query results is emphasized recently, little attention has been paid to aggregating scores of an individual object that are dependent (or, correlated) to each other. Previous studies, such as [3], have proposed approximation methods to provide fast answers for top-k typicality query. Despite existing studies have focused on approximation or new measures of association, our model mainly concerns efficient computation for top-k results without approximate solutions. Basically, we perform a prune-and-test method for a large number of objects 1) before aggregating exact scores by investigating an upper bound property of the correlation coefficient, and 2) by predicting unnecessary joins to avoid beforehand. We can check whether candidate objects have a potential to become top-k answers for a typicality query without computing their exact typicality scores. We further save a lot of join query processing cost to predict the typicality score by estimating the cardinality of tuples that directly matched to queries. Our methods significantly reduce unnecessary join processing time. To our knowledge, our work is first approach to compute top-k objects over relational databases on typicality measures, which are based on the correlation of individual objects.

We have conducted and performed performance study on a real data set. Extensive sets of evaluation tests are not provided in this paper because this work is still in progress. As a summary, our method, TPFilter yields average execution time that are much smaller than that of the competitive work [3] on zoology data sets.

The rest of the paper is organized as follows. In Section 2, we define the typicality query in relational databases and the typicality score based on descriptive statistics, namely correlation. Section 3, we introduce the top-k typicality query processing method, TPFilter. In Section 4, we show a brief set of evaluation results. Finally, we present concluding remarks and further study in Section 5.

2. QUERY MODEL

In this section, we formally define a typicality query model in relational databases. In Section 2.1, we introduce the notion of the typicality query. In Section 2.2, we develop a probabilistic ranking function based on a statistical model from classical statistics.

2.1. Typicality Query

We consider a set of relations $R = \{r_1, r_2, \dots, r_N\}$ and each relation r_i as a set of n tuples $\{t_{i1}, t_{i2}, \dots, t_{in}\}$. For simplicity, we use tuple t_j to represent t_{ij} when r_i is clear in the context. Given a keyword query $Q = \{k_1, k_2, \dots, k_q\}$, we would like to assign a ranking score $S(I)$ for an object I of a certain relational schema $H(R)$ defined on the relations R . The relational schema $H(R)$ contains referential relationships between relations. Figure 1(a) illustrate an example relational schema as a directed graph that has 7 vertices, corresponding to relations $R = \{r_1, \dots, r_7\}$. Directed edges represent the referential relationships between the relations. Colored vertices, r_4 and r_6 , represent relations that contain query keywords $Q = \{k_1, k_2\}$ in their tuples. We restrict our attention in this work to acyclic relational schema, which are common in database contexts. In our query model, the logical unit of the retrieval may be multiple tuples joined together based on primary key-foreign key relationships. In the example above, joining tuples of schema $H(R')$ (Figure 1(b)) represent a set of result given keyword query $Q = \{k_1, k_2\}$. It corresponds to join query expression that produce joining network of tuple set for the keyword query Q . We assign the ranking score $S(I)$ to each answer I , which is a joining network of tuple set. We define basic requirements for I as follows:

- 1) Every keyword in query Q is contained in at least one relation r_i in $H(R')$
- 2) Let t and t' be any two adjacent tuples, and assume that they are in relations r and r' , respectively. r and r' must be connected in the relational schema $H(R')$, and joining tuples, $t \bowtie t'$, must belong to $r \bowtie r'$.
- 3) No adjacent tuple can be removed if it fulfills the above requirements.

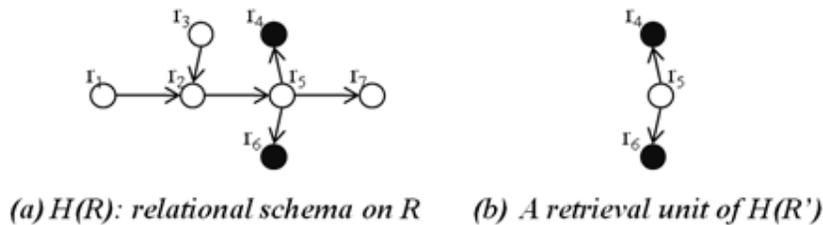


Figure 1. Directed graph of relational schema

From the requirement (2), $H(R')$ may contain the set of relations that do not include any keyword but connects others. We call tuple sets from those relations as *free tuple sets*. On the other hand, the set of tuples that satisfy requirement (1) is denoted as a *non-free tuple set*. Finding optimal answers satisfying above requirements in arbitrary queries is NP-hard problem. The focus of this paper is not on developing algorithms to efficiently compute near-optimal (or approximate) answers of relational schema $H(R')$. Rather, the objective of this paper is to introduce an effective ranking model in relational databases – that of computing a typicality measure $S(I)$ efficiently for top- k objects I . We assume that all possible $H(R')$ s for the query Q are generated.

Our typicality query model retrieves a list of objects ordered by their typicality scores. Now the typicality query is defined as follows:

Definition 1. (Typicality query) Given a keyword query $Q = \{k_1, k_2, \dots, k_q\}$ and a database $R = \{r_1, r_2, \dots, r_N\}$ with a schema $H(R)$, a *typicality query* is defined as following form.

```

SELECT *
FROM  $\forall r^K = \{r | \exists k_i \in t \wedge t \in r\}$  JOIN  $\exists r^F = \{r | r^K \leftarrow r \rightarrow r^K\} \in H(R')$ 
WHERE  $r^K.a = k_1$  AND ... AND  $r^K.a = k_q$ 
ORDER BY  $S(I \text{ of } H(R'))$ 

```

where the arrows denote the primary key-foreign key relationship, and I is an object of a relational schema $H(R')$, which produce the joining network of tuples in r^K and r^F . r^K corresponds non-free tuple sets, and r^F corresponds to free tuple sets. We call the score $S(I)$ as *typicality score* of an object I .

Proposition 1. (Typical instance) Given objects I enumerated from all possible relational schema $H(R')$ over $H(R)$, $Q = \{k_1, k_2, \dots, k_q\}$ and user specified threshold t , an instance whose score $S(I)$ is over the threshold t ($S(I) > t$) is denoted as a *typical instance*.

In a straightforward way, typicality query model process all the joins in every $H(R')$ for given queries, compute typicality score S , and then selects the most typical objects according to user specified threshold. With large databases, the total cost of query processing may be prohibitive. The computation method will be presented in Section 3.

2.2. Typicality Score

Assuming a keyword query $Q = \{k_1, k_2, \dots, k_q\}$ and relational schema $H(R')$ are given, we note that typicality query selects all objects $I = \{I_1, I_2, \dots, I_n\}$ having identical attributes $A = \{a_1, a_2, \dots, a_m\}$. We aim to assign a *typicality score* for each object I_i to order them by its occurrence distribution in database D ; it follows the general notion of typicality measure used in cognitive science. Based on the perception in [3], a typical object shares many attribute values with other objects of the same category, and few attribute values with objects of other categories. Intuitively, we can estimate the typicality score by counting common attribute values of objects given queries. Figure 2 illustrates a simple data set to compute typicality scores for eight objects, and objects $I_1 \sim I_4$ are in same category.

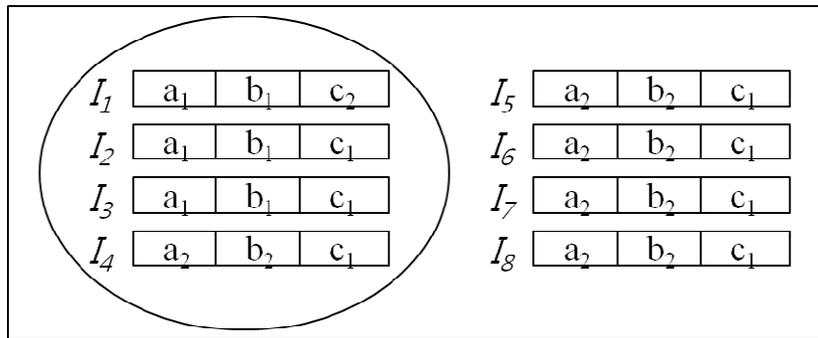


Figure 2. A single category selects four objects over a set of eight objects

Assuming the category is identified by given query Q , we can estimate each typicality score as the ratio of the number of common attribute values within given category to the number of attribute values shared with objects of other categories.

$$S(I_1) = \frac{2 \times 3}{1} = 6$$

$$S(I_2) = S(I_3) = \frac{3 \times 3}{1 \times 4} = 2.25$$

$$S(I_4) = \frac{1}{2 \times 4} = 0.125$$

Above scores are calculated by naively counting the number of occurrences to quantify the typicality of an object. The object I_1 is most typical because it shares two attribute values with the objects in the same category, but also no attribute is shared with objects in other categories. On the other hand, the objects I_2 and I_3 share an attribute value c_i with the objects in other categories. This is a simplified notion of typicality score. To define typicality score in a principled way, mutual implications on the occurrences or absences of attribute values $I.a_j$ with Q should be derived effectively. We note that the intuition is closely linked to the notion of correlation from classical descriptive statistics; correlation has been recognized as an interesting and useful type of patterns due to its ability to reveal the underlying occurrence dependency between data objects [9].

Any existing statistical measures [10] can be used to represent the extent of relationship (dependency) between elements. In this paper, we adopt *Pearson correlation coefficient* to model the interpretation from previous paragraph; but we remark that other measurements [10] can also be applied in a similar way. In our model, a binary random variable represents the absence and the presence of an attribute value given a query. In this context, the *Pearson correlation coefficient* for two random variables X and Y

(as $\frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2} \sqrt{E(Y^2) - (E(Y))^2}}$) is reduced to computational form as follows. We omit the proof due to limited space.

Given two binary random variables X and Y , the *Pearson correlation coefficient* ρ is:

$$\rho(X, Y) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}} \quad (1)$$

where n_{XY} , (for $X = 0, 1$ and $Y = 0, 1$), is the number of attribute value observations in a set of n objects, which are specified in Table 1.

Table 1. A two-way table of binary random variables X and Y

	Y=1	Y=0	Total
X=1	n_{11}	n_{10}	$n_{1.}$
X=0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Two binary random variables are considered positively associated if most of the observations fall along the right diagonal cells. In contrast, negative implication between variables is determined based on values in the left cells. Based on the correlation ρ , we can estimate the mutual implications of the occurrences of attribute values given a keyword query Q . We can specify the implication for each given query keyword $\rho(X, Y|k \in Q)$ as an aggregated score for an object I .

Definition 2. (Typicality score) Given a keyword query $Q = \{k_1, k_2, \dots, k_q\}$ and an object I with attributes $A = \{a_1, a_2, \dots, a_m\}$, a *typicality score* S of an object I is defined as following equation

$$S(I) = \sum_{j=1}^q \sum_{a_x, a_y \in A} \rho(I.a_x, I.a_y | k_j) \quad (2)$$

where $I.a_x$ and $I.a_y$ denote a pair of arbitrary attribute values of the object I . In order to estimate the typicality score of an object I , we aggregate every correlation ρ between pairs of attribute values $I.a_x$ and $I.a_y$ given query Q .

However, computing all combinations of attribute values is very expensive due to the complexity of relational databases with many attributes. In practice, it is necessary to define a practical assumption to avoid computing the correlation coefficients for an exponential number of attribute value combinations. We propose a limited independence assumption as in binary independence model as follows:

Definition 3. (Limited Independence Assumption) Given a keyword query $Q = \{k_1, k_2, \dots, k_q\}$ and an object I with attributes $A = \{a_1, a_2, \dots, a_m\}$, we assume dependence only between two specified sets of attribute values ($I.A_q$ and $I.A_{nq}$). The two sets of attributes are defined as follows: $A_q = \{a | I.a \cap Q \neq \emptyset\}$ and $A_{nq} = A - A_q$. The attribute values $I.a_i$ ($a_i \in A_q$) are assumed to be mutually independent. Analogously, the attribute values $I.a_j$ ($a_j \in A_{nq}$) are assumed to be mutually independent. We allow dependencies between $I.a_i$ ($a_i \in A_q$) and $I.a_j$ ($a_j \in A_{nq}$). Therefore, $\rho(I.a_x, I.a_y | k_j)$ is considered for typicality score.

Like most successful retrieval model (e.g., TF-IDF and BM25), our assumption between elementary values has empirically shown to be practical. Although our model defines the limited dependencies among values for our purpose, this assumption is patently significant for ranking relational data [6]. From our previous work [7], the assumption is validated to improve the retrieval performance. The assumption reduces the expression (Equation 2) to a following function, which is a simplified form:

$$S(I) = \sum_{j=1}^q \sum_{a_x \in A_q, a_y \in A_{nq}} \rho(I.a_x, I.a_y | k_j) \quad (3)$$

3. TOP-K PROCESSING OF TYPICALITY QUERY

In this section, we introduce a pruning method to efficiently remove the unpromising candidate objects before computing the actual typicality scores. By analyzing the mathematical properties of the correlation coefficients, we can derive upper bounds of typicality scores to test false positive candidates. Also, to compute top-k scores of objects, we don't need to join all the candidate tuples, but aggregate only the correlation values to calculate typicality scores. In this area, a number of top-k query processing techniques have already been proposed. However, top-k typicality query processing has crucial difference from the previous studies. Although most previous studies have focused on the ranking scores of individual objects with sorted access, our typicality score is quantified by its relationship with other objects. Therefore, classical algorithm cannot be adopted in a straightforward way. Moreover, our method is represented in a feasible form as compared to computational approaches in cognitive science.

In Section 3.1, we introduce a candidate pruning method, TPFilter, to efficiently prune the unpromising objects before computing the typicality scores. By analyzing the mathematical properties of the correlation coefficients, we derive upper bounds of typicality scores to test false positive candidate objects. In Section 3.2, we propose an efficient join query processing method, Lazy Join, to reduce the cost of join operations on multiple relations. To compute top-k scores of

objects, we don't need to join all the candidate tuples, but aggregate only the correlation values to calculate typicality.

3.1. TPFilter

Let $Pr(I_i.a_j)$ denote the ratio of the cardinality of the attribute value $I_i.a_j$ to the size of the database subset I of $(H(R'))$, which has same schema with object I_i . From section 2.2, we can transform Equation 1 by adopting observable variables $Pr(I_i.a_j)$ to Equation 3 if we specify the binary random variable X as $I_i.a_j$ (also, Y by $I_i.a_k$). For simple presentation, we use X and Y to represent $I_i.a_j$ and $I_i.a_k$, respectively and $A_q = \{I_i.a_j\}$. This is not an unusual constraint since we assume that keywords in Q are independent to each other.

$$\begin{aligned} \rho(X, Y) &= \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_1 \cdot n_0 \cdot n_0 \cdot n_1}} = \frac{(n - n_{10} - n_{01} - n_{11})n_{11} - n_{10}n_{01}}{\sqrt{n_1 \cdot n_0 \cdot n_0 \cdot n_1}} = \frac{nn_{11} - (n_{10} + n_{11})(n_{01} + n_{11})}{\sqrt{n_1 \cdot n_0 \cdot n_0 \cdot n_1}} \\ &= \frac{\frac{n_{11}}{n} - \frac{n_1 \cdot n_{11}}{n \cdot n}}{\sqrt{\frac{n_1 \cdot n_0 \cdot n_0 \cdot n_1}{n \cdot n \cdot n \cdot n}}} = \frac{Pr(X, Y) - Pr(X)Pr(Y)}{\sqrt{Pr(X)Pr(Y)(1 - Pr(X))(1 - Pr(Y))}} \end{aligned} \quad (4)$$

We propose an upper bound $\bar{\rho}(X, Y)$ for the bivariate correlation coefficient as a filter of unpromising objects.

Definition 4. (Typicality score upper bound $\bar{\mathcal{S}}(I)$) Given an object I ($I = \{I_i\}$) with attributes $A = \{a_1, a_2, \dots, a_m\}$, let $A_q = \{a_1, \dots, a_i\}$ for a keyword query Q . The upper bound of typicality score of object I is defined as follows:

$$\bar{\mathcal{S}}(I) = \sqrt{\frac{1 - Pr(I, A_q)}{Pr(I, A_q)}} \sum_{a_y \in A - A_q} \sqrt{\frac{Pr(I, a_y)}{1 - Pr(I, a_y)}} \quad (5)$$

Proof sketch. Without loss of generality, we assume $Pr(X) \leq Pr(Y)$. Then, following inequalities are to be true.

$$Pr(X, Y) \leq P(X) \leq P(Y). \quad (6)$$

$$\begin{aligned} \rho(X, Y) &= \frac{Pr(X, Y) - Pr(X)Pr(Y)}{\sqrt{Pr(X)Pr(Y)(1 - Pr(X))(1 - Pr(Y))}} \\ &\leq \frac{Pr(X) - Pr(X)Pr(Y)}{\sqrt{Pr(X)Pr(Y)(1 - Pr(X))(1 - Pr(Y))}} \\ &\leq \sqrt{\frac{Pr(Y)}{Pr(X)}} \sqrt{\frac{1 - Pr(X)}{1 - Pr(Y)}} = \bar{\rho}(X, Y) \end{aligned} \quad (7)$$

Therefore, for all attribute values $Y = I.a_i$ ($\in A$), we can aggregate each correlation upper bounds $\bar{\rho}(X, Y)$. Then we can derive typicality score upper bound $\bar{\mathcal{S}}(I)$ (Equation 4).

Basically, to calculate a typicality score of an object I_i , we have to compute the joint distribution $Pr(X, Y)$ of all attribute values in I_i . Computing all these pairs of attribute values in R' is too costly for online queries on large databases. The typicality score upper bound is determined only by the observable variables $Pr(X)$ and $Pr(Y)$ ($Y \in A_{nq}$). We note that calculating the upper bound is

much cheaper than the computation of the exact typicality score, since the upper bound can be easily computed as a function of cardinality of the joining tuples without considering the joint distributions, e.g., $Pr(X,Y)$. Storing every pairs of attribute values is inefficient for online processing. Note that the $\bar{S}(I)$ has monotone property, which is useful to filter lower scores at early stage. If both $Pr(X)$ and $Pr(X,Y)$ are fixed, then the correlation value of X and Y is monotonically decreasing with $Pr(Y)$. Therefore, we can maintain a queue of current top-k typical objects discovered so far, which is denoted as C . The objects in C are sorted in the descending order of their typicality scores. The typicality score of the k-th object in C is also denoted as $typicality_min$. For each newly candidate object I to be evaluated, its typicality score $S(I)$ should be at least $typicality_min$; otherwise, the object I is immediately removed from the set of candidates.

3.2. Lazy Join

Typicality query model must view all relations in a holistic manner in order to aggregate the tuples joined for a keyword query. While a complete evaluation of all the joins for queries is necessary for conventional selection query, we are interested in only top-k results. We propose an algorithm *LazyJoin* that perform joins without producing all the objects for relational schema $H(R')$.

We start by describing baseline method *Baseline* for top-k typicality query. *Baseline* issues a SQL expression equivalent to CN to retrieve result objects. Then, the objects from each CN are computed to derive typicality scores. We get the top-k typical objects with highest typicality scores. Candidate network generation algorithm reviewed in Section 2 cannot avoid unnecessary CN generation without evaluation on a large set of realtions.

LazyJoin computes a bound $\bar{S}(I)$ before join operations are performed. If $\bar{S}(I)$ quarantees that the instance I does not exceed the typicality scores already processed k-th instance, the instance I safely removed from further consideration. To derive $\bar{S}(I)$ without joins, we have to consider a hypothetical score of each tuple to be aggregated as $S(I)$. Similarly, we can calculate a typicality score of each tuple set. However, joining tuples make redundant tuples. Typicality scores are multiplied by the number of tuple connections, that is, primary key-foreign key relationship. We estimate the number of connections to predict final typicality scores for joining network of tuples. Let $TS(t)$ denote a partial score of a participating tuple $t \in r^K$ in $H(r) \in H(R')$. We calculate $TS(t)$ by counting the number of join tuples determined by t . This can be easily retrieved by a single scan of database.

4. EXPERIMENTAL EVALUATION

In our experimental study, we use a zoology database from the UCI Machine Learning Database Repository. All tuples are classified into 7 categories (*mammals, birds, reptiles, fish, amphibians, insects and invertebrates*). All the experiments are conducted on a PC with MySQL Server 5.0 RDBMS, AMD Athlon 64 processor 3.2 GHz PC, and 2GB main memory. Our methods are implemented in JAVA, connected to the RDBMS through JDBC. Due to a lack of space, the algorithm codes of the database probing modules and the index construction are not provided in this paper. We proactively identify all of the correlations between attribute values using an SQL query interface. The interface computes all pair-wise correlation by single table scan and stores the results in the auxiliary tables.

We have computed the typicality scores to evaluate the correspondence of our typicality model for real-world semantics. Several measures in cognitive science are adapted to test the

effectiveness of *categorization* and *specification*. However, the extensive set of the evaluation study on the quality of our model is incomplete and is still in progress. While computational studies in cognitive science rely on manual surveys, we perform the quality evaluation based on the classical measures in information science, e.g., precision and recall. The average precision is up to 0.715, which is a competitive result compared to [3]. As we consider every relation is identified at static time, the comparative study with [3] is feasible. To evaluate the performance of our top-k computation method, we measure the execution time of top-k results with various query sets ($Q_1 \sim Q_{10}$, fixed $k=3$) and various the parameter k (1~6, fixed query Q_4). The parameter, typicality_min t is determined as 0.4. Query sets are constructed by randomly selected keywords from the data sets. Our method greatly improves the *Baseline* (in Section 3) in query execution time, and reasonably yields better performance in time compared to the previous work [3]. From the above results, we find that our basic premise, that the prune-and-test method is very efficient for top-k retrieval. It is premature to conclude that our query model is effective for every context in structured data because this work is still in early stage. In the evaluation, we would like to introduce the potential impact of the topic, typicality analysis for ranking data.

Table 2. Query execution time (varying query sets) in msec

	Baseline	Hua et al. [3]	TPFilter
Q₁	1790	205	102
Q₂	2990	340	190
Q₃	5010	401	310
Q₄	8506	489	353
Q₅	10809	550	450
Q₆	17609	610	531
Q₇	21002	721	608
Q₈	30002	795	689
Q₉	59725	860	765
Q₁₀	96094	903	833

Table 3. Query execution time (varying k) in msec

k	Baseline	Hua et al. [3]	TPFilter
1	1702	1259	690
2	4420	1542	830
3	7520	1605	999
4	10290	1701	1480
5	28892	5020	3012
6	44205	10450	7895

5. CONCLUSIONS

In this paper, we introduced a novel ranking measure, *typicality*, based on the notions from cognitive science. We proposed the typicality query model and the typicality score based on the correlation measure, *Pearson correlation coefficient*. Then, we propose an efficient computation method, *TPFilter*, that efficiently prunes unpromising objects based on a tight upper bound, and avoid unnecessary joins. Experimental results show that our method works successfully for the real data set. Although the detail discussions of several parts are omitted, this paper proposed a promising tool for ranking structured data.

Further study is required to develop different types of typicality analysis in various applications. We would like to explore the potential of typicality analysis in data mining, data warehousing and

other emerging application domains. For example, for social networks, it would be required to identify typical users in the network, which will represent certain communities or groups. Also, ranking user nodes and user groups considering typicality would be an interesting topic in social network analysis.

ACKNOWLEDGEMENTS

This research was funded by the MSIP(Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.

REFERENCES

- [1] Rein, J., Goldwater, M., Markman, A.: What is typical about the typicality effect in category-based induction?. *Memory & Cognition*, Vol. 38 (3), pp. 377--388. (2010).
- [2] Yager, R.: A note on a fuzzy measure of typicality. *International Journal of Intelligent Systems*, Vol. 12 (3) pp. 233--249. (1997).
- [3] Hua, M., Pei, J., Fu, A., Lin, X., Leung, H.: Efficiently answering top-k typicality queries on large databases. In: *VLDB*, pp. 890--901. (2007).
- [4] Ilyas, I., Markl, V., Haas, P., Brown, P., Aboulnaga, A.: CORDS: automatic discovery of correlations and soft functional dependencies. In: *SIGMOD*, pp. 647--658. (2004).
- [5] Xiong, H., Shekhar, S., Tan, P., Kumar, V.: TAPER: a two-Step approach for all-strong-pairs correlation query in large databases. *TKDE VOL. 18(4)*, pp. 493--508. (2006).
- [6] Chaudhuri, S., Das, G., Hristidis, V., Gerhard, W.: Probabilistic ranking of database query results. In *VLDB*, pp. 888--899. (2004).
- [7] Park, J., Lee, S.: Probabilistic ranking for relational databases based on correlations. In *PIKM*, pp. 79--82. (2010).
- [8] Hristidis, V. and Papakonstantinou, Y. 2002. DISCOVER: keyword search in relational databases. In *VLDB*, pp. 670-681. (2002).
- [9] Ke, Y., Cheng, J., Yu, J.: Top-k Correlative Graph Mining. In *SDM*, pp 493--508 (2009).
- [10] Tan, P, Kumar, V., Sririvastava, J.: Selecting the right interestingness measure for association patterns. In: *SIGKDD*, pp. 32--41, (2002)..

AUTHORS

Jaehui Park received his Ph.D. degree in Department of Computer Science and Engineering from Seoul National University, Korea, in 2012 and his B.S. degree in Computer Science from KAIST, Korea, in 2005. Currently, he is a research engineer of Electronics and Telecommunications Research Institute, Korea. His research interests include keyword search in relational databases, information retrieval, semantic technology, and e-Business technologies.



AUTHOR INDEX

- Abdelkader Belkhir 199
Alaa M. Abbas 261
Ali Erkan 27
Alok Jain 241
Amany A. Kandeel 261
Amina Serir 271
Andrzej Kasinski 215
Antón-Rodríguez M 101
Bahram Lavi Sefidgari 361
Basilio Bona 339
Ben Khayut 369
Benmohammed Mohamed 53
Bhattacharya Uttam 155
Bo Sun 141
Cem Rifki Aydin 27
Cerezo-Sánchez J M 293
Chkiwa Mounira 353
Claude Duvalet 75
De Sujoy 155
Dhirendra K. Swami 241
Díaz-Pernas F.J 101
Emna Bouazizi 75
Faiez Gargouri 353
Gholamreza ShahMohammadi 183
González-Ortega D 101
GUO Yue 11
Haissam El-Aawar 323
Hamid Khemissa 199
Herman Akdag 305
Hidayet Takçi 27
Hsiao-Shan Wong 225
Huijie Chen 141
Hussein Bakri 323
Inma Mohino-Herranz 249
Jaehui Park 391
Jedidi Anis 353
Jing-Chen 39
Jiulin Hu 141
Jung-San Lee 225
Keedong Yoo 381
Khaled SAHNOUN 113, 119
Khatibi S 279
Lakehal Elkhamssa 53
Lawrence Nderu 305
León-del Rosario S 293
Lina Fabri 369
Li-qing Qiu 39 Ludovico Russo 339
Malek Ben Salem 75
Manuel Rosa-Zurera 249
Marcin Michalak 01
Martínez-Zarzuela M 101
Matteo Matteucci 339
Maya Abukhana 369
Md. Shiplu Hawlader 127
Mohamed A. Hagal 163
Mohamed Ahmed-nacer 199
Mohammed Abufouda 173
Mohd Sapiyan 311
Mohiy M. Hadhoud 261
Mustapha Delassi 271
Nicolas Jouandeau 305
Noureddine BENABADJI 113, 119
Pedraza-Hueso M 101
Preety D. Swami 241
Raffaella Gentilini 45
Rafik Bouaziz 75
Rahut Amit Kumar 155
Roberto Gil-Pita 249
Sagrario Alonso-Diaz 249
Sahand Pourhassan Shamchi 361
Saifuddin Md. Tareeq 127
Sang-goo Lee 391
SHEN Xuelian 11
Siddiqui J.R 279
Stefano Rosa 339
Taher Ali 311
Tawfig M. Abdelaziz 163
Tsuneshi Isomura 89
Tunga Güngör 27
Vega-Fuentes E 293
Vega-Martínez A 293
Wojciech Bieganski 215
Wu Wu 141
Xiaofang Huang 141
Yasmeen.N.Zada 163
Yasuhiro Matsuda 89
Yi-Hua Wang 225
Yong-quan Liang 39
Young-Hoon Ko 65
Zeiad El-Saghir 261
ZHU Zhanfeng 11
Ziad Najem 311