Computer Science & Information Technology 21

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan (Eds)

Computer Science & Information Technology

Second International Conference on Computational Science and Engineering (CSE-2014)
Dubai, UAE, April 04 ~ 05 - 2014

AIRCC

Volume Editors

Dhinaharan Nagamalai
Dept of Computer Engineering, Faculty of Engineering
KTO Karatay University, Turkey
E-mail: dhinthia@yahoo.com

Sundarapandian Vaidyanathan,
R & D Centre,
Vel Tech University, India
E-mail: sundarvtu@gmail.com

ISSN : 2231 - 5403
ISBN : 978-1-921987-30-4
DOI : 10.5121/csit.2014.4401 - 10.5121/csit.2014.4423

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

Preface

Second International Conference on Computational Science and Engineering (CSE-2014) was held in Duabai, UAE, during April 04~05, 2014. Second International Conference on Instrumentation and Control Systems (CICS-2014), Second International Conference on Database and Data Mining (DBDM-2014), Second International Conference of Soft Computing (SCOM-2014), Second International Conference of Artificial Intelligence & Fuzzy Logic (AIFL-2014), First International Conference on Computer Networks & Communications (CCNET-2014), First International Conference on Signal Processing (CSIP-2014) were collocated with the CSE-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 CSE-2014, CICS-2014, DBDM-2014, SCOM-2014, AIFL-2014, CCNET-2014, CSIP-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, CSE-2014, CICS-2014, DBDM-2014, SCOM-2014, AIFL-2014, CCNET-2014,CSIP-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 CSE-2014, CICS-2014, DBDM-2014, SCOM-2014, AIFL-2014, CCNET-2014, CSIP-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.

Dhinaharan Nagamalai
Sundarapandian Vaidyanathan

Organization

General Chairs

David C. Wyld	Southeastern Louisiana University, USA
Natarajan Meghanathan	Jackson State University, USA

Steering Committee

Abdul Kadhir Ozcan	The American University, Cyprus
Brajesh Kumar Kaushik	Indian Institute of Technology - Roorkee, India
Dhinaharan Nagamalai	Wireilla Net Solutions PTY Ltd, Australia
Eric Renault	Institut Telecom–Telecom SudParis, France
John Karamitsos	University of the Aegean, Samos, Greece
Khoa N. Le	University of Western Sydney, Australia

Program Committee Members

A Vadivel	National Institute of Technology Trichy, India
A.G.Ananth	R.V. College of Engineering-Bangalore, India
A.Kannan	K.L.N. College of Engineering, India
Abdellatif BERKAT	Abou-Bekr Belkadd University (Tlemcen), Algeria
Achhman Das Dhomeja	University of Sindh, Pakistan
Ajay K Sharma	Dr B R Ambedkar NIT, India
Alejandro Regalado Mendez	Universidad del Mar. Mexico
Alvin Lim	Auburn University, USA
Amandeep Singh Thethi	Guru Nanak Dev University Amritsar, India
Asghar gholamian	Babol University of Technology, Iran
Ashok kumar Sharma	YMCA Institute of Engineering, India
Ayad salhieh	Australian College at Kuwait, Kuwait
Azween Bin Abdullah	Universiti Teknologi Petronas, Malaysia
Balaji Raj N	JJ College of Engineering and Technology, India
Binod Kumar Pattanayak	Siksha O Anusandhan University, India
Buket Barkana	University of Bridgeport, USA
Carlos E. Otero	The University of Virginia's College at Wise, USA
Ch.V.Rama Rao	Gudlavalleru Engineering College, India
Choudhari	Bhagwati Chaturvedi College of Engineering, India
D.Minnie	Madras Christian College, India
Deepak Laxmi Narasimha	University of Malaya, Malaysia
Denivaldo LOPES	Federal University of Maranhao - UFMA, Brazil
Dinesh Chandrajain	University of RGPV, India
Ferdin Joe J	Prathyusha Institute of Tech. & Management, India
G.M. Nasira	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 Instituite 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 Acadmy 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	Jawaharlal Nehru University, India

Sherif S. Rashad	Morehead State University, USA
Shin-ichi Kuribayashi	Seikei University, Japan
Shrirang.Ambaji.Kulkarni	National Institute of Engineering, India
Sundarapandian V	Vel Tech Dr. RR & Dr. SR Technical University, India
T Venkat Narayana Rao	Hyderabad ITM , India
Tien D. Nguyen	Coventry University, UK
Tuli Bakshi	Calcutta Institute of Technology(WBUT), India
Utpal Biswas	University of Kalyani, India
V.Radha	Avinashilingam University, India
Vijayanandh. R	Bharathiar Univ, India
Wichian Sittiprapaporn	Mahasarakham University, Thailand
wided oueslati	l'institut superieur de gestion de tunis, Tunisia
Zuhal Tanrikulu	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

Second International Conference on Computational Science and Engineering (CSE-2014)

Second International Conference on Database and Data Mining (DBDM-2014)

First International Conference on Computer Networks & Communications (CCNET-2014)

Second International Conference of Artificial Intelligence & Fuzzy Logic (AIFL-2014)

Second International Conference of Soft Computing (SCOM-2014)

Second International Conference on Instrumentation and Control Systems (CICS-2014)

First International Conference on Signal Processing (CSIP-2014)

INFLUENCE OF QUANTITY OF PRINCIPAL COMPONENT IN DISCRIMINATIVE FILTERING

Kenny V. dos Santos[1],Luiz Eduardo S. e Silva[2] and Waldir S. S. Junior[3]

[1]DEQ/PPGEE/CETELI, Federal University of Amazonas,
Manaus, Amazonas, Brazil
kennyvinente@ufam.edu.br
[2]PPGEE/CETELI, Federal University of Amazonas, Manaus, Amazonas, Brazil
luiz.edu.sales@gmail.com, pedrodonadio7@gmail.com,
[3]DTEC/PPGEE/CETELI, Federal University of Amazonas, Manaus, Amazonas,
Brazil
waldirjr@ufam.edu.br

ABSTRACT

Discriminative filtering is a pattern recognition technique which aim maximize the energy of output signal when a pattern is found. Looking improve the performance of filter response, was incorporated the principal component analysis in discriminative filters design. In this work, we investigate the influence of the quantity of principal components in the performance of discriminative filtering applied to a facial fiducial point detection system. We show that quantity of principal components directly affects the performance of the system, both in relation of true and false positives rate.

KEYWORDS

Pattern Recognition, Discriminative Filtering, Principal Component Analysis & Fiducial Points Detection.

1. INTRODUCTION

Facial fiducial points detection can be understood as a pattern recognition problem. Currently, there are several approaches that attempt to solve this problem. In general, these approaches propose a system for fiducial points detection and try to solve the problem via pattern recognition techniques. Some of these systems can be viewed in [1], [3], [4], [5], [6].

The papers published in [2], [7] proposes a pattern recognition technique that uses linear filtering and can be applied to fiducial points detection. Recently, a robust filter was design to fiducial points detection [1]. Those filters, called Discriminative Filters with Principal Component Analysis (FD-PCA), are designed using the theory of principal component analysis [10]. In this approach, the filters are designed to detect the principal components of higher variance associated with patterns of interest. In a facial fiducial points detection system, the quantity of principal components used directly impacts the system performance.

In this paper, we propose investigate the influence of the quantity of principal components in the performance of a facial fiducial points detection system. The experimental procedure is performed

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 01–10, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4401

using 11 fiducial points from a subset of 503 images from the BioID database [15]. From the results obtained, we can determine the number of principal components thats make satisfactory the system performance, using as criterion for evaluating the true positives (TP) and false positives (FP) rate.

This paper as organized as follows: Section II presents a review of concepts associated of discriminative filtering with principal component analysis. Section III presents the proposal of this work, constituted by the facial fiducial points detection system and experiments and results. And finally, the conclusions are commented in Section IV.

2. DISCRIMINATIVE FILTERING WITH PRINCIPAL COMPONENT ANALYSIS

2.1. Discriminative Filtering

The goal in discriminative filtering method is design an optimal linear filter **Θ** which detects a pattern of interest **U** existing in an evaluation signal **G**. An important feature of this method is the fact that it uses signal filtering for detection. The metric used to evaluate the signal **C** is the $DSNR_2$ expressed by:

$$\mathrm{DSNR}_2(m_0, n_0) = \frac{c(m_0, n_0)^2}{\sum_{i=0}^{\mathrm{M}-1} \sum_{j=0}^{\mathrm{N}-1} c(i,j)^2 - c(m_0, n_0)^2} \qquad (1)$$

In [7], the authors propose a closed-form solution using an impulse restoration approach, which can be obtained as follows: given an array *g(m, n)* which contains the pattern of interest $u(m - m_0, n - n_0)$ located at position (m_0, n_0), and other signals that can be interpreted as an additive noise *b(m, n)*. Thus, we have:

$$g(m,n) = u(m,n) * \delta(m - m_0, n - n_0) + b(m,n) \qquad (2)$$

Developing the Equation (2) and using the matrix notation, we can find the formulation of the impulse restoration problem as follows: given the signal **g** and an array **F** with dimensions N ×N, we must to find the best linear estimative of the vector $\hat{\boldsymbol{\delta}} = \mathbf{Ag}$. Considering the case where the noise **b** is gaussian, with zero mean and covariance matrix equal to $\frac{1}{\mathbf{N}}\mathbf{C_b}$ 1NC b with dimensions N × N, the vector $\hat{\boldsymbol{\delta}}$ shall be expressed by:

$$\hat{\boldsymbol{\delta}} = \mathbf{F}^T[\mathbf{F}\mathbf{F}^T + \mathbf{C}_\mathrm{b}]^{-1}\mathbf{g} \qquad (3)$$

where the superscript ^{T} is the Hermitian. The discriminative filter Θ can be obtained by inspection of the linear estimator **A** [7].

2.2. Discriminative Filtering with Principal Components Analysis

In [1], the authors suggested the design of robust discriminative filters. In this approach, the discriminative filters are designed using principal components [10] of the set of matrices formed by occurrences of the pattern of interest. Mathematically, we can obtain the discriminative filters

as follows: suppose a random variable U_{NxI} with M realizations equal to the vectors $\mathbf{u_1}, \ldots, \mathbf{u_M}$. The principal components $\mathbf{\Phi} = [\phi_1, \ldots, \phi_N]$ and their eigenvalues $\lambda_1, \ldots, \lambda_N$ can be obtained using the solution of the eigenvalues problem [10] described below:

$$\mathbf{\Lambda} = \mathbf{\Phi}^T \mathbf{\Sigma}_{\mathcal{U}} \mathbf{\Phi},$$

where $\mathbf{\Sigma}_U$ is the covariance matrix from U.

The two-dimensional discriminative filters $\mathbf{\Theta}\phi_1, \ldots, \mathbf{\Theta}\phi_S$ are designed for S principal components $\phi_1, \ldots, \phi_S$ with associated eigenvalues $\lambda_1, \ldots, \lambda_S$, according the Equation (3). Thus, the equation of the estimator is given by:

$$\mathbf{A}_{\phi_i} = \mathbf{F}_{\phi_i}^H \left[\mathbf{F}_{\phi_i} \mathbf{F}_{\phi_i}^H + \mathbf{C}_{\mathbf{b}i} \right]^{-1}. \qquad (4)$$

Finally, the authors interpret $\mathbf{C_{bi}}$ how an orthogonal subspace from the principal component of interest. In this case, the covariance matrix $\mathbf{C_{bi}}$ associated with ϕ_i can be written as:

$$\mathbf{C}_{\mathbf{b}i} = \sum_{\substack{j=1 \\ j \neq i}}^{N} \psi_j \mathbf{F}_{\phi_j} \mathbf{F}_{\phi_j}^H, \qquad (5)$$

where ψ_j are constants that indicate the statistical noise, and $F\phi_j$ is the circular matrix by blocks obtained from the component ϕ_j.

3. EXPERIMENTS AND RESULTS

3.1. Facial Fiducial Points Detection System

The performance evaluation of robust discriminative filters was performed using a supervised facial fiducial points detection system. This system has two stages: training and test (Figures 1 and 2). In both cases, we have a pre-processing block for the images. This block has the configuration presented in [1]: a Viola-Jones face detector [11], a scaling block to 220 × 220 resolution, an illumination correction block [12] and a Gaussian Prior Model [1].

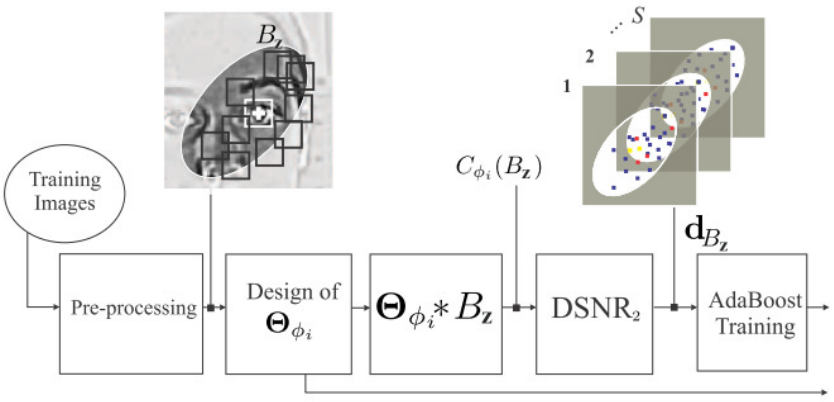


Figure 1. Training procedure of the fiducial points detection system.

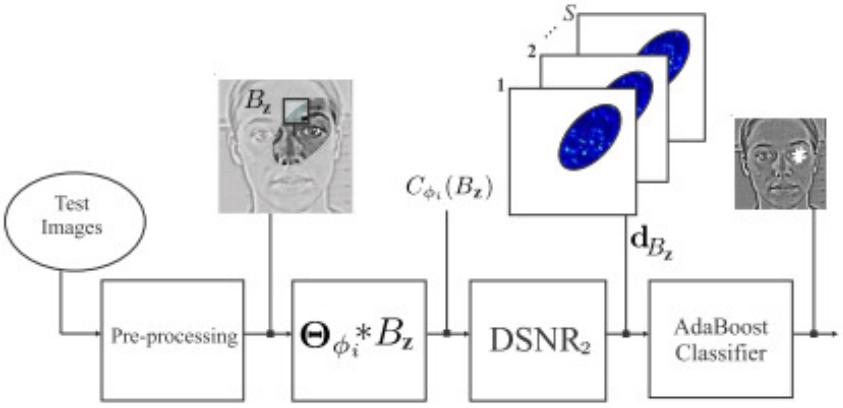


Figure 2. Training procedure of the fiducial points detection system.

The training procedure (Figure 1) can be described as follows: first, we perform a pre-processing at the image. Then, the S discriminative filters $\mathbf{\Theta}\phi_i$ will be designed for each of the principal components. Through a sliding window, each B_z block belonging to the elliptical region of interest is filtering by all S filters $\Theta\phi_i$, generating S matrices $C\phi_i(B_z)$. Using the Equation (1) we obtain the $DSNR_2$ values of the matrices $C\phi_i(B_z)$. So, each block B_z has a $DSNR_2$ associated vector named $\mathbf{d}_{Bz}$ with dimensions $1 \times S$. The vectors $\mathbf{d}_{Bz}$ will be used for training of the AdaBoost classifier using the GML AdaBoost toolbox [14]. The output of the training stage consist of $\mathbf{\Theta}\phi_i$ filters, the AdaBoost classifier and the mean block of the patterns of interest, called $\boldsymbol{\mu}_U$.

The test procedure, presented at Figure 2, can be described by: first, the input image is pre-processed. Then, using a sliding window, we process each B_z. block. The matrices $C\phi_i(B_z)$. are obtained by filtering between B_z, subtracting by $\boldsymbol{\mu}_U$ and each of the S filters $\mathbf{\Theta}\phi_i$. After, we calculate the $DSNR_2$ for each $C\phi_i(B_z)$, resulting in a $DSNR_2$ associated vector named $\mathbf{d}_{Bz}$. The AdaBoost classifier will categorized $\mathbf{d}_{Bz}$ as positive (the center of B_z is the fiducial point) or negative (the center of B_z isn't the fiducial point).

3.2. Experiments and Results

To evaluate the effect of the quantity of principal components in the facial fiducial points detection system which uses discriminative filters, we use a total of 11 fiducial points and a subset of 503 images from the BioID database [15]. The fiducial points and your numeration are presented at the Figure 3. In all experiments, we use cross-validation with 7 folds [13]. We used 6/7 of the total images and use in the training step and 1/7 for the test step. For this experiment, we varied the quantity of principal components (S) used in the proposed system as follows: S = [8, 13, 23, 33, 43, 53, 63, 73, 83, 93, 100].

The system's performance is measured using the intraocular distance. This distance, designated $\tilde{d}_o$, is obtained as follows:

$$\tilde{d}_o = ||\widetilde{\mathbf{OE}} - \widetilde{\mathbf{OD}}||,$$

where: $\widetilde{\mathbf{OE}}$ corresponds to the coordinates of the manual label from the left pupil and $\widetilde{\mathbf{OD}}$ are the coordinates of the manual label from the right pupil. For validation, we use the true and false positive rates, presented in [1]. We consider a candidate of fiducial point any mark which have distance from the manual fiducial point less then 10% $\tilde{d}_o$.

Figures 4 to 9 show the results obtained as a function of the quantity of principal components for fiducial points 0, 1, 2, 6, 7 and 9. The blue curves relate the true positives rate and the red curve relating the false positives rate. Due to the symmetry of the face, we present only the curves of the fiducial points located at the left side of the face. For comparison, the best results are summarized in Table I. In this table, we compared the best results of our proposed method with the state-of-art method Support Vector Machines. To perform this comparison, we performed the experiments using two approaches: the linear SVM (SVM-L) and polynomial SVM (SVM-P) [8], [9].

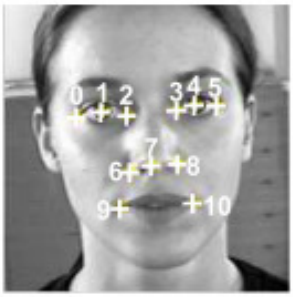


Figure 3.Fiducial points used in the experiments.

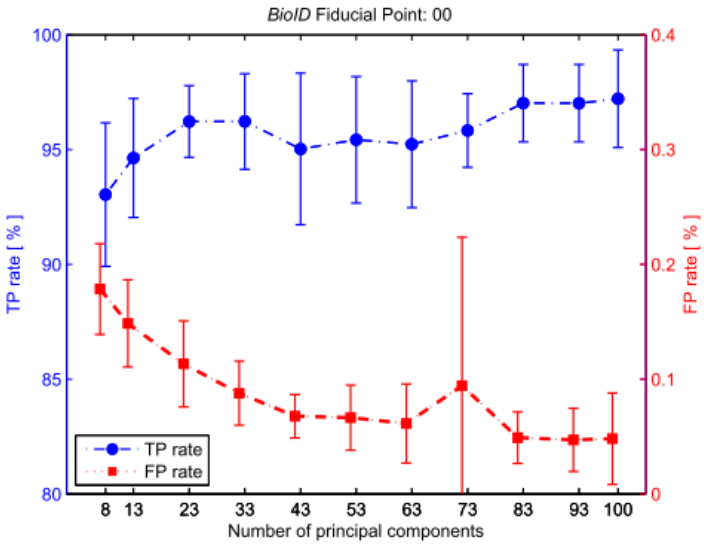


Figure 4. TP and FP curves for fiducial point 00.

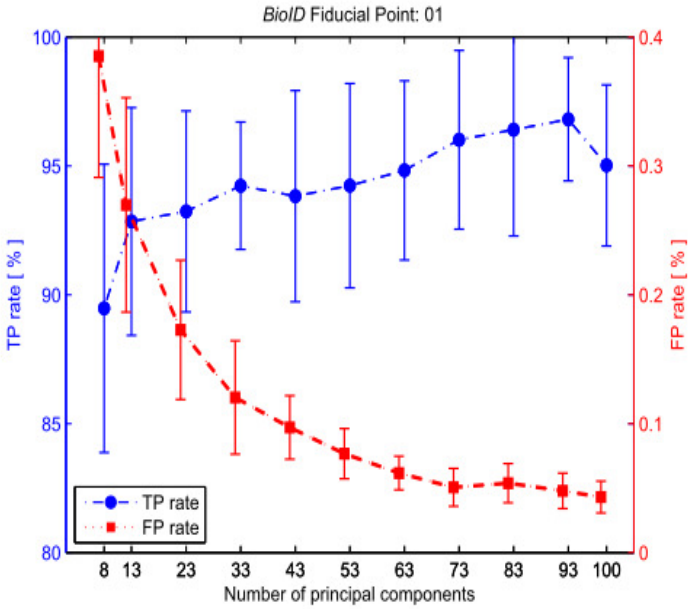


Figure 5. TP and FP curves for fiducial point 01.

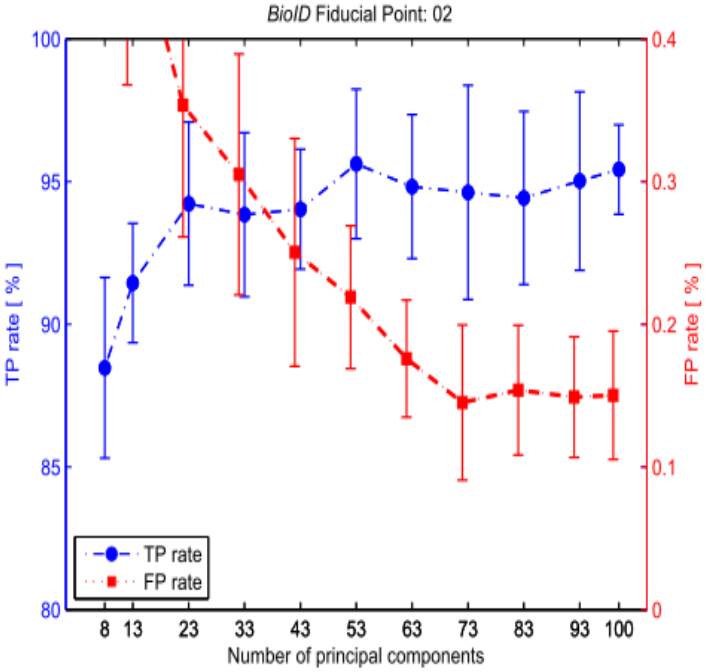


Figure 6. TP and FP curves for fiducial point 02.

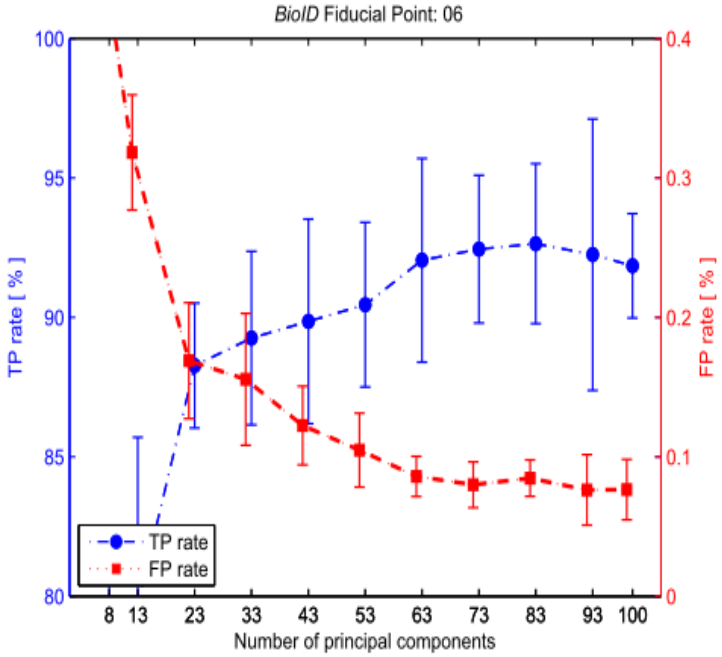


Figure 7. TP and FP curves for fiducial point 06.

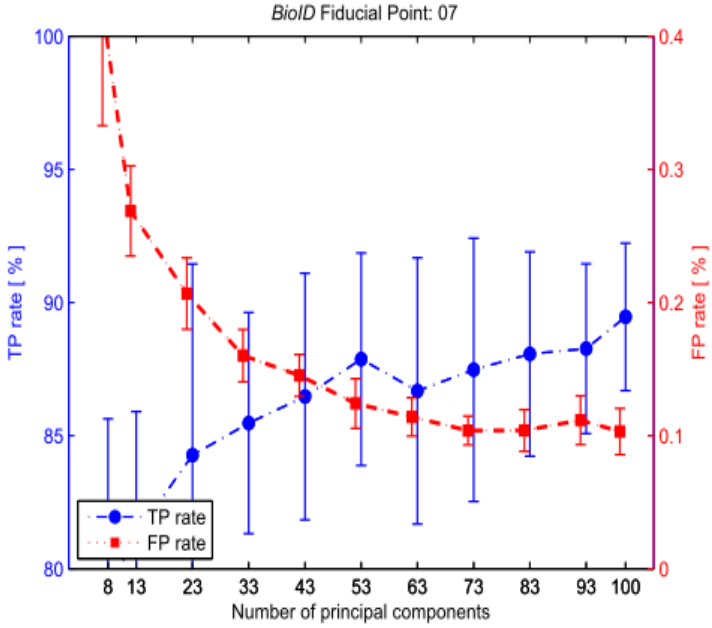


Figure 8. TP and FP curves for fiducial point 07.

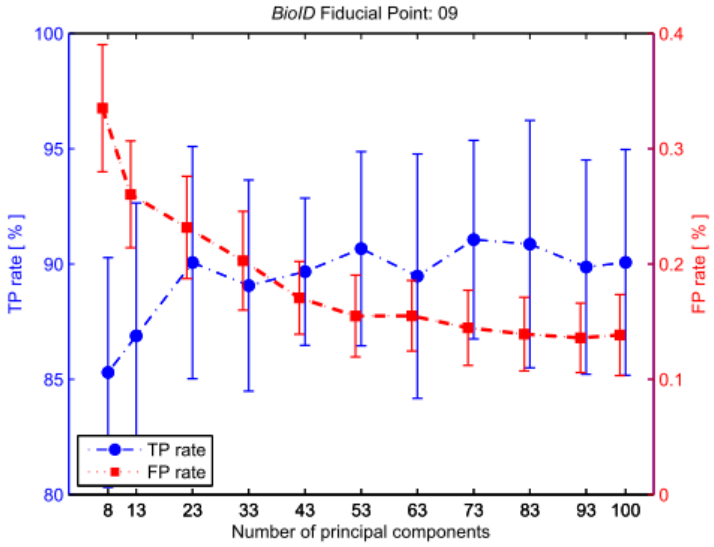


Figure 9. TP and FP curves for fiducial point 09.

From the curves presented at Figures 4 to 9, we can conclude that increasing the quantity of principal components used improves the system performance. In relation to Table I, we find that proposed method is better than SVM-L at the fiducial points 0 and 2. We also observed that for the most of points, our method is slightly lower than SVM-L and SVM-P in terms of true positives rate and superior in terms of false positives rate.

Table I. Results obtained to 11 fiducial points from BioID database.

Fiducial Point	FD-PCA (proposed)					SVM-L				SVM-P			
	TP	σ	FP	σ	NC	TP	σ	FP	σ	TP	σ	FP	σ
0	97,0	1,68	0,05	0,02	83	92,4	9,8	1,09	0,74	99,6	0,7	0,39	0,07
1	96,8	2,39	0,05	0,01	93	99,6	1,0	0,32	0,09	98,8	1,2	0,04	0,02
2	95,4	1,57	0,15	0,04	100	93,0	14,3	1,58	0,59	94,2	2,7	0,15	0,03
3	95,4	2,07	0,13	0,03	93	96,0	1,7	0,46	0,21	99,8	0,5	0,33	0,09
4	96,0	0,96	0,05	0,02	100	97,2	4,2	2,49	4,76	98,8	1,5	0,08	0,05
5	96,8	2,36	0,03	0,01	100	98,8	3,1	2,27	0,62	99,6	0,7	0,42	0,09
6	92,6	2,86	0,08	0,01	83	99,6	1,0	1,40	1,44	99,6	0,7	0,31	0,05
7	89,4	2,76	0,10	0,01	100	47,1	9,6	0,06	0,03	99,6	0,7	0,60	0,05
8	93,6	2,63	0,09	0,01	93	95,4	3,9	0,40	0,09	99,0	1,1	0,24	0,03
9	91,0	4,30	0,14	0,03	73	96,8	5,5	2,91	5,12	98,8	1,7	0,39	0,06
10	85,8	4,80	0,07	0,04	63	98,6	1,9	1,47	0,49	99,6	0,7	0,35	0,04

4. CONCLUSION

In this work we investigated the influence of the quantity of principal components in the performance of a facial fiducial points detection system. We performed the experiments in the BioID database, using cross-validation, splitting the total of images in two segments: training set (6/7 of total) and test set (1/7 of total). In this experiment, we varied the quantity of principal components and studied the influence of this quantity of components in the system performance, using as criteria the true and false positives rate. The results showed that quantity of principal components used determine the system performance. A low number of components will generate underperforming systems. We also observed that as of 43 components we obtain systems with

good performance and from 73 components the false positives rate stabilizes. Finally, we conclude that the proposed method can compete with the state-of-art method SVM.

ACKNOWLEDGEMENT

We would like to thanks Samsung for financial support. This research was supported by FAPEAM and CAPES agency.

REFERENCES

[1] W. S. da Silva Júnior, G. M. Araújo, E. A. B. da Silva and S. K. Goldenstein, "Facial Fiducial Points Detection Using Discriminative Filtering on Principal Components". In: *Proceedings of the IEEE International Conference on Image Processing*, September, pp. 2681-2684, 2010.

[2] A. P. Mendonça, E. A. B. Silva, "Two-Dimensional Discriminative Filters for Image Template Detection" In: *Proceedings of the International Conference on Image Processing*, pp. 680-683, 2001.

[3] B. Martinez, M. F. Valstar, X. Binefa, M. Pantic, "Local Evidence Aggregation for Regression Based Facial Point Detection" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1149 -1163, 2013.

[4] F. Jerome, H. Trevor and R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting", ˙In: *Annals of Statistics*, vol. 28, pp. 2000, 1998.

[5] T. Cootes, G. J. Edwards and C. J. Taylor, "Active Appearance Models", ˙In: *Proceedings of the European Conference on Computer Vision*, vol.2, pp. 484-498, 1998.

[6] G. M. Araujo, W. S. da Silva Júnior, E. A. B. da Silva, and S. K. Goldenstein, "Facial Landmarks Detection based on Correlation Filter", In: *Proceedings of the IEEE International Telecommunication Symposium*, Manaus, AM, Brazil, October 2010.

[7] A. P. Mendonça, E. A. B. Silva, "Closed-Form Solutions for Discriminative Filtering using Impulse Restoration Techniques", *IEE Electronics Letters*, vol. 38, n. 22, pp. 1332-1333, 2002.

[8] D. Wu, F. Cao, "Learning rates for SVM classifiers with polynomial kernels". In: *Proceedings of the Eighth International Conference on Machine Learning and Cybernetics*, Baoding, July, 2009.

[9] G. L. Prajapati, A. Patle, "On Performing Classification Using SVM with Radial Basis and Polynomial Kernel Functions". In: Third International Conference on Emerging Trends in Engineering and Technology, 2010.

[10] M. Kirby and L. Sirovich, "Application of the Karhunen-Loève Procedure for the Characterization of Human Faces", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, 1990.

[11] Viola, P. and Jones, M., "Robust Real-Time Object Detection". *International Journal of Computer Vision*, vol. 57, n. 2, pp. 137-154, 2001.

[12] Xiaoyang, T. and Triggs, B., Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. *IEEE Transactions on Image Processing*, vol. 19, n. 6, pp. 1635-1650, 2010.

[13] Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", In: *Proceedings of the International Joint Conference on Artificial Intelligence*, August, pp. 1137-1143, 1996.

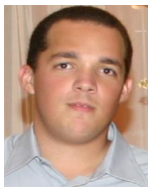
[14] "The GML AdaBoost Matlab Toolbox", [Last access in August 2013]. [Online]. Avaliable in: http://graphics.cs.msu.ru/en/science/research/machinelearning/adaboosttoolbox .

[15] "Bioid Database", [Last access in October 2013]. [Online]. Avaliable in: http://www.bioid.com/

Authors

Kenny V. Santos is a graduate student in Electrical Engineering from Federal University of Amazonas. He received his bacharel degree in Electrical Engineering at Federal University of Amazonas in 2008. Currently he works in the fields of computer vision and pattern recognition, where his main interest is applications of digital signal processing in pattern recognition.

Luiz Eduardo S. e Silva received the B.S. degree in electrical engineering at the Federal University of Amazonas (UFAM), Manaus, AM, Brazil, in 2011. Currently, he is a M.Sc. student at the same University and he is working with computer vision. Support Vector Machine is a technique applied in his pattern recognition researches, which try to classify elements of interest. He is Professor of Nokia Foundation in Electrical Circuits e Telecommunications. Automatic Control, Digital Signal Processing, Machine Learning and Mathematical Morphology are also his interest areas.



Waldir S. S. Júnior received the B.S. degree in electrical engineering at the Federal University of Amazonas (UFAM), Manaus, AM, Brazil, in 2000, and the M.S. degree in electrical engineering at the Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, RJ, Brazil, in 2004 and Ph.D. degree in electrical engineering at the Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, RJ, Brazil, in 2010. Since 2006, he has been with the Federal University of Amazonas, as Full Professor. His research interests are in the fields of data compression, as well as in mathematical morphology, pattern recognition and digital signal processing in general.

ANDROID MAPPING APPLICATION

Abdalwhab Bakheet[1] Ahmed Abd Almahmoud[2] and Wigdan Ahmed[3]

Department of Electrical and Electronic Engineering,
University of Khartoum, Khartoum, Sudan
abdalwhab.bakheet@gmail.com
eng.ahmedabdelmahmoud@gmail.com
wigdan.mohamed@gmail.com

ABSTRACT

Location-aware and mapping applications have gone from a desirable feature to an essential part of any smart phone. Whether a user is checking into a social network, looking for a pharmacy in the middle of the night, or located in somewhere and needs help, the key is always the same: location.

In this project, an Android mapping application is developed. The application is able to display the map of the whole world while online or, display a pre-downloaded map while offline, track the user's location, display a compass to determine north, send the user's location to others in case of emergency using SMS, receive and interpret received location from the message, display it on the map, and notify the user by the reception of the location.

The application was developed using agile methodology. It, met its objectives and successfully passed 91% of the final system test, recording that some limitations were discovered, the application needs further testing and can be implemented for particular company or university using their own maps or editing the maps in OSM (open street maps).

KEYWORDS

Android, mapping application, mobile development

1. INTRODUCTION

Mobile phones and their applications have become an essential part of our lives. They are not only connecting us with friends and families, but also they can now tell us where we are, where to go, what to do, and how to do it. The Internet is chock full of applications that can entertain us and make our lives easier.

Location-aware and mapping applications have gone from a desirable feature to an essential part of any smart phone. Companies, universities, airports and organizations of today are now providing maps either as part of their websites or as stand-alone applications to give direction services. Users provided with information about how to reach the specific organization, and how to navigate inside that organization.

Due to this massive need, scores of companies and developers have developed customized mapping applications and mapping APIs (application programming interface). There is no need to provide a detailed rundown of the available mapping applications and their features, because we

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 11–22, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4402

believe that the reader is already familiar with those applications. Instead this paper describes the development of such an application, its phases and problems encountered, in the hope that it will help anyone developing a similar application.

The original purpose to develop this application was to provide a custom mapping application for the University of Khartoum middle complex. The goal was to allow visitors, students and staff of the University to navigate in the complex using their smart-phones. However, after starting the research and development of the application, it was found that it is better to provide a general solution that will be interesting and useful for any user (not just for the University of Khartoum) to increase its market value.

2. APPLICATION FINAL SPECIFICATION

The application:

1. Displays a detailed map of the University of Khartoum, since the application was originally developed for the University of Khartoum.
2. Detects and displays the user's location on the map.
3. Enables the user to zoom the map in and out.
4. Allows the user to choose between two modes: an online mode displaying the map of the whole world and an offline mode displaying a pre-downloaded or auto-cached map that does not need an Internet connection.
5. Tracks the user location online and offline (without using the Internet).
6. Provides a compass that shows the direction towards north even in areas where signal is missing.
7. Provides a scale bar showing the relation between distances in the map and real distances.
8. Can send the user location to saved emergency numbers.
9. Receives the location sent by another user and displays it on the map.
10. Alerts the user of the reception of a message that contains location sent by another user.
11. Allows registering and edit of emergency numbers.

3. METHODOLOGY

Being the first time an Android application and maybe the first mobile application ever to be developed at the University of Khartoum it was not possible to foresee and plan for the whole application development life cycle. Therefore an agile software development methodology was adopted, tacking a small piece of requirement (only one or two new features), implementing and testing them separately from the application and then integrate them with the application, testing the application and repeating this sequence for the next feature.

Any mapping application includes two main features: displaying a map and determining the user's location.

A. Display a map

The application should display a map that is detailed enough to allow the user to find a particular place of interest.

The map is displayed from a map server. It can be displayed directly from the server or stored in the handset for offline use. Displaying the map directly from the server has the disadvantage that the loading of the map tiles will be dependent on the availability and speed of the Internet

connection i.e. in some places there is no signal, and in others there is signal but the data rate is very small.

The level of details of the map displayed by the application will depend on the quality of the map on the server. Maps qualities vary greatly from one server to another, and even for the same server the level of detail varies for different regions.

Many map servers allow the community of users to edit maps and contribute additional details, and then they review the edits and display them in the server. After that the developers can access them by the application or download them.

To display maps from a particular server, developers use the application programing interface (API) provided by the server that provides classes needed to access the maps located in the server.

B. Determine user location

To help users find a destination, or pinpoint their locations, mapping applications include the ability to determine the user's location. The required accuracy depends on the purpose of the application, but for most applications the degree of accuracy is sufficient if the application can determine the location accurate enough for the user to be able to know what street or building he is in, this degree of accuracy will be fine.

At the same time the accuracy of the location provided by the application depends on the method used to determine the location, and the accuracy and the availability of that method in the particular region where the user is.

Android supports determining the user's location using global position system (GPS) that is built in the handset, or using information from a nearby transmission tower. Moreover, if the user is connected to a Wi-Fi network, it can be used to determine the location. It must be mentioned that all these location providers are unreliable to some extent. For example, GPS signals do not reach inside buildings [1].

3.1. Application development iterations

3.1.1. Determine how to do it

Like with any project, the first phase was information gathering and analysis, trying to understand how the project can be done, what the various methods are, together with their respective merits.
Any mapping application involves two parts; a client-side part and a server-side part. The client-side part provides a user interface and accesses a map server. On the server-side part developers prepare the maps in a well-defined format and provide an API to access the map located on the server, then developers can use the API to develop applications that access the maps and use additional features provided by the API.

Some servers publish their API for public access and use for free, others require an API key, and some raise a fee for the use of the API.

Therefore the first phase of the project was to study the various APIs available on the Internet, their functionality, and their merits.

There are so many mapping APIs available, each of which belonging to and accessing a particular server. Some APIs target indoor mapping, while others target outdoor mapping.

The most famous API is Google's mapping API, but there are many other mapping APIs like MapQuest, Nutiteq, ArcGIS, Decarta, Guidbee, Mapsfore, OpenLayers, and Osmdroid.

We studied the merits of each one, and even tested some of them by accessing and displaying maps using them, finally settle for outdoor mapping, because outdoor mapping is more general and more suitable for the University of Khartoum complex.

In this paper we will not discuss all APIs available, because there are many APIs, and they are always changing, new APIs are introduced and the available ones are constantly improved. The first step for anyone trying to develop a similar application is to look and examine the available APIs, their features and license agreements.

One interesting API is Nutiteq mapping API besides supporting the basic features: showing interactive online map, and map overlays, it also supports many base maps options OpenStreetMap, Bing, MapQuest, MapBox, AND, CloudMade, where Google supports only Google Maps, and aerials.

It also supports online routing, geocoding, and unlike Google's API it supports offline maps and offline routing [2].

Google's mapping API, has the advantages of wide spread, it is well documented, and easy to use. We didn't use it because for free license the application must be free and publicly available, also the number of requests per day is limited, it only access Google maps server, and we couldn't grantee that if we edit the maps on their server that they will approve it (because of the political problem with Sudan).

Osmdroid was selected because of its flexible license agreement, and its features: Osmdroid mapping API is free, and does not even require a mapping API key. It accesses Open street maps which is a collaborative project to create a free editable map of the world. For these two reasons we had chosen it. In addition Osmdroid allows developer to change the server of the maps; it can be used with other APIs to provide routing service, and several other features that we used part of in the application.

3.1.2. Adding features

The features of the application's final specification were developed one at a time, with each development iteration adding a new feature: displaying the map directly from open street maps server, determining the user's location, supporting offline mapping, editing open street maps to add more details about the University of Khartoum middle complex, tracking user location both online and offline, and arguably the most interesting feature of all: sending the user location in case of emergency. This last feature is the one we will discuss in a bit more details.

Emergency service:

A user may save up to five numbers to send his or her location to using SMS in case of emergency by two clicks. When the message is received by the recipient's mobile the application receives the message, and extracts the location information from it, then opens displaying the received location and notifies the user by the reception of the location.

Some screen-shots for the application are displayed in the results section.

4. DESIGN

As shown in (Figure 1) there is sex main activities (classes that extend Android activity class which corresponds and manage one user interface).

“OurMapping” class is responsible for managing the application main window.

“OsmMap” class was written because many activities need to display map, so they all need common classes like “MapView” which displays maps, “MapController” to control the map. Also they need “MyLocationOverlay” to display a location overlay on the map, “ScaleBarOverlay” to display scale bar.

Therefore any other class that needs to display a map will just extends “OsmMap”, no need to rewrite code, and reduced maintenance overhead.

“SetNumbers” class is responsible for setting the emergency numbers; the class uses a “SharedPreferences” and “SharedPreferences.Editor” to store the numbers.

“SendMyLocation” is responsible for sending the user location; it uses “LocationManager” to get the user location, and “SharedPreferences” to get the stored telephone numbers.

Help class displays help information about the application, where “About” class displays information about the version of the application.

(Figure 2) shows the five activities that extend “OsmMap”, where each one of them is responsible for managing one user interface.

(Figure 2) also shows “SMSReceiver” class which is the only class in the application that is not an activity, it does not need a user interface, it runs in the background that is why it extends Android “BroadcastReceiver”, which allows it to listen to the reception of new messages and activate “MessageReceived” class.

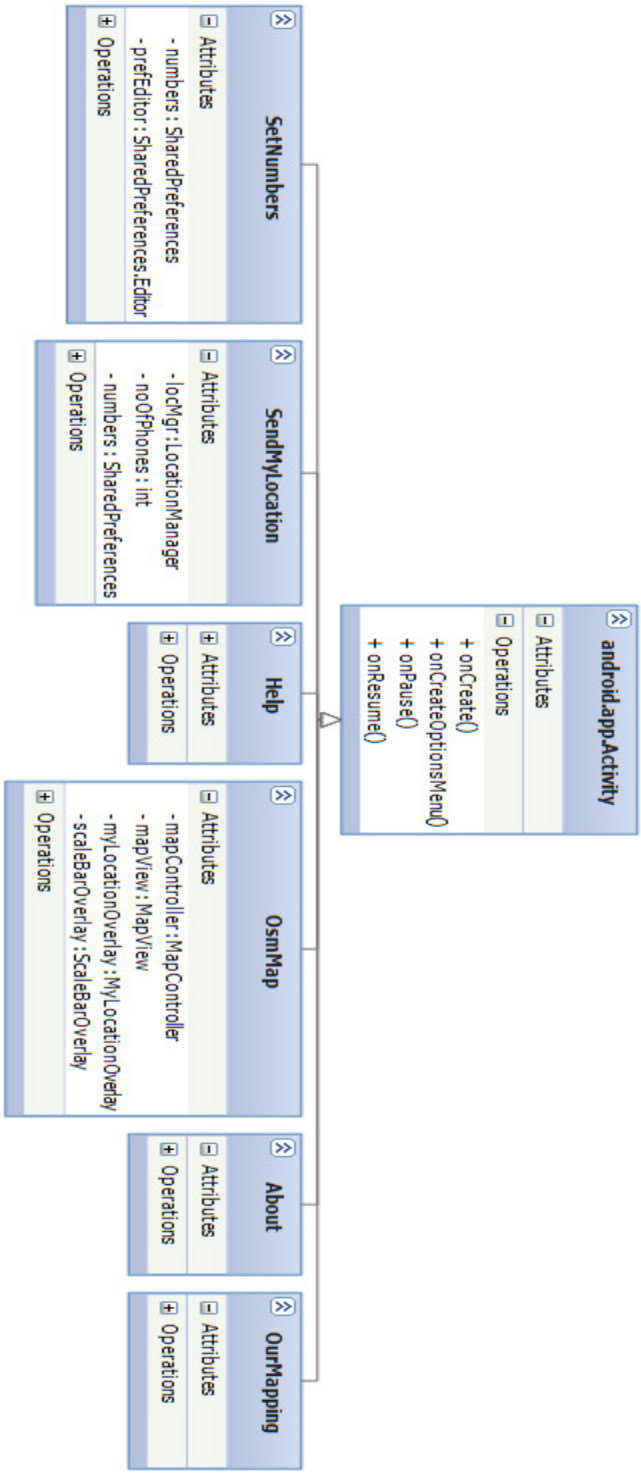


Figure 1. Program basic structure part1 using unified modeling language

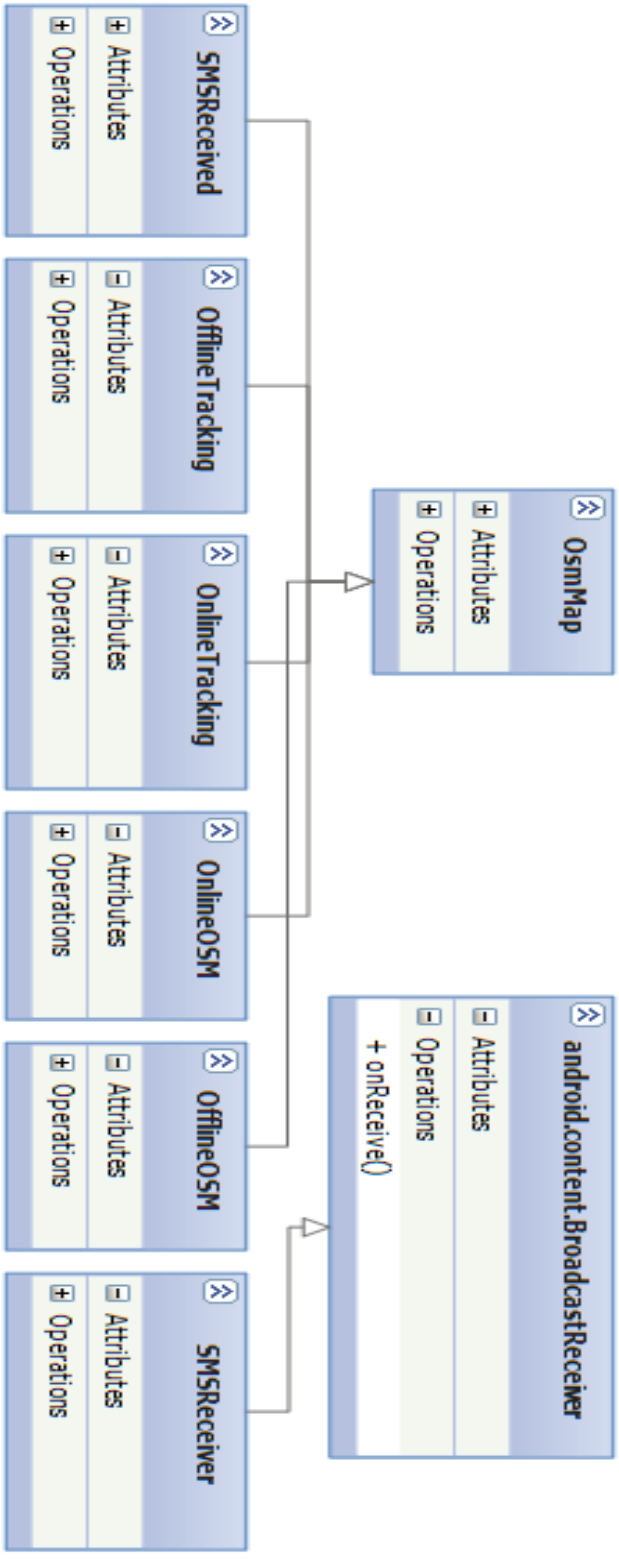


Figure 2. Program basic structure part2 using unified modeling language

5. RESULTS

Here are some screen shots for the application:

Like shown in (Figure 3) when the application starts it displays the main menu, showing four options: online, offline, help and about, allowing the user to choose any one of them.

Figure 3. Main menu

(Figure 4) shows the application online mode, where the application downloads maps from the server. The application shows a scale bar, a compass, zoom controls, and online icon is highlighted in green indicate that this mode is active, while offline and location icons are inactive. The location icon is highlighted in green if tracking the user location is enabled.

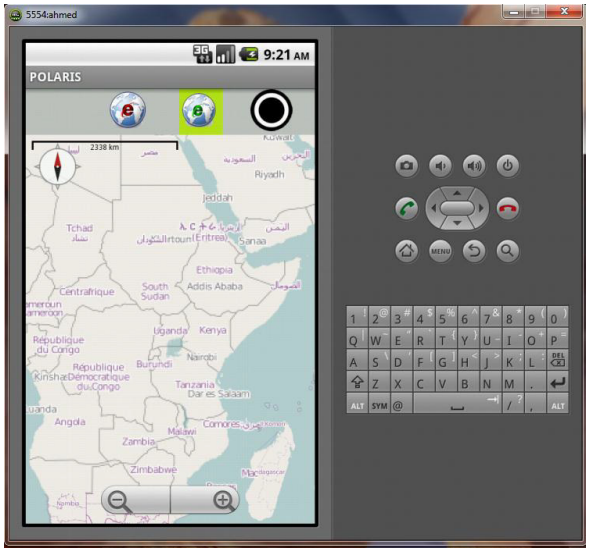


Figure 4. Application on the online mode

(Figure 5) shows the application offline mode, where maps are not loaded from the server and only the maps that were cached or pre-downloaded before are displayed, that why it is noticeable that when the map is zoomed in no details appeared, just like an image, also it is noticeable that the offline icon is highlighted in green while the online and location icons are not.

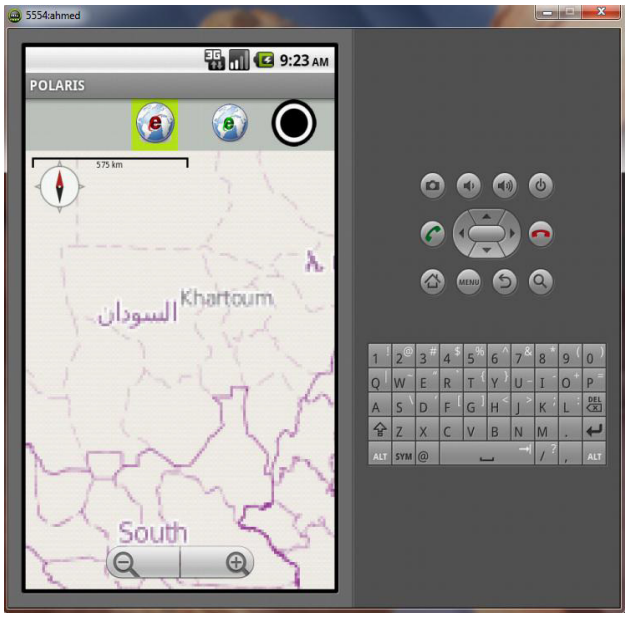


Figure 5. Application on the offline mode

(Figure 6) shows the tracking and displaying of the user's location on the map, also the location button is displayed in green which means it is active.

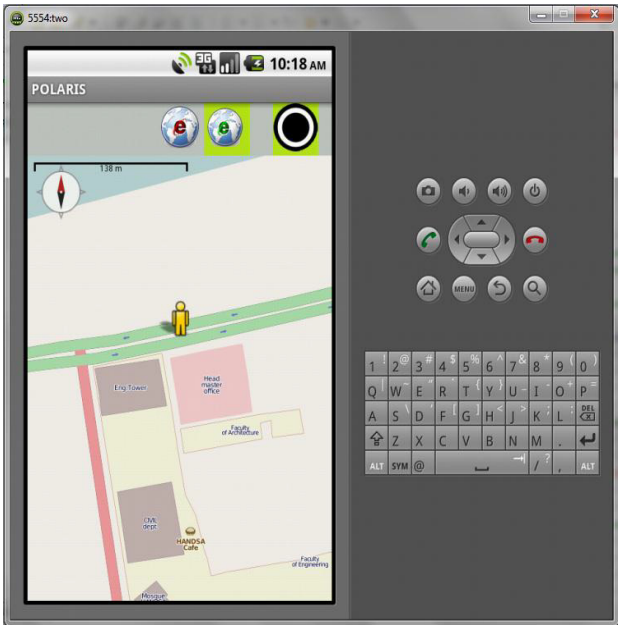


Figure 6. Tracking user location

(Figure 7) Shows the options to register the emergency numbers, to send the user's location and to return to the main menu.

Figure 7. Options to set the emergency numbers and send the user's location

(Figure 8) shows the window to register the emergency numbers, where the user can add and save up to five numbers.

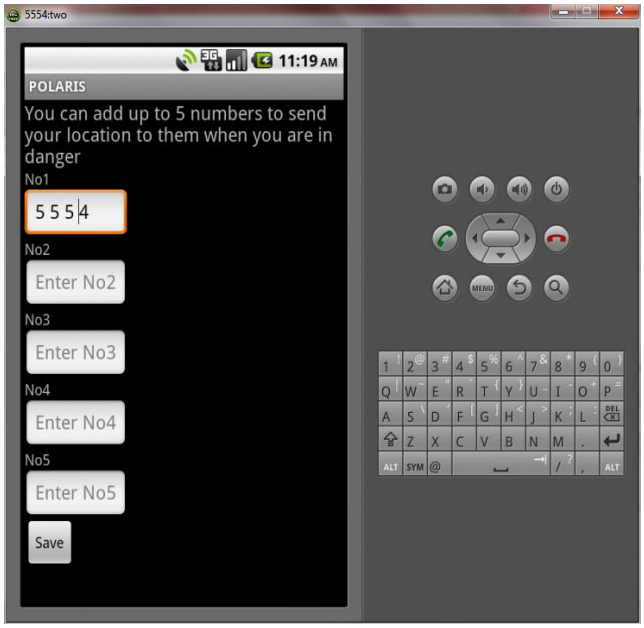


Figure 8. Window to set the emergency number

(Table 1) shows a comparison between the application main features and couple of major android application that use open street maps.

Table 1. Comparison between our application and other OSM mapping applications [3]

Name	**Display map**	**Navigate**	**Make track**	**Price**
AFTrack GPS-Tracking	Yes	Yes	Yes	3.90€;4.95$;free
AlpineQuest GPS Hiking	Yes	No	Yes	3€
AndNav2	Yes	Yes	Yes	Free
AndRoad	Yes	Yes	Yes	Free
Ape@map	Yes	No	Yes	Free
ARnav	Yes	Yes	No	Free
Our application	Yes	Yes	No	---------

The table shows a couple examples with comparison based on main features only, because this is not the right place to do a complete, deep comparison between available applications. This paper is more focused on how similar applications can be developed. However from this comparison we can see that there are many available applications that can differ in the features they provide.

6. CONCLUSION

The application was successfully developed and achieved its requirements, but it requires more testing to increase stability, especially the user's tracking feature. The application needs to be

tested on real physical devices. All the testing of the application was done using emulators due to the lack of a real handset for testing.

The application can be implemented to any particular company or university by editing the maps at open street maps. Further research can be done to change the map server to a server specified by the customer.

ACKNOWLEDGMENT

Thanks to Allah the Almighty without His help we would not be able to get this project done.
We wish to thanks Wigdan Ahmed lecturer at the University of Khartoum, faculty of engineering for being our supervisor in this project.

Thanks to everyone who helped us in the development of the project specially Alla Ahmed, Ahmed Abdallah, Habab Siddig.

Thanks to Elamin Altayeb and Abdallah Ulber for helping us revising the paper.

REFERENCES

[1] (2012) Android developer website. [Online]. Available: http://developer.android.com/guide/topics/location/index.html.
[2] (2013) Nutiteq website. [Online]. Available: http://www.nutiteq.com/nutiteq-sdk/comparison/
[3] (2013) Open Street Map website. [Online] Available: http://wiki.openstreetmap.org/wiki/Android

Authors

1. Abdalwhab Bakheet Mohamed

B.Sc. Electrical and electronic engineering (Software Engineering major) Currently: Working as Software test engineer at Banan IT, Sudan

2. Ahmed Abd Elmahmoud Elhaj Salih.

B.Sc. Electrical and electronic engineering (Telecommunication Engineering major) Currently: Technical manager at Furgan FM, Sudan

3. Wigdan Ahmed

M.SC. Computer Science and Entrepreneurship CCNP, CCDP, MCSE, ITIL V3 Foundation, CDRE Currently: Senior Infrastructure Engineer at EBS, Sudan

FIDUCIAL POINTS DETECTION USING SVM LINEAR CLASSIFIERS

Luiz Eduardo S. e Silva[1], Pedro Donadio de T. Júnior[1], Kenny V. dos Santos[2] and Waldir S. S. Junior[3]

[1]PPGEE/CETELI, Federal University of Amazonas, Manaus, Amazonas, Brazil
luiz.edu.sales@gmail.com, pedrodonadio7@gmail.com,
[2]DEQ/PPGEE/CETELI, Federal University of Amazonas,
Manaus, Amazonas, Brazil
kennyvinente@ufam.edu.br
[3]DTEC/PPGEE/CETELI, Federal University of Amazonas,
Manaus, Amazonas, Brazil
waldirjr@ufam.edu.br

ABSTRACT

Currently, there is a growing interest from the scientific and/or industrial community in respect to methods that offer solutions to the problem of fiducial points detection in human faces. Some methods use the SVM for classification, but we observed that some formulations of optimization problems were not discussed. In this article, we propose to investigate the performance of mathematical formulation C-SVC when applied in fiducial point detection system. Futhermore, we explore new parameters for training the proposed system. The performance of the proposed system is evaluated in a fiducial points detection problem. The results demonstrate that the method is competitive.

KEYWORDS

Pattern Recognition, SVM, Fiducial Points & Parameter Grid

1. INTRODUCTION

Currently, the interest of scientific community and industry in methods to solve fiducial point detection problem is increasing (salient points [1], for example, tip of the nose) in human faces [1].

Typically, recent approaches can be divided in two main categories: local and global. In local methods, the fiducial points are detected and processed individually and no additional information is used. The global methods are characterized by detecting more fiducial points using deformable models, less susceptible to pose and illumination variations than local methods.

The classifier design is the most important stage of a fiducial point detection (supervised). In this stage, several machine learning techniques can be used [1] e [2]. Particularly, some studies [3] use a classification method called Support Vector Machine [4]. The mathematical formulation of SVM is obtained by optimization problem with restrictions [5].

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 23–31, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4403

There are many recent papers about this theme. For example, the authors propose in [2] a face recognition subsystem framework that uses fiducial points detection. The detection of these points combines two techniques. The first uses Gabor filters coefficients for local detecting and the second uses a combination of human face anthropometric measurements. The system proposed in [8] explores the same problem. The authors use classifiers based on IPD (Inner Detector Product) correlation filters. In [1], the authors developed robust filters to investigate the same problem. These filters are designed using the principal components.

In this article, we study the performance of SVM mathematical formulation called C-SVC (Support Vector Classification) [5], for fiducial point detection. The main contribution of this paper is the investigation of parameters that have not been used before. For performance evaluation, the classifiers were designed based on C-SVC. The training algorithm was performed using eleven (11) fiducial points in a database composed of 503 images (base subset BioID [7]). The performance of the system with C-SVC classifier was compared with similar methods, two of them using discriminating and filtering and one using linear SVM [1]. The results demonstrate the proposed method has competitive performance, and, in some cases, the proposed method outperforms the others.

The remainder of this paper is organized as follows. In next section, we present a SVM description. The system proposed is shown in Section 3, which is emphasized in the design of classifiers, parameters optimization of different types of SVM and the chosen parameters, in training and testing, results, and comments are made about these results. Finally, conclusions are drawn in Section 4.

2. SUPPORT VECTORS MACHINES (SVM)

The SVM, proposed by Vapnik [4], is used to solve classification and regression problems. The SVM algorithm provides an optimal separating hyperplane with maximum margin. The SVM can be formulated to consider mislabelled samples. In this case, the technique is called soft margin [5]. The use of soft margin provides a generalization of the SVM method to deal with problems of separating classes. The mathematical formulation can be written as follows. First, consider the hyperplane:

$$y_i(\mathbf{w}^t\,\mathbf{x}_i + b) \geqslant 1 - \xi_i, \text{ to } i = \{1, \ldots, n\}. \qquad (1)$$

where: ξ_i , is the non-negative slack variable, y_i is the class labels, n is the number of total class elements and the superscript ^{t} is the transpose of the vector. To maximize the class-separating margin and obtain the optimal hyperplane, it is necessary to find a solution to the optimization problem given by:

$$\begin{array}{ll} \textbf{Minimize} & \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right). \\ \textbf{Restrict to} & y_i(\mathbf{w}^t\,\mathbf{x}_i + b) \geqslant 1 - \xi_i, \\ & \text{to } i = \{1, \ldots, n\}. \end{array} \qquad (2)$$

In Equation (2), C represents the weight imposed on the soft margin. The implementation of this optimization problem is known as C-SVC [5]. There are others SVM optimization problems that differ in some aspects such as: the formulation of the optimization problem and the used parameters [5]. The performance depends on the parameters choice and the search for the best set of parameters is extremely important. In this work, we use classifiers C-SVC due to its linear structure. The linear SVM classifiers have low computational cost and therefore we can say that the categorization of patterns is fast.

3. FIDUCIAL POINT DETECTION SYSTEM USING C-SVC CLASSIFIERS

3.1. Design of Classifiers C-SVC

As mentioned in Section II, we can find several SVM formulation of the optimization problem. In our research, we employ the optimization problem denominated C-SVC to design classifiers for a fiducial points detection system. An important aspect is investigate which set of parameters produces the best performance of the C-SVC classifiers. In this work, we explore the following parameters:

- Parameter of regularization C with values equal to 2^n with $n = \{-12, -9, -6, -3, 0, 3\}$;
- Maximum number of iterations to convergence: 10000; 18000; 26000; 34000; 42000; 50000.
- Proportional weighting of the number of class elements (positive: negative): (1:1), (1:5), (2:10), (10:500), (5:1), (10:2) and (500:10).

The range of parameter values were obtained empirically by preliminary results. In these cases, we observe the range have a significant influence in the performance of the fiducial points detection system. An important point of view is that each parameter and its ranges represent one experiment which must be explored. This approach is called parameters grid [5]. In our research, we have an extensive grid with 252 parameter combinations. We use ξ_i equal to 0.1.

3.2. Fiducial Points Detection System using C-SVC Classifiers

The fiducial points detection system has two stages: training and testing. In both, all images are pre-processed using a common step. The pre-processing step is composed by two stages. Firstly, we apply the Viola-Jones face detection algorithm [11] in the image. After that, the face is scaled to 200x200, and then we perform illumination correction [12].

In the training stage (see Figure 1), first, the image is pre-processed. Next, we use a method to select an elliptical region of interest with high probability of containing the fiducial point. The mathematical formulation can be seen in [1]. We can view the elliptical region of interest in the block output of the Gaussian Priori Model (Figure 1).

Following, we choose the B_z blocks, with dimensions 13×13, that define the positive and negative classes, where **z** represents the coordinates of the center of the block. We select a restricted number of negative blocks due to its high amount. This selection is done using an algorithm to generate random two-dimensional coordinates called Salt and Pepper [13]. The positive and negative ratio is 160/1. Next, we transform each block B_z (array) in a vector $v(B_z)$ by concatenating the rows of the transposed matrix. Finally, we use the blocks of positive and negative classes to design the C-SVC classifier. It is important comment that all images used in the test system are different from the images used in the training system, i.e., the sets of training and testing images are mutually exclusive.

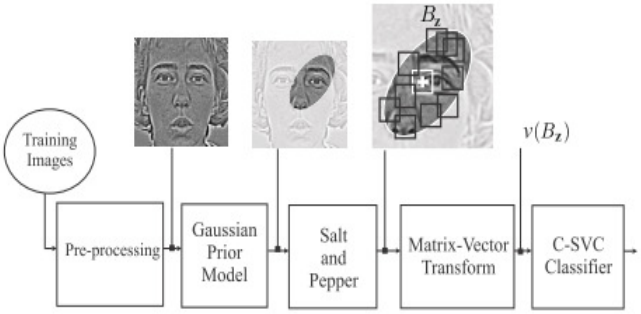


Figure 1. Block diagram for training the fiducial points detection system using C-SVC classifiers.

In the testing stage, shown in Figure 2, can be described as follows: first, we pre-processing the image. Next, we use the same elliptical region of interest determined in the training stage. Following, each block B z is processed using a sliding window. The vector $v(B_z)$ is obtained through the transformation matrix B_z in a vector (identically to the transformation performed in the training). Finally, we use the C-SVC classifier, found in the training, to label the vector $v(B_z)$ as positive or negative.

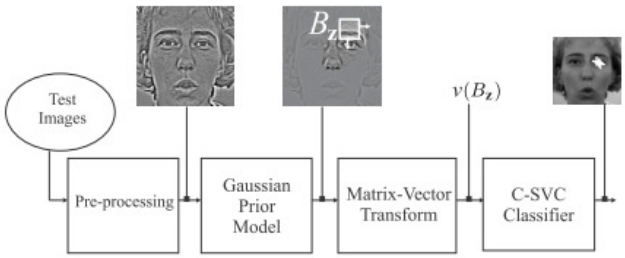


Figure 2. Block diagram for test the fiducial points detection system using C-SVC classifiers.

3.3. Experiments

The C-SVC classifiers of training and testing procedures (see Sections 3.1 and 3.2) were designed using OpenCV (Open Source Computer Library) [14]. OpenCV is written in C and C++ and can be operated in Windows, Linux and MAC OS. OpenCV was designed for computational efficiency and has specific mathematical functions and methods, for example, in signal processing and machine learning areas.

In order to evaluate the proposed method, we use 11 fiducial points as illustrated in Figure 3. We use the system described in Section 3.2 for each fiducial point. Thus, we have 11 different fiducial points detection systems. The tests were made using cross-validation [15] with 7 partitions (value generally found in the literature [16]). In each experiment, we used two different sets for training and test. That total numbers of images in the database, we used 6/7 for verification and training parameters and 1/7 to validate the performance of the algorithm.

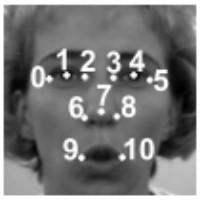


Figure 3. Fiducial points and a human face considered in the experiments.

The experiments were performed by cluster computer with 30 cores, 4GB RAM memory and 2.26 GHz processor. In a brief description, BioID database [7] have 1521 images with frontal human faces in grey scale from 23 different people. The image resolution is 384*x*286 pixels. The faces have change of scale (for instance, some are close to the camera and), lighting, and small rotations. Some people wear glasses, others have beard and/or moustache. The images in this database have manual labels for 20 fiducial points. In this work, we use a BioID subset with 503 images. We used only the frontal face images whose individuals do not wear glasses and do not have moustaches or beards. The faces have scale and lighting variations, and small rotations. Were performed 8316 experiments for C-SVC classifier parameter design and 11 fiducial points.

3.4. Results

The performance of classifier is assessed computing the True Positive (TP) and False Positive rates. We consider a candidate as a true positive when the distance between automatic and manual annotations is smaller than 10% of face intraocular distance (distance between the pupils).

In Figures 4, 5, 6, 7 and 8 are shown the simulations results of the points 04, 05, 07, 08 and 10. The graphics present the average TP, FP and standard deviation rates of all folds of all parameters grid. Each number in grid axis represents one parameter combination of set described in Section 3.1. To assess in order to evaluate the graphs, we consider the best performing regions those with higher TP and smaller FP rates jointly. Note that, for all graphs, the 84 first combinations are best performance of the parameters grid.

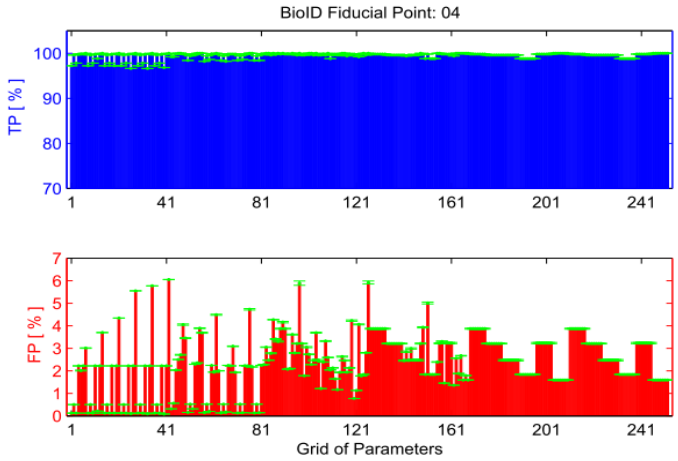


Figure 4. Grid of parameters for the point 04 (left eye pupil) from BioID dataset.

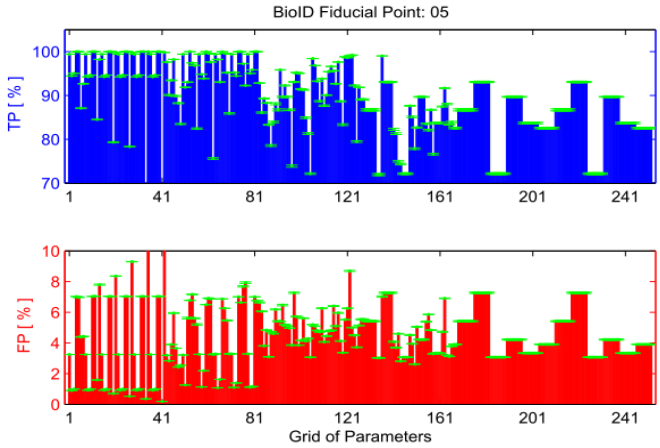


Figure 5. Grid of parameters for the point 05 (outer corner of left eye) from BioID dataset.

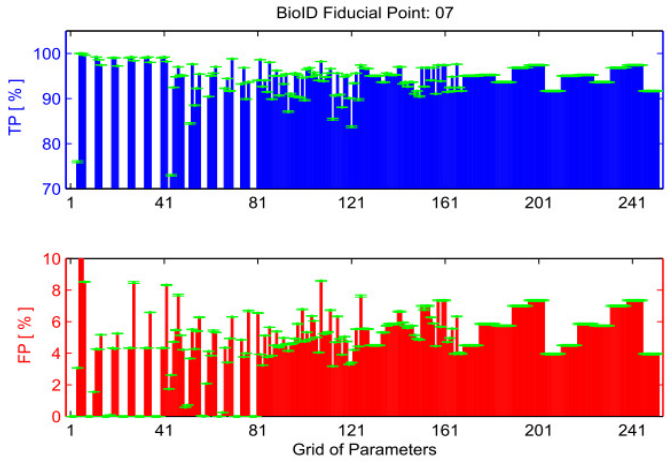


Figure 6. Grid of parameters for the point 07 (tip of nose) from BioID dataset.

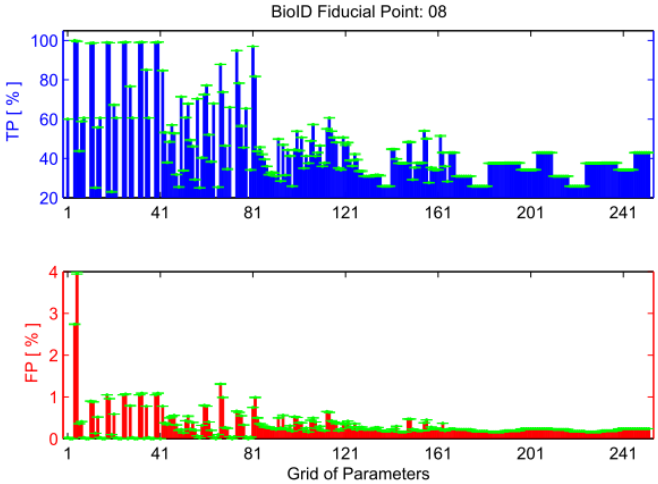


Figure 7. Grid of parameters for the point 08 (left nostril) from BioID dataset.

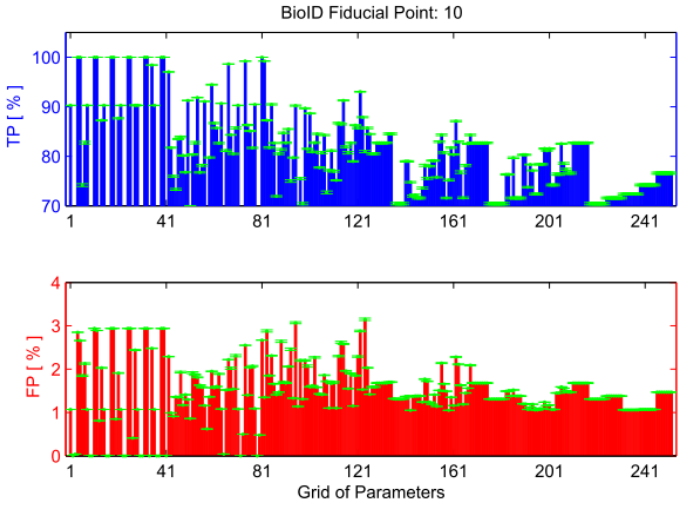


Figure 8. Grid of parameters for the point 10 (left mouth corner) from BioID dataset.

In Table I, we show the results of the proposed method with the best performance of parameters set for 11 fiducial points. The systems used in the methods SVM-L, DF and DF-PCA, are similar those presented in [1]. These results differ due to the method of post-processing. In our research, the post-processing is done by considering a region of 10% around the fiducial point for the calculation of true and false positives (TP and FP). We can see that, in Table I, the proposed method outperforms DF for all fiducial points, except in point 07 and outperforms linear SVM in points 0, 1, 2, 4, 5, 9 and 10. In points 0, 1, 3, 4, 5, 8, 9 and 10, TP rate outperforms DF-PCA method. We consider FP rates less than 2% satisfactory, this fact can be observed in the proposed method, except in fiducial point 09.

Table I. Performance of the proposed system for 11 points using BioID database.

PF	DF-PCA				DF				SVM L				C-SVC (Proposed)			
	FP	σ	TP	σ	FP	σ	TP	σ	FP	σ	TP	σ	FP	σ	TP	σ
0	0.05	0.03	96.6	2.2	0.22	0.03	21.1	4.7	1.09	0.74	92.4	9.8	0.76	0.07	99.2	0.74
1	0.10	0.05	95.4	2.8	0.30	0.04	59.2	6.8	0.32	0.09	99.6	1.0	0.40	0.03	99.8	0.52
2	0.24	0.14	95.0	2.1	0.39	0.03	53.7	6.1	1.58	0.59	93.0	14.3	0.76	0.17	92.3	4.2
3	0.08	0.03	95.0	2.5	0.30	0.03	65.6	13.7	0.46	0.21	96.0	1.7	0.92	0.40	97.6	1.7
4	0.11	0.04	97.8	2.8	0.40	0.02	65.0	9.6	2.49	4.76	97.2	4.2	0.50	0.11	99.9	0.67
5	0.05	0.02	97.4	2.5	0.23	0.04	70.3	12.4	2.27	0.62	98.8	3.1	0.31	0.14	99.6	0.68
6	0.14	0.03	93.2	1.6	0.25	0.04	56.7	9.3	1.40	1.44	99.6	1.0	1.95	0.79	90.5	11.74
7	0.18	0.02	95.8	3.7	0.20	0.01	54.1	8.1	0.06	0.03	47.1	9.6	0.58	0.28	55.1	12.74
8	0.17	0.02	94.8	2.6	0.23	0.01	54.9	5.0	0.40	0.09	95.4	3.9	0.75	0.15	97.0	2.32
9	0.18	0.04	92.8	3.0	0.22	0.01	43.5	5.5	2.91	5.12	96.8	5.5	2.08	0.33	99.3	1.09
10	0.18	0.02	87.4	2.9	0.11	0.04	22.1	6.2	1.47	0.49	98.6	1.9	1.34	0.50	99.2	0.74

4. CONCLUSION

In this paper, we propose the performance evaluation of mathematical formulation SVM, called C-SVC employed in 11 fiducial points detection. In this research, we use parameters of C-SVC unexploited in the literature. The supervised system was implemented using the library in C++ called OpenCV. In this work, considering all the parameters of the system, we performed 8316 experiments. The performance of parameters grid in 5 fiducial points was shown and the best performance was compared with other similar systems in the literature. From this set, two classifiers were used for discriminating filter with and without PCA and one using a linear SVM. Evaluating the performance in terms of TP and FP rates using cross-validation with 7 folds, we identify the proposed system presents good results. Finally, we comment the relevance of this

research in methods that provide fiducial points detection on human faces due to large number of applications.

ACKNOWLEDGEMENT

We would like to thanks Samsung for financial support. This research was supported by FAPEAM and CAPES agency.

REFERENCES

[1] W. S. da Silva Júnior, G. M. Araújo, E. A. B. da Silva and S. K. Goldenstein, "Facial Fiducial Points Detection Using Discriminative Filtering on Principal Components". In: Proceedings of the IEEE International Conference on Image Processing, September, pp. 2681-2684, 2010.

[2] S. Jahanbin, H. Choi and A.C. Bovik, "Passive Multimodal 2-D+3-D Face Recognition Using Gabor Features and Landmark Distances". IEEE Transactions on Information Forensics and Security, vol.6, pp.1287-1304, 2011.

[3] Du, Chunhua and Yang, Jie and Wu, Qiang and He, Xiangjian, "Locating Facial Landmarks by Support Vector Machine-based Active Shape Model". International Journal of Intelligent Systems Technologies and Application, vol. 10, n. 2, pp. 151-170, 2011.

[4] V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag New York, Inc., 1998.

[5] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines". ACM Transactions on Intelligent Systems and Technology,vol.2, n. 3, 2011.

[6] K. Wu and S. Wang "Choosing the Kernel Parameters for Support Vector Machines by the Inter-cluster in the Feature Space". Journal on Pattern Recognition, vol. 42, pp. 710-717, 2009.

[7] "Bioid Database", [Last access in May of 2013]. [Online]. Available in: http://www.bioid.com/

[8] G. M. Araujo, W. S. da Silva Júnior, E. A. B. da Silva, and S. K. Goldenstein, "Facial Landmarks Detection based on Correlation Filter", In: Proceedings of the IEEE International Telecommunication Symposium, Manaus, AM, Brazil, October 2010.

[9] "Cognitec Systems", [Last access in May of 2013]. [Online]. Available in: http://www.cognitec-systems.de/

[10] A. Pradhan, "Support Vector Machine - A Survey". International Journal of Emerging technology and Advanced Engineering, vol. 2, pp. 82-85, 2012.

[11] Viola, P. and Jones, M., "Robust Real-Time Object Detection". International Journal of Computer Vision, vol. 57, n. 2, pp. 137-154, 2001.

[12] Xiaoyang, T. and Triggs, B., Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Transactions on Image Processing, vol. 19, n. 6, pp. 1635-1650, 2010.

[13] G. R., Arce, J. L., Paredes and J. Mullan, "Nonlinear Filtering for Image Analysis and Enhancement". In: A. L. Bovik (Ed.), Handbook of Image & Video Processing, Academic Press, 2000.

[14] "Open Computer Vision Library - OpenCV", [Last access in May of 2013]. [Online]. Available in: http://opencv.org/

[15] Reeves, S., "A Cross-Validation Framework for Solving Image Restoration Problems". Journal of Visual Communication and Image Processing, vol. 3, pp. 433-445, 1992.

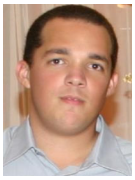
[16] Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", In: Proceedings of the International Joint Conference on Artificial Intelligence, August, pp. 1137-1143, 1996.

Authors

Luiz Eduardo S. e Silva received the B.S. degree in electrical engineering at the Federal University of Amazonas (UFAM), Manaus, AM, Brazil, in 2011. Currently, he is a M.Sc. student at the same University and he is working with computer vision. Support Vector Machine is a technique applied in his pattern recognition researches, which try to classify elements of interest. He is Professor of Nokia Foundation in Electrical Circuits e Telecommunications. Automatic Control, Digital Signal Processing, Machine Learning and Mathematical Morphology are also his interest areas.

Pedro Donadio de T. Júnior is a researcher at Federal University of Amazonas – Brazil. He received his Eletrical Engineer degree in 2009 at Federal University of Amazonas. Currently, he is a M.Sc. student at the same University and he is working with computer vision. Discriminative filtering is a technique applied in his pattern recognition researches, which try to improve the filters performance by parallel processing. Automatic Control and noise cancellation are also his interest areas.

Kenny V. Santos is a graduate student in Electrical Engineering from Federal University of Amazonas. He received his bacharel degree in Electrical Engineering at Federal University of Amazonas in 2008. Currently he works in the fields of computer vision and pattern recognition, where his main interest is applications of digital signal processing in pattern recognition.



Waldir S. S. Júnior received the B.S. degree in electrical engineering at the Federal University of Amazonas (UFAM), Manaus, AM, Brazil, in 2000, and the M.S. degree in electrical engineering at the Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, RJ, Brazil, in 2004 and Ph.D. degree in electrical engineering at the Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, RJ, Brazil, in 2010. Since 2006, he has been with the Federal University of Amazonas, as Full Professor. His research interests are in the fields of data compression, as well as in mather gtay

INTENTIONAL BLANK

A REAL-TIME H.264/AVC ENCODER&DECODER WITH VERTICAL MODE FOR INTRA FRAME AND THREE STEP SEARCH ALGORITHM FOR P-FRAME

Dr. Mohammed H. Al-Jammas[1] and Mrs. Noor N. Hamdoon[2]

[1]Deputy Dean/College of Electronics Eng./University of Mosul/Mosul-Iraq
dr_mohammed_al_jammas@uomosul.edu.iq
[2]Department of Electrical Engineering /College of Engineering/Mosul–Iraq
Noornawaf81@yahoo.com

ABSTRACT

The video coding standards are being developed to satisfy the requirements of applications for various purposes, better picture quality, higher coding efficiency, and more error robustness. The new international video coding standard H.264 /AVC aims at having significant improvements in coding efficiency, and error robustness in comparison with the previous standards such as MPEG-2, H261, H263,and H264. Video stream needs to be processed from several steps in order to encode and decode the video such that it is compressed efficiently with available limited resources of hardware and software. All advantages and disadvantages of available algorithms should be known to implement a codec to accomplish final requirement. The purpose of this project is to implement all basic building blocks of H.264 video encoder and decoder. The significance of the project is the inclusion of all components required to encode and decode a video in MatLab .

KEYWORDS

H264/AVC , Intra frame (I-frame) , Inter frame (P-frame)

1. INTRODUCTION

A Digital video compression is an important techniques that enables efficient transmission bandwidth and storage space of multimedia. The H.264/AVC is a standard video coding that developed to achieve significant improvements, in the compression performance, over the existing standards. In fact, the high compression performance comes mainly from the prediction techniques that remove spatial and temporal redundancies. To remove spatial redundancy, H.264/AVC intra prediction supports many prediction modes to make better prediction. Inter prediction is enhanced by motion estimation (ME) to remove more temporal redundancy. However, the H.264/AVC coding performance comes at the price of computational complexity[1].

H.264/AVC intra encoding achieve higher compression ratio and picture quality compared with the latest still image coding standard JPEG2000].intra prediction is the first process of advanced video coding standard. It predicts a macro block by referring to its previous macro blocks to

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 33–44, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4404

reduce spatial redundancy. Intra prediction supports nine modes for 4x4 block and four modes for 16x16 blocks[2].

H.264 is an open, licensed standard that supports the most efficient video compression techniques available today. Without compromising image quality, an H.264 encoder can reduce the size of a digital video file by more than 80% compared with the Motion JPEG format and as much as 50% more than with the MPEG-4 Part 2 standard. This means that much less network bandwidth and storage space are required for a video file. Or seen another way, much higher video quality can be achieved for a given bit rate[3].

And also entered as an adjunct of the kind of evolutionary in public services such as video storage on the Internet and telecommunications companies and surveillance cameras used in industrial plants , and due to accept this kind of decoder over a wide range of frames during the second (60/30/25 (fps)) has been expanding in applications control of highways, airports , Most of the controversy over the techniques used to process coded information is how fast and accurate images and video after the code process it possible to re-information fully if it is reduced to half?. Will be answered by this research, which includes the representation of the coded H264 and decode of the file with a high level using the MATLAB simulation software to achieve complete system of coded data and return it in the same efficient used for the original system.

2. THE ENCODING PROCESS

H.264 encoder works on the same principles as that of any other codec. Figures (eg, Figure 1) shows the basic building blocks of H.264 video codec.

The input to the encoder is generally an intermediate format stream, which goes through the prediction block; the prediction block will perform intra and inter prediction (motion estimation and compensation) and exploit the redundancies that exist within the frame and between successive frames. The output of the prediction block is then transformed and quantized. An integer approximate of the discrete cosine transform is used (DCT) for transformation. It uses 4x4 or 8x8 integer transform, and outputs a set of coefficients each of which is a weighted value for a standard basis pattern. The coefficients are then quantized i.e. each coefficient is divided by an integer value. Quantization reduces the precision of the transform coefficients according to the quantization parameter (QP). Typically, the result is a block in which most or all of the coefficients are zero, with a few non-zero coefficients. Next, the coefficients are encoded into a bit stream. The video coding process creates a number of parameters that must be encoded to form a compressed bit stream[4]. These values include:

- Quantized transform coefficients.
- Information to re-create prediction.
- Information about the structure of compressed data and the compression tools used under encoding.

These parameters are converted into binary codes using variable length coding or arithmetic coding. Each of these encoding methods produces an efficient, compact binary representation of the information. The encoded bit stream can now be transmitted or stored.

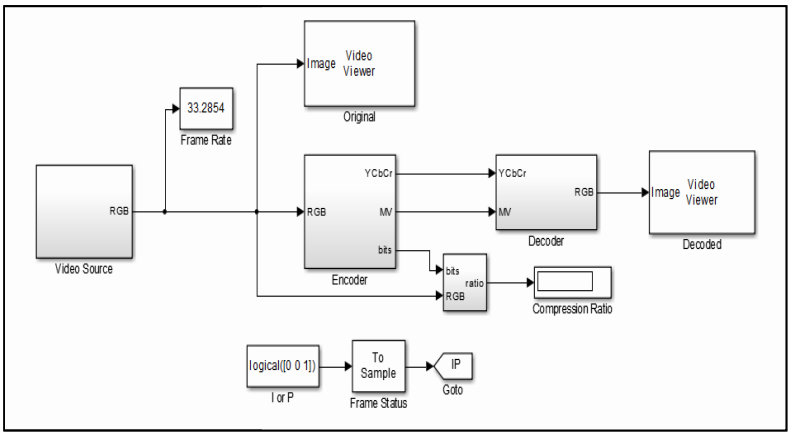


Figure 1. Encode and Decode circuit

3. VIDEO COMPRESSION WORKS

Video compression is about reducing and removing redundant video data so that a digital video file can be effectively sent and stored. The process involves applying an algorithm to the source video to create a compressed file that is ready for transmission or storage. To play the compressed file, an inverse algorithm is applied to produce a video that shows virtually the same content as the original source video. The time it takes to compress, send, decompress and display a file is called latency. The more advanced the compression algorithm, the higher the latency, given the same processing power. A pair of algorithms that works together is called a video codec (encoder/decoder).

Video codec's that implement different standards are normally not compatible with each other; that is, video content that is compressed using one standard cannot be decompressed with a different standard. For instance, an MPEG-4 Part 2 decoder will not work with an H.264 encoder. This is simply because one algorithm cannot correctly decode the output from another algorithm but it is possible to implement many different algorithms in the same software or hardware, which would then enable multiple formats to be compressed.

Different video compression standards utilize different methods of reducing data, and hence, results differ in bit rate, quality and latency. Results from encoders that use the same compression standard may also vary because the designer of an encoder can choose to implement different sets of tools defined by a standard. As long as the output of an encoder conforms to a standard's format and decoder, it is possible to make different implementations.

This is advantageous because different implementations have different goals and budget. Professional non-real-time software encoders for mastering optical media should have the option of being able to deliver better encoded video than a real-time hardware encoder for video conferencing that is integrated in a hand-held device. A given standard, therefore, cannot guarantee a given bit rate or quality. Furthermore, the performance of a standard cannot be properly compared with other standards, or even other implementations of the same standard, without first defining how it is implemented. A decoder, unlike an encoder, must implement all the required parts of a standard in order to decode a compliant bit stream. This is because a standard specifies exactly how a decompression algorithm should restore every bit of a compressed video [3]>

4. H.264 LEVELS

The Group focused joint development in determining work H.264 to find a solution is simple and flexible could include various applications through the use of a single standard, as is the case in video standards, etc., and is this flexibility in the provision of facilities for several profiles (represented groups of algorithms for the pressure data) and levels (level private suite of applications). The standard includes the following seven sets of capabilities, which are referred to as profiles, targeting specific classes of applications

- Baseline profile (BP): It is primarily for applications with low cost resources with physical entity and this file is used widely in applications of mobile devices and applications to transfer data, such as video conversations.
- Main Profile (MP): This file is used for broadcast applications and storage, and the importance of this file vanished when placing a high level to include those applications.
- Extended Profile (XP): means the image streaming video, this file has the ability of a relatively high pressure, and some additional possibilities to avoid data loss.
- High Profile (HiP): The primary file used for broadcast applications and disk storage, particularly in the application of high-definition television and, for example, applications in the HD DVD and Blu-ray .
- High 10 Profile (Hi10P): This file adds support for the previous file in decryption process where they can processing 10 bits per sample of image resolution to decode the data.
- High 4:2:2 Profile (Hi422P): used in professional applications that use interlaced video, this file is used in the previous file basis (Hi10P) and added his coordination shorthand chromatography with 4:2:2 sample while it uses 10 bits per sample per unit decoding and gives the same quality.
- High 4:4:4 Predictive Profile (Hi444PP): used in the applications non are for the loss of data was added to the previous file format support reduction chrome quality of 4:4:4 and support for samples up to 14 bits per sample, and also supports encoding videos and pictures with tri-color separate.

5. TYPES OF FRAMES

H.264 consists of several different types of frames, such as (I-P-B), and can be used for encryption to get the required efficiency below illustrate the theoretical formula for each quality of frames.

- (I-intra frame) Is an autonomous framework which can encrypt and decrypt independently without need for another picture as a source of information retrieval, the first image of the video is for this type of frame, and the (I-frame) is the starting point for the video display as well as his importance in information retrieval synchronization if any damage in transport stream bit (bit stream), the flaw in this window that consumes the largest possible number of bits for encryption because it takes the window image full but on the other hand, the error rate is low. Encryption method for this type of window has two properties, depending on the method of dividing the cluster either ((16x16) or (4x4)) but in General is to convert the frame version (RGB)format (YCbCr) and separated from the other components of the final representation and is treated with a single image, so the representation of video format ((4:2:0) YCbCr) is to reduce the sensitivity of the eye where the eye responds to brightness by colors so the component (Y) represents the symbol of brightness luminance while (CbCr) represents the color (chrominance) taken the element Y with full size while the rest of the elements are reduced by deciding to half

the amount of action in the element size (Y) is (16x16) , the rest of the elements are the size of (8x8), this means that embedded type of encryption key encryption process. Encryption process as previously mentioned it is dependent on frame division, divides the frame into multiple blocks of size (16x16) and has (4) types of encryption as shown in Figures (eg, Figure 2).

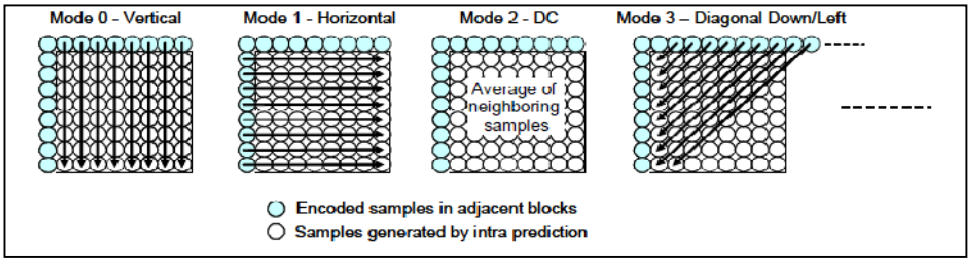


Figure 2. The types of patterns to divide block (16x16)

But the case of the split window to (4x4) , it has (9) types, as in Figures (eg, Figure 3)

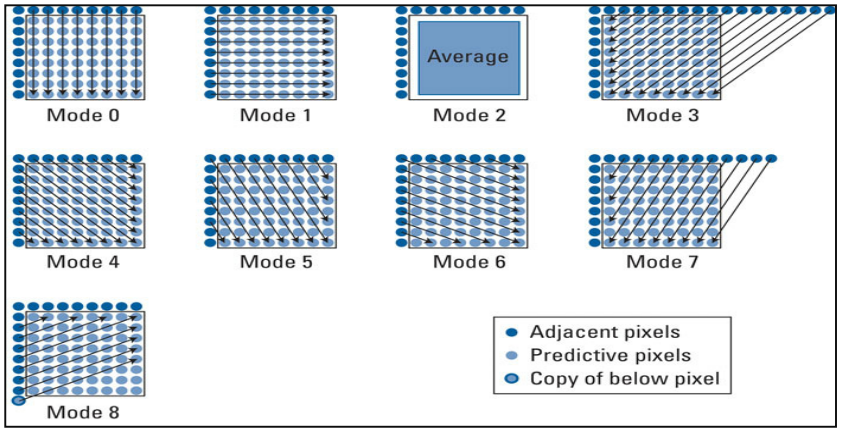


Figure 3. Types of Patterns to divide block (4x4)

Choose a style for the adoption application and competence required for the encryption process and the admissibility of the error rate, in most cases the amount of the error rate of the video is of a higher flexibility compared to the error rate in the case of a single image.

- (P-Inter Frame) Predictive Inter Frame: is derived from the current frame to the video sequence frame by reducing the time between frames increase unlike previous quality work only within the space of pixels, the principle of its work essentially compare the block of
- the current window with the block of the previous frame and the centre of block is search for match, this called (matching block), all theories have one and is the best possible match and this is called motion estimation (ME), after finding the best match, we put the block of the original block and the remaining known as compensation (motion compensation), the link between the location of the current block with original block is the transmission (motion vector (MV)) shown in Figures (eg, Figure 4).

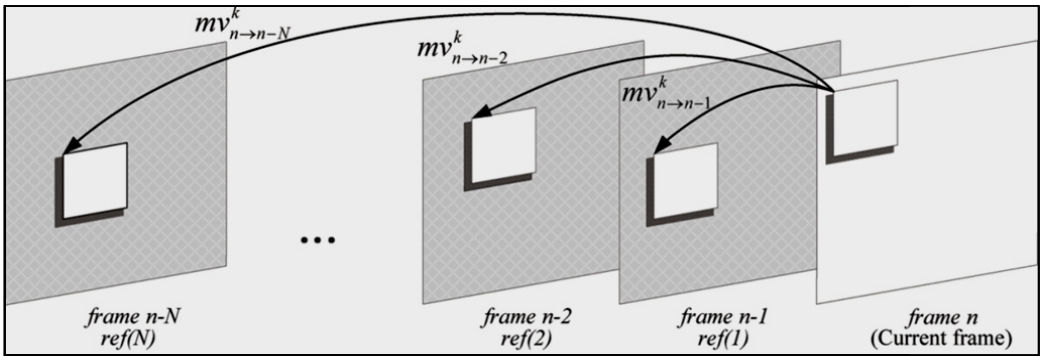


Figure 4. The basic idea to represent the predictive inter frame (P-frame)

- B-frames (Bi-predictive inter frame), this type of frame be intermediate between (I, and B frames) used at high levels for perfect efficiency but complex where the highest of qualities as follows based on the comparison between more than one source for block, meaning most forecasts from original source and source is expected as in Figures (eg, Figure 5).

When retrieving the information in the decryption process is (I-frame), the former to decrypt followed by (B-P frames) if used, the decryption depend one upon the other in the information retrieval of the original frame. H264 has several ways in the encryption that uses encryption (I-frame) or use (I&P ,(I&B&P)) and each method has its qualities, if you use the first method, the quantity of bits encoded be high compared with other cases but the error rate low because all the encrypted individually without relying on the previous window, this method is used in some applications that need high resolution cameras also in prisons and banks to get clearer picture during the up seizing process as in Figures (eg, Figure 6) .

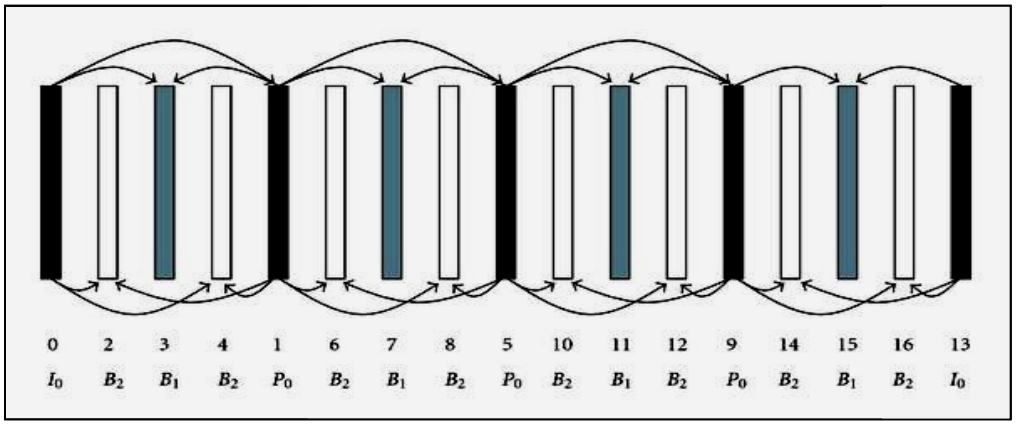


Figure 5. The representation frame (B)

Figure 6. Video encoding using spatial locations for pictures

But if you use the second method as shown in Figures (eg, Figure 7) they have characteristics that they reduce the number of bits encoded and the error rate is acceptable and this method is used in video compression in general and cameras , In the third grade are more complex than both methods but with a reduced data encrypted and contains a higher delay method because search matches more than one source.

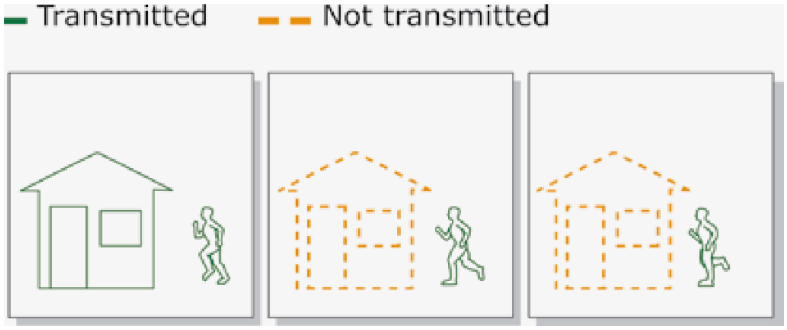


Figure 7. The first image is a frame (Frame-I) and encodes a single, the second and third picture only encrypts mobile part

6. SIMULATE ENCRYPTION AND DECRYPTION USING MATLAB

Previously we mentioned about the quality of the frames and how to process encoded in H.264 encoding process is divided into two ways:

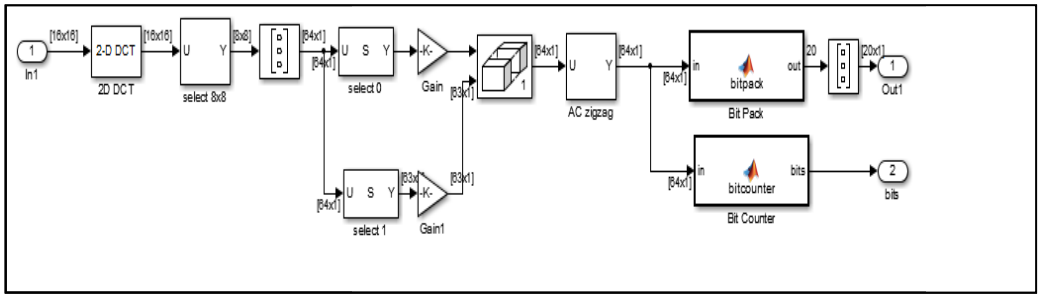
6.1 Encryption Process (I-frame)

I-frame is spatially encoded with a specific kind of styles in this research is the use of block size (16 x16), divided into (8x8), Tables (eg, Table 1) shows real-time to encrypt and decrypt the frame interconnection pattern within the first mode (Vertical mode) has been achieved through Figures (eg, Figure 8). Each block is handled independently of and separately from other 8x8 block are taken separately and applied the first style (vertical mode) and enters the cosine (DCT). This conversion generates a representation of each block of (8x8) in the frequency domain rather than the spatial domain. The resulting values of DCT process usually consists of a few large values and many small values represent the relative sizes of these transactions how important each block of information in the decryption stage of min after this operation involved transactions to bring it shown in Figures (eg, Figure 8).

Table 1. Time encode and decode for I-frame

Video / 30 fps	Time code /(sec/frame)	Time decode (sec/frame)
Forman(288x352)	0.0383	0.034
Vipmen(160x120)	0.00052	0.00104
Bride(720x1280)	0.468	0.159
Piano(352x480)	0.054	0.04

Figure 8. The encrypted representation service



6.2 Encryption Process (P-frame)

We mentioned already that this type of encryption is encryption And this means that it uses the correlation between the current frame and the frame (or frames) to achieve compression. Tables (eg, Table 2) Shows real-time to encrypt and decrypt for P-frame, Temporal encryption is achieved using vector motion (motion vectors.) the basic idea is to match each block in the current window size (16x16) pixel in the frame of reference, the match here can be calculated in many ways, but in H.264 uses a simple and more common is the sum of absolute differences (SAD) offset from the current transaction to the last movement is represented in reference to two values (horizontal, vertical vectors vector) are searched to find the best tankers in the selected area to search only when a less value of the total difference match between blocks this process is carried out under several algorithms one three-step algorithm to search for blocks to neighbourhoods Match blocks beginning with predictive search steps greater than half the search steps used in the previous methods and mechanism of action as follows:

- Compared with nine points in each step, three score and six points on the sides.
- Under search by one space after each step and the search stops when the size of the search space by one pixel.
- At each step a new search moving space research to find the best match for the Center block from the previous search, blue circles of number one as in Figures (eg, Figure 9) represent the first step of three steps when creating less differences between the current and previous bloc begin the second step of the green circles and also look on the differences is less up to the level of a single pixel of Orange and last circles represent the end of the search for this theory.

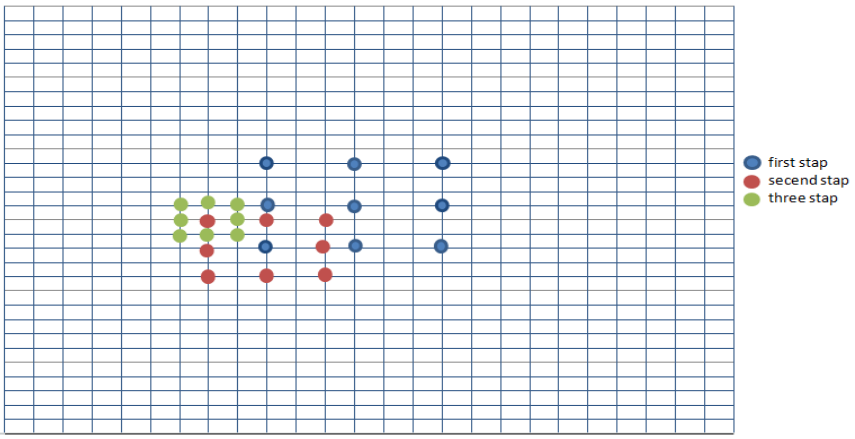


Figure 9. The theory of the three steps

Motion vector is a simple way to move a lot of information as shown in Figures (eg, Figure 10), but not always give an exact match to give the best quality, taking output subtraction between the original frame and block the cluster framework forecast output encrypted as shown in Figures (eg, Figure 11), the encryption process for residual image is similar to the encryption (I-frame), but the difference is in the process of rounding and through practical results as in Figures (eg, Figure 12) and found that the compression ratio is 70%of the original size.

Table 2. Time encode and decode for P-frame

Video 30 fps	Time code (sec/frame)	Time decode (sec/frame)
Forman(288x352)	0.0013	0.0011
Vipmen(160x120)	0.00032	0.001
Bride(720x1280)	0.0161	0.0054
Piano(352x480)	0.00186	0.00173

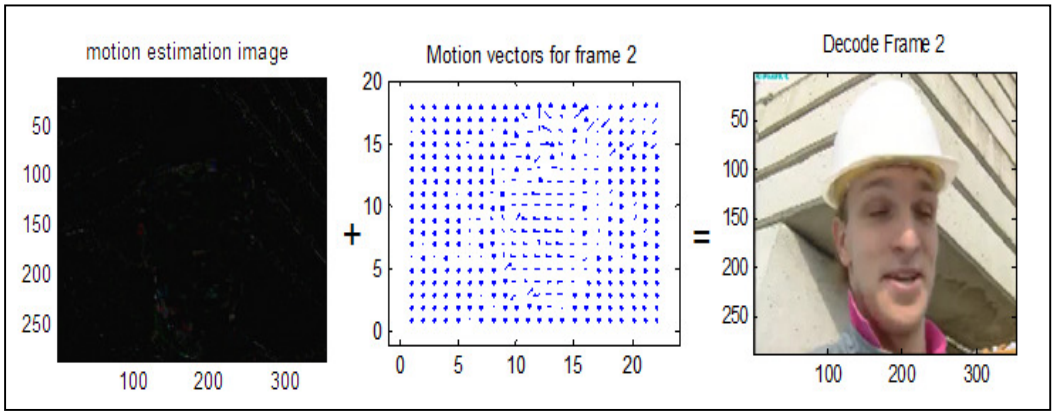


Figure 10. The motion vectors for Residual Image

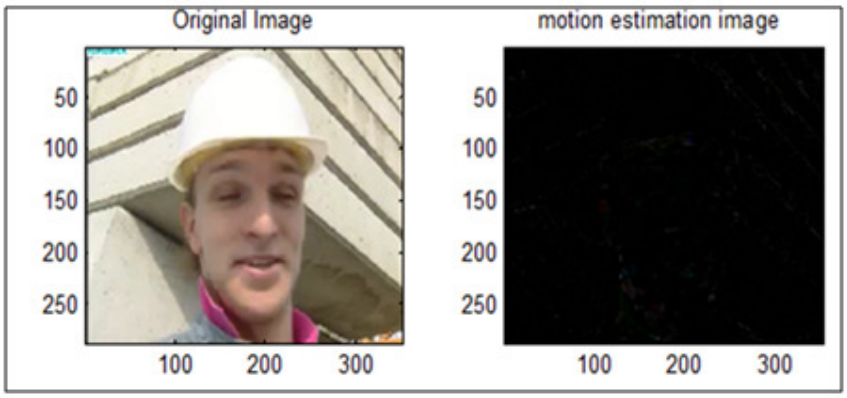


Figure 11. The Residual Image

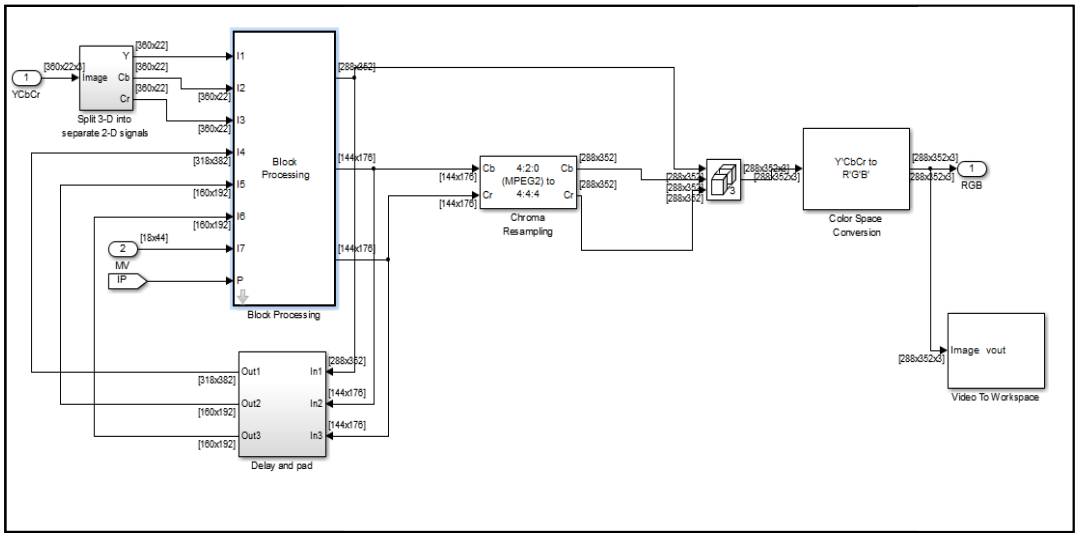


Figure 12. The Decode Circuit

In Figures (eg, Figure 13) shows the encryption and decryption of video sequences Forman and note that in the case of image encryption and decryption have same properties.

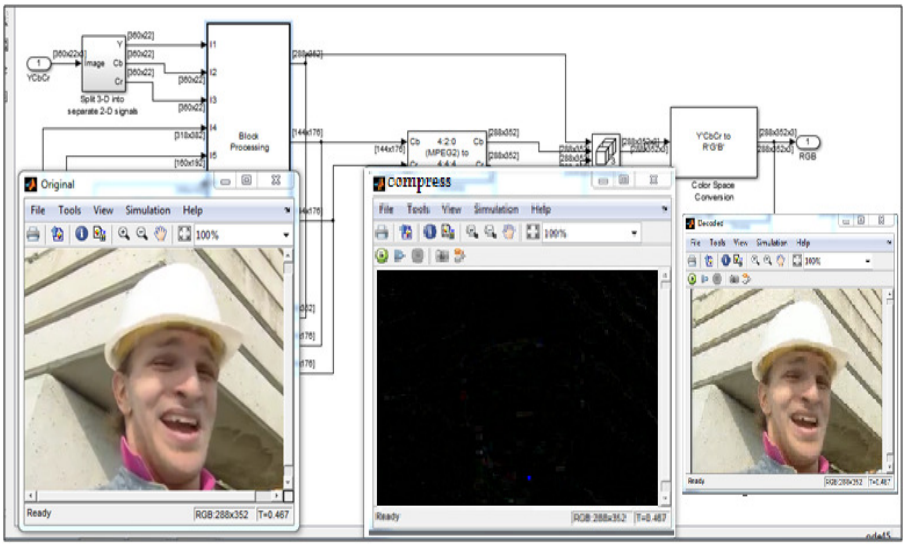


Figure 13. The Original Foreman Sequence and Return Within Three-Step algorithm

7. CONCLUSIONS

DisplaysH.264 A major step forward in the field of video compression technology. and provides techniques which enable better compression efficiency, due to more accurate forecasting capabilities, as well as improving the ability to minimize errors. It provides new possibilities for creating video encoders that managed to get high quality video and high frame rates at per second and higher resolution at bitrates (compared to the preceding criteria), and through the practical results of the MATLAB simulation was found that data compression during real time up to(70%) of the video size Original part time implementation 261.89 s and time clk (4ns) and highest value can be obtained for a reference to the noise up to (45 db) also shown in Figures (eg, Figure 14).

Simulink Profile Report: Summary

Report generated 14-Aug-2013 16:15:47

Total recorded time:	261.89 s
Number of Block Methods:	623
Number of Internal Methods:	5
Number of Model Methods:	9
Number of Nonvirtual Subsystem Methods:	71
Clock precision:	0.00000004 s
Clock Speed:	2501 MHz

To write this data as untitlednoorProfileData in the base workspace click here

Function List

Name	Time		Calls	Time/call	Self time	
simulate(untitlednoor)	261.89447880	100.0%	1	261.8944788000000	0.00000000	0.0%
simulationPhase	248.94639580	95.1%	1	248.9463958000000	4.10282630	1.6%
untitlednoor.Outputs.Major	242.37875370	92.5%	302	0.8025786546358	0.28080180	0.1%
untitlednoor/Encoder/Block Processing (AtomicSubSystem.Outputs.Major)	133.25605420	50.9%	301	0.4427111435216	0.04680030	0.0%
untitlednoor/Encoder/Block Processing/Block iterator (ForIteratorSubSystem.Outputs.Major)	133.20925390	50.9%	301	0.4425556607973	51.55833050	19.7%
untitlednoor/Decoder/Block Processing (AtomicSubSystem.Outputs.Major)	107.21948730	40.9%	301	0.3562109212625	0.00000000	0.0%
untitlednoor/Decoder/Block Processing/Block iterator (ForIteratorSubSystem.Outputs.Major)	107.20388720	40.9%	301	0.3561590936877	44.44468490	17.0%

Figure 14. Simulink profile for H264/AVC

H.264 It is expected to replace other compression standards and methods used today, and form became H.264 more widely available species in network cameras, video encoders and video management software, designers of systems at the present time, the network video products support both H.264 and Motion JPEG is perfect for maximum flexibility and possibilities of integration.

REFERENCE

[1] A. Ben Atitallah, H. Loukil , and N. Masmoudi, FPGA DESIGN FOR H.264/AVC ENCODER, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.1, No.5, October 2011, pp 119-138.

[2] Manjanaik.N, Dr.Manjunatha.R, Development of Efficient Intra Frame Coding in Advanced Video Standard Using Horizontal Prediction Mode, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 2, February 2013, pp 192-196.

[3] H.264 video compression standard, New possibilities within video surveillance, Axis Communications, White paper, 2008.

[4] Amruta Kiran Kulkarni, Implementation of Fast Inter-Prediction Mode Decision in H.264/AVC Video Encoder, Master Thesis, May 2012.

Authors

Mohammed H. AL-Jammas (Jun'02) born in 1966 in Mosul-Iraq. He awarded BSc in Electronic and Communication Engineering from the University of Mosul, Mosul-Iraq in 1988. Next, he awarded the MSc in Communication from the University of Mosul, Mosul-Iraq in 1994, and PhD in Computer Engineering from the University of Technology, Baghdad-Iraq in 2007. From 2002-2006, Dr. Mohammed worked with the University of Technology in Baghdad. From 2007, he acts as an Assistance dean of the College of Electronics Engineering at the University of Mosul.

Through his academic life he published over 7 papers in field of computer engineering, and information security.

Noor N. AL-Sawaf (August'08) born in 1988 in Mosul-Iraq. she awarded BSc in Electronics Engineering from the University of Mosul, Mosul-Iraq in 2010. Next, she awarded the MSc in Electrical from the University of Mosul, Mosul-Iraq in 2014.

Through her academic life she published 1 papers in field of image and video compress .

ESTIMATING THE EFFORT OF MOBILE APPLICATION DEVELOPMENT

Laudson Silva de Souza[1] and Gibeon Soares de Aquino Jr.[1]

[1]Department of Informatics and Applied Mathematics,
Federal University of Rio Grande do Norte, Natal, Brazil
{laudyson,gibeon}@gmail.com

ABSTRACT

The rise of the use of mobile technologies in the world, such as smartphones and tablets, connected to mobile networks is changing old habits and creating new ways for the society to access information and interact with computer systems. Thus, traditional information systems are undergoing a process of adaptation to this new computing context. However, it is important to note that the characteristics of this new context are different. There are new features and, thereafter, new possibilities, as well as restrictions that did not exist before. Finally, the systems developed for this environment have different requirements and characteristics than the traditional information systems. For this reason, there is the need to reassess the current knowledge about the processes of planning and building for the development of systems in this new environment. One area in particular that demands such adaptation is software estimation. The estimation processes, in general, are based on characteristics of the systems, trying to quantify the complexity of implementing them. Hence, the main objective of this paper is to present a proposal for an estimation model for mobile applications, as well as discuss the applicability of traditional estimation models for the purpose of developing systems in the context of mobile computing. Hence, the main objective of this paper is to present an effort estimation model for mobile applications.

KEYWORDS

Software Engineering, Software Quality, Estimating Software, Systematic Review, Mobile Applications, Mobile Computing

1. INTRODUCTION

Computing is becoming increasingly present in people's lives and currently in a much more intense and accelerated way due to the rise of the use of mobile technologies in the world, such as mobile phones, smartphones and tablets, all connected to mobile networks, which are increasingly more present in many places and with better speeds. We are facing a new technological scenario that is changing old habits and creating new ways for the society to access information and interact with computer systems [1], [2] and [3].

The ITU (International Telecommunication Union) estimates that there are more than 6 (six) billion mobile clients worldwide. According to Gartner, 1.75 billion people own mobile phones with advanced capabilities; he also foresees further growth in the use of this technology in the upcoming years [4]. There is a global trend towards the increase of the number of users connected to the network via mobile devices which, consequently, will create an increasing demand for information, applications and content for such equipments. New ways to use existing information

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 45–63, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4405

systems are emerging. In particular, systems that were once accessed via web interfaces through personal computers physically located in offices, universities or homes are providing new ways to access from mobile devices which, in turn, have different requirements and capabilities than the personal computers.

Thus, we realize that traditional information systems are undergoing a process of adaptation to this new computing context. Current developments, including the increase of the computational power of these new devices, in addition to the integration of multiple devices on a single one and lined up with the change of the users' behavior, actually create a new environment for the development of computing solutions. However, it is important to note that the characteristics of this new context are different. They present new resources and, thereafter, new possibilities [5], [6], [7] and [8], as well as introduce non-existing restrictions in conventional systems [9] and [10].

The fact is that this new technological scenario that is emerging with new requirements and restrictions requires a reevaluation of current knowledge about the processes of planning and building software systems. These new systems have different characteristics and, therefore, an area in particular that demands such adaptation is software estimation. The estimation processes, in general, are based on characteristics of the systems, trying to quantify the complexity of implementing them. For this reason, it is important to analyze the methods currently proposed for software projects estimation and evaluate their applicability to this new context of mobile computing.

Hence, the main objective of this paper is to present a proposal for an estimation model for mobile applications, as well as discuss the applicability of traditional models used in estimation of information systems for the purpose of the development of systems in the context of mobile computing. In this work, the main estimation methods that exist now will be analyzed, the specific characteristics of mobile systems will be identified and an adaptation of a estimation method that exists in this context will be proposed. For this, the paper is organized as follows: Section II discusses the main estimation methods with the purpose of identifying those that address development in this new scenario of mobile computing. Section III summarizes the main articles that discuss and point out the particular characteristics of mobile computing systems. In Section IV, the summary of the results of a survey on the characteristics of mobile development will be presented. In Section V, the problem addressed will be discussed. In Section VI, the objective of the work will be presented. Finally, in Section VII, the final considerations resulting from this research.

2. MAIN ESTIMATION METHODS

In order to identify how the traditional estimation methods could address the characteristics of the systems, a literature review on the main estimation methods was performed. The methods identified in the survey can be seen in Table 1.

Table 1. Main Estimation Methods.

Year	Method	Author
1979	Function Point Analysis (FPA)	Albrecht [11]
1981	COnstructive COst MOdel (COCOMO)	Barry W. Boehm 's [12]
1982	DeMarco's Bang Metrics	Tom DeMarco [13]
1986	Feature Points	Jones [14]
1988	Mark II FPA	Charles Symons [14]
1989	Data Points	Harry Sneed [15]
1990	Netherlands Software Metrics Users Association (NESMA) FPA	The Netherlands Software Metrics Users Association [16]
1990	Analytical Software Size Estimation Technique-Real-Time (ASSET-R)	Reifer [17]
1992	3-D Function Points	Whitmire [18]
1993	Use Case Points UCP	Gustav Karner [19]
1994	Object Points	Banker et al. [20]
1994	Function Points by Matson, Barret and Mellichamp	Matson, Barret e Mellichamp [21]
1997	Full Function Points (FFP)	University of Quebec in cooperation with the Software Engineering Laboratory in Applied Metrics [18]
1997	Early FPA (EFPA)	Meli, Conte et al. [22]
1998	Object Oriented Function Points – (OOFPs)	Caldiera et al. [23]
1999	Predictive Object Points – (POPs)	Teologlou [24]
1999	Common Software Measurement International Consortium (COSMIC) FFP	Common Software Measurement International Consortium (COSMIC) [25]
2000	Early & Quick COSMIC-Full Function Points (E&Q COSMIC FFP)	Meli et al. [26]
2000	Kammelar's Component Object Points	Kammelar [27]
2001	Object Oriented Method Function Points – (OOmFP)	Pastor and his colleagues [28]
2004	Finnish Software Metrics Association FSM	The Finnish Software Metrics Association (FiSMA) [29]

Table 1 displays in chronological order the main estimation methods, showing the year of creation, the name of the method and the author of it. For each of the methods identified, a summary of the description and characteristics was developed, as can be seen in the following items.

- Function Point Analysis (FPA) is a method of measurement of size in function points based on what the user notices, taking into account the implemented functionality. This method is independent of technology and was designed to estimate business information systems. Its characteristics are: estimates the functionality requested by the user, calculating its size and cost; estimates projects of development and maintenance of softwares, regardless the technology used; estimates the functionality received by the user, after its development, in order to check if its size and costs are in accordance with what was estimated [11].
- COnstructive COst Model (COCOMO) is a cost model for the process of planning and implementation of software projects. Its characteristics are: a framework for communication of business decisions between everyone involved in the project, making it possible to estimate the effort and cost and follow up on the schedule of the project [12].

- DeMarco's Bang Metrics is a method which consists in assigning an independent measure of functionality based on the structured analysis of the project and design notation. Its characteristics are: estimates the size of the software from the structured description of the components during the process of gathering requirements [13].
- Feature Points is an adaptation of Albrecht's FPA. Its characteristics are: changes the weights of the components of the original function point through an additional algorithm. This algorithm is based on a set of rules created to solve a computational problem. But due to the lack of standardization for the use of the algorithm, it has fallen into disuse [14].
- Mark II FPA is a method certified by ISO as an international standard, designed to estimate business information systems. It was created to try to remedy the shortcomings identified in the FPA method. Its characteristics are: estimates the processing of information, in which it treats a software as a set of logical operations and examines its functional size, counting the input data types, types of data entity referenced and output data elements types for each logical transaction [14].
- Data Points is a method similar to the function point. Its characteristics are: it was designed for object-oriented development and is applied based on what the user sees [15].
- Netherlands Software Metrics Users Association (NESMA) is a method certified by ISO as an international standard, designed for counting function points of the software, similar to the FPA method. Its characteristics are: External Input, External Output, External Consultation, Internal Logical File and External Interface File. The difference is that NESMA FPA gives more concrete guidelines, due to its three types of counts: detailed, estimated and indicative [16].
- Analytical Software Size Estimation Technique-Real-Time (ASSET-R) is a method designed to estimate the size of data processing in real-time systems. Its characteristics are: is similar to FPA, but takes into consideration factors, process interfaces and execution modes [17].
- 3-D Function Points is a method designed to estimate the size of real-time systems. Its characteristics are: independent of technology, is similar to the FPA but uses two new approaches, the transformations and the transitions. However, its use is impracticable in the initial design phase, for it requires a greater degree of detailing of the system [18].
- Use Case Points UCP is a method developed to measure projects in its initial phase, during the survey of requirements based on use cases of the system being developed. It was designed to measure object-oriented software projects and, to determine some factors, it takes into account the experience of the developers [19].
- Object Points is a method similar to the FPA, but it counts the objects instead of the functions. Its characteristics are: better defines the complexity adjustment factor and also adds to its count the percentage of reused code [20].
- Function Points by Matson, Barret and Mellichamp is an adaptation of the Albrecht's FPA. Its characteristics are: a linear combination of five basic components of the software, inputs, outputs, master files, interfaces and surveys [21].
- Full Function Points (FFP) is a method designed to estimate real-time and embedded systems. Its characteristics are: when the measurement is being made, he adds six more types of function in comparison to the FPA [18].
- Early FPA (EFPA) is a method similar to the FPA. Its characteristics are: estimates the functional size of the software quickly using a standard of rules different than the FPA's, which allows it to identify software objects at different levels [22].
- Object Oriented Function Points - (OOFPs) is an adaptation of the FPA, used to estimate object oriented softwares. Its characteristics are: instead of using logical files and operations like the FPA, it uses classes and methods. Its counting also takes into account the reuse of code [23].
- Predictive Object Points is a specific method for object-oriented softwares. Its characteristics are: estimates taking into account the classes, the behaviors of these classes and the effects of these behaviors on the rest of the software [24].

- Common Software Measurement International Consortium (COSMIC) FFP was designed based on the FPA, but directed to applications in real time and multi-platform softwares. Its characteristics are: estimates the development effort, evolves software quality, compares specified systems in different languages, in terms of productivity and cost maintenance, considering concepts such as functional requirements of users, software users, layers and boundaries [25].
- Early & Quick COSMIC - Full Function Points (E&Q COSMIC FFP) is an estimation method for the size of the software, which was based on COSMIC FFP. Its characteristics are: classification of the types of processes into functional process, general process or macro-process [26].
- Kammelar's Component Object Points was designed based on the FPA but was directed to object-oriented systems. Its characteristics are: takes into account two types of counting elements, users' domain elements (users' functional requirements) and system elements, including services, classes, operations and transformations [27].
- Object Oriented Method Function Points - (OOmFP) was designed based on the rules of the FPA, but was directed to object-oriented systems. Its characteristics are: the classes are considered internal logical files and concepts such as inheritance and aggregation are also relevant to the estimation [28].
- Finnish Software Metrics Association FSM, or just FiSMA, is a method of sizing the software. Its characteristics are: it is service-oriented rather than process-oriented, in other words, all services are identified to calculate the functional size of the software [29].

At first glance, one realizes that the main existing methods were not designed to consider the requirements of mobile applications. Indeed, the very creation of most of them precedes the emergence of mobile devices as we know today. This suggests that the use of these methods to estimate the effort of the development of projects involving systems or applications for mobile devices would cause a possible failure to quantify the complexity of some features and, therefore, would not produce adequate estimates.

3. CHARACTERISTICS OF MOBILE APPLICATIONS

In order to identify characteristics that are inherent to systems and mobile applications, a surveying of the characteristics of these types of software was accomplished through a systematic review. Conducting a systematic review is relevant because most searches begin with some kind of review of the literature, and a systematic review summarizes the existing work fairly, without inclinations. So the surveys were conducted according to a predefined search strategy, in which the search strategy should allow the integrity of the research to be evaluated. The planning and accomplishment of the methodology discussed were directed by Procedures for Performing Systematic Reviews [30].

3.1. Planning The Systematic Review

In the context of research questions, the following research question was formulated: “What are the characteristics of Mobile Applications?”, based on the issue about the proposed study.
Search Strategies - The search strategy was divided into three parts: sources, keywords and search strings.
Sources - the researches were directed to the following databases: ACM DL Digital Library (http://dl.acm.org/), Google Scholar (http://scholar.google.com.br/) and IEEE Xplore Digital Library (http://ieeexplore.ieee.org/Xplore/).

Keywords - the keywords were defined and based on the research question elicited previously and on their synonyms, as follows: Mobile; Applications; Computing; Features; Characteristics; Attribute; Aspect; Property; Factors; Individuality; Differential; Detail; Software; System;

Search string - based on the keywords defined previously and according to the sources to be used, the following search string was prepared: "((“Mobile Applications”) OR (“Mobile Computing”) OR (“Mobile System”) OR (“Mobile Software”)) AND (Features OR Characteristics OR Attribute OR Aspect OR Property OR Factors OR Individuality OR Differential OR Detail)".

The results obtained through the researches made with the string search string defined previously in the three databases mentioned above were analyzed according to the following criteria:

- Inclusion Criteria: The returned result should be available in English or Portuguese; The returned result should be available in PDF or HTML format; The returned result should answer the research question directly;
- Exclusion Criteria: The returned result has already been found in previous research; The returned result has not been published in conferences, books, newspapers or magazines; The returned result has no relation to the research question; The access to the result is not available through agreements with CAPES or UFRN; The returned result was not published between 2002 and 2013;

Procedures for The Evaluation of the Articles: the articles will be analyzed considering its relation with the issues addressed in the research questions, inclusion criteria and exclusion criteria, and their respective situation will be assigned with either “Accepted” or “Rejected”. The evaluation will follow the following procedure: read the title and abstract and, should it be related with the research question, also read the whole article.

3.2. Implementation of the Systematic Review

The implementation of the systematic review was performed almost in line with its planning, except for the need to adjust the syntax of the proposed search string due to the particularities of the research bases. 234 articles were analyzed, of which 40 were selected and considered “Accepted” according to the inclusion criteria; 194 were considered “Rejected” according to the exclusion criteria. The list with all the articles can be accessed at the following address: http://www.laudson.com/sr-articles.pdf. The 40 articles that were accepted were fully read, thus performing the data extraction. All the characteristics found during this extraction phase were described in the following subsection.

3.3. Completion of Systematic Review

Given the results extracted from the systematic review, it's is possible to identify 29 kinds of characteristics in 100% of the articles evaluated and considered accepted in accordance with the inclusion criteria. However some of these are a mixture of characteristics of mobile devices and characteristics of mobile applications, such as the characteristic called “Limited Energy”, which is a characteristic of the device and not the application, however the articles that mention this type of characteristic emphasize that in the development of a mobile application, this “limitation” must be taken into account since all the mobile devices are powered by batteries, which have a limited life, depending completely on what the user operates daily. Applications requiring more hardware or software resources will consume more energy. In Figure 1, the 23 types of characteristics mentioned the most in the selected articles can be observed.

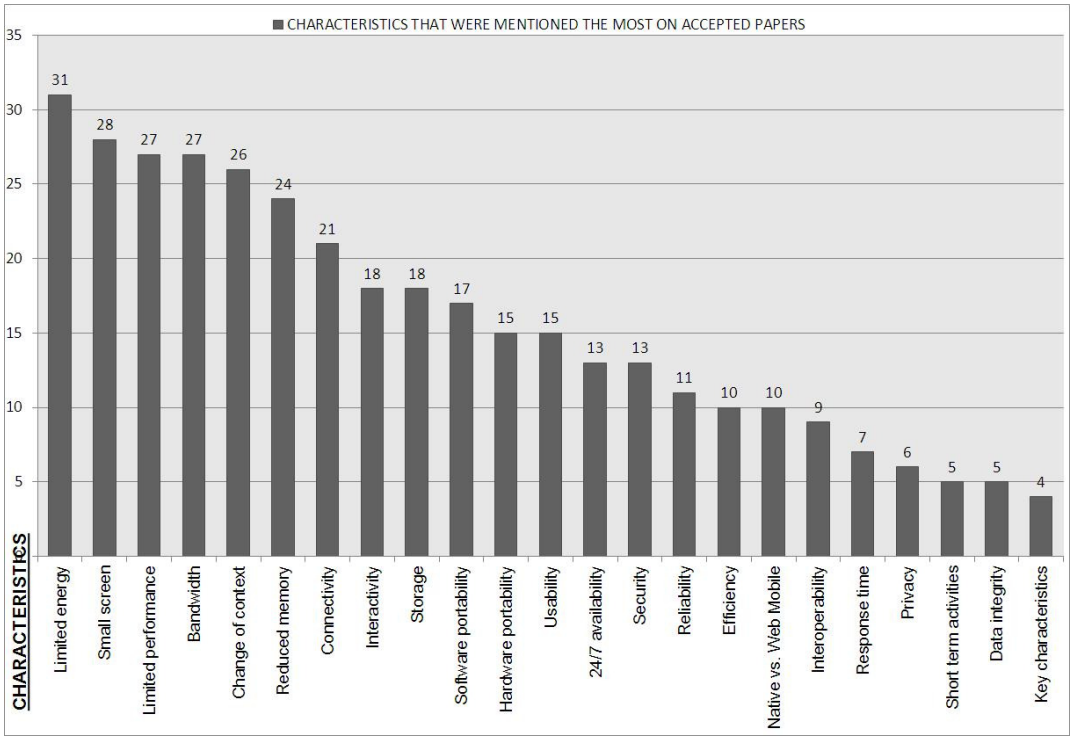


Figure 1. Characteristics that were mentioned the most on accepted papers

The other six types of characteristics identified are mentioned only three times, which are: “Complex integration of tasks in real time” and “Constant interruption of activities”; and, finally, are mentioned only once, which are: “Functional area”, “Price”, “Target public” and “Type of provider Type”.

Following, there is a description of each characteristic identified in the review:

- Limited energy: every mobile device is powered by battery and, because of this, it has a certain lifetime period [31].
- Small screen: mobile device screens are pretty small and, because of this, interface design is limited [31].
- Limited performance: due to its size and technological advancement all mobile devices, even the most advanced in its class, have limitations of specific resources such as processing power, memory and connectivity. Because of this, the performance is limited [32].
- Bandwidth: given an application that requires the maximum, the minimum or a reasonable bandwidth, one must consider its enormous variation [32].
- Change of context: the change of context occurs in accordance with the environment [32].
- Reduced memory: due to its size and technological advancement, all mobile devices, even the most advanced in its class, have limitations of specific resources, including the size of its memory [31].
- Connectivity: the kind of connectivity that the application will use, such as 3G, bluetooth, infrared and Wi-Fi [33].
- Interactivity: what will be the type of input that the user will use to interact with the application [32].
- Storage: the applications have to take into consideration how it is going to be done [32].

- Software portability: the application should be performed on all types of operating systems [32].
- Hardware portability: the application should be performed on all types of devices [32].
- Usability: is a set of attributes which affect the effort needed for the use, in which it must be intuitive and as natural as possible to make or receive a call or text message [33].
- 24/7 availability: the application must be available to access anywhere, anytime [33].
- Security: must prevent accidental or deliberate unauthorized access to the applications and data [33].
- Reliability: is a set of attributes which affect the application's ability to maintain its level of performance under stated conditions for a stated period of time [34].
- Efficiency: is a set of attributes that relate to the relationship between the application's level of performance and the amount of resources used, under stated conditions [35].
- Native vs. Web Mobile: it must be defined if the application will be designed to be installed on the device itself, which is known as native applications, or used on the web [33].
- Interoperability: the application should be able to interact with other specific systems. In other words, it must have interoperability with other services [32].
- Response time: the applications must be initialized and finalized immediately [36].
- Privacy: the application must demonstrate to the user how his or her personal information are being collected, used and shared, and let the user exercise his or her choice and control over their use [33].
- Short term activities: activities in mobile applications tend to have a short duration, ranging from several seconds to several minutes [37].
- Data integrity: making sure that in an accidental shutdown of the application or of the device itself, the application will ensures data integrity [36].
- Key characteristics: mobile applications tend to be more focused or, in other words, they have specific key characteristics rather than offer the exploratory environment commonly used [37].
- Complex integration of real-time tasks: mobile applications should provide integration between application of different sources (native or web) [38].
- Constant interruption of activities: when using a mobile application, the activities are constantly interrupted, like when you receive a call, lose connection or have a low battery, which are examples of such interruptions [37].
- Functional area: data, collaboration and communication services, information services and productivity services such as business and office applications [39].
- Price: free, less than five euros and more than five euros [39].
- Target audience: applications for final private consumer or business applications [39].
- Provider type: businesses, professionals or other service providers [39].

After this survey, a refinement was made and a mix of characteristics was elicited with the purpose of defining which characteristics would be emphasized. Of a total of 23 types of characteristics that were most mentioned in the selected articles, a common denominator of 13 characteristics was reached, some of which had their names redefined, like “Interactivity”, which became “Input Interface”.

4. CHARACTERISTICS OF MOBILE APPLICATIONS SURVEY

With the conclusion of the systematic review, a survey was carried out among experts in mobile development with the purpose of ratifying the characteristics previously raised and to prove their respective influence on mobile development. The disclosure of the survey was conducted in more than 70 locations, among them universities and businesses, through e-mails, study groups and social groups.

In general, of all 117 feedbacks received through the survey, 100% of the experts confirmed the characteristics; among them, an average of 72% indicated a greater effort and complexity regarding the characteristics during development, an average of 12% indicated less effort and complexity and, finally, an average of 16% indicated they did not perceive any difference in mobile development, even though they confirmed the presence of the characteristics.

5. PROBLEM ADDRESSED

As noted in Section II, there is no estimation method developed for mobile applications projects. Moreover, some of the characteristics elicited in Section III aggravate the complexity and, thereafter, the effort in the development of mobile applications.

From the analysis that follows, with the characteristics of applications on mobile devices elicited in Section III, it is clear that they are different from the characteristics of traditional systems and directly influence its development. A clear example, which is different from the information or desktop systems, is the characteristic that the mobile devices have “Limited Energy”. As mobile devices are powered by battery, which have a limited lifetime period, the applications must be programmed to require the minimal amount of hardware resources possible, since the more resources consumed, the greater amount of energy expended. This characteristic makes it necessary for the solution project to address this concern, generating a higher complexity of development and, thereafter, a greater effort and cost.

Another specific characteristic of this context is the “Graphical Interface”. Due to the reduced screen size, the interface design is limited. Therefore, a greater complexity and, thereafter, a larger effort is required in the development of the graphical interface. Another characteristic related to the screen is the “Input Interface”, which defines how the user will interact with the application, in other words, if the user will interact via keypad, stylus, touch screen or voice and image recognition. The latter makes the task of developing applications that offers all these interaction options more complex, thus requiring a bigger effort.

Regarding connectivity, the characteristic “Bandwidth” was identified, wherein a mobile application might have the maximum band at times and the minimum in other moments. Some types of applications need to realize this and act differently in each situation. Another related feature is the “Connectivity Type”. Mobile applications can be developed to support different types of connectivity such as 3G, bluetooth, infrared, Wi-Fi, Wireless, NFC and others. In addition, a single application can support multiple types of connectivity simultaneously. These behaviors directly affect the complexity of the software and therefore require a larger development effort.

The “Change in Context” is also another characteristic inherent in mobile applications, which should take into account not only the data entries explicitly provided by users, but also the implicit entries concerning the physical and computational context of the users and the environments that surround them. In addition, the “Constant Interruption of Activities” is a much more common characteristic in this context, as well as the need for some applications to be developed to work offline and therefore be able to synchronize. Mobile applications should be prepared for different scenarios because the activities are interrupted constantly. Receiving a call, lack of connection and low battery are examples of such interruptions, which makes the applications become much more complex.

Despite the advances related to the computational ability of these devices, their hardware must still be considered as limited, especially when compared to desktops and servers. Two characteristics related to this issue are “Limited Performance” and “Reduced Memory”. Besides

these, a characteristic inherent to the use of mobile devices is the "Response Time", that is directly related to the power of "Processing". Mobile applications must be initialized and finalized immediately, in other words, any development should be focused in the time variable. These characteristics require the applications to be developed with a possible resource optimization for a better efficiency and response time, requiring more effort.

The "Portability" is also a required characteristic of these applications. It can be divided into two characteristics: the "Hardware Portability" and the "Software Portability". Regarding the first one, nowadays there is a large number of different devices with different capabilities and resources. A mobile application should be able to run on the largest number of devices possible. This requires an increased effort in the development. Moreover, a greater effort in testing this kind of portability is required. Regarding "Software Portability", it is necessary to develop specific applications for each existing platform should the application be native. With this, more effort is required for replications of the same software product, including the tests.

Finally, mobile applications can be separated into two types: "Native or Web Mobile". The first one has higher performance and easiness in accessing the hardware, while the second has lower performance since it is web, but it is easier to achieve portability. In addition, there are some applications that are considered hybrids. Depending on the type of application, the issues that must be considered and the complexity can be different, requiring different development efforts.
From the survey of the most popular estimation methods cited in Section III, it was found that these characteristics are not covered by the current estimation methods for two explicit reasons: first, none of the existing methods was designed to perform project estimation in mobile applications development; and second, all the characteristics discussed in this section are exclusive to mobile applications, with direct interference in their development, thereby generating a greater complexity and, thereafter, a greater effort. However, to consider any of the existing estimation methods to apply to the process of development of mobile applications is to assume that this kind of development is no different than the project of developing desktop applications, in other words, an eminent risk is assumed.

6. PROPOSAL: Estimation in Mobile Application Development Project

A solution to solve this problem would be to create a new estimation method or even to adapt some existing estimation method, in which would be added all of the characteristics identified that directly affect the mobile application development project, taking into account whatever is needed to reach the minimum efficiency in the estimates.

The approached proposed is an adaptation of an existing method, based exclusively on methods recognized as international standards by ISO. Among the most popular estimation methods mentioned in Section III, the method used to base the proposal below on is known as "Finnish Software Metrics Association (FiSMA)". The model is one of the five methods for measuring software that complies with the ISO/IEC 14143-1 standard, is accepted as an international standard for software measuring [29] and nowadays over 750 software projects are completed being estimated by FISMA. However, the difference between this and other methods that are in accordance with the above standard, which are the Common Software Measurement International Consortium Function Points (COSMIC FP) [25], the International Function Point Users Group (IFPUG) FPA [11], MarkII FPA [14] and the Netherlands Software Metrics Association (NESMA) WSF [40], is that the method used is based in functionality but is service-oriented. It also proposes in its definition that it can be applied to all types of software, but this statement is lightly wrong since in its application, the method does not take into account the characteristics elicited in Section IV.

The COMISC FP [25], the MarkII FPA [14] and the NESMA [40] were created based on the FPA [11], in other words, they assume the counting of Function Point (FP), but considering the implemented functionality from the user's point of view. With this, it is clear that the methods mentioned above do not take into account the characteristics of mobile applications because they are not noticed by the user. The methods are independent of the programming language or technology used. And, unlike FISMA, they do not bring in their literature the information that they can be applied to all types of software.

Overall, the FISMA method proposes that all services provided by the application are identified. It previously defines some services, among which stands out the user's interactive navigation, consulting services, user input interactive services, interface services for other applications, data storage services, algorithmic services and handling services. Finally, after identifying all the services, the size of each service is calculated using the same method and thus obtaining a total functional size of the application by adding the size of each service found [41].

6.1. Approaching the Chosen Model

The FiSMA method in its original usage proposes a structure of seven classes of the Base Functional Component or BFC (Base Functional Component) type, which is defined as a basic component of functional requirement. The seven classes used to account for the services during the application of the method are [41]: interactive navigation of the end user and query services (q); interactive input services from end users (i); non-interactive outbound services for the end user (o); interface services for another application (t); interface services for other applications (f); data storage services (d) and algorithmic manipulation services (a).

The identification for each class name BFC previously mentioned, with a letter in parenthesis, is used to facilitate the application of the method during the counting process, because each of the seven classes BFCs are composed of other BFC classes which, at the time of calculating, these BFCs "daughter" classes are identified by the letter of their BFC "mother" class followed by a numeral, as can be seen in Figure 2.

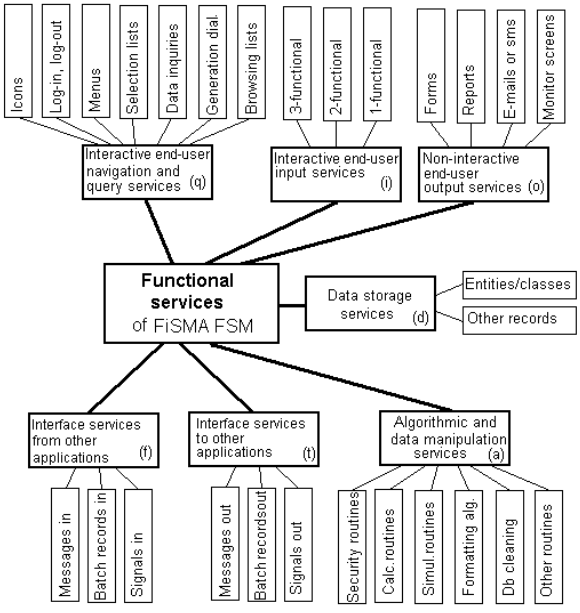


Figure 2. Types of BFCs classes of the base model [41]

The unit of measurement is the point of function with the letter "F" added to its nomenclature to identify the "FiSMA", resulting in FfP (FiSMA Function Point) or Ffsu (FiSMA functional size unit). The measurement process generally consists of measuring the services and end-user interface and the services considered indirect [41], as can be seen in Figure 3.

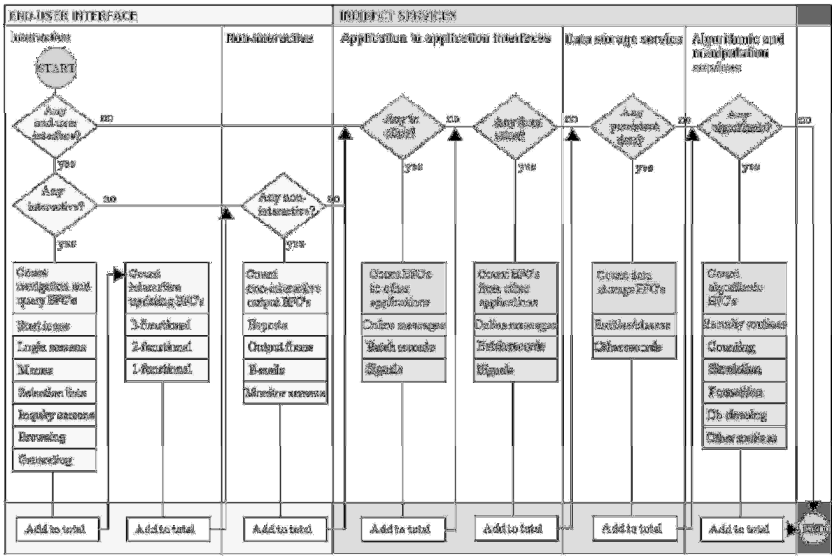


Figure 3. Representation of the measurement process of the base model [41]

Figure 3 shows the process of measuring the base model, in which it defines each step and sum of each BFC class of the model. Briefly, the process of counting should be done as follows. Identify: 1-How many types of BFCs does the software have? 2-Which are they? (identify all) 3-What are they? (provide details of each BFC identified)
After doing this, it is necessary to add each BFC root using the formulas pre-defined by the method and their assignments. Finally, the formula of the final result of the sum is the general sum of all the BFCs classes.

6.2. Applying the Chosen Model

The FiSMA method can be applied manually or with the aid of the Experience Service[1] tool, which was the case, provided by FiSMA itself through contact made with senior consultant Pekka Forselius and with the chairman of the board Hannu Lappalainen.

When using the tool, it is necessary to perform all the steps of the previous subsection to obtain the functional size. Figure 4 shows the final report after the implementation of the FiSMA on a real system, the Management of Academic Activities Integrated System (Sigaa) in its Mobile version, developed by the Superintendence of Computing (SINFO) of the Federal University of Rio Grande do Norte (UFRN).

[1] http://www.experiencesaas.com/

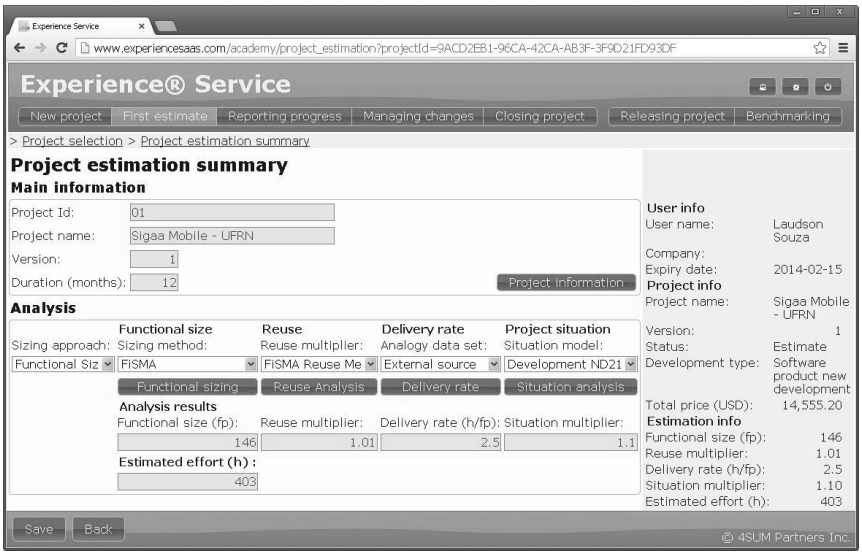


Figure 4. Final Report of FiSMA applied to Sigaa Mobile

After the application of FiSMA, the functional size of the software is obtained and from this it is possible to find the effort using the formula: Estimated effort (h) = size (fp) x reuse x rate of delivery (h/fp) x project status; the latter is related to productivity factors that are taken into account for the calculation of the effort. However, of the factors predefined by the FiSMA regarding the product, only 6 (six) are proposed, in which the basic idea of the evaluation is that "the better the circumstances of the project, the more positive the assessment". The weighting goes from - - to + +, as follows:

Caption:

- "+ +" = [1.10] Excellent situation, much better circumstances than in the average case
- "+" = [1.05] Good situation, better circumstances than in the average case
- "+ / -" = [1.0] Normal situation
- "-" = [0.95] Bad situation, worse circumstances than in the average case
- "- -" = [0.90] Very bad situation, much worse circumstances than in the average case

Productivity factors:

Functionality requirements: compatibility with the needs of the end user, the complexity of the requirements.

- (- -) Complex and critical application area (thousands of FPs), multiple users and multicultural system.
- (-) Interoperable application area with some complex characteristics, requiring special understanding from users and developers.
- (+ / -) Partly automated, integrated application area and a medium size application (between 600 and 1000 FPs) with standard security requirements.
- (+) Application area mostly automated and application with less than 5 interfaces with other systems; there are specific security requirements.

- (+ +) Very mature application area, simple and easy, a small stand-alone application (less than 200 FPs) for a small group of users.

Reliability requirements: maturity, tolerance to faults and recovery for different types of use cases.

- (- -) Malfunctions may put in danger human lives and cause significant economic or environmental losses.}
- (-) The software is part of a large real-time system where all the failures of operation will cause problems to many other applications.}
- (+ / -) Not more than 2 hours of downtime is acceptable, but the system recovery routines are appropriate.
- (+) Need for non-continuous operation, but daily.
- (+ +) Need for periodic operation. Pausing for a few days will not cause any damage to the organization.

Usability requirements: understandability and easiness to learn the user interface and workflow logic.

- (- -) A large number of different types of end users around the world.
- (-) 2 or 3 different types of users with different skills.
- (+ / -) A large number of end users with equal abilities.
- (+) No more than tens or hundreds of homogeneous users in perhaps more than one location.
- (+ +) Only a few users, all located on one site.

Efficiency requirements: effective use of resources and adequate performance in each use case and under a reasonable workload.

- (- -) Complex database with millions of data records and transactions per day, thousands of simultaneous end users.
- (-) Large database, hundreds of simultaneous end users, critical response most of the time.
- (+ / -) Large database, less than millions of data records and less than hundreds of simultaneous end users.
- (+) Medium database in volume and structure, simple and predictable data requests from some simultaneous end users.
- (+ +) Simple and small database without simultaneous end users or complex data requests.

Maintainability requirements: lifetime of the application, criticality of fault diagnosis and test performance.

- (- -) Very large strategic software (over 20 years of lifetime) in a volatile area of business, with frequent changes in laws, regulations and business rules.
- (-) Large software (10-20 years of lifetime), and frequent changes in laws, regulations and business rules.

- (+ / -) Medium size software (5-10 years of lifetime), monthly changes in laws, regulations and business rules.
- (+) Small software, rarely changes (2 to 5 years of lifetime).
- (+ +) Temporary software (less than 2 years of lifetime), without modifications.

Portability requirements: adaptability and instability to different environments, to the architecture and to structural components.

- (- -) Software users are located in many types of organizations, with various platforms (hardware, browsers, operating systems, middleware, protocols, etc), various versions and various update frequencies.
- (-) The software must operate on some different platforms (hardware, browsers, operating systems, middleware, protocols, etc) and in various versions of each of them.
- (+ / -) Each version of the software must run on multiple versions of a given platform (hardware, browser, operating system, middleware, protocols, etc), and the frequencies of update of the users are quite predictable.
- (+) The software must run on a given platform (hardware, browser, operating system, middleware, protocols, etc), but the use of system-level services is limited because the upgrade process is partial.
- (+ +) Software must be run on a particular platform (hardware, browser, operating system, middleware, protocols, etc), but the upgrade process is completely controllable.

Among the productivity factors mentioned above, only the “Portability Requirement” factor fits in harmony with the “Portability” characteristic regarding both hardware and software. However, none of the other factors discusses the characteristics of mobile application, in other words, after obtaining the functional size of the software and applying the productivity factors related to the product to estimate the effort, this estimate ignores all of the characteristics of mobile applications, judging that the estimate of traditional information systems is equal to the mobile application. However, with the proposal of the creation of new productivity factors, which would be the specific characteristics of mobile applications, this problem will be solved, as presented below.

Performance Factor:

- (-) The application should be concerned with the optimization of resources for a better efficiency and response time.
- (+ / -) Resource optimization for better efficiency and response time may or may not exist.
- (+) Resource optimization for better efficiency and response time should not be taken into consideration.

Power Factor:

- (-) The application should be concerned with the optimization of resources for a lower battery consumption.
- (+ / -) Resource optimization for lower battery consumption may or may not exist.
- (+) Resource optimization for a lower battery consumption should not be taken into consideration.

Band Factor:

- (-) The application shall require the maximum bandwidth.
- (+ / -) The application shall require reasonable bandwidth.
- (+) The application shall require a minimum bandwidth.

Connectivity Factor:

- (-) The application must have the maximum willingness to use connections such as 3G, Wi-fi, Wireless, Bluetooth, Infrared and others.
- (+ / -) The application must have reasonable predisposition to use connections such as 3G, Wi-Fi and Wireless.
- (+) The application must have only a predisposition to use connections, which can be: 3G, Wi-fi, Wireless, Bluetooth, Infrared or others.

Context Factor:

- (-) The application should work offline and synchronize.
- (+ / -) The application should work offline and it is not necessary to synchronize.
- (+) The application should not work offline.

Graphic Interface Factor:

- (-) The application has limitations due to the screen size because it will be mainly used by cell phone users.
- (+ / -) The application has reasonable limitation due to the screen size because it will be used both by cell phone and tablet users.
- (+) The application has little limitation due to the screen size because it will be mainly used by tablet users.

Input Interface Factor:

- (-) The application must have input interfaces for touch screen, voice, video, keyboard and others.
- (+ / -) The application must have standard input interfaces for keyboard.
- (+) The application must have any one of the types of interfaces, such as: touch screen, voice, video, keyboard or others.

The proposed factors take into account the same weighting proposed by FiSMA, but only ranging from - to +, in other words:

- "+" = [1.05] Good situation, better circumstances than in the average case
- "+ / -" = [1.0] Normal Situation
- "-" = [0.95] Bad situation, worse circumstances than in the average case

The functional size remains the same, thus affecting only the formula used to obtain the effort, which will now consider in its "project situation" variable the new productivity factors specific for mobile applications.

7. CONCLUSION

Given the results presented, based on the literature review of estimation methods and on the systematic review of the characteristics of mobile applications, it was observed that this sub-area of software engineering still falls short. Basically, it's risky to use any existing estimation method in development projects for mobile applications, as much as there are some models already widespread in industry, such as the Function Point Analysis, the Mark II and the COSMIC-FFP, which are even approved by ISO as international standards. They all fall short by not taking into account the particularities of mobile applications, which makes the method partially ineffective in this situation.

With the common emergence of new systems, experts always find a barrier when using one of the current methods of software measurement. This barrier can be on the effectiveness of the method, on what type of method should be used, when it comes to a software that is considered unconventional and, mostly, when it is required to apply it in completely atypical scenarios. This whole situation is aggravated further when it comes to mobile applications.

Based on this study, it is concluded that the proposal presented in this work is entirely appropriate and viable and that this proposal should take into account all the peculiarities of such applications, finally creating a belief that there actually are considerable differences in the development project for mobile applications.

Regarding future work, it is possible to accomplish the same systematic review with other purposes, increasing the number of research questions and, thereby, increasing the number of selected articles. Finally, exhaustively applying the presented proposal in several development projects for mobile devices in order to attain a perfect validation and a solid foundation in the industry.

REFERENCES

[1] L. Naismith, M. Sharples, G. Vavoula, P. Lonsdale *et al.*, "Literature review in mobile technologies and learning," 2004.

[2] G. Macario, M. Torchiano, and M. Violante, "An in-vehicle infotainment software architecture based on google android," in *Industrial Embedded Systems, 2009. SIES '09. IEEE International Symposium on*, 2009, pp. 257–260.

[3] T. Liu, H. Wang, J. Liang, T.-W. Chan, H. Ko, and J. Yang, "Wireless and mobile technologies to enhance teaching and learning," *Journal of Computer Assisted Learning*, vol. 19, no. 3, pp. 371–382, 2003.

[4] I. GARTNER. (2013) Gartner says worldwide mobile phone sales declined 1.7 percent in 2012. egham, uk: Gartner, 2013. [Online]. Available: http://www.gartner.com/newsroom/id/2335616

[5] C.-C. Yang, H.-W. Yang, and H.-C. Huang, "A robust and secure data transmission scheme based on identity-based cryptosystem for ad hoc networks," in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, ser. IWCMC '10. New York, NY, USA: ACM, 2010, pp. 1198–1202, aCM. [Online]. Available: http://doi.acm.org/10.1145/1815396.1815670

[6] I. Ketykó, K. De Moor, T. De Pessemier, A. J. Verdejo, K. Vanhecke, W. Joseph, L. Martens, and L. De Marez, "Qoe measurement of mobile youtube video streaming," in *Proceedings of the 3rd workshop on Mobile video delivery*, ser. MoViD '10. New York, NY, USA: ACM, 2010, pp. 27–32. [Online]. Available: http://doi.acm.org/10.1145/1878022.1878030

[7] S.-Y. Yang, D. liang Lee, and K.-Y. Chen, "A new ubiquitous information agent system for cloud computing - example on gps and bluetooth techniques in google android platform," in *Electric Information and Control Engineering (ICEICE), 2011 International Conference on*, 2011, pp. 1929–1932.

[8] R. Lowe, P. Mandl, and M. Weber, "Context directory: A context-aware service for mobile context-aware computing applications by the example of google android," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, 2012, pp. 76–81.

[9] N. Husted, H. Sadi, and A. Gehani, "Smartphone security limitations: conflicting traditions," in *Proceedings of the 2011 Workshop on Governance of Technology, Information, and Policies*, ser. GTIP '11. New York, NY, USA: ACM, 2011, pp. 5–12. [Online]. Available: http://doi.acm.org/-10.1145/2076496.2076497

[10] A. Shabtai, Y. Fledel, U. Kanonov, Y. Elovici, S. Dolev, and C. Glezer, "Google android: A comprehensive security assessment," *Security Privacy, IEEE*, vol. 8, no. 2, pp. 35–44, 2010.

[11] S. Oligny, J.-M. Desharnais, and A. Abran, "A method for measuring the functional size of embedded software," in *3rd International Conference on Industrial Automation*, 1999, pp. 7–9.

[12] B. Boehm, R. Valerdi, J. Lane, and A. Brown, "Cocomo suite methodology and evolution," *CrossTalk*, vol. 18, no. 4, pp. 20–25, 2005.

[13] C. Jones and T. C. Jones, *Estimating software costs*. McGraw-Hill New York, 1998, vol. 3.

[14] C. Symons, "Come back function point analysis (modernized)–all is forgiven!)," in *Proc. of the 4th European Conference on Software Measurement and ICT Control, FESMA-DASMA*, 2001, pp. 413–426.

[15] M. Lother and R. Dumke, "Points metrics-comparison and analysis," in *International Workshop on Software Measurement (IWSM 2001), Montréal, Québec*, 2001, pp. 155–172.

[16] J. Engelhart, P. Langbroek *et al.*, *Function Point Analysis (FPA) for Software Enhancement*. NESMA, 2001.

[17] D. J. Reifer, "Asset-r: A function point sizing tool for scientific and real-time systems," *Journal of Systems and Software*, vol. 11, no. 3, pp. 159–171, 1990.

[18] M. Maya, A. Abran, S. Oligny, D. St-Pierre, and J.-M. Desharnais, "Measuring the functional size of real-time software," in *Proc. of 1998 European Software Control and Metrics Conference, Maastricht, The Netherlands*, 1998, pp. 191–199.

[19] S. Kusumoto, F. Matukawa, K. Inoue, S. Hanabusa, and Y. Maegawa, "Estimating effort by use case points: method, tool and case study," in *Software Metrics, 2004. Proceedings. 10th International Symposium on*, 2004, pp. 292–299.

[20] R. Banker, R. Kauffman, C. Wright, and D. Zweig, "Automating output size and reuse metrics in a repository-based computer-aided software engineering (case) environment," *Software Engineering, IEEE Transactions on*, vol. 20, no. 3, pp. 169–187, 1994.

[21] J. Matson, B. Barrett, and J. Mellichamp, "Software development cost estimation using function points," *Software Engineering, IEEE Transactions on*, vol. 20, no. 4, pp. 275–287, 1994.

[22] R. Meli, "Early and extended function point: a new method for function points estimation," in *Proceedings of the IFPUG-Fall Conference*, 1997, pp. 15–19.

[23] M. Morisio, I. Stamelos, V. Spahos, and D. Romano, "Measuring functionality and productivity in web-based applications: a case study," in *Software Metrics Symposium, 1999. Proceedings. Sixth International*, 1999, pp. 111–118.

[24] G. Caldiera, G. Antoniol, R. Fiutem, and C. Lokan, "Definition and experimental evaluation of function points for object-oriented systems," in *Software Metrics Symposium, 1998. Metrics 1998. Proceedings. Fifth International*, 1998, pp. 167–178.

[25] C.-C. S. M. I. Consortium *et al.*, "The cosmic functional size measurement method-version 3.0 measurement manual (the cosmic implementation guide for iso/iec 19761: 2003)," 2007.

[26] R. Meli, A. Abran, V. T. Ho, and S. Oligny, "On the applicability of cosmic-ffp for measuring software throughout its life cycle," in *Proceedings of the 11th European Software Control and Metrics Conference*, 2000, pp. 18–20.

[27] J. Kammelar, "A sizing approach for oo-environments," in *Proceedings of the 4th International ECOOP Workshop on Quantitative Approaches in Object-Oriented Software Engineering*, 2000.

[28] S. Abrahão, G. Poels, and O. Pastor, "A functional size measurement method for object-oriented conceptual schemas: design and evaluation issues," *Software & Systems Modeling*, vol. 5, no. 1, pp. 48–71, 2006.

[29] P. Forselius, "Finnish software measurement association (fisma), fsm working group: Fisma functional size measurement method v. 1.1." 2004.
[30] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, p. 2004, 2004.
[31] J.-H. Sohn, J.-H. Woo, M.-W. Lee, H.-J. Kim, R. Woo, and H.-J. Yoo, "A 50 mvertices/s graphics processor with fixed-point programmable vertex shader for mobile applications," in *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, 2005, pp. 192–592 Vol. 1, google.
[32] H. Mukhtar, D. Belad, and G. Bernard, "A model for resource specification in mobile services," in *Proceedings of the 3rd international workshop on Services integration in pervasive environments*, ser. SIPE '08. New York, NY, USA: ACM, 2008, pp. 37–42, aCM. [Online]. Available: http://-doi.acm.org/10.1145/1387309.1387318
[33] H. Feng, "A literature analysis on the adoption of mobile commerce," in *Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on*, 2009, pp. 1353–1358, iEEE.
[34] A. Kumar Maji, K. Hao, S. Sultana, and S. Bagchi, "Characterizing failures in mobile oses: A case study with android and symbian," in *Software Reliability Engineering (ISSRE), 2010 IEEE 21st International Symposium on*, 2010, pp. 249–258, iEEE.
[35] J. Al-Jaroodi, A. Al-Dhaheri, F. Al-Abdouli, and N. Mohamed, "A survey of security middleware for pervasive and ubiquitous systems," in *Network-Based Information Systems, 2009. NBIS '09. International Conference on*, 2009, pp. 188–193, iEEE.
[36] K. Hameed *et al.*, "Mobile applications and systems," 2010, google.
[37] I. R. D. E. A. G. Alekhya Mandadi, Deepti Mudegowder, "Mobile applications: Characteristics & group project summary," *Mobile Application Development*, 2009, google.
[38] M. Hayenga, C. Sudanthi, M. Ghosh, P. Ramrakhyani, and N. Paver, "Accurate system-level performance modeling and workload characterization for mobile internet devices," in *Proceedings of the 9th workshop on MEmory performance: DEaling with Applications, systems and architecture*, ser. MEDEA '08. New York, NY, USA: ACM, 2008, pp. 54–60, aCM. [Online]. Available: http://-doi.acm.org/10.1145/1509084.1509092
[39] A. Giessmann, K. Stanoevska-Slabeva, and B. de Visser, "Mobile enterprise applications–current state and future directions," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2012, pp. 1363–1372, google.
[40] C. Gencel, R. Heldal, and K. Lind, "On the conversion between the sizes of software products in the life cycle."
[41] F. S. M. A. FiSMA. (2004) Fisma functional size measurement method version 1-1. [Online]. Available: http://www.fisma.fi/in-english/methods/

Authors

Laudson Silva de Souza is masters student the Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, Brazil.

Gibeon Soares de Aquino Jr. is PhD teacher the Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, Brazil.

INTENTIONAL BLANK

A BLIND ROBUST WATERMARKING SCHEME BASED ON SVD AND CIRCULANT MATRICES

Noui Oussama[1] and Noui Lemnouar[2]

[1]Department of Computer science, Faculty of Science University of Batna
os.noui@gmail.com
[2] Department of Mathematics, Faculty of Science University of Batna
nouilem@yahoo.fr

ABSTRACT

Multimedia security has been the aim point of considerable research activity because of its wide application area. The major technology to achieve copyright protection, content authentication, access control and multimedia security is watermarking which is the process of embedding data into a multimedia element such as image or audio, this embedded data can later be extracted from, or detected in the embedded element for different purposes. In this work, a blind watermarking algorithm based on SVD and circulant matrices has been presented. Every circulant matrix is associated with a matrix for which the SVD decomposition coincides with the spectral decomposition. This leads to improve the Chandra algorithm [1], our presentation will include a discussion on the data hiding capacity, watermark transparency and robustness against a wide range of common image processing attacks.

KEYWORDS

Digital image watermarking, Singular value Decomposition, circulant matrix, ownership protection.

1. INTRODUCTION

Due to improvements in the digital image technology and growing availability and usability of internet during the past several years, demands for storage and transmitting of digital images have seen a distinct increase, Unfortunately, the problem of illegal piracy is increasingly serious. Protection of digital multimedia content has become an increasingly important issue for content owners and service providers. Encryption data was a way to ensure only the owner to view the content, there are still ways for illegal using of the content after decryption [2, 3], that lead us to a new method for protection. Watermarking is the process of embedding data called a watermark into the multimedia object such that watermark can be detected or extracted later to make an assertion about the object. The object may be an audio, image or video and even 3-D models [4, 5]. Watermarking algorithms fall into two categories. The first form of watermarking was a spatial watermarking technique work with the pixel values directly. Generally, spatial domain watermarking is easy to implement from a computational point of view, but too fragile to resist numerous attacks [6, 7]. In spatial domain, the watermark is directly embedded into the specific pixels of the host image, but in transform domain which our proposed scheme based in the watermark is embedded into the transform coefficients of the host image after applying DWT, DFT, DCT or SVD transform, and this called the frequency domain watermarking. Because of the

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 65–77, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4406

weakness of the spatial domain watermarks, watermarking in the frequency domain became more attractive as a result of its higher robustness against attacks compared to the spatial domain watermarking. To this aim a number of robust methods based on the SVD transform were introduced but these methods didn't offer good transparency and robustness against geometric attacks. Starting with Liu and Tan [8] an image watermarking method based on SVD, this method is robust against some attacks, and it is a non-blind method and it has a weak imperceptibility, Chandra et al. [1] also introduced a digital image watermarking method. This method is based on moderate modifying of the singular values of the host image. This method is weak against geometric attacks. Ganic et al. [9] proposed a method based on SVD in discrete wavelet transform (DWT) domain. The insertion procedure concerns the modifying of the singular values of the wavelet transformed sub-bands with the singular values of the mark. This scheme is a non-blind and the transparency of the watermarked image is weak. Makhloghi et al.[10] also proposed a scheme based on singular value decomposition in wavelet domain for copyright protection but his method lacks to robustness. Lin et al. [11] presented a full-band DWT domain image watermarking method based on SVD. This method has good robustness against common attacks but its drawback is that the original image is required in extraction procedure. Also the quality of the watermarked image is not good. The watermarking algorithms described in [1-8] are semi or non blind. Soumya Mukherjee and Arup Kumar Pal [12] proposed a robust watermarking scheme which employs both the Discrete Cosine Transform (DCT) and the Singular Value Decomposition (SVD), It starts with transformation of an original image into a transformed image using block based DCT. From each transformed block, the middle band DCT coefficients are selected to form a reduced transformed image and then the watermark is embedded into the constructed reduced transformed image after a suitable SVD operation. This method has good robustness against different attacks but it has a weak imperceptibility. In this paper, a new SVD-based method is proposed which gives the variety in creating the watermark under (1, 2, ..., n) blocks depending on the initiate coefficients and using the circulant's matrix properties to make Chandra algorithm [1] blind instead of non blind and turn it robust against different geometric and non geometric attacks. Organization of the paper is as follows: Section 2 explains the concept of SVD and Circulant matrices while Section 3 presents the proposed method. Section 4 throws light on the experimental results and a comparative analysis of our scheme and other schemes is given, whereas the summary of results and the conclusion is presented in Section 5.

2. SINGULAR VALUE DECOMPOSITION AND CIRCULANT MATRICES

2.1. Singular value decomposition SVD

Every real matrix A can be decomposed into a product of three matrices :

$$A = U \times S \times V^t \tag{1}$$

where U and V are orthogonal matrices such that $U \times U^t = I$ and $V \times V^t = I$ where I is the Identity matrix and S is the diagonal matrix, $S = diag(\partial_1, \partial_2 \cdot\cdot)$ with $\partial_1 \geq \partial_2 \geq \cdot\cdot \geq 0$ The diagonal entries ∂_i of S are called the singular values of A, they are the eigenvalues of $A \times A^t$ or $A^t \times A$. The columns of U are the left singular vectors of A, they are eigenvectors of $A \times A^t$ and the columns of V called the right singular vectors of A and they are eigenvectors of $A^t \times A$.

2.2. Circulant matrices

The circulant matrix $C = cir(c)$ associated to the vector $c \in R^n$ is the $n \times n$ matrix whose rows are given by iterations of the shift operator T acting on c, its K^{th} row is $T^k c$, $k = 1,..,n$
For example if $c = (c_1, c_2, c_3, c_4)$, the 4×4 circulant matrix

$$C = cir(c) \text{ is giving by } \begin{pmatrix} c_1 & c_2 & c_3 & c_4 \\ c_4 & c_1 & c_2 & c_3 \\ c_3 & c_4 & c_1 & c_2 \\ c_2 & c_3 & c_4 & c_1 \end{pmatrix} \tag{2}$$

3. PROPOSED METHOD

We consider a circulant matrix $c = cir(c_1, c_2, c_3, c_4)$
The matrix $CC^t = C^t C$ is positive symmetric matrix so its spectral decomposition coincides with its SVD decomposition that is $CC^t = U_0 diag(\delta_1, \delta_2, \delta_3, \delta_4) {U_0}^t$ with

$$\begin{aligned} \delta_1 &= (c_1 + c_2 + c_3 + c_4)^2 \\ \delta_2 &= (c_1 - c_2 + c_3 - c_4)^2 \\ \delta_3 &= \delta_4 = (c_1 - c_3)^2 + (c_2 - c_4)^2 \end{aligned} \tag{3}$$

are the singular values and U_0 is the constant matrix :

$$U_0 = \begin{pmatrix} 1/2 & -1/2 & 0 & -\sqrt{2}/2 \\ 1/2 & 1/2 & -\sqrt{2}/2 & 0 \\ 1/2 & -1/2 & 0 & \sqrt{2}/2 \\ 1/2 & 1/2 & \sqrt{2}/2 & 0 \end{pmatrix} \tag{4}$$

If A is an image of size $4m \times 4m$, to every vector $c = (c_1, c_2, c_3, c_4)$ is associated a 4×4 circulant matrix $C_1 = cir(c_1)$ and a watermark as $4m \times 4m$ matrix with one block

$$W_1 = \begin{pmatrix} C_1 C_1^t & 0 & . & 0 \\ 0 & 0 & & . \\ . & & . & . \\ 0 & . & . & 0 \end{pmatrix} \tag{5}$$

To obtain a watermark W_k with k blocks

$$W_k = \begin{pmatrix} C_1C_1^t & 0 & . & . & . & 0 \\ 0 & C_2C_2^t & & & & . \\ . & & . & & & . \\ . & & & C_kC_k^t & & . \\ . & & & & 0 & . \\ 0 & . & . & . & . & 0 \end{pmatrix} \quad (6)$$

3.1. Watermark insertion procedure

To watermark a given original image A of size $4m \times 4m$ we will use a watermark with one block as following:

1. Take $c = (c_1, c_2, c_3, c_4)$ such that $\partial_4 \geq \partial_3 \geq \partial_2 \geq \partial_1$
2. Construct the watermark of size $4m \times 4m$

$$W_1 = \begin{pmatrix} C_1C_1^t & 0 & . & 0 \\ 0 & 0 & & . \\ . & & . & . \\ 0 & . & . & 0 \end{pmatrix} \quad (7)$$

3. Apply SVD on A:

$$A = U \times S \times V^t \text{ with } S = diag(S_i) \quad (8)$$

4. Apply SVD on W_1:

$$W_1 = \begin{pmatrix} U_0 & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \partial & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} {U_0}^t & 0 \\ 0 & I \end{pmatrix} \quad (9)$$

With $\partial = diag(\delta_1^1, \delta_2^1, \delta_3^1, \delta_4^1)$ and I is $4(m-1) \times 4(m-1)$ identity matrix.

Put

$$Y_i = S_i + \alpha \partial_i^{'} \quad (10)$$

with $\partial_1^{'} = \delta_1^1$, $\partial_2^{'} = \delta_2^1$, $\partial_3^{'} = \delta_3^1$, $\partial_4^{'} = \delta_4^1$ and $\forall i > 4 \ \partial_i^{'} = 0$

So

$$A^* = U \times diag(Y_i) \times V^t \quad (11)$$

is the watermarked image.

3.2. Watermarking detection and extraction procedure

We don't require the original image A to detect the watermark, we only require the watermarked image A^*, the scaling factor α and the key $S_i = (S_1, S_2, S_3, S_4)$ formed by the first four values of S.

1. Apply SVD to A^*

$$A^* = U^* \times S^* \times V^{*t} \tag{12}$$

2. Calculate

$$x_i = \frac{S_i^* - S_i}{\alpha} \tag{13}$$

for the first four elements.

If $x_3 = x_4$ then the mark is detected else the watermark is not present on the image.
To extract the mark we compute:

$$W^* = \begin{pmatrix} U_0 & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} X & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_0{}^t & 0 \\ 0 & I \end{pmatrix} \tag{14}$$

where $X = diag(x_1, x_2, x_3, x_4)$ and I is $4(m-1)\times 4(m-1)$ identity matrix.

Remarks:

1. If we use a watermark W_k with k blocks, to detect or extract the watermark we only require the scaling factor α and a key $K_2 = (S_1, ..., S_{4k})$ of length 4k which contains the first 4k values of S. In this case the sequence X = (x_i) is of length 4k and the mark is detected if $x_{4i-1} = x_{4i}$ for $i = 1, ..., k$.
2. In Chandra algorithm [1], to extract the watermark (W), U_w, V_w are required, in the proposed scheme $U_w = V_w$ is a constant matrix and independent of the watermark, thus our proposed algorithm is blind.

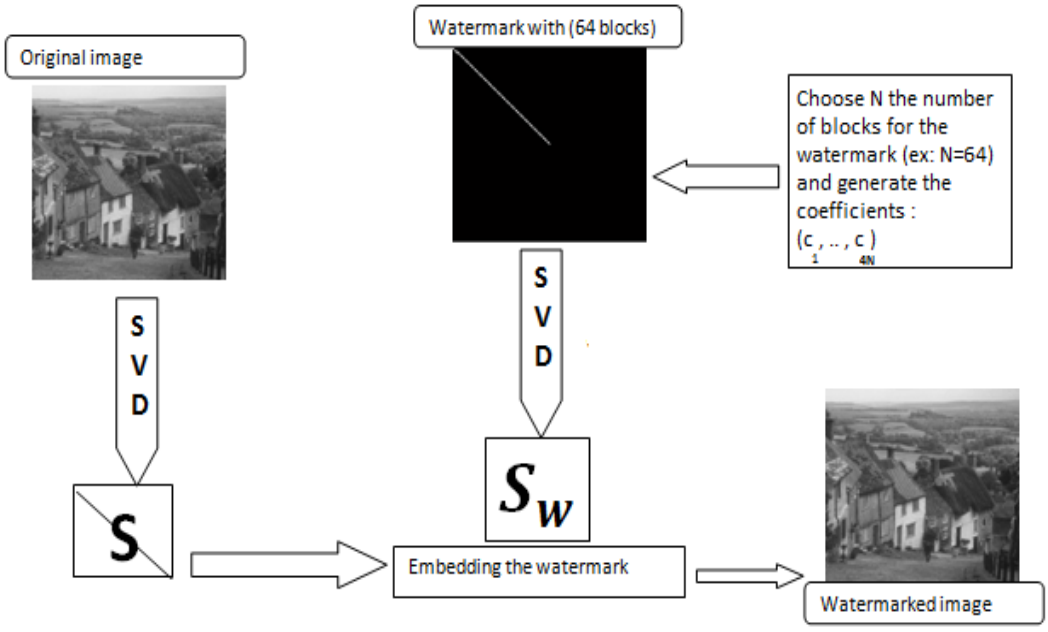


Figure 1. The Proposed watermarking scheme.

Our proposed watermarking method can be concluded in (Figure 1). As it is shown to construct the watermark we first choose N the number of blocks of the watermark then we generate the

coefficients ($c_1,..,c_{4N}$) and by following the steps mentioned above we embed the watermark into the original image.

4. EXPERIMENTAL RESULTS

The proposed scheme is implemented using MATLAB. Six 512×512 images Lena , Goldhill, Baboon, Barbara, Peppers, Boat were used in the simulation (figure 2). The signature is a 512 × 512 image composed with N blocks in its diagonal.

Figure 2. Original test images

To evaluate the quality of the watermarked images we use the PSNR measure defined as:

$$PSNR = 10\log_{10}(\frac{255^2}{MSE})db \quad (15)$$

With MSE is the mean squared error between the original and watermarked image. The PSNR values of the watermarked images by our method indicate that our method in general achieves very good quality as it shown in (Table 1).

Table 1. Relationship of number of blocks of the watermark and PSNR of different watermarked images. (dB) Sf sets to 0.06

Nb blocks / Image name	1	3	5	10	30	64	80	100	128
Lena	56.6994	55.5982	55.7382	55.0831	55.8089	56.0630	55.4090	55.4366	55.3805
Goldhill	57.2825	55.5499	55.4955	55.1460	54.6665	54.6762	52.9705	52.8710	52.8454
Baboon	55.6458	54.6902	55.0781	55.1058	51.6568	50.4369	49.4199	49.1576	49.2020
Barbara	56.9139	55.5102	57.1737	55.2064	53.3403	51.5035	50.2628	50.1065	50.0764
Peppers	55.9094	55.2634	55.1153	53.5932	54.5946	55.4747	56.2183	56.2190	56.2759
Boat	57.1737	55.6458	55.2064	54.6902	51.5035	52.9705	52.8710	52.8454	52.8710

Its clear from (Table1) that the proposed method preserves good transparency for the watermarked images.

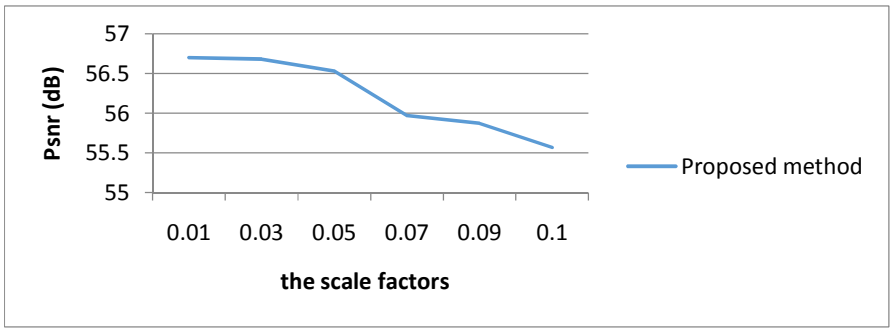


Figure 3. Relationship of the scale factors (α) and PSNR for the proposed method

(Figure 3) shows a relation between α and transparency in terms of the PSNR value. We can notice that our method has a good PNSR values and it reach its optimal value when the scale factors (α) is between 0.01 and 0.03.

Figure 4. Watermarked images

(Figure 4) shows the watermarked images used in the simulation, it can be seen that there is no perceptual difference between original and watermarked images that is also supported by good PSNR value. It shows that the good imperceptibility is obtained by proposed technique.
We use normalized correlation (NC) to evaluate the quality of the extracted watermark, and this measure is defined by:

$$NC(W,W') = \frac{1}{W_h \times W_w} \sum_{i=0}^{W_h-1} \sum_{j=0}^{W_w-1} W(i,j) \times W'(i,j) \tag{16}$$

with W_h and W_w are the height and width of the watermarked image ,respectively. $W(i,j)$ and $W'(i,j)$ denote the coefficients of the inserted watermark and the extracted watermark respectively.

As we can see in (Table 2) our proposed method is robust against different geometric and non-geometric attacks, and we can notice also that the relationship between the number of blocks of the watermark and the resulted *NC* generally are inversely proportional except for the rotation attack.

Table 2. Resulted NC for different image attacks using variation of watermarks.

Attacks / Nb blocks	Jpeg			speckle	imsharpen	Smooth	Rotation		salt & pepper	FFT	Filtre median	translate	Gaussian filter
	50	60	90	0.04			3°	5°	0.02			[20 35]	hsize = [5 5] sigma = 2
1	0.9941	0.9941	0.9941	0.9985	0.9996	0.9993	0.9684	0.8477	0.9998	0.9941	1.0000	0.9989	0.9994
3	0.9577	0.9577	0.9578	0.9952	0.9982	0.9982	0.9517	0.8560	0.9997	0.9581	0.9998	0.9941	0.9974
5	0.9236	0.9236	0.9238	0.9928	0.9972	0.9975	0.9471	0.8634	0.9996	0.9246	0.9993	0.9906	0.9947
10	0.8596	0.8598	0.8603	0.9900	0.9931	0.9941	0.9454	0.8822	0.9993	0.8623	0.9962	0.9882	0.9756
30	0.7551	0.7559	0.7584	0.9885	0.9830	0.9489	0.9582	0.9195	0.9985	0.7656	0.9615	0.9889	0.8280
64	0.7235	0.7249	0.7280	0.9867	0.9796	0.8864	0.9711	0.9423	0.9978	0.7373	0.9226	0.9911	0.7058
80	0.7102	0.7123	0.7143	0.9825	0.9782	0.8459	0.9740	0.9465	0.9961	0.7243	0.8979	0.9927	0.6360
100	0.7082	0.7107	0.7121	0.9807	0.9781	0.8389	0.9755	0.9487	0.9952	0.7220	0.8939	0.9929	0.6243
128	0.7072	0.7100	0.7115	0.9803	0.9781	0.8371	0.9757	0.9491	0.9945	0.7213	0.8924	0.9928	0.6214

Table 3. Comparison of PSNR for Lai et al. [13], Tsai et al [14] and our algorithm.

Method	the scale factors				
	0.01	0.03	0.05	0.07	0.09
Lai et al .[13]	51.14	51.14	50.89	49.52	47.49
Tsai et al [14]	47	37	33	28	about 25
Proposed method	56.70	56.68	56.53	55.97	55.87

(Table 3) shows the comparison of PSNR for two other algorithms and our algorithm. The values of the scale factors, (SFs) are carried out with constant range from 0.01 to 0.09 with an interval of 0.02. Size of host images are 256×256 for Lai et al. [13], and 512×512 for Tsai et al. [14] and our scheme.

Table 4. The comparison of robustness and imperceptibility (dB) for our scheme and Soumya et al. [12] under various image processing attacks.

Attack	**NC of the extracted watermark (Proposed Scheme)**	**PSNR of the Attacked watermarked image (Proposed Scheme)**	**NC of the extracted watermark [12]**	**PSNR of the Attacked watermarked image [12]**
Gaussian Lowpass Filtering (3×3)	**0.9991**	**56.1734 dB**	0.9974	41.2799dB
Average Filtering	**0.9946**	**52.6980 dB**	0.8253	35.7368 dB
Image Noising by salt-and-pepper noise	**0.9927**	**50.1413 dB**	0.9009	24.7876 dB

image enhancement by histogram equalization	0.9122	**40.0745 dB**	**0.9254**	12.1774 dB
Center-cropped attack (*64 ×64* pixels) and filled with pixel value *0*	0.9178	**41.4920 dB**	**0.9417**	20.8188 dB
Center-cropped attack (*64 ×64* pixels) and filled with pixel value *255*	**0.9582**	**24.7561 dB**	0.8983	15.0842 dB
Center-cropped attack (*128 ×128* pixels) and filled with pixel value 0	0.8793	**26.3998 dB**	**0.8979**	15.3449 dB
Center-cropped attack (*128 ×128* pixels) and filled with pixel value *255*	**0.9577**	**13.8987 dB**	0.8743	8.8648 dB
JPEG Compression (QF=25)	**0.9979**	**55.4146 dB**	0.9281	35.9800 dB
JPEG Compression (QF=50)	0. 9979	**55.4132 dB**	**1**	38.7407 dB

The bold values indicates the best values comparing with the others.

(Table 4) presents the comparison of robustness and imperceptibility (dB) for our scheme and Soumya et al. [12] under various image processing attacks, besides the robustness results the proposed scheme achieved high imperceptibility compared with Soumya et al. [12

.

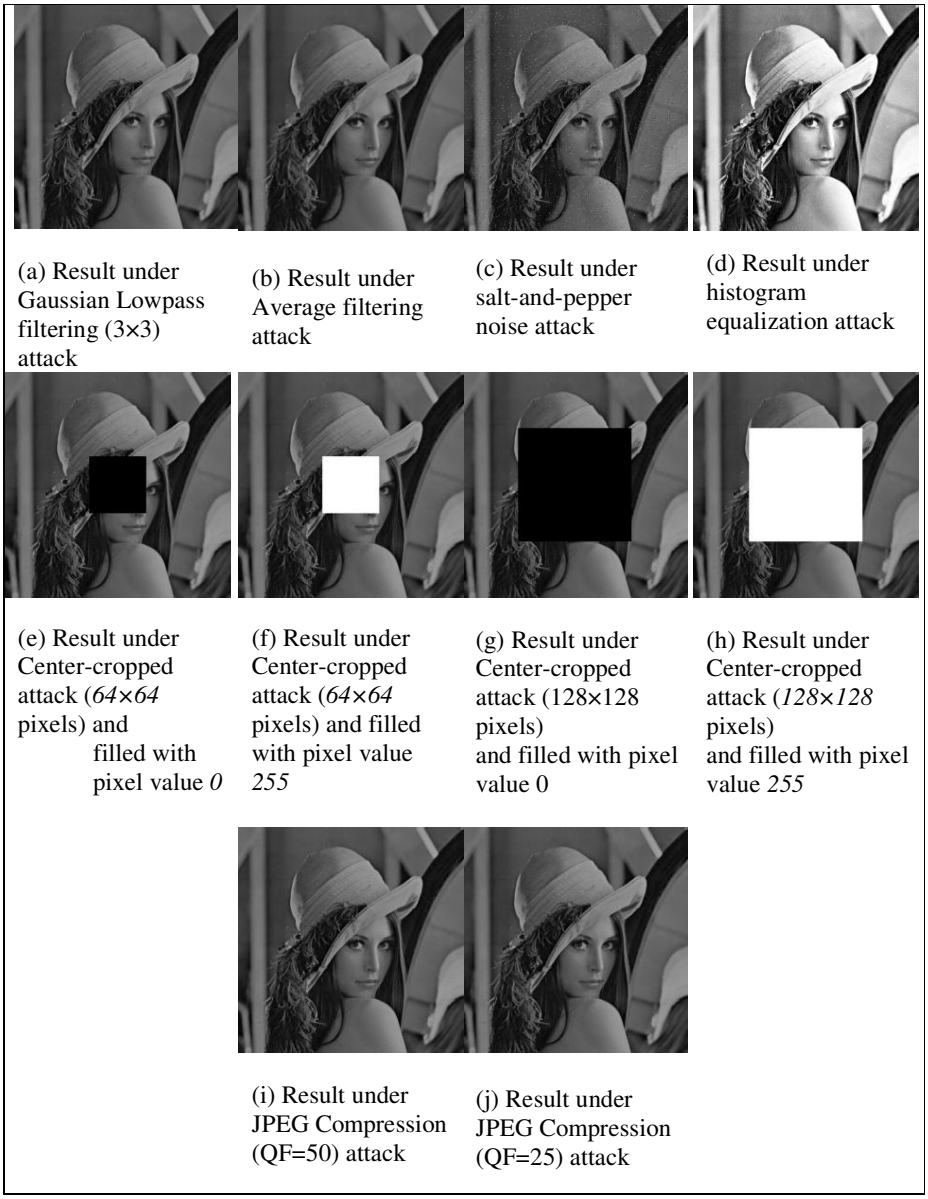


Figure 5. The attacked watermarked image under various image processing attacks in the comparison with Soumya et al. [12]

(Figure 5) shows the attacked images of the comparison with the scheme of Soumya et al.[12].

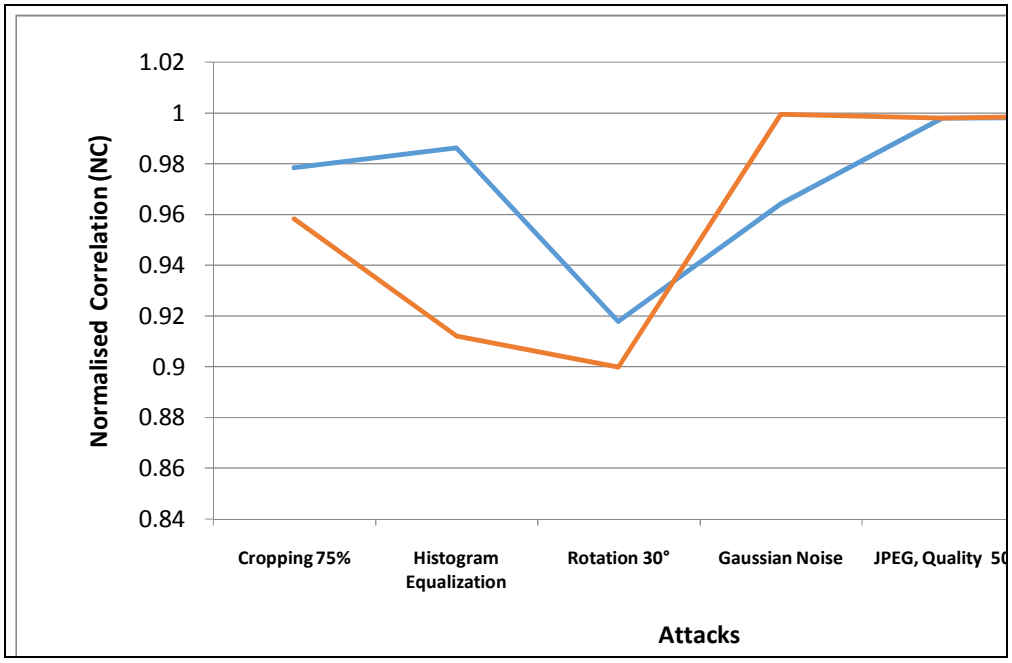


Figure 6. Comparison of our algorithm and Musrat et al [15] in term of *NC* values

(Figure 6) presents a comparison of robustness between the proposed method and Musrat et al [15] method, it shows that Musrat's scheme is more robust against such attacks as cropping 75% , histogram equalization and rotation 30° compared to our method but ours still achieves a good NC values after applying those attacks, and the minimum value of NC was 0.9 which means its robust, and for the other attacks (Gaussian noise, JPEG compression and Translation) our method perform better than Musrat's scheme.

Table 5**.** The comparison of robustness for our scheme, Nasrin et al [17] , Lai et al. [13] and Rastegar et al. [16].

Attack	Proposed scheme	Nasrin et al [17]	Lai et al .[13]	Rastegar et al.[16] [a]	Rastegar et al.[16] [b]
Pepper & salt noise (0.3)	**0.9927**	0.8926	–	0.7515	0.8258
Speckle noise (var=0.01)	**0.9950**	0.952	–	0.9609	0.9667
Gaussian noise (M=0,var=0.5)	**0.9210**	0.8935	–	0.7926	0.82
Gaussian filtering (3 ×3)	**0.9990**	0.987	–	0.8023	0.9843
Median filtering (3×3)	**0.9885**	0.982	0.9597	0.7534	0.9706
Wiener filtering (3×3)	0.9826	**0.984**	–	0.9824	0.9569
Sharpening	**0.9966**	0.932	–	0.9687	0.9511
Histogram equalization	0.9122	**0.990**	0.9862	0.9648	0.9628
Gamma correction (0.7)	0.9887	0.9935	**0.9982**	–	–
Gamma correction (0.8)	0.9890	**0.9950**	–	0.7203	0.9217
JPEG compression Q = 30	**0.9937**	0.987	–	–	–
JPEG compression Q = 10	**0.9915**	0.972	0.9772	0.9824	0.9843
JPEG compression Q = 5	**0.9907**	0.952	–	0.8532	0.9354
Scaling zoom(out = 0.5, in = 2)	**0.9772**	0.948	–	0.5127	0.953
Rotation (angle = 2°)	0.9648	**0.981**	–	0.5068	0.9628
Rotation (angle =–30°)	0.8532	**0.9823**	0.9780	–	–

'–' means the attacks are not done.
The bold values indicates the best values comparing with the others.
a Indicates the first scheme in Rastegar et al. [16]
b Indicates the second scheme in Rastegar et al.[16]

(Table 5) shows the proposed results with Rastegar's schemes results and Lai's scheme results and Nasrin et al's scheme when scaling factor is 0.05. Rastegar scheme (a) represents the embedding in all sub-bands while Rastegar scheme (b) represents the embedding in LH and HL only. Our scheme performed better than Nasrin's , Lai's and Rastegar's schemes as shown in (Table 5) .

5. CONCLUSIONS

This paper presents a blind robust digital image watermarking scheme based on singular value decomposition and on circulant matrix for copyright protection, using the circulant matrix's properties we improved the algorithm of Chandra and turned it to a blind watermarking algorithm after it was a non blind in addition to the augmentation of its robustness. Simulation results indicate that the proposed method achieves higher robustness compared to other known watermarking methods. The proposed method is robust against a wide range of common image processing attacks.

REFERENCES

[1] D. S. Chandra, "Digital image watermarking using singular value decomposition," Proceedings of the 45th mid-west Symposium on Circuits , Systems (MndWSCAS'02), vol.3, 2002, pp.264–267.

[2] B. WR,D.Gruhl,N.Morimoto,A.Lu,Techniques for data hiding ,IBM Systems Journa l35(1996)313–336.

[3] C.S. Lu,Multimedia Security :Stenography and Digital Watermarking Techniques for Protection of Intellectual Property ,PA Pub: Idea Group Pub,Hershey157–172.

[4] M.D.Swanson,M.KabayashiA.H.Tewfik ,Multimedia data embedding and watermarking technologies ,in:IEEE,vol.86,no.6, June 1998,pp.1064–1087.

[5] F.Hartung M. Kutter, Multimedia watermarking techniques,in: IEEE, vol. 87, no.7,July1999,pp.1079–1107 .

[6] W. H. Lin, Y. R. Wang, and S. J. Horng, "A Blind Watermarking Scheme Based on Wavelet Tree Quantization," The Second International Conference on Secure System Integration and Reliability Improvement, 2008, pp. 89-94.

[7] M. Ouhsain and A. B. Hamza, " Image watermarking scheme using non-negative matrix factorization and wavelet transform," Journal of Expert Systems with Applications, vol. 36, pp. 2123-2129, 2009.

[8] R. Liu and T. Tan, "A SVD-based watermarking scheme for protecting rightful ownership," IEEE Transactions on Multimedia, vol. 4, no. 1, pp. 121-128, 2002.

[9] E. Ganic and A. M. Eskicioglu, "Robust DWT-SVD domain image watermarking: embedding data in all frequencies," ACM Multimedia and Security Workshop 2004, Germany, 2004, pp. 20-21.

[10] Makhloghi, M.; Akhlaghian, F.; Danyali, H., "Robust digital image watermarking using singular value decomposition," Signal Processing and Information Technology (ISSPIT), 2010 IEEE International Symposium on , vol., no., pp.219,224, 15-18 Dec. 2010

[11] C.H. Lin, J.C. Liu, and P.C. Han, "On the security of the full-band image watermark for copyright protection," IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, 2008, pp. 74-79.

[12] Soumya Mukherjee and Arup Kumar Pal. 2012. A DCT-SVD based robust watermarking scheme for grayscale image. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI '12). ACM, New York, NY, USA, 573-578.

[13] C. Lai and C. Tsai, "Digital Image Watermarking Using Discrete Wavelet Transform and Singular Value Decomposition," IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 11, November 2010.

[14] H.-hsu Tsai, Y.-jie Jhuang, and Y.-shou Lai, "An SVD-based image watermarking in wavelet domain using SVR and PSO," Applied Soft Computing Journal, vol. 12, no. 8, pp. 2442-2453, 2012.

[15] Musrrat Ali and Chang Wook Ahn. 2013. An optimized watermarking technique employing SVD in DWT domain. In Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication (ICUIMC '13). ACM, New York, NY, USA, , Article 86 , 7 page

[16] Rastegar S, Namazi F, Yaghmaie K, Aliabadian A. Hybrid watermarking algorithm based on singular value decomposition and radon transform. Int J Electron Commun (AEU) 2011;65(7):658–63.

[17] Nasrin M. Makbol, Bee Ee Khoo, Robust blind image watermarking scheme based on Redundant Discrete Wavelet Transform and Singular Value Decomposition, AEU - International Journal of Electronics and Communications, Volume 67, Issue 2, February 2013, Pages 102-112, ISSN 1434-841

INTENTIONAL BLANK

ANOTHER PROOF OF THE DENUMERABILITY OF THE COMPLEX NUMBERS

J. Ulisses Ferreira

Trv. Pirapora 36 Costa Azul 41.770-220
Salvador, Brazil
ulisses@ferreira.mat.br

ABSTRACT

This short paper suggests that there might be numerals that do not represent numbers. It introduces an alternative proof that the set of complex numbers is denumerable, and also an algorithm for denumerating them.

KEYWORDS

Enumeration, Georg Cantor, diagonal process

1. INTRODUCTION

In 2004 [1] and 2005 [2], I demonstrated that C, i.e. the set of the complex numbers, is denumerable. Before that, it was already known that Q, i.e. the rational numbers, is denumerable, and that repetitions of the same numbers, such as 1/2, 2/4 and 3/6 in any denumeration such as

0, 1/1, -1/1, 1/2, -1/2, 2/1, -2/1, 1/3, -1/3, 2/2, -2/2, 3/1, -3/1...

caused no problem. The order of the numbers does not matter either. Furthermore, I observed that the Cantor's diagonal process was based on numerals, instead of numbers.

One cannot arbitrarily invent any set of new numbers. Historically, the zero and negative integers were discovered because there was some need: the subtraction operation. Then, the rational numbers were discovered since there was the need for dividing numbers, and so on. Thus, except for the natural numbers, the notion of number expresses quantity but it is also the result from a sequence of arithmetical operations. Pi appeared because of the geometry, and so on. Thus, there do exist irrational numerals between 0 and 1 that do not represent numbers, possibly most of them, if we can talk in this way. This means that using randomizing functions such as the well known **randomize** and **rand** in a programming language, even if the computer memory was infinite, will produce numerals based on decimal digits, but not necessarily numbers. Since the memory is finite, they do not even produce irrational numbers in such a representation. Not even 10/3 can be represented on a computer using decimal digits only.

Cantor's diagonal process just seems to work because it was a deduction over symbols.

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 79–82, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4407

In [1,2], I also proved that not only R but also C are denumerable. In this short paper, I introduce a proof that is easier to be understood.

Note that MMXII (Roman) and 2012 (Arabic) are two different numerals that represent the same number, and that this is much older than Cantor and our culture. Whether there are or not numerals in decimal Arabic notation that are supposed to represent some irrational number but do not represent any number, the following proof works.

2. THE PROOF

Here, I use the coding and decoding system previously used by Kurt Gödel in the proof of his historic theorem of incompleteness, published in 1931 [3], which, in its turn, was inspired by Whitehead and Russell´s work entitled *Principia Mathematica* [4]. The used numbers are better known as "Gödel Numbers".

As a professional originally from the area of programming languages design and implementation, whenever I do not have suitable means for representing any arithmetic expression, I simply write the expression as if I was using a subset of some full programming language, such as BASIC, Pascal, Java, Haskell or any other, taking into consideration that I use the symbol "^" to represent power like in BASIC, instead of "**" like in FORTRAN and that some languages do not provide the power function. In this way, those are the *numerals* that I chose for representing the set C of the complex numbers: for instance, I represent the constant *e* by simply "e"; For the irrational number π, I simply represent it by "pi"; The irrational number which is the square root of -1, I represent by "sqrt(-1)", but as I wish to also represent all complex numbers, I simply write "i" for representing the same number, and so on. A few examples of numerals of mine are "10/3", "sqrt(2)", "1+3/(sin(30)*2)", "log(2,10)-pi", "e^32", "2+6*i", each of them representing its corresponding complex number. Furthermore, since they are arithmetic expressions only, not logical propositions nor predicates nor function definitions, there is no self-references or paradoxes involved.

Given a coding table for characters such as ASCII, the next step consist in calculating the Gödel numbers that correspond to those strings, using a function that I would naturally define having G as its identifier. Let *c* be the function that receives a character and results in its ASCII code (or, alternatively, any more suitable code from a shorter table), and, skipping 1, let us use the same sequence of the first smaller prime numbers such as the one used by Gödel while proving that theorem of his, I obtain the following examples:

G("e") = 2^c('e').

G("pi") = 2^c('p') * 3^c('i').

G("10/3") = 2^c('1') * 3^c('0') * 5^c('/') * 7^c('3').

G("2+6*i") = 2^c('2') * 3^c('+') * 5^c('6') * 7^c('*') * 11^c('i').

One can observe that there are infinite Gödel numbers that do not represent any complex number, but this is not a problem since both sets are infinite and any subset of the natural numbers is denumerable.

Since the function G always receives a complete expression as a string and results in a natural number, and, using the numerals that I chose as strings containing arithmetic expressions, I can represent any complex number, the set of the complex numbers is denumerable.

Remark: Note that, as in any coding function using Gödel numbers, there is an inverse function that, from any natural number obtained from the above explained manner, results in the string that contains the original arithmetic expression.

Theorem 2. *There does exist some algorithm that denumerates the set of the complex numbers.*
Proof. I will prove constructively, showing my algorithm: first, I define an alphabet, a formal language for arithmetic expressions and write a parser for expressions of this language, which, in its turn, contains only the alphabet defined for constant expressions, operators and symbols such as "(", "," and ")" used in expressions, identifiers of all arithmetic functions such as "log" and "sin", and reserved words of constant values, such as "e", "pi" and "i", but that language does not contain any variable. One typically uses BNF for doing so.

A question is to know what the first number of my denumeration is. Thus, I have to set an order. Let G' be the inverse function that receives a Gödel number and results in the original arithmetic expression. Starting from the natural number n = 2, we apply G' and verify to know whether the resulting string denotes any well-formed arithmetic expression, in accordance with the formed language previously defined. If true, this is the first numeral of my denumeration. Otherwise, we can try using n = 3, then n = 4, then n = 5, and so on, until we find the first natural number n whose G'(n) application results in some arithmetic expression well formed. The obtained string is the first numeral, which, in its turn, represents the first number of my denumeration.

Moreover, we obtain the other numbers in order using the following algorithm, in Algol 60 or Pascal style:

```
var n, ok: integer; s: string;
n := 1; ok := 1000;
while ok = ok do
begin
   n := n + 1;
   (* Gprime means G' as above *)
   s := Gprime(n);
   if parser(s) = ok then
     writeln(s)
end;
```

In the end, what is written by the above algorithm is a sequence of complex numbers, i.e. a denumeration of them.

3. CONCLUSIONS

There are no transfinite numbers, and there is no set with cardinality greater than infinite. Cantor, Post and others collapsed the older notions of numbers and numerals, and that was a mistake. There is no need for the notion of "computable numbers" either, since what was called "computable number" is just a number whose numeral in the Arabic decimal notation can represent another number up to some given precision.

REFERENCES

[1] Ferreira, Ulisses, (2004) "The real set can be seen as denumerable", In Hamid R. Arabnia, Iyad A. Ajwa, and George A. Gravanis, editors, *Post-Conference Proceedings of the 2004 International Conference on Algorithmic Mathematics & Computer Science,* CSREA Press, Las Vegas, Nevada, USA, pp523-526.

[2] Ulisses Ferreira (2005) The sets of real and complex numbers are denumerable. *ACM SIGACT News* 36(2): pp126-130.

[3] Gödel, Kurt (1931) Uber formal unentscheidbare sätze der principia mathematica und verwandter system i. *Monatschefte Mathematik und Physik.* 38, pp173-98.

[4] Whitehead, Alfred & Russell, Bertrand (1910) *Principia Mathematica*, Cambridge University Press, 3 volumes, 13.

Author

Ferreira was born in Salvador in 1961, studied and did some research work in computer science at UFBA, UFPB, Sussex University, Edinburgh University and Trinity College in Dublin.

USING RELATIONAL MODEL TO STORE OWL ONTOLOGIES AND FACTS

Tarek Bourbia and Mahmoud Boufaida

LIRE Laboratory, University Constantine 2, Algeria
{bourbia_tarek, boufaida_mahmoud}@yahoo.fr

ABSTRACT

The storing and the processing of OWL instances are important subjects in database modeling. Many research works have focused on the way of managing OWL instances efficiently. Some systems store and manage OWL instances using relational models to ensure their persistence. Nevertheless, several approaches keep only RDF triplets as instances in relational tables explicitly, and the manner of structuring instances as graph and keeping links between concepts is not taken into account. In this paper, we propose an architecture that permits relational tables behave as an OWL model by adapting relational tables to OWL instances and an OWL hierarchy structure. Therefore, two kinds of tables are used: facts or instances relational tables. The tables hold instances and the OWL table holds a specification of how the concepts are structured. Instances tables should conform to OWLtable to be valid. A mechanism of construction of OWLtable and instances tables is defined in order to enable and enhance inference and semantic querying of OWL in relational model context.

KEYWORDS

Relational Model, Database, OWL, Instance, Fact, Ontology

1. INTRODUCTION

A database model defines the logical structure of database and determines in which manner data can be stored, organized and manipulated. It is still evolving and new models are being considered, especially in the semantic aspects. Despite this evolution, the relational model is still the most used and none other model even made the end of the dominance of relational databases. Its simplicity and its performances motivate research to design new models on the relational engine by operating, mapping and transformation mechanisms, as is achieved -for instance- by object-relational, XML[2], RDF models [6] [7].

Besides, the field of knowledge representation [20] defines the syntax of ordered symbols that are readable and interpretable by the machine. In the same context the semantic web [9] aims to ensure for computer to analyze data contents and their relationships. It enables structuring and storing the web knowledge. One of the motivations considered by semantic web, through its models, is the semantic search. In this context, ontological languages are proposed to formulate knowledge bases to improve the retrieval in documents repositories.

The field of database model offers the best way to store data efficiently and guarantees its persistence; the knowledge representation languages provide a rich vocabulary and a power of

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 83–98, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4408

expressiveness. The combination of those two fields allows providing solutions to manage knowledge semantically and efficiently with great performances.

The Web Ontology Language (OWL) [1] supplies for developers more than eXtensible Markup Language (XML), Resource Description Framework (RDF) [6]. With a wide formal vocabulary, OWL [5] is a formalism to build a knowledge base. However, if the quantity of the knowledge is very important and the size of the OWL domain is voluminous, it will be complex to treat them by managing a document in a simple file format. Thus, it would be beneficial to embed these ontologies on a database management system and to manage them as databases. The proposed approach, allows the storage of metadata and ontology instances into a persistent and optimal way, while preserving semantics and constraints at an abstract level.

It is important in engineering to implement database based ontology onto an existing database system. It is also interesting to integrate ontology features with an existing database model. Using an existing infrastructure does not deprive the ontologies-based database to be native.

OWL contains three sublanguages [1]: OWL Lite, characterized by a hierarchy and a simple mechanism of constraints; OWL DL, based on the description logic with offering a high degree of expressiveness and ensures computational completeness and decidability; OWL Full, has a big capacity of expressiveness, without ensuring completeness and decidability. OWL DL language has been chosen here to express the knowledge base. It contains all the structures of OWL with some limitations such as the separation of types (a class cannot be at the same time an individual and a property), an important thing is to not confuse the data and the metadata.

A semantic web application requires storing and manipulating enormous data. Storing an ontology and its instances and processing them by a user application need some mechanisms to cope, on the one hand, with huge numbers and relations between concepts and, on the other hand, with large amounts of instances. Relational database is a good support as repository to ensure the persistence of ontology instances, due to its experience, performance and features. For this reason, it has been chosen in this paper to embed OWL DL knowledge bases.

To take benefit of the structure expressed in OWL and at the same time of the database system of which embeds the ontology instances, this paper provides a solution to explicit a graph structured document as OWL into relational tables with preserving constraints and relationships. It aims at ensuring the persistence of a huge number of OWL concepts and a large amount of instances in a native way without using the classical way of triples embedded into relational table of three columns [17]. The objective of this work is to present a mechanism of how to store OWL concepts and instances in relational tables while preserving the semantic and without losing data. The main advantage of this solution is to have a specification and conception under the OWL-DL language, a logical modelling under relational model, and a physical storage into the relational database. Therewith the distance is close between the conceptual model and the logical model in matters of preserving data, link and constraint. For that purpose, it could be said that this work enters in the field of the Ontology-Based Database.

This paper is organized as follows: The following section presents the works carried out on this subject, section (3) shows the principles of the proposed approach of how storing OWL ontology into relational tables and section (4) details the mechanism of storing OWL ontology concepts and facts, followed by a conclusion and some perspectives.

2. SOME RELATED WORKS

In order to manage ontologies efficiently and to ensure a robust persistence, several knowledge management systems have been proposed. Each work in this context uses its own strategies to embed ontologies in a persistent data model.

Pierra et al. have proposed a database architecture model called "Ontology-Based Database" (OntoDB) [10]. The latter defines separately the implementation of the ontology and those of data. The solution consists of two parts: representing the primitive of the ontology model in a meta-schema and defining, once and for all, the physical storage of ontologies from RDF triplets by using the relational model and the object-relational one. ONTOMS [3] developed by Myung et al. is another OWL database management system architecture, whereby the data is physically stored in classes modeled into relational tables. It also performs complex operation as inverseOf, symmetric, transitive and reasoning for instances. ONTOMS also evaluates the requests expressed by OWL-QL [4]. M. Shoaib has developed ERMOS [11], a solution that provides an efficient transformation of the OWL concepts to relational tables to store ontology and to allow easy and fast knowledge retrieval with semantic query. JENA [18] is a framework for building Semantic Web applications. It allows processing RDF and OWL ontology by providing a container for collections of RDF triples. The implementation of SPARQL [8] is used in JENA to query RDF triples. Although JENA provides interaction with OWL ontology, its OWL instances, as well as for RDF, are based on RDF triples.

The two architectures ONTOMS and ERMOS adopt a representation in OWL, which are richer in concepts and semantics by report of what is chosen by other work presented above. The challenge in this kind of work is to make it possible to store ontologies in a right, coherent, scalable and efficient way in order to retrieve knowledge by semantic inference. In the literature, the near total of work aiming at the management of ontologies as a database use the relational model as logical and physical model to shelter the ontology-based data.

All the precedent systems are founded on an external middleware layer that is added on top of a database management system engine. These systems keep users far from ontology instances and do not provide ability to interact with stored ontology facts directly.

The triplet-based approach has been used by several systems, which aim to ensure persistence of OWL ontologies. It solves, somehow, the issue of the integration of semantics in the field of databases, and improves the way of storing and dealing with big amounts of OWL instances. However, the triplet-based approach presents a number of drawbacks such as:

- Losing the power of expressiveness of OWL ontology.
- Storing all triples extract from OWL ontology in the same table. This make self-join complex and with less performance when retrieving instances
- Scanning all table triples to reach the needed triples.

Several works based on NoSQL [15] models use the structure offered by the XML language to store data and to locate them [16]. XML represents a significant evolution of the concept of database to store large volumes of data or documents. An XML database defines a logical model in a XML document, and stores and retrieves the documents according to that model. The structure in graph of the ontology language, inspired from XML, is far from the tables of the relational model. For this purpose, there exists some works that deal with the mapping or the correspondence of ontology model with XML tree [2]. It seems more relevant than its

correspondence with relational table, by the fact that XML is the first pillar to build assertional languages and ontology. Nevertheless, XML databases are built into relational engines and applying mapping under "mapping" could lead to degrade the performances and cause data loss.

Besides, there exist other works, in the opposite direction, that aim to ensure the extraction of ontology from relational database schema [12]. These works construct a local ontology from data that already exist in a relational database. Their goal is not to solve problems of huge ontologies and instances, but to deal with data stored in relational model semantically.

Summarizing, storing ontology instances using a file system makes querying those instances very difficult and with poor performances. To cope with huge ontology, most of the works in the literature use mapping mechanisms from the ontology into relational or objects databases. And that would not be fluent and work perfectly in terms of performance due to large ontology vocabulary and complexity of graph. The proposed approach in this paper allows storing OWL concepts and instances in relational tables without losing semantic, nor data, and enables users to interact with data directly. Compared with existing methods [13] [14], the proposed approach keeps faithfully a class hierarchy and relationship between concepts. In addition, it provides capabilities to interact directly with instances, unlike other existing methods.

3. OVERVIEW OF THE ARCHITECTURE

The architecture of storing OWL ontology into relational tables resulted after combining some aspects of the field of databases and those of the semantic web and knowledge representation. This architecture is designed to allow both the efficient management of the ontology concepts and instances, and to ensure semantic search and update regardless of the physical data structure.

The essential functions of the proposed system are (see Figure 1):

- Storing and manipulating OWL Ontology in relational model. A meta-concept table named “OWLtable” represents the classes and relationships between classes and the properties ensure predicates between classes and individuals. In function of this table, the OWL ontology will be constructed and represented by a graph or an OWL expression. In fact, the unique OWL ontology domain that exists is that stored in relational table into “OWLtable”. This way allows approaching the semantic conceptual model and the logical model, and avoiding losing concepts that suffer the architectures based on transformation mechanism and mapping solution.

- Storing and manipulating ontology instances and facts. A set of relational tables represents classes and some properties and holds individuals of the OWL ontology as tuples. The meaning of tuples is preserved, and each instance is associated with an ontological concept stored into “OWLtable” and referenced by both table name and attribute. Instances tables and instances have to be conforming to the structure of OWL concepts and match with constraints. In other words, any operation in the set of instances tables should be checked and validated with concepts structure and constraints stored in “OWLtable”.

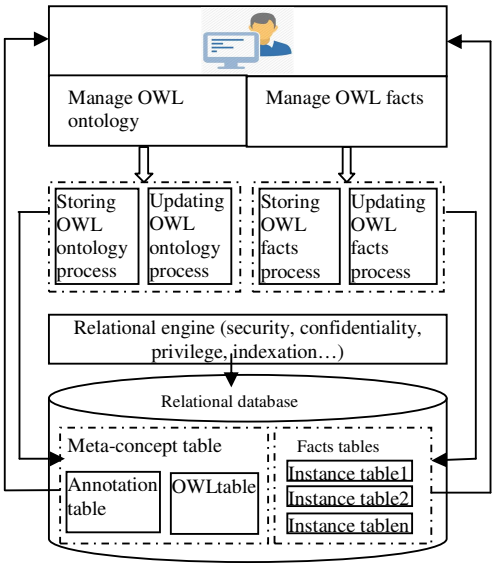


Figure 1:Storing ontology architecture.

The proposed architecture consists of a native OWL database engine built on the system of a relational database and integrated with the relational engine (definition and manipulation language). This approach allows storing, accessing and searching on the OWL using relational tables to ensure the persistence of data linked with OWL concepts and properties.

In fact, neither OWL document, nor OWL instances exist; what exists is their equivalent in relational database that are “OWLtable” and instances tables. The purpose of “OWLtable” is to define concepts of OWL ontology, and so instances tables will be an instance document to describe an OWL individuals and facts that conforms to “OWLtable”.

The “OWLtable” contains kinds of elements like: class, object property, data type property and constraints which give the content nature of instances tables and the hierarchy of element.

This approach allows manipulating OWL ontology in relational tables. Furthermore, accessing, reasoning and querying OWL ontology are performed in persistence layer.

The next section details the part of how using relational tables to store ontology classes, properties and individuals to make them ready for any actions later on.

4. STORING OWL ONTOLOGY INTO RELATIONAL TABLES

The mechanism adopted to reach the objective of managing OWL ontology in relational engine, contains a number of actions, as is shown in Figure 2. Those actions allow having: a meta-concept table named “OWLtable” that acts like OWL ontology; and a set of relational tables resulted from ontological concept which embed the facts and instances of OWL ontology. For this purpose, an ordered steps process is conceived here to create a set of relational tables that represent OWL concepts and facts without losing hierarchy of concept, links, constraints and facts. In other words, this set of ordered steps process ensures the correspondence between the semantic structure of the ontology as a graph with relationships, and the syntactical logical structure in the form of relational tables to ensure persistence of instances with the best performance possible.

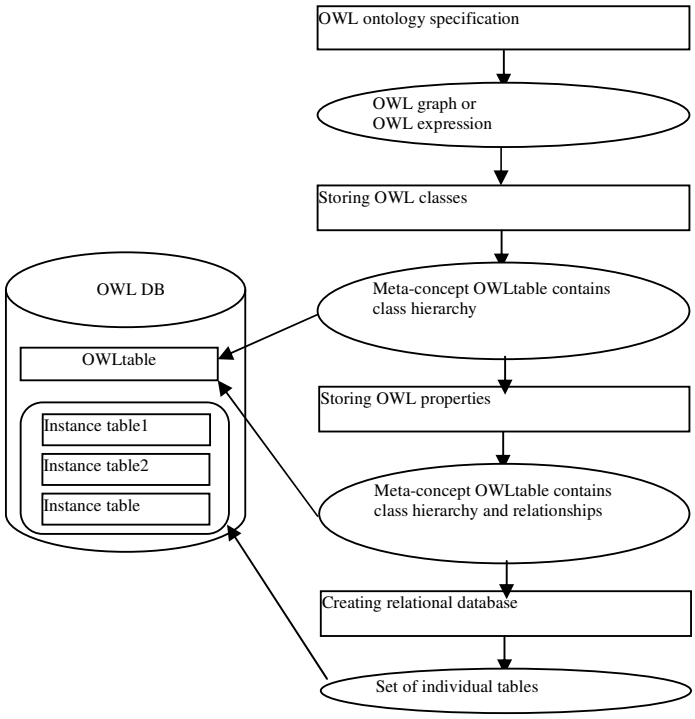


Figure 2:A Mechanism of storing OWL ontology into relational tables.

As shown in Figure (2), the creation of the meta-concept table "OWLtable" is the first step, then filling it with class hierarchy and properties, the next step is to create the database that is a set of instance tables obtained by parsing classes, data type properties and object properties. The set of instances tables hold facts.

In the following subsections, a number of processes are developed with rules to detail how storing classes and properties into a meta-concept table that keep the hierarchy and the relationships between concepts; and how storing facts and instances into relational tables that hold the database of OWL knowledge base. To illustrate that, an example of OWL ontology domain is used.

4.1. Storing OWL Class into Relational Table

The basic concept of ontology is the class. So, the most important step and rule is to perform the correspondent relational table to embed ontology classes. It is around class that any object property or data type property is defined. So the first correspondence applies to the classes to support the other concepts of OWL later on.

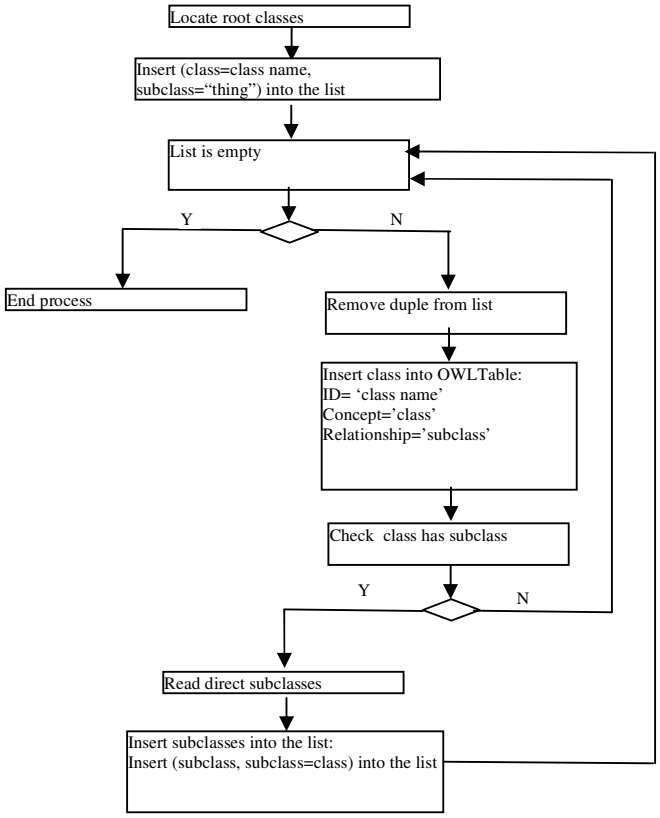


Figure3:The process of storing OWL classes into relational table.

The process of storing OWL classes into relational table that under the title "OWLtable" focuses on locating every OWL class, regardless of its nature and its position in OWL graph and on saving it with some information in order to keep hierarchical structure and type of every relationship with another class, as is shown in Figure (3). This process is based on the following rule:

RULE 1: for each class in OWL ontology, a record in "OWLtable" is inserted. This record defines: that is a class concept and the super class in relationship with this class. The super class of root classes is considered in this process as "thing", the super class of all classes in OWL specification [1].

EXAMPLE: the following classes presented in OWL syntax:

<owl:Class rdf:ID="Human"> …</owl:Class>
<owl:Class rdf:ID="Man"><rdfs:subClassOf rdf:resource="#Human"/></owl:Class>
<owl:Class rdf:ID="Woman"><rdfs:subClassOf rdf:resource="#Human"/></owl:Class>
<owl:Class rdf:ID="Country"> …</owl:Class>
<owl:Class rdf:ID="Town">…</owl:Class>

are inserted in OWLtable as shown in Table(1).

Table 1:OWLtable contains hierarchical of classes.

ID	Concept	Relationship
Human	Class	Thing
Man	Class	Human
Woman	Class	Human
Country	Class	Thing
Town	Class	Thing

At the end of the process of storing OWL classes into relational table, a hierarchy of classes is obtained from records of "OWLtable", and it will be considered as a meta-concept table for knowledge base. The next section is devoted to enrich the meta-concept "OWLtable" by another important ontological concept which is "OWL property".

4.2. Storing OWL Property into Relational Table

OWL property gives a way of describing a kind of relationship of individuals to individuals and of individuals to data values [1]. To complete OWL graph in relational field, on the other hand, to enrich class hierarchy already stored in the meta-concept "OWLtable", linked classes by predicates are treated by the process shown in Figure (4) to have a complete OWL graph in the meta-concept table. The process deals with a number of ontological concepts: object property, data type property, cardinality constraint.

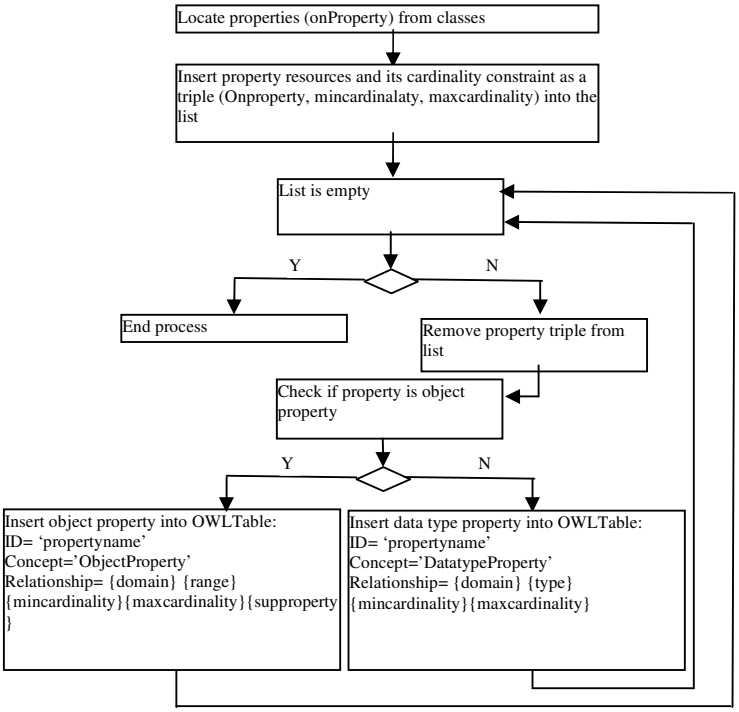


Figure 4:The process of storing OWL property into relational table.

The process of storing OWL predicates into "OWLtable" affects the two types of relationship of individuals: Object Property and DatatypeProperty. A number of restrictions that define property are preserved in "OWLtable", that are: domain, range and cardinalities. In the case of

ObjectProperty the generalization is preserved by sup property in meta-concept table. The property definition is formulated by the expression:

Relationship= {domain} {range} {mincardinality} {maxcardinality} [{supproperty}], as is shown in Figure (4). This process is based on these three rules:

RULE 1: for each onProperty in OWL class, a record in "OWLtable" is inserted. This record defines: that is an onProperty concept; the relationship that details predicates by a set a restriction: domain, range, cardinalities.

RULE 2: the type of onProperty is procured from property statement by checking each ID. There are two types: ObjectProperty and DatatypeProperty.

RULE 3: The super property of ObjectProperty is considered in this process and is added at the end of relationship expression:

Relationship= {domain} {range} {mincardinality} {maxcardinality}{supproperty}.

As an example, consider the following set of statements about properties presented in OWL syntax:

<!--**classes definition**-->

<owl:Classrdf:ID=" Human ">
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource="#HasFriend">
<owl:mincardinality rdf:datatype="&xsd:int">0< /rdfs:mincardinality>
</owl:Restriction>< /rdfs:subClassOf>
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource="#HasFather">
<owl:cardinality rdf:datatype="&xsd:int">1< /rdfs:cardinality>
</owl:Restriction>< /rdfs:subClassOf>
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource="#HasSpouse">
<owl:maxcardinality rdf:datatype="&xsd:int">1< /rdfs:maxcardinality>
</owl:Restriction>< /rdfs:subClassOf>
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource="#name">
<owl:cardinality rdf:datatype="&xsd:int">1< /rdfs:cardinality>
</owl:Restriction>< /rdfs:subClassOf>
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource="#Mail">
<owl:mincardinality rdf:datatype="&xsd:int">0< /rdfs:mincardinality>
</owl:Restriction>< /rdfs:subClassOf></owl:Class>……

<!--**object properties definition**-->

<owl:ObjectProperty rdf:ID=" HasFriend ">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="#Human"/></owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="HasFather">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="#Man"/></owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="HasSpouse">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="#Human"/></owl:ObjectProperty>……

<!--**data type properties definition**-->

<owl:DatatypeProperty rdf:ID="name">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="&xsd:string"/></owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="mail">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="&xsd:string"/></owl:DatatypeProperty>…….

Those properties are inserted in OWLtable as shown in Table(2).

Table 2:OWLtable contains hierarchical of classes.

ID	Concept	Relationship
Human	Class	Thing
Man	Class	Human
Woman	Class	Human
Country	Class	Thing
Town	Class	Thing
HasFriend	ObjectProperty	{Human}{Human}{0}{unlimited} {does not exist}
HasFather	ObjectProperty	{Human}{Man}{1}{1} {does not exist}
HasSpouse	ObjectProperty	{Human}{Human}{0}{1} {does not exist}
Name	DatatypeProperty	{Human}{String}{1}{1}
Mail	DatatypeProperty	{Human}{String}{0} {unlimited}

At the end of the process of storing OWL property into meta-concept “OWLtable”, a complete OWL graph is embedded in relational database with its class hierarchy and all predicates. The next section presents the way of storing facts and defines instances relational tables that are conform to “OWLtable”.

4.3. Storing OWL Facts into Relational Tables

OWL data type properties give “facts” that link individuals to data values [1]. Facts typically are statements indicating class membership of individuals and property values of individuals. In relational model [19], a table is a set of tuple (individual) that have the same attributes (table membership). A tuple usually represents an object and information about that object. All the data referenced by an attribute are in the same domain and conform to the same constraints. So by analogy, for each class that has individuals and property values of individuals, a relational table, given the same name of the class with attributes which are no longer than OWL data type properties, is the suitable container to ensure persistence of OWL instances in relational models, as is shown in Figure (5). Besides, to keep relationships and predicates between classes a relationship table, given the same name of the property and foreign keys attribute will be necessary to ensure the relationship between relational tables obtained from parsing OWL data type properties, as is shown in Figure (6).

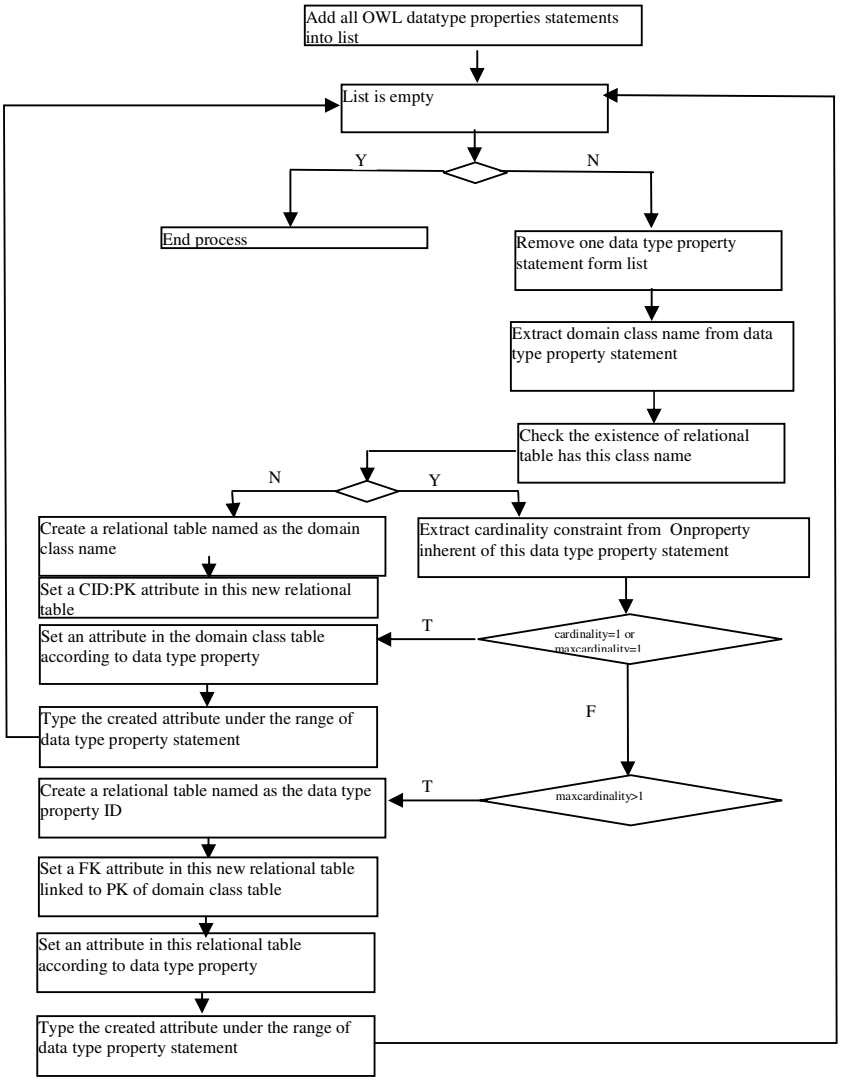


Figure 5:Process of creating relational tables to embed data type properties.

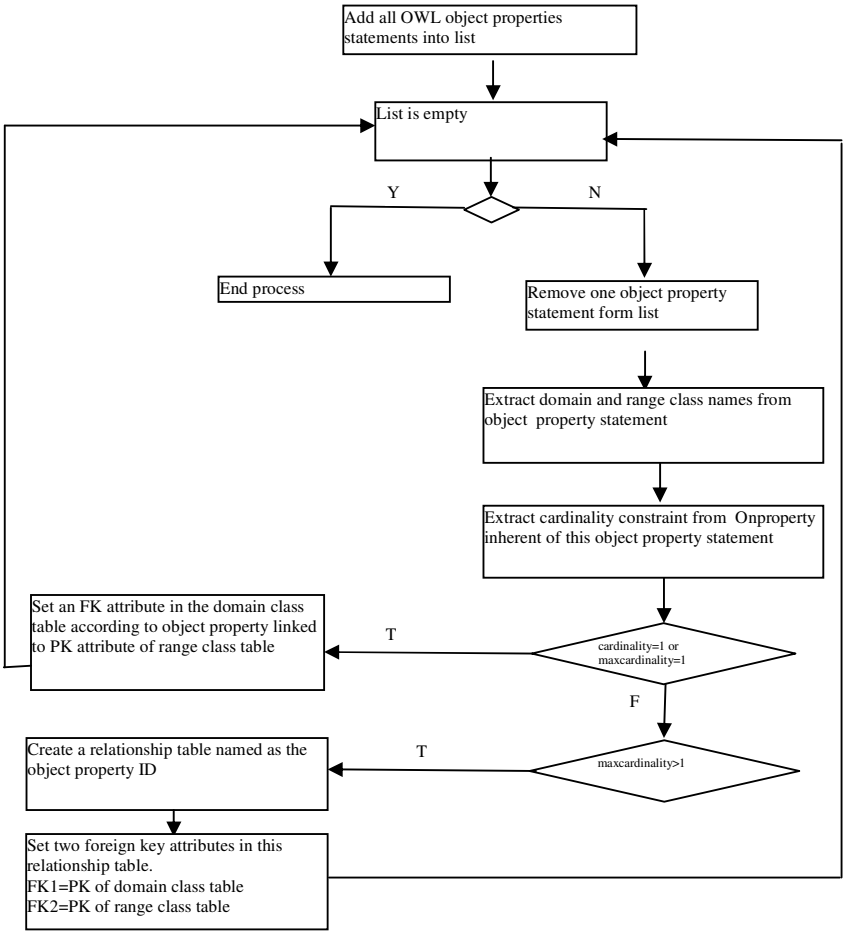


Figure 6:Process of creating relational tables to embed object properties.

After building the meta-concept table "OWLtable" that describe the ontology, the process of storing facts into relational table is founded on how OWL data type properties are grouped then embedded in relational table as attributes (Figure 5); and how OWL object properties are embedded in relational table as attributes with foreign key feature (Figure 6). The outcome relational tables are the repository of each part of fact. The semantic of each part is preserved in "OWLtable" and its correspondent attribute in tables. The mechanism, proposed here, to guarantee to have the appropriate environment to store and preserve OWL instances and facts in relational domain is based on six rules:

RULE 1: for each data type property statement, a relational table is created according to the domain class name if it has not been created by another data type property.

RULE 2: OWL classes must be unambiguously identifiable when using relational tables to store their facts. A class identifier CID for each OWL class is added as a primary key in its correspondent relational table.

The relational table obtained is shown in Table3:

Table 3:Domain class table.

Humain (table name)	
Attribute	Data type
CID	PK
.....	.....

RULE 3: If a cardinality constraint restricts an instance of a class to have at most one semantically value for a data type property, an atomic elementary attribute is set in the relational table that corresponds a domain class of this data type property statement.

EXAMPLE:

<owl:Classrdf:ID="Human">
<rdfs:subClassOf><owl:Restriction>
<owl:OnProperty rdf:resource="#name"/>
<owl:cardinality rdf:datatype="&xsd:int">1<rdfs:cardinality />
<owl:Restriction/><rdfs:subClassOf />……..
<owl:DatatypeProperty rdf:ID="name">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="&xsd:string"/></owl:DatatypeProperty>…..

The obtained relational table is shown in Table (4).

Table 4:Domain class table.

Human (table name)	
Attribute	Data type
CID	PK
Name	String
.....	.....

RULE 4: If a cardinality constraint restricts an instance of a class to have semantically more than one value for a data type property, a relational table is created and named as the data type property ID. A Foreign Key attribute is added in this new relational table linked to Primary Key (CID) of domain class table.

EXAMPLE:

<owl:Classrdf:ID="Human">
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource=”#Mail”>
<owl:mincardinalityr df:datatype="&xsd:int">0< /rdfs:mincardinality>
</owl:Restriction>< /rdfs:subClassOf></owl:Class>……
<owl:DatatypeProperty rdf:ID="mail">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="&xsd:string"/></owl:DatatypeProperty>…..

The relational tables obtained are shown in Figure (7).

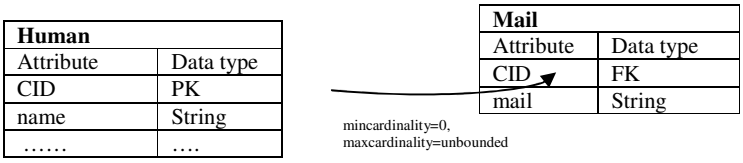


Figure 7:Multi value data type property

RULE 5: If a cardinality constraint restricts individuals of a class to have semantically more than individual for an object property, a relationship table named as the object property ID, is created. Two foreign key attributes are added in this relationship table to link individuals of domain class table to ones of range class table.

EXAMPLE:

<owl:Classrdf:ID="Human">
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource=”#HasFriend”>
<owl:mincardinalityrdf:datatype="&xsd:int">0< /rdfs:mincardinality>
</owl:Restriction>< /rdfs:subClassOf></owl:Class>……
<owl:ObjectProperty rdf:ID=" HasFriend ">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="#Human"/></owl:ObjectProperty>
The obtained relational tables are shown in Figure (8).

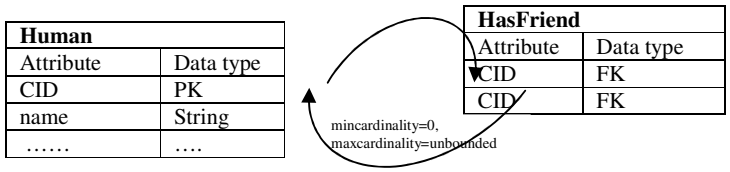


Figure 8:Relationship table

RULE 6: If a cardinality constraint restricts the individuals of a class to have at most one semantically individual for an object property, a foreign key attribute is added in the domain class table according to object property linked to the primary key attribute of range class table.

EXAMPLE:

<owl:Classrdf:ID="Human">
<rdfs:subClassOf><owl:Restriction><owl:onPropertyrdf:resource=”#live”>
<owl:cardinality rdf:datatype="&xsd:int">1< /rdfs:cardinality>
</owl:Restriction>< /rdfs:subClassOf></owl:Class>……
<owl:ObjectProperty rdf:ID="live">
<rdfs:domain rdf:resource="#Human"/>
<rdfs:range rdf:resource="#Town"/></owl:ObjectProperty>

The obtained relational tables are shown in Figure (9).

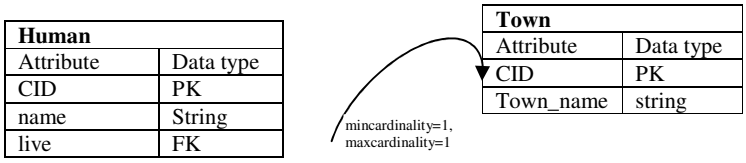


Figure 9:Object property foreign key

5. CONCLUSIONS

In this paper, we have presented an approach that aims at storing a huge OWL ontology and big amounts of instances in the relational engine efficiently without losing concepts neither instances. This approach allows the use of the expressiveness power of the knowledge representation language "OWL" as well as the facilities, and the efficiency that offer relational model. We have defined the meta-concept table "OWLtable" that embeds faithfully class hierarchy and OWL concepts with their relationships and constraints. A set of relational tables are created according to class and property features in order to store OWL instances. In fact, only "OWLtable" and a set of instances table are effective to represent natively the whole ontology. That is why users are able to access, manage and manipulate instances directly. More complex OWL concepts such as subproperty, value constraints and other class and property description [1] could be added into meta-concept table to enrich the OWL specification in relational model.

In order to enable querying ontology instances that are stored in a relational database, our future work will be focused on the refinement of the proposed architecture to integrate the mechanism of searching both OWL instances and concepts with the mechanism of updating. The principle of these mechanisms is based on parsing the meta-concept table "OWLtable" before any query run on OWL instances tables. This leads to check on the one hand, the well-formed and the validity of OWL instances tables according to "OWLtable", and on the other hand, to check constraints and locate properties that give a semantic dimension to the query.

REFERENCES

[1] S.Bechhofer, F.V.Harmelen, J.Hendler, I.Horrocks, D.L.McGuinness, P.F.Patel-Schneider & L.A.Stein, (2004) "OWL Web Ontology Language Reference, W3C Recommendation", http://www.w3.org/TR/owl-ref.

[2] T.Bray, J.Paoli, C.M.Sperberg-McQueen, E.Maler & F.Yergeau, (2008) "Extensible Markup Language (XML) Version 1.0. (fifth edition), W3C Recommendation", http://www.w3.org/TR/REC-xml.

[3] P.Myung, L.Ji-Hyun, L.Chun-Hee, L.Jiexi, O.Serres & C.Chin-Wan, (2007)"ONTOMS : An Efficient and Scalable Ontology Management System", Springer Advances in Databases: Concepts, Systems and Applications, Vol. 4443/2007, pp 975-980.

[4] R.Fikes, P.Hayes & I.Horrocks, (2004) "OWL-QL-A Language for Deductive Query Answering on the Semantic Web", Journal of Web Semantics 2(1), pp 19–29.

[5] F.Gandon, C.F.Zucker & O.Corby, (2012) Le web sémantique: Comment lier les données et les schémas sur le web?, Dunod, France, pp 83-131.

[6] G.Klyne, J.J.C aroll & B.McBride, (2004) "RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation", http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[7] D.Brickley & B.McBride, (2004) "RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation", http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

[8] E.Prud'hommeaux & A.Seaborne, (2008) "SPARQL Query Language for RDF, W3C Recommendation", http://www.w3.org/TR/rdf-sparql-query.

[9] T.Berners-Lee, J.Hendler & O.Lassila, (2001) "The Semantic Web", Scientific American, 284 (5), pp 34-44.

[10] H.Dehainsala, G.Pierra & L.Bellatreche, (2007) "Ontodb: An ontology-based database for data intensive applications". InProceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA'07), Lecture Notes in Computer ScienceSpringer, pp 497-508.
[11] M.Shoaib & A.Basharat, (2010) "ERMOS: An Efficient Relational Mapping for Ontology Storage". 2010 IEEE International Conference on Advanced Management Science (ICAMS 2010), Chengdu, China, pp 399 – 403.
[12] H. A.Santoso, S.C.Haw and Z.T.Abdul-Mehdi, (2011) "Ontology extraction from relational database: Concept hierarchy as background knowledge". Knowledge-Based System 24 Elseiver , pp 457-464.
[13] E.Vysniauskas & L.Nemuraite, (2006) "Transforming Ontology Representation from OWL to Relational Database". Information Technology And Control, Kaunas, Technologija, Vol. 35A, No. 3, pp 333 - 343.
[14] E.Vyšniauskas, L.Nemuraitė & A.Šukys, (2010) "A hybrid approach for relating OWL 2 ontologies and relationaldatabases". In: P. Forbrig, H. Gunther (Eds.): Perspektives in Business Informatics Research. Proceedings of the 9th international conference, BIR 2010, Rostock, Germany, September 29 - October 1. Berlin-Heidelberg-New York, Springer, 2010, pp 86–101.
[15] J.Han, E.Haihong, G.Le & J.Du, (2011) "Survey on nosql database," in Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on. IEEE, pp 363–366.
[16] T.Bourbia & M.Boufaida, (2013) "Extension of Databases by semantics: XML Schema for embedding OWL-DL Ontology", KMIKS' 13 International Conference on Knowledge Management, Information and Knowledge Systems, 18-20 Avril 2013, Hammamet, Tunisia, pp 219-232.
[17] K.Wilkinson, C.Sayers, H.Kuno & D.Reynolds, (2003) "Efficient rdf storage and retrieval in jena2". HP Laboratories Technical Report HPL-2003-266, pp 131–150.
[18] J.J.Carroll, I.Dickinson, C.Dollin, D.Reynolds, A.Seaborne & K.Wilkinson, (2004) "Jena: implementing the semantic web recommendations". In WWW Alt. '04: Proceedings of the 3th international World Wide Web conference on Alternate track papers & posters, NY, USA. ACM Press, pp 74–83.
[19] E.F.Codd, (1970) "A Relational Model of Data for Large Shared Data Banks". Communications of the ACM 13 (6), pp 377–387.
[20] F. S.John, (2000) Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, USA.

Authors

Tarek Bourbia is affiliated to LIRE laboratory of the university Constantine 2 in Algeria. He received his magister degree in 2009 in database and web semantic domain. He is preparing his PhD thesis in the field of ontological database

Mahmoud Boufaïda is a full professor in the Computer Science department of the university Constantine 2 in Algeria. He heads the research group 'Information Systems and Knowledge Bases'. He has published several papers in international conferences and journals. He has managed and initiated multiple national and international level projects including interoperability of information systems and integration of applications in organizations. He has been program committee member of several conferences. His research interests include cooperative information systems, web databases and software engineering

A LINK-BASED APPROACH TO ENTITY RESOLUTION IN SOCIAL NETWORKS

Gergo Barta[1]

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics,
Magyar tudosok krt. 2. H-1117 Budapest, Hungary
barta@tmit.bme.hu

ABSTRACT

Social networks initially had been places for people to contact each other, find friends or new acquaintances. As such they ever proved interesting for machine aided analysis. Recent developments, however, pivoted social networks to being among the main fields of information exchange, opinion expression and debate. As a result there is growing interest in both analyzing and integrating social network services. In this environment efficient information retrieval is hindered by the vast amount and varying quality of the user-generated content. Guiding users to relevant information is a valuable service and also a difficult task, where a crucial part of the process is accurately resolving duplicate entities to real-world ones. In this paper we propose a novel approach that utilizes the principles of link mining to successfully extend the methodology of entity resolution to multitype problems. The proposed method is presented using an illustrative social network-based real-world example and validated by comprehensive evaluation of the results.

KEYWORDS

Link mining, Entity Resolution, Social Networks

1. INTRODUCTION AND RELATED WORK

Undoubtedly Facebook is the most popular and extensively used social network currently. While Google+ is gaining ground and others like LinkedIn, Orkut and Badoo also have their respective market share, these solutions tend to specialize their services on a narrow group of users instead of taking on Facebook eye to eye.

Pages on Facebook are edited and maintained by users only, there is no central administration. Some editors are closely related to the topic of the page like the owner of a service or manufacturer of a product, other individuals are mere enthusiastic fans of the given topic. There is no objective way to tell these pages apart, nor there is to rank or rate the value and utility of pages based on their content. Due to the social philosophy of Facebook there is absolutely no restriction to creating duplicates of the same topic, any supporter of a specific subject can create his own page in the topic and start discussion and information sharing. The result is a vast number of duplicate pages with a little extra flavor added to each variant.

In this environment any analysis has to be preceded by troublesome data cleaning to work efficiently and accurately. Entity resolution is the process of deduplicating data references that refer to the same underlying real-world entities. It is particularly important when performing data

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 99–107, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4409

cleaning or when integrating data from multiple sources. Besides corporate master data management for customer and product information an obvious example is price engines, e.g. Shopzilla or PriceGrabber. These services offer easy selling price comparison of a single product offered by multiple vendors. A real-world example of duplicated records is shown in Table 1. While this action is relatively easy for humans and is carried out on-the-fly, machines however, without the proper understanding of information encapsulated in each record, tend to stumble in resolving the issue. The high importance and difficulty of the entity resolution problem has triggered a large amount of research on different variants of the problem and a fair amount of approaches have been proposed, including predictive algorithms [1], similarity functions [2], graphical approaches [3] and even crowdsourcing methods [4].

Table 1. A real-world price engine example

ID	Product Name	Price
1	Apple iPad 4 (16 GB) Tablet	$589.00
2	iPad 2 16 GB with WiFi in	$313.45
3	Apple iPad Air WiFi 32GB	$379.99
4	Sealed Apple Ipad4 Wi-fi 16gb	$379.99
5	Apple iPad 2, 3, 4 Aluminum	$31.95
6	Apple iPad 64 GB with WiFi	$897.52
7	Apple iPad Air (64 GB) Tablet	$857.87
8	Apple iPad 2 16 GB with WiFi	$488.99

Two main varieties were presented in the past by researchers of the subject: similarity-based and learner-based methods.

Similarity-based techniques require two inputs, a similarity function S that calculates entity distance and also a threshold T to be applied. The similarity function $S(e_1,e_2)$ takes a pair of entities as its input, and outputs a similarity value. The more similar the two entities are according to the function, the higher the output value is. The basic approach is to calculate the similarity of all pairs of records. If $S(e_1,e_2) \geq T$ is true for the given $\{e_1,e_2\}$ entity pair, they are considered to refer to the same entity. Gravano et al. show in [5] that cosine similarity and other similarity functions are efficiently applicable where data is mainly textual.

Learner-based techniques handle entity resolution as a type of classification problem and build a model M to solve it. They represent a pair of entities $\{e_1,e_2\}$ as a $[a_1,\ldots,a_n]$ vector of attributes, where a_i is a similarity value of the entities calculated on a single attribute. This approach, as any learner-based method, requires a training dataset Tr in order to build M. A suitable training set contains both positive and negative feature vectors corresponding to matching pairs and non-matching pairs respectively. The classifier built this way can later be applied to label new record pairs as matching or non-matching entities. Formally the entity pair $\{e_1,e_2\}$ resolve to the same real-world entity if $M(Tr, e_1,e_2, [T/CM])$ delivers true. In this case providing threshold T is optional, as it can be calculated based on a cost matrix CM as well. Sehgal et al. [6] successfully implement learners like logistic regression and support vector machines to enhance geospatial data integration. While Breese et al. [1] used decision trees and Bayesian networks for collaborative filtering. From now on $e_1 \leftrightarrow e_2$ denotes both $S(e_1,e_2)$ and $M(Tr, e_1,e_2, [T/CM])$.

The rest of the paper is organized as follows: in Section 2 an intriguing application is presented for motivational reasons then problem formulation and experimental setup is discussed. In Sections 5 and 6 we go into detail on data preparation and ER methods to be applied, while results are presented in Section 7. Finally, we conclude in Section 8.

2. A MOTIVATING EXAMPLE

We will now motivate the problem of entity resolution in a specific domain, social networks, and highlight some of the issues that surface using an illustrative example from the telecasting domain. Consider the problem of trying to construct a database of television programs broadcasted on major channels and their respective "fanpage", a place for sharing information and discussion, collected from Facebook, one of the most popular social networks.

Table 2. Example of an EPG record (a) and its respective fanpage (b)

(a)

Id	Day	Title	Category	Start	Stop	Channel	Subtitle	Description
605	21-10-2013	Britain from Above	Nature	13:00	14:00	BBC HD	null	Documentary series in which broadcaster Andrew Marr…

(b)

Id	Name	Link	Category	Likes	Website	Talking about
109…813	Britain From Above	www.facebook.com/pages/Britain-From-Above/109…813	Tv show	282	www.bbc.co.uk/britainfromabove	3

While Facebook enlists an extensive amount of extremely useful and informative pages, the lack of central regulation and multitude of duplicate topics make information retrieval cumbersome. Finding a way to isolate the most suitable page for each television title is not only a practical service for users, but also an algorithmically difficult task. Apart from the actual content, fanpages have very few descriptive attributes, and the same goes for television programs acquired from an EPG[1]. Some attributes of a television show extracted from an EPG and its respective fanpage to be recommended are shown in Table 2, note some marginal variables are omitted. Performing ER in such an information poor environment is a tolling task.

For any given television show there is a number of suitable fanpages. A fanpage is considered suitable if its content is closely related to the show. The definition of "the most suitable" page is strictly subjective, although there are several ground rules to be considered during fanpage selection.

- *Rule 1*: Under no circumstances should an unrelated page be recommended.
- *Rule 2*: When more than one page is eligible, the most suitable is selected.

Violating *Rule 1* would lead to user distrust, while the violation of *Rule 2* degrades the quality of service. In context to *Rule 2* the most suitable fanpage is the one most interwoven with the subject and also with the most intensive discussion going on.

Entity resolution in general is a delicate, human input intensive error prone process. In contrast to our example generic ER methods work on a single entity type only. As we have seen in Section 1 generic methods use attribute similarity functions to calculate entity-level similarity. Since there are two types of entities discussed here (television program and fanpage) attribute matching is not

[1] Electronic Program Guide

so straightforward. Human input is crucial to identify corresponding attributes, and find a way to establish *hooks* between the rest, a way of connecting entities. This knowledge is also very application-specific, there is no proper way to automatically establish the *hooks*.

3. PROBLEM FORMULATION

As opposed to generic ER this example enlists two distinct entity types; type A and B corresponds to television programs and social network fanpages respectively. Consider the graph example *G* displayed in Figure 1 where the two distinct entity types form the vertices. To identify related vertices edges (*E*) are added to the graph, this information is provided by the blocking method (see Section 4 for more). Optionally the edges can also be weighted based on the confidence level of the similarity function chosen. The *G(A,B,E)* graph is bipartite, meaning there are no edges between identical types. $A \leftrightarrow A$ similarity (comparing TV titles) is out of the scope of this paper, while $B \leftrightarrow B$ similarity (between fanpages) is only observed as a function of their ties with their respective type A entity. For example if both $A_3 \leftrightarrow B_7$ and $A_3 \leftrightarrow B_8$ delivers a positive answer then due to the transitivity of the $\leftrightarrow$ operator also $B_7 \leftrightarrow B_8$ resolves to the same entity weighted by the two edges involved. This approach, called link mining, has been researched extensively (refer to the survey by Getoor et al. [7]). The achievements of link mining and link-based ranking are used in algorithms like PageRank and HITS, and are consequently utilized in the way we select the suitable candidates in the post-processing step.

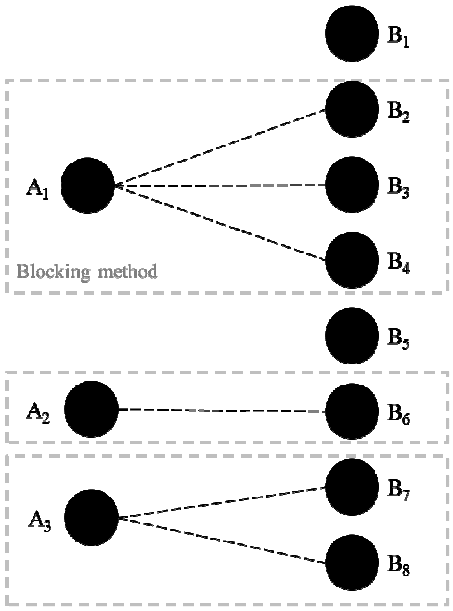


Figure 1. The telecasting example as a bipartite graph

4. EXPERIMENTAL SETUP

Our experiments were based on data gathered by XMLTV[2], a publicly available backend to download television listings. We opted for selecting the schedule of four major channels (BBC News, BBC HD, Al Jazeera English and Animal Planet), and the time period was set to October 2013. The query returned 1110 records to be used in our experiments, due to the innate program structure of television channels, there is substantial amount of repetition both on a daily and

[2] See http://sourceforge.net/projects/xmltv/

weekly basis. Since the Facebook API takes only a single query expression, the title of the television program, to avoid unnecessary querying and processing caused by recurring titles, all electronic program guide records are aggregated on *{Channel, Title, Duration}* level. It is safe to assume that the *{Channel, Title, Duration}* tuple is a unique identifier for all television shows. All other query parameters are being set independently of the respective television program.

Aggregating television programs to avoid unnecessary queries results in substantial compression of the experimental database, 163 unique television shows are extracted of the original 1110 titles. As of 2014 there are 54 million fanpages accessible on Facebook[3], comparing all titles with all available pages would be utterly unfeasible. To avoid the computational explosion caused by comparing all available candidates, generic ER approaches use a blocking method, a heuristic to identify the most probable candidates in order to cut search space significantly. In our case querying the Facebook search engine is a convenient way of blocking, as it delivers the most probable pages based on the internal context of Facebook (see Fig. 1).

Executing subsequent Facebook queries for every unique show title presents us with a total 258 fanpage results. That gives us the number of record pair comparisons to make after applying the blocking method. Figure 2 shows the number of results returned and their respective frequency. Keep in mind that the number of results per query is limited to 10 as part of the blocking mechanism. Although the high number of queries with a single result is promising, we should, by no means, automatically accept and recommend those results as it could violate *Rule 1* presented in Section 2. In any case, evaluation of query results is required.

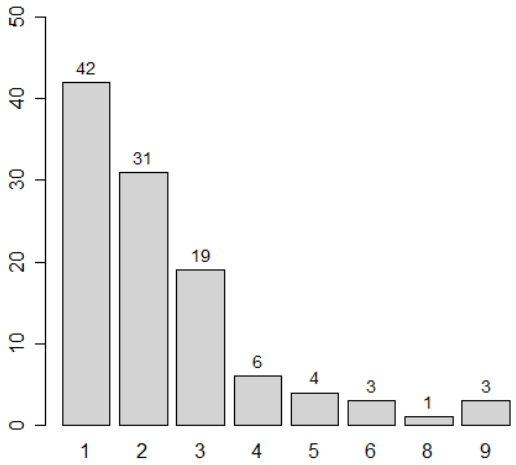


Figure 2. Distribution of result frequency

5. ENTITY SIMILARITY MEASURES

As we have seen in Section 2 in order to successfully identify matching entities of distinct types certain connections or *hooks* are to be set up. This step requires substantial human input. Also Table 2 shows that very few attributes are available to do so. In case of string attributes (*Title* and *Name*) edit distance and related string functions can be used. Long-winded strings containing whole paragraphs (e.g. *Description*) are to be processed to isolate a set of proper nouns and

[3] See http://www.statisticbrain.com/facebook-statistics

compare those only. Polynomial variables like *Category* need preprocessing as well, since they are from different sources distinct variable values are to be matched first.

As we have seen before the low number of entity attributes account for an information poor environment. Utilizing any additional data source can boost the performance of the ER process greatly. Since the whole process is run on-line, as the resolving method relies heavily on Facebook, other on-line services can be included as well (e.g. Google Search to find the URL of each channel). With the help of these additional services new variables can be computed.

As the variable *Link* is available, incidentally the main output of the whole process, the content of the fanpage can be queried (the same goes for *Website*). Page content then can be extensively used to generate *hooks*, in this case these are flags (yes-no questions) regarding the actual page content and especially the descriptive About section (since that is publicly available). New variables include whether the *Channel* variable or its URL is explicitly mentioned on the page, if the website link referenced by the fanpage is the actual website of the channel, whether the linked website is referencing the *Channel* and its URL and *Title* at all.

There are a total of 12 new descriptive features created. Flags have a value of either 0 or 1, while distance like features such as *Title* edit distance have continuous values on the [0,1] interval. Numeric variables are normalized in the same interval as well.

6. RESOLVING ALTERNATIVES

In Section 1 we discussed two well-known options for solving ER problems. In our work we explored the possibility of extending both of them to 2-type ER problems (see Figure 3).

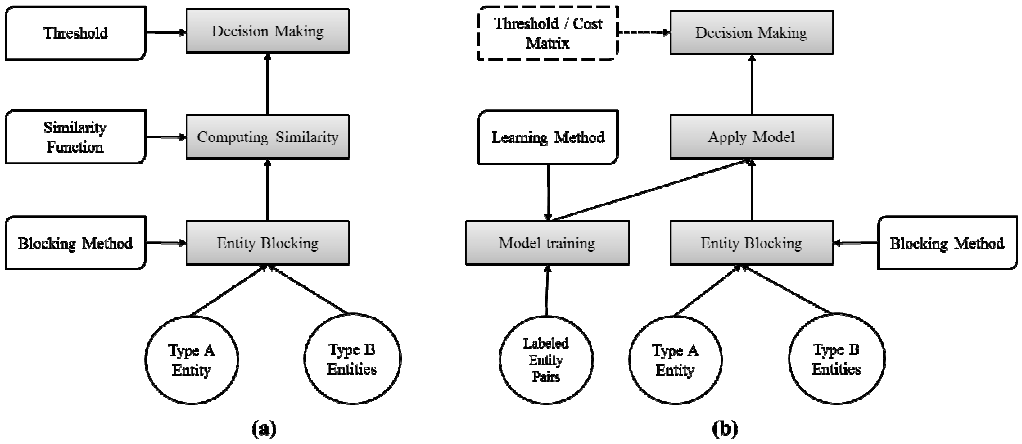


Figure 3. Similarity-based (a) and learner-based (b) techniques extended to multi-type ER

We opted for one of the most common implementation for the similarity-based method by using a simple score model. This is quite straightforward since all our features are already transformed into the same interval. On some group of features a maximum was computed first in order to avoid bias as they referenced the same underlying attributes only in slightly different aspects. Otherwise by a simple addition the final score is calculated (on a 10-point scale) and a threshold can be applied. For this setup threshold was manually set to 5 points, which covered the majority of the examples (see Figure 4 for details).

The learner-based approach treats ER as classification problem using the feature vector generated in the previous section. We explored several different classifier models implemented in

RapidMiner (RM) a free data mining toolset[4]. Out of the many possibilities 3 were selected based on their performance and penetration; Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF). Both the LR and SVM implementation in question was done by Stephen Rueping called myKLR and mySVM respectively (see [8]) RM also offers automatic threshold adjustment based on model confidence, costs and ROC analysis.

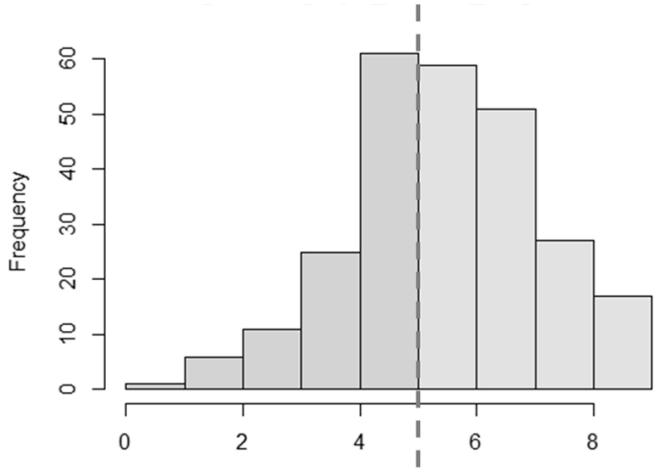


Figure 4. Similarity-based approach: respective score frequency and threshold application.

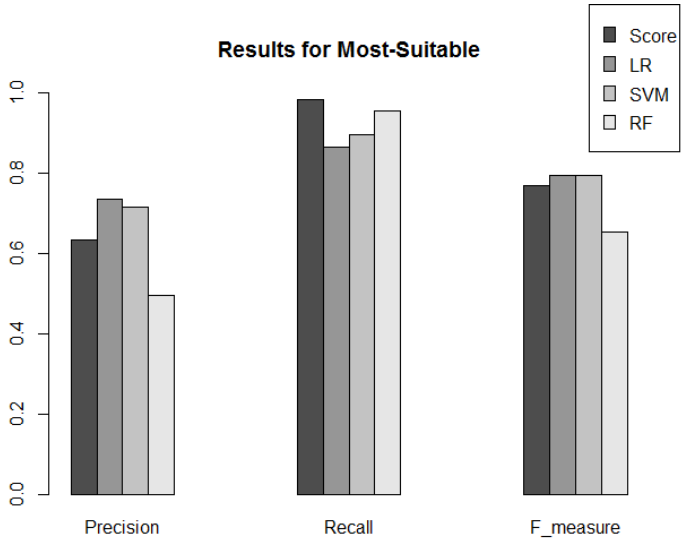
7. RESULTS AND DISCUSSION

Our experiments were carried out on the telecasting dataset using the four different decision making method described in the previous section. A dataset was prepared by labelling entries manually for training and testing purposes. For the sake of result comparability we elected three common known performance measures in machine learning; recall, precision and F-measure. In case of learner-based methods training was done using 10-fold cross-validation with stratified sampling and a cost matrix was provided with a 20% extra penalty for the so-called type I misclassifications (false positive examples).

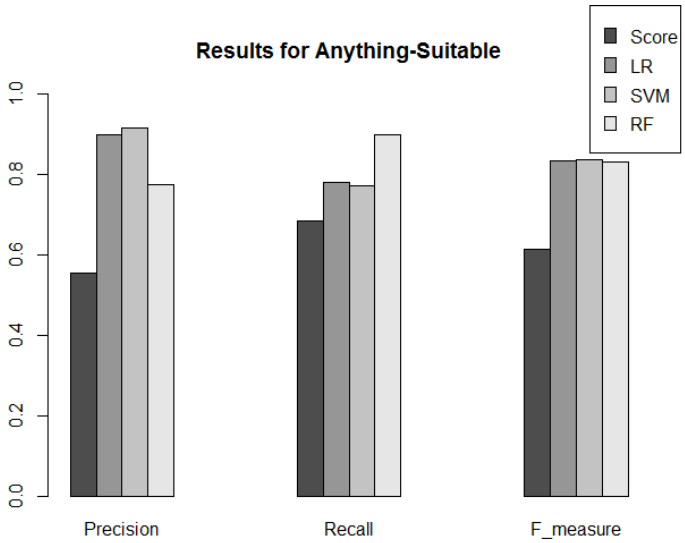
As seen in Figure 5a all approaches delivered comparable results when looking for “the most suitable” fanpage. Learner-based methods performed very similarly, with RF falling behind in precision only. The score model dominated in Recall suggesting a relatively low threshold, but F-measure compared fairly well to learning methods.

Tests were also carried out after relaxing *Rule 2*; so that all related fanpages would be acceptable. In Figure 5b all methods but RF show a substantial drop in recall caused by the foreseeable increase of false negatives. Compensated by increased precision F-measure increased for all classification models as well, yielding an almost identical performance. The more rigid approach of the similarity-based scoring method however failed to satisfy the needs of the modified environment.

[4] See http://www.rapidminer.com



(a)



(b)

Figure 5. Comparing the results of the "most suitable" (a) and "anything suitable" (b) methods

8. CONCLUSIONS

In this paper we proposed a method to successfully extend entity resolution to multitype problems through an interesting application domain. This particular solution eases information retrieval in social networks, an increasingly growing field for analysis. The procedure utilizes the achievements of link mining to effectively bound entities of distinct types. We performed an extensive evaluation of the proposed methods on an illustrative real-world problem and conclude

that the promising results can lead to growing scientific interest and may contribute to valuable services in the future.

REFERENCES

[1] John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (UAI'98), Gregory F. Cooper and Serafín Moral (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, -52.

[2] Brizan, D.G., Tansel, A.U.: A survey of entity resolution and record linkage methodologies. Communications of the IIMA 6(3), 41–50 (2006), http://www.iima.org/CIIMA/8%20CIIMA%206-3%2041-50%20%20Brizan.pdf

[3] Zhaoqi Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra. 2007. Adaptive graphical approach to entity resolution. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07). ACM, New York, NY, USA, 204-213. DOI=10.1145/1255175.1255215 http://doi.acm.org/10.1145/1255175.1255215

[4] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. 2012. CrowdER: crowdsourcing entity resolution. Proc. VLDB Endow. 5, 11 (July 2012), 1483-1494.

[5] Gravano, L., Ipeirotis, P., Koudas, N., Srivastava, D.: Text joins for data cleansing and integration in an rdbms. In: Data Engineering, 2003. Proceedings. 19th International Conference on. pp. 729 – 731 (March 2003)

[6] Vivek Sehgal, Lise Getoor, and Peter D Viechnicki. 2006. Entity resolution in geospatial data integration. In Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems (GIS '06). ACM, New York, NY, USA, 83-90. DOI=10.1145/1183471.1183486 http://doi.acm.org/10.1145/1183471.1183486

[7] Lise Getoor and Christopher P. Diehl. 2005. Link mining: a survey. SIGKDD Explor. Newsl. 7, 2 (December 2005), 3-12. DOI=10.1145/1117454.1117456 http://doi.acm.org/10.1145/1117454.1117456

[8] Rueping, S.: mySVM-Manual. University of Dortmund, Lehrstuhl Informatik 8 (2000), URL http://www-ai.cs.uni-dortmund.de/software/mysvm/

Authors

Gergo Barta received the MSc degree in computer science in 2012 from the Budapest University of Technology and Economics, Hungary, where he is currently working toward the PhD degree in the Department of Telecommunications and Media Informatics. His research interests include spatial data analysis, data mining in interdisciplinary fields, and machine learning algorithms.

INTENTIONAL BLANK

MULTI-WORD TERM EXTRACTION BASED ON NEW HYBRID APPROACH FOR ARABIC LANGUAGE

Meryeme Hadni[1], Abdelmonaime Lachkar[2] and Said Alaoui Ouatik[1]

[1]L.I.M, FSDM, USMBA, FEZ, MOROCCO, [2]L.S.I.S, ENSA, USMBA, FEZ, MOROCCO

[1]meryemehadni@gmail.com,s_ouatik@yahoo.com,
[2]abdelmonaime_lachkar@yahoo.fr

ABSTRACT

Arabic Multiword Term are relevant strings of words in text documents. Once they are automatically extracted, they can be used to increase the performance of any text mining applications such as Categorisation, Clustering, Information Retrieval System, Machine Translation, and Summarization, etc. This paper introduces our proposed Multiword term extraction system based on the contextual information. In fact, we propose a new method based a hybrid approach for Arabic Multiword term extraction. Like other method based on hybrid approach, our method is composed by two main steps: the Linguistic approach and the Statistical one. In the first step, the Linguistic approach uses Part Of Speech (POS) Tagger (Taani's Tagger) and the Sequence Identifier as patterns in order to extract the candidate AMTWs. While in the second one which includes our main contribution, the Statistical approach incorporates the contextual information by using a new proposed association measure based on Termhood and Unithood for AMWTs extraction. To evaluate the efficiency of our proposed method for AMWTs extraction, this later has been tested and compared using three different association measures: the proposed one named NTC-Value, NC-Value, and C-Value. The experimental results using Arabic Texts taken from the environment domain, show that our hybrid method outperforms the other ones in term of precision, in addition, it can deal correctly with tri-gram Arabic Multiword terms.

KEYWORDS

Multiword Term extraction,Part Of Speech, Categorisation, Clustering, Information Retrieval, Summarization.

1. INTRODUCTION

Like many other languages such as English, European, Chinese, Hindi Languages, etc, the term in Arabic Language may be a single term composed by one word, or a multiple words named Multiword Term. Note that, a multiword term may carry more meaning than a single-word term and can represent documents more currently. Therefore, once they are automatically extracted, they can be used to increase the performance of any Text mining applications such as Categorisation, Clustering, Information Retrieval System, Machine Translation, and Summarization, etc. Automatic Multiword term (MWT) extraction has gained the interest of

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 109–120, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4410

many researchers and has applications in many kinds of NLP tasks. The aim of Extraction term is to automatically extract relevant terms from a given corpus.

There are three approaches for MWTs: Linguistic Approach, Statistical Approach and Hybrid Approach. In the Linguistic approach, there exists a variety of previous researches based on morphological, syntactic or semantic information implemented in language-specific rules or programs. These methods are limited by the experience of the specialists who manually select the grammatical patterns. As examples of tools based on this approach we can cite ACABIT [8], Nomino [9] OntoLearn [10] and Lexter[11]. Many researches on MWT focus on methods that are based on Statistical Filters. The methods of T-Score [5], Log-Likelihood Ratio (LLR) [6], FLR [7], Mutual Information (MI) [1] and C-Value [4] are the widely used. Note that, the Mutual Information, Log-Likelihood Ratio and T-Score are proposed to be used in order to measure the Unithood from the strength of inner unity. While the C-Value has been used to measure the Termhood from the strength of marginal variety.

From the above presented approaches, we can conclude that linguistic and statistical approaches present some drawbacks and weakness when they are used alone: On one hand, the statistical approach is unable to deal with low-frequency of MWTs. On the other hand, the linguistic one is language dependent and not flexible enough to cope with complex structures of MWTs.

To avoid the weaknesses of the two filters a commonly recognized solution is to propose a hybrid approach that combines statistical calculus and linguistic Filters [13, 14, and 15]. The T-Score, C-Value and Part-of-speech tags are used as features for compound extraction.

In this paper we present a hybrid Arabic Multi-Word Term extraction method based on two main filters. In the Linguistic Filters we used the method which has been proposed by A. Taani [10], it consists of three levels: the Lexicon Analyzer, the Morphological Analyzer and the Syntactic Analyzer. The Statistical Approach, we adopted to use a new method based on the Unithood and the Termhood measure. The Unithood is to estimate whether a string is a complete lexical unit, and it is measured by the strength of inner unity and marginal variety. The Termhood is to investigate whether the lexical unit is used to refer to a specific concept in a specific domain. we take into account the combination between Termhood and Unithood measures, where we introduce a novel statistical measure, the NTC-Value, that unifies the contextual information and both Termhood and Unithood measure. This measure is applied to another language such as English, French. But not used by Arabic Language.

The remainder of this paper is organized as follows. In the next section, we present the related work. Section 3 describes the proposed method to extract MWTs. In section 4, we present the experimental result. Section 5 concludes this work and presents some perspectives.

2. RELATED WORKS

A lot of works has been done to extract MWT in many languages. These latter have been proposed by using linguistic filter, statistical methods, or both as a hybrid approach. Attia et al. [12] presented a pure linguistic approach for handling Arabic MWTs. It is based on a lexicon of MWTs constructed manually. Then the system tries to identify other variations using a morphological analyzer, a white space normalize and a tokenized. Precise rules allow taking into account morphological features such as gender and definiteness to extract MWTs. The MWTs structures are described as trees that can be parsed to identify the role of each constituent. However some types of MWTs are ignored such as substitution compound nouns. Besides on, the relevance of the extracted candidates is not computed because the lack of statistical measures.

However, the majority of the recently proposed MWT extraction systems have adopted the hybrid approach, because it has given better results than using only linguistic filters or statistical methods [11]. Bouleknadel et al. [13] have adopted the hybrid approach to extract Arabic MWTs. The first step of their system is extraction of MWT-like units, which fit the follow syntactic patterns:{noun adjective, noun1 noun2} using available part of speech tagger. In the second step is ranking the extract MWT-like units using association measures, these measures are: Log-Likelihood Ratio, FLR, Mutual Information, and T-Score. The evaluation process includes applying the association measures to an Arabic corpus and calculating the precision of each measure using a collected reference list of Arabic terms.

Bounhas et al. [14] have followed a hybrid method to extract compound nouns. In the linguistic side, they combined two types of linguistic approaches discussed above. In the one hand, they detect compound noun boundaries and identify sequences that are like to contain compound nouns. On the other hand, they use syntactic rules to handle MWTs. These rules are based on linguistic information: morphological analyzer and a POS tagger. In the statistical side, they applied the LLR method. In the evaluation step, they used almost the same corpus and reference list which have been used in [13]. Their results were promising especially with bigram MWTS [14]. Recently, another system has been proposed by Khalid El-Khatib et al. [15] based on Linguistic and Statistical Filters to extract Arabic MWTs. (i) The Linguistic Filter, where propose new patterns for syntactic patterns based on definite and indefinite types of nouns. Secondly the extraction of the candidate MWTs takes account the sequence of nouns, as well sequences of nouns that connected by a preposition.(ii) In the statistical filter, the Unithood measure was considered by choosing LLR measure because it gives good results with Arabic MWT extraction [14]. For the Termhood they adopted C-Value measure because it has a wide acceptance as a valuable method to rank candidate MWTs. LLR method can be used efficiently as significance of association measure between the two words in the bigram.

Note that, the most recent work in our knowledge, that has been done by our research team [20], this latter consists to combine the linguistic method that used a part-of-speech (POS) tagger named AMIRA to extract candidate MWTs based on syntactic patterns. It propose a novel statistical measure, the NLC-value, that unifies the contextual information and both Termhood and Unithood measures.

The most proposed previous works present some drawback and weakness that can be summarized as follow: the method proposed in [13], many critics can be addressed to this approach. First, the approach does not include a morphological analysis step. The used POS tagger [16] is unable to separate affixes, conjunctions and some prepositions from nouns and adjectives. The lack of a morphological analysis step obliged the authors to identify in a second step- variant of the already identified MWT. Thus, they identify graphical variants, inflectional variants, morph syntactic and syntactic variants. Second, POS tagging does not allow taking into account many features while defining MWT patterns. For example, we cannot impose constraints about the gender and/or the number of the MWT constituents. Third, this approach does not deal with syntactic ambiguities. In [12], the relevance of the extracted candidates is not computed because the lack of statistical measures. Other work [14] produces results that were promising but only using bi-grams MWTs.

The most hybrid methods presented previously are suitable to use only bi-grams. They have been evaluated the top-ranked does not exceed 100 real terms.

In this investigation, we propose a new method for MWT based ob hybrid approach extraction that can be deal with the previous problems. This proposed method composed of two main stages: the linguistic Filters and the statistical Filter. The linguistic filters operate on the POS–tagged, making use of different kinds of linguistic analyze. The POS–tagged text, obtained with the tagger described in Taani (2009), is searched for on the basis of a set of rules. Specifically, for

each multi–word term to be identified in texts, it passes through three levels of analysis. The lexical analyzer, morphological analyzer and a syntax analyzer. As a statistical filter, we proposed a new method based en C-Value, NC-Value and T-Score. The C-Value method aims at bringing out those terms which tend to occur as nested terms, then; the NC-Value incorporates context information to the C-Value, aiming at improving term extraction in general. The T-Score is used to measure the adhesion between two words in a corpus. The new measure is NTC-Value.

The main novelty of the proposed approach lies in the fact that, differently from previous studies, we incorporate contextual information for each words and frequency analysis on words association. The advantages of this idea, is to reduce the execution time and the size of the vector.

3. PROPOSED APPROACH

In this section we present our proposed multi-word term extraction system based hybrid approach. The system includes two components (fig.1): A Linguistic Filter which uses Part Of Speech (POS) Tagger and Sequence identifier to extract candidate MWTS. The Statistical Filters which unifies the contextual information and both Termhood Estimation and Unithood Estimation

.

3.1. Linguistic Filter

We decide to adopt the approach which has been proposed by Ahmad Taani and al [10], there are two reasons for that. First, this approach is simple and accurate. Therefore, it is able to keep one of the metrics of our syntactic patterns, which is the simplicity. Second, this approach has a morphological analyzer phase. The architecture of adopted approach for words classification contains three main phases. He first phase is the lexicon analyzer.

In this phase a lexicon of stop lists in Arabic language is defined. This lexicon includes prepositions, adverbs, conjunctions, interrogative particles, exceptions and interjections. All the words have to pass this phase, if the word is found in the lexicon, it is considered as tagged to one of the previous closed lists.

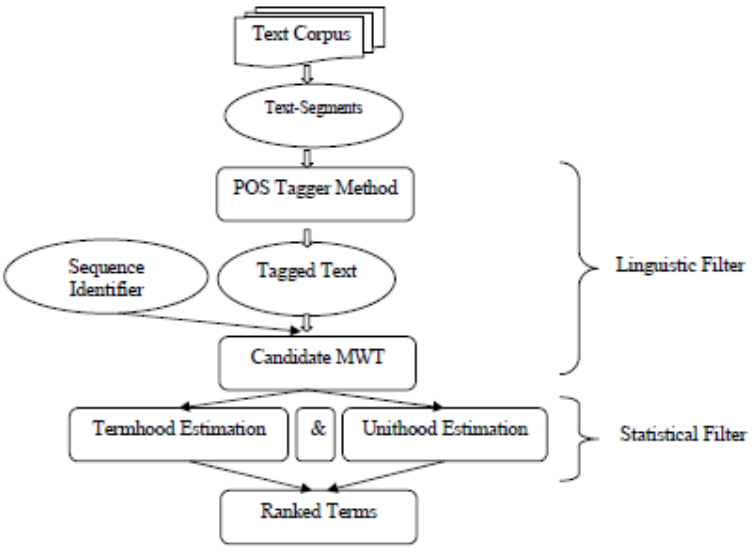


Figure1: Proposed Multiword Term Extraction System

The next phase is the morphological analyzer. Each word which has not been tagged in the previous phase will immigrate to this phase. In this phase, firstly, the affixes of each word are extracted, the affix is a set of prefixes, suffixes and infixes. After that, these affixes and the relation between them are used in a set of rules to tag the word into its class. It is important to say that this phase is the core of the system, since it distinguishes the major percentage of untagged words into nouns or verbs. The last phase is the syntax analyzer. This phase can help in tagging the words which the previous two phases failed to tag. It is consisting of two rules: sentence context and reverse parsing.

The sentence context rule is based on the relation between the untagged words and their adjacent. Where Arabic language has some types of relations between adjacent words. These relations can help in tagging the words into its corresponding class. The reverse parsing rule is based on Arabic context-free grammar. There are ten rules, which are used frequently in Arabic language.

The second component of linguistic filter is the sequence identifier, using the list of syntactic patterns as follows:

-Noun Prep Noun.
-Noun Noun.

The linguistic filtering performs a morphological analysis and takes into account several types of variations. We followed the typology suggested by [13].

3.1.1. Graphical Variants

By graphical variants, we mean the graphic alternations between the letters ي and ى. Table1 shows some examples of graphic alternations.

Table 1: Graphical variats

Variant	Arabic MWT	Translation
ي/ى	التلوث الكيميائي /التلوث الكيميائى	Chemical pollution

3.1.2. Inflectional Variants

Inflectional variants include the number inflection of nouns, the number and gender inflections of adjectives, and the definite article that is carried out by the prefixed morpheme (Al).Table 2 shows some examples of inflectional variants.

Table 2: Inflectional Variants

Variant	Arabic MWT	Translation
Number	تلوث المحيطات/ تلوث المحيط	Ocean pollution
Definitude	تلوث هوائي /التلوث الهوائي	The Air pollution

3.1.3. Morphosyntactic and Syntactic Variants

Morphosyntactic variants refer to the synonymy relation-ship between two MWTs of different structures. The example below shows synonymic terms of N1 PREP N2 structures (Table3).

Table 3: Morphosyntactic Variants

Variant	Arabic MWT	Translation
N1prepN2	بئر من النفط/بئر نفطي	Oils wells

The syntactic variants modify the internal structure of the base-term, without affecting the grammatical categories of the main item which remain identical. We distinguish modification and coordination variants.

Table 4 shows some examples of syntactic variants.

Table 4: Syntactic Variants

Variant	Arabic MWT	Translation
Insertion	الغلاف الجوي للأرض/ الغلاف للأرض	Atmosphere of Earth
Postposition	الغلاف الجوي المتحرك/ الغلاف الجوي	the atmosphere moving
Expansion	تلوث المحيط و البيئة/تلوث البيئة	pollution of Ocean and environment
Tête	المخاطر و الوقاية من التلوث/المخاطر من التلوث	Risks and prevention of pollution

The next step to extract the candidate MWTs is extraction of sequence of nouns and verb. In this step, we consider each sentence as a separated unit, and using the word's classification approach to extract sequences of nouns. However, this is the last step before using the statistical method to rank the terms.

Specifically, we identify sequences of patterns in order to cover most of the Arabic multi–words structures, using the following pattern: N1N2, N1prepN2. The term candidates are passed to the second step.

3.2. Statistical Filter

In a statistical filter, a term is evaluated using two types of feature: Termhood and Unithood [8]. In C-NC method, the features used to compute the term weight are based on Termhood only. In this paper, we introduce a Unithood feature, T-Score, to the C-NC method.

3.2.1. T-Score

The T-Score is used to measure the adhesion between two words in a corpus. It is defined by the following formula [19]:

$$TS(w_i, w_j) = \frac{P(w_i, w_j) - P(w_i).P(w_j)}{\sqrt{\frac{P(w_i, w_j)}{N}}} \quad (1)$$

Where,

$P(w_i, w_j)$ Is the probability of bi-gram w_i, w_j in the corpus, $P(w)$ is the probability of word w in the corpus, and N is the total number of words in the corpus. The adhesion is a type of Unithood feature since it is used to evaluate the intrinsic strength between two words of a term.

3.2.2. The C-Value/NC-Value Method

The NC-Value measure [4] [6], aims at combining the C-Value score with the context information. A word is considered a context word if it appears with the extracted candidate terms. The first part, C-value enhances the common statistical measure of frequency of occurrence for term extraction, making it sensitive to a particular type of multi-word terms, the nested terms. The second part, NC-value, gives: 1) a method for the extraction of term context words (words that tend to appear with terms), 2) the incorporation of information from term context words to the extraction of terms.

- ✓ *C-Value*

The C-Value calculates the frequency of a term and its sub-terms. If a candidate term is found as nested, the C-Value is calculated from the total frequency of the term itself, its length and its frequency as a nested term; while, if it is not found as nested, the C-Value, is calculated from its length and its total frequency.

$$\text{CValue}(a) = \begin{cases} \log_2|a|.f(a) & \text{if a is not nested} \\ \log_2|a|.\left(f(a) - \frac{1}{\text{P}(\text{T}_a)}\sum_{b\in T_a} f(b)\right) & \text{otherwise} \end{cases} \quad (2)$$

Where, f (a) is the frequency of term a with |a| words, T_a is the set of extracted candidate terms that contain a and P(T_a)is the total number of longer candidate terms that contain a . The formula $\frac{1}{P(T_a)}\sum_{b\in T_a} f(b)$ will have value 0 when T_a is empty.

- ✓ *NC Value*

The NC-Value measure [6] aims at combining the C-Value score with the context information. A word is considered a context word if it appears with the extracted candidate terms. The algorithm extracts the context words of the top list of candidates (context list), and then calculates the N-Value on the entire list of candidate terms. The higher the number of candidate terms with which a word appears, the higher the likelihood that the word is a context word and that it will occur with other candidates. If a context word does not appear in the extracted context list, its weight for such term is zero. Formally, given w as a context word, its weight will be:

$$weight(b) = \frac{\text{t(b)}}{\text{n}} \quad (3)$$

Where t (b) is the number of candidate terms b appears with, and n is the total number of considered candidate terms; hence, the N-Value of the term t will be

$$\text{NValue} = \sum_{b\in C_a} f_a(b) * \text{weight(b)} \quad (4)$$

where $f_a(b)$ is the frequency of b as context word of a, and C_a is the set of distinct context words of the term t. Finally, the general score, NC-Value, will be:

$$\text{NCValue}(a) = 0.8.\text{CValue}(a) + 0.2.\text{NValue}(a) \quad (5)$$

From the above formula, we find that NC-Value is mainly weighted by C-Value .It treats the term candidate as a linguistic unit and evaluates its weight based on characteristics of the Termhood,

i.e. frequency and context word of the term candidate. The performance can be improved if feature measuring the adhesion of words within the term is incorporated.

3.2.3. The NTC-Value

Theoretically, the C/NC method can be improved by adding Unithood feature to the term weighting formula. Based on the comparison of [18], we explore T-Score, a competitive metric to evaluate the association between two words, as a Unithood feature.

Our idea here is to combine the frequency with T-Score, a Unithood feature. Taking the example in Table 5, the candidates have similar rank in the output using C/NC Termhood approach.

Table 5. Example of context MWT

MWT	Translation
وزارة التعليم العالي	Ministry of Higher Education
التعليم العالي بالمغرب	Higher Education in Morocco
سلامة التعليم العالي	the Safety of Higher Education
التعليم العالي الجامعي	the Higher Education University

Example for Environmental domain:

Table 6. Example of context MWT

MWT	Translation
ملوثات الغلاف الجوي	Atmospheric pollutants
الغلاف الجوي للأرض	Earth's atmosphere
الغلاف الجوي و الطقس	the atmosphere and the weather
توازن الغلاف الجوي	The balance of the atmosphere
غازات الغلاف الجوي	Atmospheric gases

To give better ranking and differentiation, we introduce T-Score to measure the adhesion between the words within the term. We use the minimum T-Score of all bi-grams in term a, minTS(a), as a weighted parameter for the term besides the term frequency.

For a term $a = w_1 . w_2 \ldots v$, the minTS(a) is defined as :

$$minTS(a) = \min\{TS(w_i, w_{i+1})\}, i = 1 \ldots (n -$$

Table 7. Term with Minimum T-Score value

MWT	Translation	minTS(MWT)
وزارة التعليم العالي	Ministry of Higher Education	3.53
التعليم العالي بالمغرب	Higher Education in Morocco	2.64
سلامة التعليم العالي	the Safety of Higher Education	9.78
التعليم العالي الجامعي	the Higher Education University	1.73

Table8. Term with Minimum T-Score value

MWT	Translation	minTS (MWT)
ملوثات الغلاف الجوي	Atmospheric pollutants	9.65
الغلاف الجوي للأرض	Earth's Atmosphere	3.74
الغلاف الجوي و الطقس	the atmosphere and the weather	6.28
توازن الغلاف الجوي	The balance of the atmosphere	1.72
غازات الغلاف الجوي	Atmospheric gases	3.54

Table 7 and Table 8 shows the minTS(MWT) of the different terms in table 5 and Table 6 respectively. Since $minTS(a)$ can have a negative value, we only considered those terms with $minTS(a) > 0$ and combined it with the term frequency. We redefine C-Value to TC-Value by replacing f (a) using F(a), as follows:

$$F(a) = \begin{cases} f(a) & if\ minTS(a) \leq 0 \\ f(a) * \ln(2 + minTS(a)) & if\ minTS(a) > 0 \end{cases} \quad (6)$$

$$TCValue(a) = \log_2|a|.\left(F(a) - \frac{1}{P(T_a)}\sum_{b \in T_a} F(b)\right) \quad (7)$$

The final weight, defined as NTC-Value, is computed using the same parameter as NC-Value.

$$NTCValue(a) = 0.8.TCValue(a) + 0.2.NValue(a) \quad (8)$$

4. EXPERIMENT AND RESULT

4.1. The Corpus Collection

The lake of Arabic specialized domain corpora forced the research to build new corpora to evaluate their approaches. The texts are taken from the environment domain and are extracted from the web site "Al-Khat Alakhdar"[1]. The corpus contains 1.013 documents and 470.175 words.

4.2. Evaluation

Evaluation of MWT approaches is a complex task, there are no specific standards for evaluate and compare different MWT approaches. However, the most of the approaches have used one of two evaluation steps: reference list and validation. In the first step, we attest that a term is relevant to the environment domain if it has already been listed in existing terminology database AGROVOC[2]. The second methods, if the term not exists in AGROVOC we search his translation in database IATE[3] (InterActive Terminology for Europe).

Table 9 shows the comparison result of the origin C-value, NC-value and NTC-value on the ranking for the MWT candidates. We evaluate the performance based on the k best candidates from 100-500 at intervals of 100.

[1] http://www.greenline.com.kw
[2] www.fao.org/agrovoc/
[3] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load

We attested that a term is relevant if it has been listed in existing database AGROVOC and IATE.

Table 9. Precision for C-Value, NC-Value, and NTC-Value

Top terms	*C-Value*	*NC-Value*	*NTC-Value*
100	66,0%	74,0%	86,0%
200	63,0%	69,0%	75,0%
500	58,0%	66,0%	71,0%

Furthermore, the combination of the context information and the C-Value improves the performance of the process of MWT extraction because the NC-Value outperforms the C-Value for each considered MWT list. The Unithood feature NTC-Value outperforms the C-Value/NC-Value as expected from previous studies. Figure 2 illustrates the precision obtained for the C-Value/NC-Value and the NTC-Value.

Figure 2 expresses the same information as table 9, as a graph. In the horizontal axis, the number of candidate term for the three methods are shown, while in the vertical axis, the precision for number of these intervals is provided.

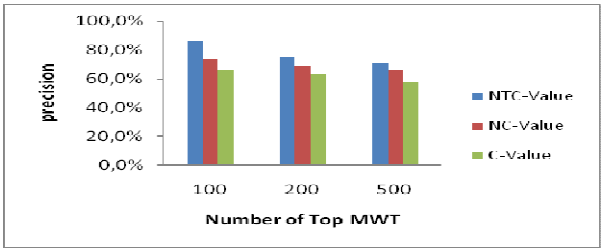


Figure 2. Precision Obtained for NC-value, C-value and NTC-value

The integration of contextual information and the T-score Unithood measure to the C-Value improves the performance of MWT acquisition, since the NTC-Value has better precision than the C-Value\NC-Value, as illustrated in Figure 2.

In figure3, we present some obtained results for Arabic MWTs extraction using the three methods: C-Value, NC-Value and NTC-Value

Figure 3. Sample of extracted MWT using: C-Value, NC-Value and NTC-Value

Arabic MWT	Translation	NTC-Value	NC-Value	C-Value
الخط الأخضر	Green Line	99.2	99.2	89.26567206138039
الطاقة الشمسية	Solar Energy	99.2	89.44804089867734	38.21052858969426
تلوث المنتجات	Contamination of products	99.2	50.88460312771606	3.806022677747762
المناطق المناخية	Climatic zones	99.2	76.76391432416152	76.17764754947734
مستوى تذبذبات	Level fluctuations	99.2	76.59618564676952	22.97811394253883
المسطحات الطينية	Mudflats	99.2	76.5961856467652	22.97811394253883
التوريد والتوزيع	Supply and distribution	99.2	50.91157110023308	3.1699250014423126
فصيلة الفيلة	Platoon elephants	99.2	50.91157110023308	3.806022677747762
الطاقة المتجددة	Renewable energy	99.2	50.91157110023308	23.071735871528716
الأمم المتحدة	United Nations	99.2	23.63345998892698	11.200000000000001
البترول الوطنية	National Petroleum	99.2	76.30461513232898	89.60000000000001
بلدية الكويت	Kuwait Municipality	99.2	76.17764754947734	5.065613907861519

محتوى المعادن	Metal content	96.80000000000001	4.754887502163469	2.53727333451947
المتجددة بالصين	China's renewable	96.80000000000001	80	38.06022677747762
التمثيل الضوني	Photosynthesis	89.60000000000001	50.85613907861519	89.44804089867734
الانبعاثات الغازية	Emissions	89.26567206138039	38.21385730048871	76.59618564676952
النظام البيني	Ecosystem	89.26567206138039	38.21385730048871	4.800000000000001
الغازات السامة	Toxic gases	89.26567206138039	95.09775004326938	5.188460312771606
مخاطر التلوث	The risk of contamination	89.26567206138039	50.85613907861519	23.200000000000003
ملوثات الهواء	Air pollutants	89.26567206138039	38.222773488520145	89.60000000000001

are performed for bi-grams and tri-grams on an Arabic Texts taken from the environment domain. In conclusion, the efficiency of our proposed method for AMWTs extraction has been tested and compared using three different association measures: the proposed one named NTC-Value, NC-Value, and C-Value. The experimental results show that our hybrid method outperforms the other ones in term of precision; in addition, it can deal correctly with tri-grams Arabic Multiword terms.

In the future work we are considering to integrate evaluation by an expert, because there's words that not exist in AGROVOC or in IATE and there are correct. For example; "فصيلة الفيلة" "A platoon elephants" and "توريد والتوزيع" "Supply and distribution". Then study the impact of POS tagging on AMWTs extraction.

REFERENCES

[1] B. Daille, (1994) " Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques ", doctoral thesis, University of Paris 7.

[2] K. Church & W. Gale & P. Hanks,&D. Hindle(1991), "Using statistics in lexical analysis," in Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. U. Zernik, pp. 115–164.

[3] Hiroshi Nakagawa,& Tatsunori Mori.(2002)." A Simple but Powerful Automatic Term Extraction Method". 2nd International Workshop on Computational Terminology,ACL.

[4] Katerine& T. Frantzi,& Sophia Ananiadou, & Junichi Tsujii,(1998)." The C-Value/NC-Value Method of Automatic Recognition for Multi-word terms". Journal on Research and Advanced Technology for Digital Libraries.

[5] Hideki Mima& Sophia Ananiadou(2001)." An Application and Evaluation of the C/NC-Value Approach for the Automatic Term Recognition of Multi-Word Units in Japanese". International Journal on Terminology.

[6] Spela Vintar.(2004)." Comparative Evaluation of C-value in the Treatment of Nested Terms", Memura 2004 –Methodologies and Evaluation of Multiword Units in Real-World Applications. Proceedings of the International Conference on Language Resources and Evaluation 2004, pp. 54-57.

[7] E. Milios& Y. Zhang& B. He,&L. Dong. (2003). "Automatic Term Extraction and Document Similarity in Special Text Corpora". Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics (PACLing'03), Halifax, Nova Scotia, Canada, pp. 275-284.

[8] Kyo Kageura (1996)." Methods of Automatic Term Recognition - A Review". Terminology, 3(2): 259 – 289.

[10] A.T Al-Taani&&S. Abu-Al-Rub(2009)," A rule-based approach for tagging non-vocalized Arabic words". The International Arab Journal of Information Technology, Volume6 (3): 320-328,

[11] M. Tadi&, K. Sojat(2003)," Finding multiword term candidate in Croatian". In the Proceeding of IESL2003 Worksop, pp. 102-107.

[12] Attia& M.A(2008),"Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a view to Machine Translation", doctoral thesis, University of Manchester, Faculty of Humanities.

[13] S. Bouleknadel& B.Daille & D. Aboutajdine(2008),"A multi-word term extraction program for Arabic language", In the 6th international Conference on language resources and evaluation LREC, pp. 1485-1488.

[14] I. Bounhas& Y. Slimani,(2009),” A hybrid approach for Arabic multi-word term extraction”, NLP-KE 2009. International Conference on Language Processing and knowledge Engineering, vol., no., pp.1-8, 24-27.

[15] K. El Khatib& A. Badarenh. (2010).” Automatic Extraction of Arabic Multi-word Term”. Proceedings of the International Multiconference on Computer Science and Information Technology, pp.411-418.

[16] M. Diab& K. Hacioglu & D. Jurafsky,(200’), "Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks", in the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04), Boston, Massachusetts, May 2-7.

[17] C. Manning & H. Schuetze. (1999).” Foundations of StatisticalNatural Language Processing”. MIT Press Cambridge, Massachusetts.

[18] Evert& S. & B. Krenn. (2001). “Methods for Qualitative Evaluation of Lexical Association Measures”. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 369 – 381.

[19] Vu Thy& Ai Ti Aw& Min Zhang(2008),” Term extraction throught unithood and termhood unification”. In proceeding of the 3rd International Joint Conference on Natural Language Processing.

[20] A.EL Mehdaoui&S. EL Alaoui Ouatik & E. Gaussier,(2013),” A Study of Association Measures and their Combination for Arabic MWT Extraction”, published in "Terminology and Artificial Intelligence, Paris : France .

A SURVEY ON ELLIPTIC CURVE DIGITAL SIGNATURE ALGORITHM AND ITS VARIANTS

Greeshma Sarath[1] , Devesh C Jinwala[2] and Sankita Patel[3]

[1,2,3]Department of Computer Engineering, SVNIT, Surat
greeshmasarath88@gmail.com, dcjinwala@gmail.com,
sankitapatel@gmail.com

ABSTRACT

The Elliptic Curve Digital Signature Algorithm (ECDSA) is an elliptic curve variant of the Digital Signature Algorithm (DSA). It gives cryptographically strong digital signatures making use of Elliptic curve discrete logarithmic problem. It uses arithmetic with much smaller numbers 160/256 bits instead of 1024/2048 bits in RSA and DSA and provides the same level of security. The ECDSA was accepted in 1999 as an ANSI standard, and was accepted in 2000 as IEEE and NIST standards. It was also accepted in 1998 as an ISO standard. Many cryptologist have studied security aspects of ECDSA and proposed different variants. In this paper, we discuss a detailed analysis of the original ECDSA and all its available variants in terms of the security level and execution time of all the phases. To the best of our knowledge, this is a unique attempt to juxtapose and compare the ECDSA with all of its variants.

KEYWORDS

Network Protocols, Wireless Network, Mobile Network, Virus, Worms &Trojon

1. INTRODUCTION

ECC is one of the most advanced and promising techniques in the field of Public key cryptography. It offers many advantages over other cryptographic techniques which uses Integer factorization or discrete logarithmic approach. The hardest problem in which ECC is built upon is Elliptic Curve Discrete Logarithmic Problem (ECDLP). ECDLP is based on the infeasibility in computing discrete logarithms on elliptic curves over finite fields. It gives Elliptic curve cryptography a greater strength-per-key-bit. It uses arithmetic with much shorter numbers 160,256 bits instead of 1024,2048 bits and provides same level of security. Elliptic Curve Digital Signature Algorithm was first proposed in 1992 by Scott Vanstone in response to NIST's proposal of DSS [1][2]. It was later accepted in 1998 as an ISO standard (ISO 14888-3), as an ANSI standard (ANSI X9.62) in 1999, and as an IEEE standard (IEEE 1363-2000) and as a NIST standard (FIPS 186-2) in 2000.

However, it has disadvantages too. It is conceptually more difficult to understand and finding secure curves in set up phase is more difficult. ECDSA based on elliptic curve discrete logarithmic problem and is the most secure digital signatures scheme [4]. Many researches are developing different variants of ECDSA each having its own advantages and disadvantages and many cryptologist are trying to find weaknesses in ECDSA variants. This paper analyzes and

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 121–136, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4411

describes different variants of ECDSA, their pros and cons and the attacks possible on each of the variants.

This section gives a brief introduction about the paper. Section 2, elaborates elliptic curve arithmetic operations and Elliptic curve discrete logarithmic problem. Section 3 gives a detailed description of original ECDSA scheme, its security proofs and an attack possible on original ECDSA scheme. Section 4 describes a variant of ECDSA suitable for signer with limited computation capability and Section 5 a variant suitable for a verifier with limited computation capability and its security proofs. Section 6 explains a two level digital signature scheme by using two different secrets. Section 7 describes Elliptic curve German digital signature scheme with inverse calculation in key generation phase. Section 8 describes a variant of ECDSA and a forging possible on it and section 9 details its improved version. Section 10 gives a brief description of two other variants to make ECDSA secure against adaptive chosen message attack and to avoid duplicate signatures. Section 11 elaborates Elliptic curve korean certificate based digital signature algorithm.. Section 12 discusses the implications, performance results and comparison of all ECDSA variants.

2. ELLIPTIC CURVE ARITHMETIC

Elliptic curve cryptography is based on the arithmetic of points on an elliptic curve[12][13]. Elliptic curves are represented by cubic equations similar to those used for calculating the circumference of an ellipse. An elliptic curve E over a field K is defined by a equation [3]:

$$E: y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where $a_1,\ a_2,\ a_3,\ a_4,\ a_6 \in K$ and $\Delta \neq 0$, where Δ is defined as follows:

$$\Delta = -d_2^2\, d_8 - 8d_4^3 - 27d_6^2 + 9d_2d_4d_6;$$

$$where\ d_2 = a_1^2 + 4a_2\,, d_4 = 2a_4 + a_1a_3,$$
$$d_6 = a_3^2 + 4a_6 \ and\ d_8 = a_1^2a_6 + 4a_2a_6a_1a_3a_4 + a_2a_3^2a_4^2$$

Set of all points (x, y) which satisfies the above equation along with ∞, a point at infinity, are the points on the elliptic curve.

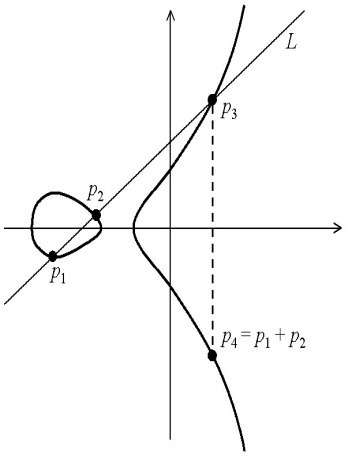


Figure 1. Point Addition

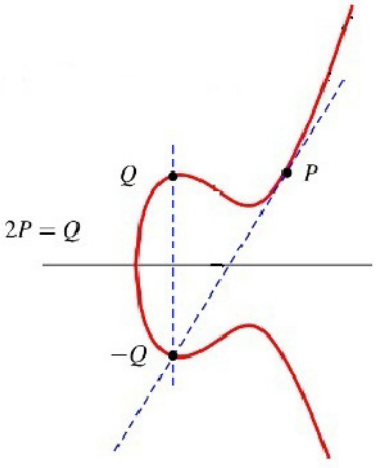


Figure 2. Point Doubling

The number of points on an elliptic curve, n, is the order of elliptic curve, (#($E(F_p)$). The set of points of E (F_p) together with addition operation forms an abelian group with point at infinity, ∞ serving as the identity element. The Equation 1 is called weierstrass equation. The condition $\Delta \neq 0$ ensures that the elliptic curve is smooth, that is, there are no points at which the curve has two or more distinct tangent lines. If the field characteristic P is not equal to 2 or 3, that is prime field, and then the admissible change of Variables

$(x, y) \rightarrow \left(\frac{x-3a_1^2-12a_2}{36}, \frac{y-3a_1 x}{216} - \frac{a_1^3+4a_1a_2-12a_3}{36}\right)$ transform E to the curve,

$$y^2 = x^3 + ax + b; \text{where } a, b \in K \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The $\Delta\ is\ 16(4a^3 + 27b^2)$.

2.1. Point Addition

Addition of points on an elliptic curve is defined by Chord and Tangent rule. Let P = (x_1, y_1) and Q = (x_2, y_2) be two distinct points on an elliptic curve E. Then the sum R, of P and Q, is defined as follows: Draw a line connecting P and Q extend it to intersect the elliptic curve at a third point. Then the sum, R is the negative of the third point. Negative of a point is defined by reflection of the point about the x-axis.

The double R, of P, is defined as follows: Draw the tangent line to the elliptic curve at P. Let it intersects the elliptic curve at a second point. Then the double R is the reflection of this point about the x-axis.

2.2. Point Multiplication

Point Multiplication (Scalar multiplication) is the arithmetic operation which computes kp where k is an integer and p is a point on elliptic curve. It is done by repeated addition. For example Q = kp means Q is obtained by adding p k times to itself (p + p + p....k times). Cryptanalysis involves determining k given P and Q. This operation dominates the execution time of elliptic curve cryptographic schemes.

2.3. Operations defined for $E(F_p)$: $y^2 = x^3+ax +b$

(1) Identity: P + ∞ = ∞ + P = P for all $P \in E(F_p)$

(2) Negatives: If P = (x,y) ∈ $E(F_P)$, then (x,y) + (x, -y) = ∞. The point (x, -y) is denoted by –P and is called negative of P. Note that P indeed is a point in $E(F_P)$.

(3) Point Addition: Let P=(x_1,y_1) ∈ E(K) and Q=(x_2,y_2) ∈ E(K) ; where P≠ ±Q, then P+Q = (x_3,y_3)

where,

$x_3 = \left(\frac{y_2-y_1}{x_2-x_1}\right)^2 - x_1 - x_2$ and $y_3 = \left(\frac{y_2-y_1}{x_2-x_1}\right)^2 (x_1 - x_3) - y_1$

(3) Point Doubling: Let P=(x_1,y_1) ∈ E(K), then 2P = (x_3,y_3) where,

$x_3 = \left(\frac{3x_1^2+a}{2y_1}\right)^2 - 2x_1$ and $y_3 = \left(\frac{3x_1^2+a}{2y_1}\right)^2 (x_1 - x_3) - y_1$

2.4. Elliptic Curve Discrete logarithm problem:

The elliptic curve discrete logarithm problem (ECDLP) is defined as follows: Given the elliptic curve domain parameters and a point $P \in E(F_p)$, find the unique integer k, $0 \leq k \leq$ n1, such that P=kG, where n1 is order of E.

ECDLP is similar to the Discrete Logarithm Problem and is the elliptic curve analogue of DLP. In the ECDLP, the subgroup Z_p^* is replaced by the group of points on an elliptic curve over a finite field. In addition, unlike the DLP and the integer factorization problem, no sub exponential-time algorithm is known for the ECDLP. ECDLP is considered to be significantly harder than DLP, thus giving elliptic curve cryptosystems a greater strength-per-key-bit than their discrete logarithmic counterparts.

3. ECDSA

Elliptic Curve Digital Signature Algorithm [6] was first proposed in 1992 by Scott Vanstone in response to NISTs proposal of DSS. It was later accepted in 1998 as an ISO standard (ISO 14888-3), as an ANSI standard (ANSI X9.62) in 1999, and as an IEEE standard (IEEE 1363-2000) and as a NIST standard (FIPS 186-2) in 2000.

Elliptic curve digital signature algorithm consists of 3 phases: 1. Key generation, 2. Signature generation, 3.Signature verification. A setup phase has to execute before the key generation phase to generate the domain parameters. Domain parameters for an elliptic curve describe an elliptic curve E defined over a finite field Fp, a base point g ∈E (Fp) (generator) with order n. The parameters should be chosen carefully so that ECDLP is resistant to all known attacks. The elliptic curve is chosen by choosing (a,b) ∈ (1, P) and substituting in equation. So the domain parameters can be defined as p, E (a, b), g, n.

3.1. Key pair generation using ECDSA:

Let A be the signatory for a message M. Entity A performs the following steps to generate a public and private key:

(1) Select a unique and unpredictable integer, d, in the interval [1, n-1]
(2) Compute Q = dg
(3) Sender A's private key is d
(4) Sender A's public key is the combination (E, g, n, Q)

3.2. Signature Generation Using ECDSA

Using A's private key, A generates the signature for message M using the following steps:

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) Compute $kg = (x_1,y_1)$, where x_1 is an integer
(3) Compute $r = x_1 \bmod n$; If r = 0, then go to step 1
(4) Compute h = H(M), where H is the SHA-512[10]
(5) Compute $s = k^{-1}(h + dr) \bmod n$; If s = 0, then go to step1
(6) The signature of A for message M is the integer pair (r, s)

3.3. Signature Verification Using ECDSA

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) Verify that values r and s are in the interval [1,n-1]
(3) Compute $w = s^{-1} \bmod n$.
(4) Compute h = H(M), where H is the same secure hash algorithm used by A.
(5) Compute u1 = hw mod n
(6) Compute u2 = rw mod n
(7) Compute $u1g + u2Q = (x_0,y_0)$
(8) Compute $v = x_0 \bmod n$
(9) The signature for message M is verified only if v = r

3.4. Security of ECDSA

Public key is generated by computing the point Q, where Q = dg. In order to crack the elliptic curve key, adversary Eve would have to discover the secret key d when Q and g are provided. The order of the Elliptic curve, E is a prime number n, then computing d given dg and g would take roughly 2n=2 operations [7] . For example, if the key length n is 192 bits (the smallest key size that NIST recommends for curves defined over GF(p)), then Eve will be required to compute about 296 operations. If Eve had a super computer and could perform one billion operations per second, it would take her around two and a half trillion years to find the secret key. This is the elliptic curve discrete logarithm problem behind ECDSA. The curve parameter should be chosen so carefully to secure Elliptic curve from well known attacks like Pollard's rho[1] and Pohlig-Hellman.

3.5. Proof of ECDSA signature Scheme

Signature send by A to B is (r, s) and s can be generated only by A because only A knows its private key d. $s = k^{-1}(h + dr) \bmod n$ on rearranging

- $K = s^{-1} (h + dr)$
- $Kg = s^{-1} (h + dr)g$
- $Kg = s^{-1}hg + s^{-1}drg$
- $r = hwg + rwdg$
- $r = u_1g + u_2Q$

3.6. A Possible Attack on ECDSA

The secret k used for signing two or more messages should be generated independent of each other. In particular, a different secret k should be used for signing different messages otherwise the private key d can be recovered. However if a secure random or pseudo-random number generator is used, then the chance of generating a repeated k value is negligible. If same secret k is used to generate signature of two different messages m1 and m2 then it will result in two signatures (r,s1) and (r, s2).

- $s1 = k^{-1}(h1 + dr)$
- $s2 = k^{-1}(h2 + dr)$; where h1 = SHA512 (m1) and h2 = SHA512 (m2).
- ks1- ks2 = h1+dr-h2-dr

- k = (h1-h2)/(s1-s2)
- d =(ks-h)/r

4. VARIANT 1

The scheme [5] is suitable for a signer who has limited computing capability like, a signer using his Smart Card which stores his secret key and signs a message on a terminal.

Key pair phase of this scheme is same as the ECDSA scheme.

4.1. Signature Generation

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) $kg \leftarrow (x_1,y_1)$, where x_1 is an integer.
(3) $r \leftarrow x_1 \bmod n$; If r = 0, then go to step 1
(4) $h \leftarrow H(M)$, where H is the SHA-512
(5) $s \leftarrow d^{-1}(rk - h) \bmod n$; If s = 0, then go to step1
(6) The signature of A for message M is the integer pair (r,s)

Here the advantage is that there is no need of calculating inverse of d in each individual signing operation. d is the private key of the signer which will remain stable for a period of time, it can be precomputed and stored in the key generation phase itself.

4.2. Signature Verification

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) Verify that values r and s are in the interval [1,n-1]
(3) $w \leftarrow r^{-1} \bmod n$
(4) $h \leftarrow H(M)$, where H is the same secure hash algorithm used by A
(5) $u1 \leftarrow hw \bmod n$
(6) $u2 \leftarrow sw \bmod n$
(7) $(x0,y0) \leftarrow u1g + u2Q$
(8) $v \leftarrow x0 \bmod n$
(9) The signature for message M is verified only if v = r

4.3. Proof of the Scheme

Signature send by A to B is (r, s) and s can be generated only by A because only A knows its private key d. $s = d^{-1}(rk - e) \bmod n$ on rearranging,

- $sd = (rk - e)$
- $dsr^{-1} = r^{-1}(rk - e)$
- $dsgr^{-1} = gk - egr^{-1}$
- $kg = egw + Qws$
- $r = u1g + u2Q$.

Attack on same k can be implemented successfully on this method as well.

5. VARIANT 2

The scheme [5] is suitable for the verifier who has limited compute apparatus. That is in this scheme the complexity of verification operation is less compared to that of the above schemes. Key pair generation phase of this scheme is same as the ECDSA scheme.

5.1. Signature Generation

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) $kg \leftarrow (x_1,y_1)$, where x_1 is an integer
(3) $r \leftarrow x_1 \bmod n$; If r = 0, then go to step 1
(4) $h \leftarrow H(M)$, where H is the SHA-512
(5) $s \leftarrow k(h + rd)^{-1} \bmod n$; If s = 0, then go to step1
(6) The signature of A for message M is the integer pair (r,s).

5.2. Signature Verification

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) Verify that values r and s are in the interval [1,n-1]
(3) $h \leftarrow H(M)$, where H is the same secure hash algorithm used by A
(4) $u1 \leftarrow hs \bmod n$
(5) $u2 \leftarrow rs \bmod n$
(6) $(x_0,y_0) \leftarrow u1g + u2Q$
(7) $v \leftarrow x_0 \bmod n$
(8) The signature for message M is verified only if v = r

In this scheme, k^{-1} is no longer be calculated, but we must calculate $(h + rd)^{-1}$ in the signing phase. But there is no need of calculating inverse in verification phase which is one of the most costlier operation in modular arithmetic. So the complexity of the verification operation is less in this scheme. Since r and k are functions, pre-calculating couldn't be used to reduce the operation amount. If an attacker want to forge a signature, he must decide a pair of (r, s) too, which must fit for the equation $R = (xR, yR) = u1g + u2Q = hsg + rsQ$. So he encounters the same difficulty as attacking original algorithm.

5.3. Proof of the Scheme

Signature send by A to B is (r, s) and s can be generated only by A because only A knows its private key d. $s \leftarrow k(h + rd)^{-1} \bmod n$ on rearranging,

- $s(h + rd) = k$.
- $s(h + rd)g = kg$
- $shg + rdsg = kg$
- $u1g + u2dg = kg$
- $u1g + u2Q = kg$.

Attack on same k can be implemented successfully on this method as well.

6. VARIANT 3

In this scheme [6], [7] two levels of digital signature are implemented by using two secrets k1 and k2. Here d cannot be determined even if the same secret (k1, k2) is repeated. The processes are more complex than original ECDSA scheme and it increases the security level. Key pair generation phase of this scheme is same as the ECDSA scheme.

6.1. Signature Generation

(1) SELECT Select two integers k1 and k2 such that $1 \leq k1,k2 \leq n-1$
(2) $k1g \leftarrow (x_1, y_1)$, $k2g \leftarrow (x_2,y_2)$ where x_1,x_2,y_1,y_2 are integers
(3) $r1 \leftarrow x1 \bmod n$; $r2 \leftarrow x2 \bmod n$ If r1,r2 = 0, then go to step 1
(4) $h \leftarrow H(M)$, where H is the SHA-512
(5) $s \leftarrow k1^{-1}(hk2 + d(r1 + r2)) \bmod n$; If s = 0, then go to step1
(6) The signature of A for message M is the integer pair (r1, r2, s)

6.2. Signature Verification

(1) Obtain signatory's public key (E, q, n, Q)
(2) Verify that values r1,r2 and s are in the interval [1,n-1]
(3) $w \leftarrow s^{-1} \bmod n$
(4) $h \leftarrow H(M)$, where H is the same secure hash algorithm used by A
(5) $u1 \leftarrow hwk2 \bmod n$
(6) $u2 \leftarrow (r1+r2)\, w \bmod n$
(7) $(x0,y0) \leftarrow u1g + u2Q$
(8) $v \leftarrow x0 \bmod n$
(9) The signature for message M is verified only if v = r1.

6.3. Proof of the Scheme

Signature send by A to B is (r, s) and s can be generated only by A because only A knows its private key d. $s = k1^{-1}(hk2 + d(r1 + r2)) \bmod n$ on rearranging,

- $k1 = s^{-1}(hk2 + d(r1 + r2))$.
- $k1g = s^{-1}(hk2 + d(r1 + r2))g$.
- $k1g = s^{-1}hk2g + s^{-1}d(r1 + r2)g$.
- $r = hwk2g + (r1 + r2)wdg$.
- $r = u1g + u2Q$

If same secret (k1,k2) is used for signing two different messages, It will generate two different signatures (r1,s1) and (r1,s2)

- $s1 = k1^{-1}(h1k2 + d(r1 + r2))$
- $s2 = k1^{-1}(h2k2 + d(r1 + r2))$ Where h1 = SHA512(m1) and h2 = SHA512(m2)
- $k1s1 - k1s2 = h1k2 + dr - h2k2 - dr$
- $k1(s1 - s2) = k2(h1 - h2)$

We cannot obtain k1, k2 from this equation and so this scheme is more secure than original ECDSA scheme.

7. VARIANT 4

This scheme is also called Elliptic Curve German Digital Signature Algorithm[8]. One of the disadvantages of ECDSA scheme is the calculation of inverse in signing phase. Calculation of inverse is one of the expensive operations in Modular Arithmetic, so avoiding it will reduce the cost and time. In ECGDSA inverse calculation is done in the key pair generation phase and not in Signing phase. A key will remain constant for a stable amount of time so signing is done more frequently than key generation. ECGDSA will save time and cost than ECDSA.

7.1. Key pair generation using ECGDSA

Let A be the signatory for a message M. Entity A performs the following steps to generate a public and private key

(1) Select a unique and unpredictable integer, d, in the interval [1,n-1]
(2) $Q \leftarrow (d^{-1} \bmod n)g$
(3) Sender A's private key is d
(4) Sender A's public key is the combination (E, g, n, Q)

7.2. Signature Generation using ECGDSA

Using A's private key, A generates the signature for message M using the following steps:

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) $kg \leftarrow (x_1,y_1)$, where x_1 is an integer
(3) $r \leftarrow x_1 \bmod n$; If r = 0, then go to step 1
(4) $h \leftarrow H(M)$, where H is the SHA-512
(5) $s \leftarrow (kr-h)\, d \bmod n$; If s = 0, then go to step1
(6) The signature of A for message M is the integer pair (r,s)

7.3. Signature Verification using ECGDSA

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) Verify that values r and s are in the interval [1,n-1]
(3) $w \leftarrow r^{-1} \bmod n$
(4) $h \leftarrow H(M)$, where H is the same secure hash algorithm used by A
(5) $u1 \leftarrow hw \bmod n$
(6) $u2 \leftarrow sw \bmod n$
(7) $(x0,y0) \leftarrow u1g + u2Q$
(8) $v \leftarrow x0 \bmod n$
(9) The signature for message M is verified only if v = r

7.4. Proof of the Scheme

Signature send by A to B is (r, s) and s can be generated only by A because only A knows its private key d. s = (kr-h) d mod n.

- s = (kr-h) d
- $s * r^{-1}d^{-1} = (kh * r^{-1})$
- $sw*d^{-1}g = kg – hw* g$
- kg = hw*g + sw*Q
- r = u1g+u2Q

If same secret (k) is used for signing two different messages, It will generate two different signatures (r,s1) and (r,s2).

- s1 = (kr-h1) d
- s2 = (kr-h2) d, Where h1 = SHA512 (m1) and h2 = SHA512 (m2)
- s1-s2= h2-h1

K cannot be determined even though same secret is used to sign two different messages. So this scheme is not vulnerable to attack on same secret.

8. VARIANT 5

In ECGDSA (Variant 4) there is no need of finding inverse in signing phase but there is a need in key generation phase. In this variant[9] there is no need in finding inverse in both key generation and signing phase. This scheme embeds the information of signature into a point on the ellipse.

8.1. Key pair generation

Let A be the signatory for a message M. Entity A performs the following steps to generate a public and private key:

(1) Select a unique and unpredictable integer, d, in the interval [1,n-1]
(2) Q ← (dg mod n)
(3) Sender A's private key is d
(4) Sender A's public key is the combination (E, g, n, Q)

8.2. Signature Generation

Using As private key, A generates the signature for message M using the following steps:

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) kg ← (x_1,y_1), where x_1 is an integer
(3) r ← x_1 mod n; If r = 0, then go to step 1
(4) h ← H(M), where H is the SHA-512
(5) s ← (kh + (r xor h)d)g mod n
(6) If s = 0, then go to step1
(7) The signature of A for message M is the pair (r, s)

8.3. Signature Verification

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) Verify that values r and s are in the interval [1,n-1]
(3) h ← H(M), where H is the same secure hash algorithm used by A
(4) w ← h^{-1} mod n
(5) u ← (r xor h) mod n
(6) (x0, y0) ← w(s - uQ)
(7) v ← x0 mod n
(8) The signature for message M is verified only if v = r

8.4. Proof of the scheme

Signature send by A to B is (r, s) and s can be generated only by because only A knows its private key d. s = (kh + (r xor h)d) g mod n

- s = (kh + (r xor h)d)g modn
- s = (kh +ud) g
- sw = kg +uwQ
- kg = sw - uwQ
- r = w (s-uQ)

8.5. A Forging possible on variant 5

An attacker T can forge the signature with the knowledge of public parameters (E, g, n, Q)

(1) Select an integer k in the interval [1,n-1]
(2) kg = (x1,y1), where x1 is an integer
(3) r = x1 mod n; If r = 0, then go to step 1
(4) h = H(M), where H is the SHA-512
(5) s = (khg + (r xor h)Q) mod n
(6) The Forged signature of A for message M is the pair (r, s)

On receipt on signature B will verify the signature as normal verification procedure of variant 4

V = w(s - uQ)
= w(s(r xor h)Q)
= w((khg + (r xor h)Q)(r xor h)Q)
= w(khg)
= kg

Then the forged signature validated, the attacker can successfully attack. Therefore this digital signature scheme is not secure. Anyone can use legitimate user's public-key to forge the signature of any information.

9. AN IMPROVED VARIANT 5

Variant 5 can be made secured by adding one more step in both the signing and verification phase[9].

9.1. Signature Generation

Using A's private key, A generates the signature for message M using the following steps

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) kg ← (x1,y1), where x1 is an integer
(3) r ← x1 mod n; If r = 0, then go to step 1
(4) h ← H(M), where H is the SHA-512
(5) s1 ← (kh + (r xor h)d) g mod n; If s1 = 0, then go to step1
(6) s2 ← s1d; If s2 = 0, then go to step1
(7) The signature of A for message M is the integer pair (r,s)

9.2. Signature Verification

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) Verify that values r and s are in the interval [1,n-1]
(3) Verify the equation s2g = s1Q If the equation was established continue with the verification process else refuse the signature
(4) h ← H(M), where H is the same secure hash algorithm used by A
(5) w ← h^{-1}modn
(6) u ← (r xor h) mod n
(7) (x0, y0) ← w(s - uQ)
(8) v ← x0 mod n
(9) The signature for message M is verified only if v = r

In this scheme after generating the signature s1, which is a point on the ellipse, it is encrypted with the private key of the signer. Firstly verifier verifies the encrypted result s2 before verifying the signature. As d, private key is known only to the signer and the encryption scheme s2 = s1d is guaranteed by elliptic curve discrete logarithm problem the improved scheme is secure. Attacker attempting to forge the signature by replacing the message cannot encrypt the signature so the improved scheme can prevent forgery. The signature generation and validation phase involves more elliptic curve point multiplication operation and hence is more complex and will take more time and cost. But as the complexity is increasing security level provided by the algorithm also increases.

10. OTHER VARIANTS

In order to make ECDSA secure against existential forgery by adaptive chosen message attack authors of [11] proposed a new variant of ECDSA named as ECDSAII. In ECDSAII instead of calculating the hash of the message they are calculating hash of, the message appended with r wher er = X-Coordinate of kg(mod n). Even if same message is signing hash generated will differ with high probability since k is randomly generated number.

Authors of the same paper improved ECDSAII so that it avoids the notion of duplicate signatures. They name this algorithm as ECDSAIII The alteration is to replace r=X-Coordinate of kg(mod n) with r = X-Coordinate of kg(mod n) + Y-Coordinate of kg(mod n). Since it is hard to find two elliptic curve points Q1 = (x1, y1) and Q2 = (x2, y2) such that one knows the respective discrete logarithms Qi =kiP and such that x1 + y1 = x2 + y2. It avoids duplicate signatures to an extent. This can only happen when the line L(t) : X + Y = t for some constant t is geometrically related to the group law linking Q1 and Q2. If L(t) is a tangent at Q1 and we know the discrete logarithm k1 then we know that L(t) intersects the curve in one other point, say Q2 = k2P, of the required form and that k2 = (2k1) (modn). Hence we need to avoid points where L(t) is a tangent. But for all possible values of t the line L(t) is only a tangent for atmost four points on any given elliptic curve.

11. ECKCDSA

A group of Korean cryptographers, in association with government-supported agencies, had developed a candidate algorithm for Korean digital signature standard, which is named KCDSA, Korean Certificate Based Digital Signature Algorithm, is a signature algorithm in which the public key is validated by means of a certificate issued by some trusted authority. The X.509-based certificate may be used for this purpose. In this case, the Cert Data can be simply the formatted certification data defined by X.509.An elliptic curve variant of KCDSA is EC-KCDSA Elliptic curve Korean Certificate Based Digital Signature Algorithm. The algorithm uses the public key PA := [d1 mod n]g and z is a hash-value of Cert Data. Cert Data denotes the signer's certification data, which should contain at least signer's distinguished identifier, public key Q and the domain parameters.

11.1. Signature Generation

(1) Select a unique and unpredictable integer k in the interval [1,n-1]
(2) kg ← (x1,y1), where x1 is an integer
(3) r ← x1 mod n; If r = 0, then go to step 1
(4) h ← H(z || M), where H is the SHA-512
(5) s ← d(k - r xor h) mod n; If s = 0, then go to step1
(6) The signature of A for message M is the integer pair (r,s)

Here the advantage is that there is no need of calculating inverse of d in each individual signing operation. d is the private key of the signer which will remain stable for a period of time, it can be precomputed and stored in the key generation phase itself.

11.2. Signature Verification

The receiver B can verify the authenticity of A's signature (r, s) for message M by performing the following:

(1) Obtain signatory A's public key (E, q, n, Q)
(2) check the validity of the signer's certificate, extracts the signer's certification data Cert Data from the certificate and computes the hash value z = h(Cert Data)
(3) Verify that values r and s are in the interval [1,n-1]
(4) h ← H(z || M), where H is the same secure hash algorithm used by A
(5) w ← r^{-1}modn
(6) u1 ← r xor h mod n
(7) u2 ← s

(8) (x0,y0) ← u1g + u2Q
(9) v ← x0 mod n
(10) The signature for message M is verified only if v = r

11.3. Proof of the Scheme

Signature send by A to B is (r, s) and s can be generated only by A because only A knows its private key d. u1g + u2Q on rearranging,

- r xor hg + sQ
- r xor hg +d(k – r xor h)d^{-1}g
- r xor hg +(k – r xor h)g
- kg

12. COMPARISON AMONG THE VARIANTS

Table 1. Comparison of ECDSA Variants

Algorithm	Signature Generation	Verification	No. of Secrets	Attack on same secret	Inverse in Signing	Inverse in Key Generation	Inverse in Verification
ECDSA	$k^{-1}(e+dr)$	u1 = es^{-1} u2 = rs^{-1}1 u1g + u2Q	1	Vulnerable	Yes	No	Yes
Variant 1	$d^{-1}(rk-h)$	u1 = $r^{-1}e$ u2 = $r^{-1}s$ u1g + u2Q	1	Vulnerable	No	Yes	Yes
Variant 2	$k(h + rd)^{-1}$	u1 = hs u2 = rs u1g+u2Q	1	Vulnerable	Yes	No	No
Variant 3	$k1^{-1}$(ek2+ d(r1 +r2))	u1 = $r^{-1}e$ u2 = $r^{-1}s$ u1g+u2Q	2	Not Vulnerable	Yes	No	Yes
Variant 4	s=(kr-e)d	u1 = $r^{-1}e$ u2 = $r^{-1}s$ u1g+u2Q	1	Not Vulnerable	No	Yes	Yes
Variant 5	s1 = (ke + (r xor e)d)g s2 = s1d	s2g=s1Q u = r xor e; e^{-1}(s -uQ)	1	Not Vulnerable	No	No	Yes
ECKCDSA	d(k – r xor h	u1 = r xor h u2=su1g+ u2Q	1	Not Vulnerable	No	Yes	No

Comparing all the variants Original ECDSA is vulnerable to attack on same secret. Variant 1 is suitable for signer with limited compute apparatus and variant 2 for a verifier with limited compute apparatus. Variant 3 requires the use of 2 variables and time taken for signature generation and verification is also more. In variant 4 it requires no inverse calculation in signing phase and time taken is less for signing phase but it takes more time in key generation phase as it needs to calculate inverse in key generation phase. In case of variant 5 there is no inverse calculation in both key generation and signing phase but it is more complex and it takes more time as it needs more point multiplication operation. As the complexity increases the security

level also increases. Table I shows the comparison of operations and possible attacks on all the variants.

Time taken for key generation signature generation and verification of all the variants are measured and detailed in Table II. Here the key size used is 192 bits and the processor used for implementation is Pentium(R) Dual-Core CPU 2.30 GHz and platform used is java.

Table 2. Time taken for ECDSA Variants in Milliseconds

Algorithm	Key Generation	Signature Generation	Verification
ECDSA	78	93	125
Variant 1	82	78	125
Variant 2	78	98	120
Variant 3	78	153	131
Variant 4	83	78	125
Variant 5	78	141	218

13. CONCLUSION

Performance and Security of ECDSA and its variants is compared and listed. Algorithm to be used can be determined according to the application and compute apparatus available for the application. Improved variant 5 uses more elliptic curve operations and the time taken in each of the phases is large compared to the other schemes. For applications which need more security and is having enough resources improved variant 5 can be used as the signature scheme. Variant 1 and variant 4 can be used for applications with signer having limited resources. Among them variant 4 is more efficient as it does not need pre computing and storage of an extra value d^{-1}.Variant 3 is resistant to attack on same k by usage of two secrets k1 and k2 but variant 4 is also vulnerable to same attack even without using a second secret.

ACKNOWLEDGEMENTS

This work was supported in part by Space Applications Center, ISRO, Ahmedabad and in part by the Research Grant from the Department of Electronics and Information Technology, Ministry of Communications and Information Technology, Govt of India, New Delhi. The First author would like to thank Mr Deval Mehta , Head SCTD and Ms Bhanu Panjwani Sci/Engr-SC ,SAC, ISRO for their valid suggestions and support throughout the work.

REFERENCES

[1] Darrel Hankerson, Alfred Menezes, Scott VanstoneH, ”Guide to Elliptic Curve Cryptography”, Springer 2004
[2] T.Elgamal, ”A public key Cryptosystem and a signature scheme based on discrete logarithms”, IEEE transactions on information theory, 1985.
[3] Aqeel Khalique ,Kuldip Singh Sandeep Soodv,”Implementation of Elliptic Curve Digital Signature Algorithm”, International Journal of Computer Applications 2010.
[4] Don B. Johnson,Alfred J. Menezes,Elliptic Curve DSA (ECDSA): An Enhanced DSA
[5] Hu Junru , ”The Improved Elliptic Curve Digital Signature Algorithm”, 2011 IEEE.
[6] M.Prabu,R.Shanmugalakshmi, ”A comparative Analysis of signature schemes in a new approach to variant on ECDSA ”, 2009 IEEE.
[7] Hung-Zih Liao, Yuan-Yuan Shen, ”On the Elliptic Curve Digital Signature Algorithm”Tunghai Science Vol. 8: 109126 July, 2006

[8] “Technical Guideline TR-0311 Elliptic Curve Cryptography Version 2.0”, Bundesamt fur Sicherheit in der Informationstechnik 2012.
[9] Qiuxia Zhang , Zhan Li , Chao Song , ”The Improvement of digital signature algorithm Based on elliptic curve cryptography”, 2011 IEEE.
[10] William Stallings ”Cryptography and Network Security”, fourth edition
[11] John Malone-Lee , Nigel P. Smart, ”Modifications of ECDSA”, Springer- Verlag Berlin Heidelberg 2003.
[12] Victor S. Miller ”Use of Elliptic Curves in Cryptography”, Springer- Verlag Berlin Heidelberg 1986.
[13] N. Koblitz,” Elliptic curve cryptosystems”, Mathematics of Computation 48, 1987, pp. 203-209

IMPROVED NEURAL NETWORK PREDICTION PERFORMANCES OF ELECTRICITY DEMAND: MODIFYING INPUTS THROUGH CLUSTERING

K.A.D. Deshani[1], Liwan Liyanage Hansen[2], M.D.T. Attygalle[3], A. Karunaratne[4]

[1, 3, 4]Department of Statistics, University of Colombo, Colombo 03, Sri Lanka
[2] School of Computing, Engineering and Mathematics, University of Western Sydney, Australia
[1]deshani@stat.cmb.ac.lk, [3]dilhari@stat.cmb.ac.lk, [4]ak@stat.cmb.ac.lk
l.liyanage@uws.edu.au

ABSTRACT

Accurate prediction of electricity demand can bring extensive benefits to any country as the forecast values help the relevant authorities to take decisions regarding electricity generation, transmission and distribution much appropriately. The literature reveals that, when compared to conventional time series techniques, the improved artificial intelligent approaches provide better prediction accuracies. However, the accuracy of predictions using intelligent approaches like neural networks are strongly influenced by the correct selection of inputs and the number of neuro-forecasters used for prediction. This research shows how a cluster analysis performed to group similar day types, could contribute towards selecting a better set of neuro-forecasters in neural networks. Daily total electricity demands for five years were considered for the analysis and each date was assigned to one of the thirteen day-types, in a Sri Lankan context. As a stochastic trend could be seen over the years, prior to performing the k-means clustering, the trend was removed by taking the first difference of the series. Three different clusters were found using Silhouette plots, and thus three neuro-forecasters were used for predictions. This paper illustrates the proposed modified neural network procedure using electricity demand data.

KEYWORDS

Clustering, Silhouette plots, Improve performance

1. INTRODUCTION

Predicting the future electricity demand is an essential task for a country, as a huge amount of money could be saved by utilizing the available electricity generation options. In this scenario, increasing the accuracy of short-term predictions is very crucial, as decisions regarding the required load, has to be taken within a short period of time. Literature regarding short-term load forecasting consists of both conventional time series models and artificial intelligent approaches from different fields mostly in the field of engineering. To develop a dynamic forecasting system, intelligent approaches yields better results than usual conventional time series techniques as they could be adapted to suit novel conditions and handle more complex patterns in data. However, the

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 137–147, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4412

accuracy of predictions using intelligent approaches like neural networks are strongly influenced by the correct selection of inputs and the number of neuro-forecasters used for prediction. This paper presents a cluster analysis, performed to group similar day types with respect to electricity demand. Even though many external causes like metrological conditions such as temperature, rainfall, humidity, wind speed and cloud cover, economic and demographic factors influence the electricity demand, this paper has considered only a single input which is day type. The main focus has been given to this input as this paper attempts to illustrate how data mining techniques can be complimented by statistical techniques to make them more efficient.

A dataset consisting of daily total electricity demands in Sri Lanka was considered for the period of 01st January 2008 to 31st December 2012. Each day was assigned to one of the predefined thirteen categories, suitable to Sri Lanka.

2. LITERATURE REVIEW

Literature related to load forecasting reveal that higher prediction accuracies could be obtained when using intelligent techniques when compared to using conventional statistical techniques (Farahat & Talaat, 2012; Barzamini, Hajati, Gheisari, & Motamadinejad, 2012; Nagi, Yap, Tiong & Ahmed, 2008). Many researchers point out the importance of using intelligent techniques in situations where quick weather changes lead to fail accurate predictions. (Seetha & Saravanan, 2007; Senjyu, Takara, Uezato, & Funabashi, 2002; Barzamini et al., 2012). Some of those popular intelligent techniques used in the literature are neural networks, fuzzy inference systems, genetic algorithms and expert systems.

Many researches had used the effect of different day types to enhance the load predictions considering their own country's situations. The literature reveals that, Soared and Medeiros (2008) had incorporated the maximum number of day types to their model, as Sunday - Saturday, holiday, working day after holiday, working day before holiday, working day between a holiday and weekend, Saturday after a holiday, working only during the mornings, working only during the afternoons and Special holidays. Another research considers seven days of the week and bank holidays as day types, and a principal component analysis had been performed accordingly and a segmentation scheme based on the first principal direction had been used to cluster similar months (Cho, Goude, Brossat, & Yao, 2013). Unlike these approaches, (Barzamini et al., 2012) had divided the weekly days into four categories based on unique load lags and had incorporated to the model. Considering the above, this research considers thirteen day types, which can be considered as different in Sri Lankan context.

Even though thirteen day types are considered, including all these day types into the model will complicate the prediction process. As such, the 'day type' will be clustered into similar day types in order to avoid complexities in the computation operations and to reduce forecasting error when training the neural networks (Barzamini et al., 2012; Seetha & Saravanan, 2007). They have discussed how accurate predictions are made when the inputs are wisely chosen to be fed into the neural network having different neuro-load forecasters to train similar featured loads. Literature also shows that in some research, similar days had been clustered based on experience of the experts of electricity supplying companies rather than performing any statistical analysis (Cho et al., 2013). Moreover, to understand energy consumption patterns in industrial parks, a cascade application of a Self-Organizing Map and a clustering k-means algorithm had been performed by Hernandez, Baladron, Aguiar, Carro & Esguevillas (2012). Even though no study has considered performing a cluster analysis, this study focuses on a statistical analysis based on k-means clustering to complement the neural network approach.

3. METHODOLOGY

3.1. Unit Root Test

Two common trend removal or de-trending procedures are first differencing and time-trend regression. First differencing is appropriate for I(1) time series and time-trend regression is appropriate for trend stationary

I(0) time series. Unit root tests can be used to determine if trending data should be first differenced or regressed on deterministic functions of time to render the data stationary. In this paper a stochastic trend could be seen in order to test for it Augmented Dickey Fuller test was used.

The ADF test tests the null hypothesis that a time series Y_t is I(1) against the alternative that it is I(0), assuming that the dynamics in the data have an ARMA structure. The ADF test is based on estimating the test regression

$$Y_t = \beta' D_t + \phi Y_{t-1} + \sum_{j=1}^{p} \psi_j \Delta Y_{t-j} + \varepsilon_t$$

where D_t is a vector of deterministic terms (constant, trend etc.), p is the lagged difference terms, ΔY_{t-j} are used to approximate the ARMA structure of the errors, and the value of p is set so that the error ε_t is serially uncorrelated.

When the null hypothesis of having a unit root cannot be rejected, the series can be made stationary by taking the first difference. One should pay special attention not to take the first difference to make a series trend stationary when there is a deterministic trend as it will introduce a non-invertible moving average component.

3.2. K-Means clustering

K-means clustering is a partitioning method. It partitions data into k mutually exclusive clusters. Unlike hierarchical clustering, k-means clustering operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters. The distinctions mean that k-means clustering is often more suitable than hierarchical clustering for large amounts of data.

Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. kmeans computes cluster centroids differently for each distance measure, to minimize the sum with respect to the measure that you specify.

Distance measure: In this paper, ‘kmeans’ function in Matlab software has been used with ‘city block’ as the distance measure. Even though there are five distance measure options, only ‘city block’ and 'sqEuclidean' were suited for the data, and the results for both the cases were almost similar. Therefore, ‘sum of absolute differences’, that is the ‘city block’ distance measure was considered.

Determining the number of clusters: To get an idea of how well-separated the resulting clusters are silhouette plot can be used. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points

that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster.

Avoiding Local Minima: Like many other types of numerical minimizations, the solution that kmeans reaches often depends on the starting points. It is possible for kmeans to reach a local minimum, where reassigning any one point to a new cluster would increase the total sum of point-to-centroid distances, but where a better solution does exist. However using 'replicates' one can overcome that problem by taking the one with the lowest total sum of distances, over all replicates as the final answer. (The MathWorks)

4. ANALYSIS AND INTERPRETATION

4.1. Trend Removal Process

Figure 1 displays fluctuations of daily total electricity demand from January 2008 to December 2012. A stochastic trend could be seen over the years.

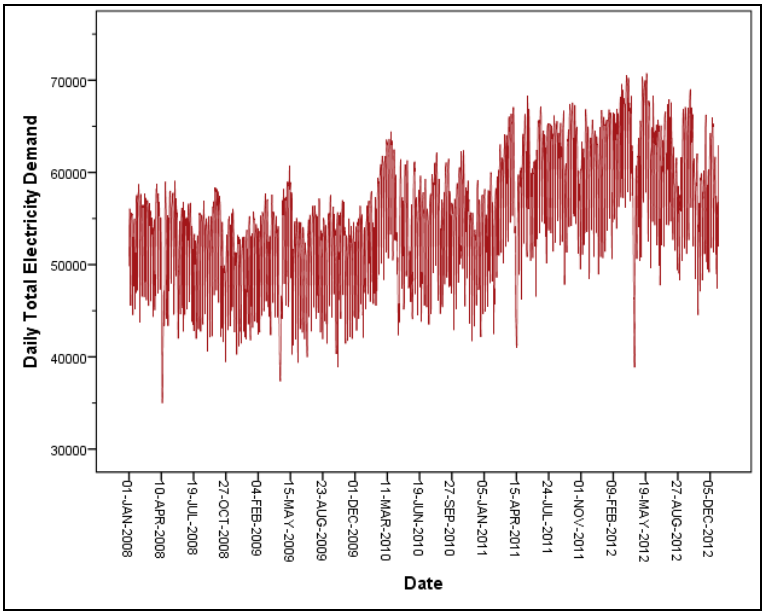


Figure 1. Daily total electricity demand over the years

Figure 1 shows that the electricity demand gradually rises during this period. This was confirmed by the Augmented Dickey-Fuller Unit Root test. When the test was performed using the original series without any additional variables, the null hypothesis of having a unit root was not rejected suggesting that the original series is not stationary. However, when a trend and an intercept terms was included into the model, the null hypothesis of having a unit root was strongly rejected and both the trend and intercept coefficients were significant in the model. When applying the test for the first difference series, the null hypothesis of having a unit root was rejected even without the trend and the intercept term. Therefore, it can be concluded that there is a trend in the series and the trend can be removed using the first difference as a difference stationary process.

4.2. Clustering Similar Day Types

After taking the first difference of the time series, the trend was removed and then the days were categorized based on the first difference of the daily total electricity demand. The trend has been removed prior to clustering, as days in the latter years tend to cluster into separate clusters where the demand is comparatively higher than the former years.

In the dataset, each day had been assigned into one of the thirteen categories; Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Poyaday, PBM Holiday (Public, Bank, Mercantile), PB Holiday (Public, Bank), Working day before holiday, Working day after holiday, Working day between a holiday and a weekend, Saturday after holiday.

In order to select the most appropriate number of clusters, silhouette plots were used based on the results of the k-means algorithm. Three clusters could be found as the most appropriate number of clusters, which resulted the maximum average Silhouette Value of 0.709384. (Figure 2). In order to avoid the iterations to end up at local minimas, each clustering procedure was replicated 5 times and considered the one with the lowest total sum of distances.

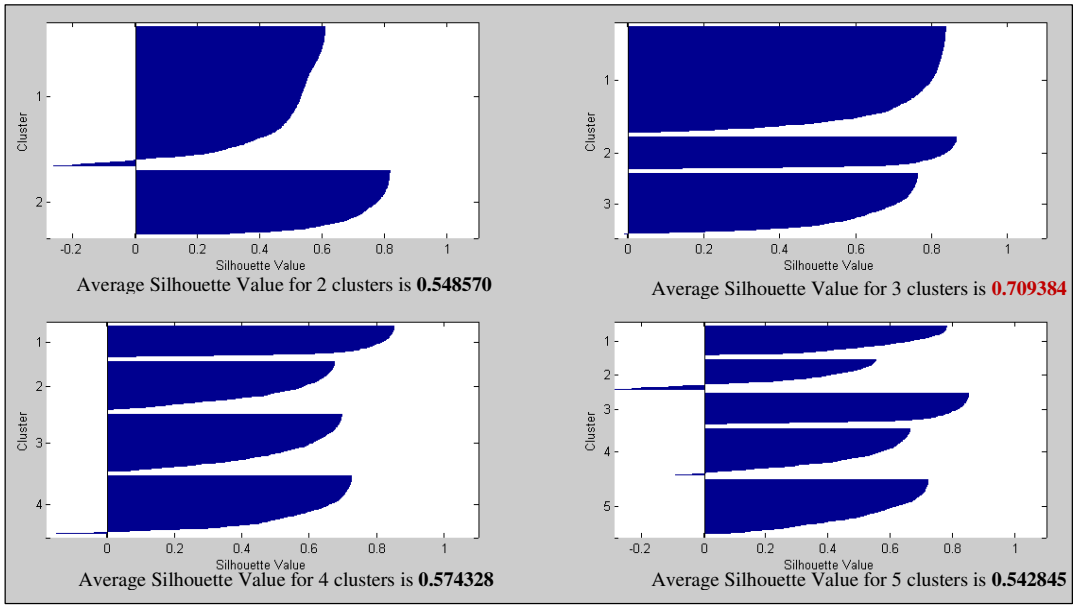


Figure 2. Silhouette Plots to determine the correct number of clusters

Based on results of the cluster analysis, the thirteen day types could be categorized into three clusters (Table 1). Figure 3 displays how the data points are scattered across three layers. A very high percentage of Tuesdays (97.2%), Wednesdays (95.3%), Thursdays (95.8%), Fridays (96.7%) and days before holidays (92.1%) were clustered into the first cluster. A 98.2% of the Mondays were classified into the second cluster. Finally, a high percentage of Saturdays (91.1%), Sundays (91.4%), Poyadays (88.7%) and PBM holidays (74.4%) were clustered into the third cluster. However, PB holidays and Saturday after holidays seemed not prominent in any of these three clusters.

Table 1. Distribution of day types across the three identified clusters

		Clusters based on the first difference series		
		1	2	3
		Row N %	Row N %	Row N %
Day Type	Monday	1.8	98.2	0.0
	Tuesday	97.2	1.9	0.9
	Wednesday	95.3	0.5	4.3
	Thursday	95.8	0.5	3.8
	Friday	96.7	0.0	3.3
	Saturday	8.0	0.9	91.1
	Sunday	8.6	0.0	91.4
	Poyaday	11.3	0.0	88.7
	PBM Holiday	25.6	0.0	74.4
	PB Holiday	39.1	17.4	43.5
	Working day before a holiday	92.1	1.6	6.3
	Working day after a holiday	22.2	77.8	0.0
	Working day between a holiday and weekend	26.9	73.1	0.0
	Saturday after holiday	58.8	11.8	29.4

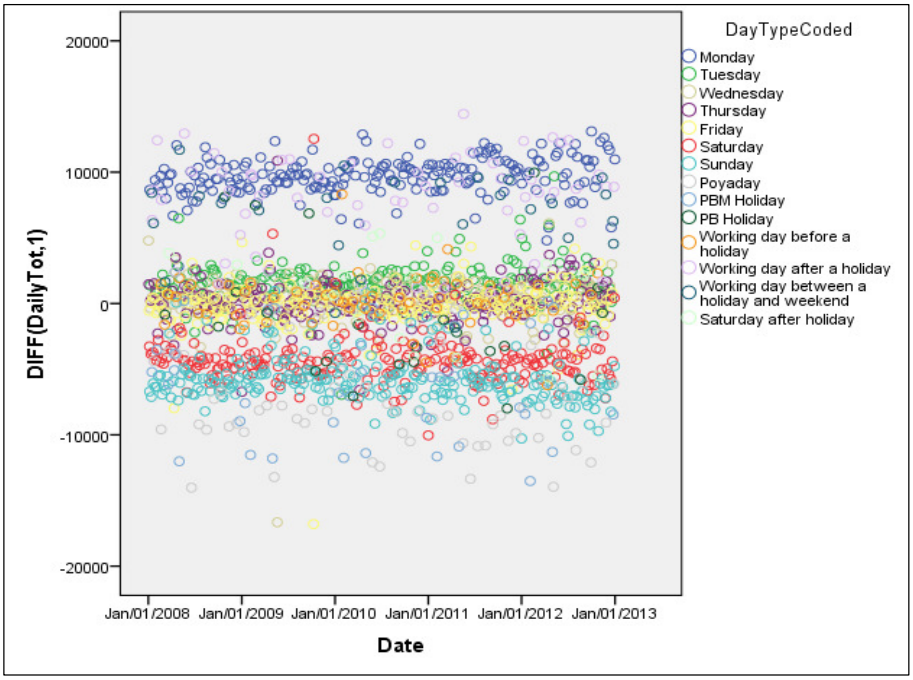


Figure 3. Spreading of different day types across the three clusters

<u>Working day after a holiday</u> (Cluster 1- 22.2%, Cluster 2 – 77.8%, Cluster 3 – 0.0%)

There were 15 'day after holiday's classified into cluster 1 where normal Tues-Friday was included. The rest of the days were similar to Mondays contributing 77.8% of the total 'day after holidays' to cluster 2. It was interesting to find out that all these observations clustering into the

first cluster were either '***a day after a PB Holiday in a weekday***' or '***a working day after a new year day***'. Therefore, a new category was created as 'Workingday after a PB holiday in weekday or 'a working day after a new year day'.

Table 2. Distribution of 'working day after a holiday's after modification

		Clusters based on the first difference series		
		1	2	3
		Row N %	Row N %	Row N %
Day Type	Working day after a holiday I: (NOT PB Holidays)	0.0	100.0	0.0
	Working day after a holiday II: (Working day after PB holiday in weekday OR Day after New Year)	100.0	0.0	0.0

<u>Poya day</u> (Cluster 1- 11.3%, Cluster 2 – 0.0%, Cluster 3 – 88.7%)

It could be clearly seen that Poya days behaved like Saturdays and Sundays, except the 11.3% clustered under cluster 1 which behaved as normal working days like Tuesdays, Wenesdays, Thursdays and Fridays. From a detailed analysis of Poya days, it was found that, if a Poya day is a Monday there is a tendency to cluster those days into cluster 1. On the other hand, if a poya day is in May, it is the Wesak festival and if a poya day is in June, it is the Poson festival, which are exceptional to any other poya day. Therefore, a new category was introduced as 'Poyaday on Monday except in May and June'

Table 3. Distribution of poya days after modification

		Clusters based on the first difference series		
		1	2	3
		Row N %	Row N %	Row N %
Day Type	**Poyaday I :** Poyadays not in Mondays and Monday Poyadays in May & June	1.8	0.0	98.2
	Poyaday II: Poya on Monday Except May & June	100.0	0.0	0.0

<u>PBM Holiday</u> (Cluster 1- 25.6%, Cluster 2 – 0.0%, Cluster 3 – 74.4%)

When considering PBM holidays, all days following Wesak Full Moon poyadays had been classified to cluster I behaving differently than other PBM holidays. Therefore a new category was introduced as 'Day Following Wesak Full Moon Poyaday'.

Table 4. Distribution of PBM holidays after modification

		Clusters based on the first difference series		
		1	2	3
		Row N %	Row N %	Row N %
	PBM Holiday (Except day following Wesak)	14.7	0.0	85.3
	Day following Wesak	100.0	0.0	0.0

Saturday after Holiday (Cluster 1- 58.8%, Cluster 2 – 11.8%, Cluster 3 – 29.4%)

This day type did not have prominent clustering percentages like other day types. Main problem was identified as the small number of data points. However, the percentages could be modified when the Saturday after holidays were classified into two new categories as 'Saturday after PB Holiday' and 'Saturday after PB Holiday OR Poyaday'.

Table 5. Distribution of 'Saturday after holiday's after modification

		Clusters based on the first difference series		
		1	2	3
		Row N %	Row N %	Row N %
	Saturday after PB holiday	33.3	0.0	66.7
	Saturday after PBM holiday OR Poyaday	66.7	25.0	8.3

PB Holiday (Cluster 1- 39.1%, Cluster 2 – 17.4%, Cluster 3 – 43.5%)
PB holidays could not be further categorized as there were a small number of observations for the subcategories.
Newly identified day types and how they are distributed among the three clusters are presented in Table 6.

5. CONCLUSION

From the K-means clustering algorithm, day types are categorized into three clusters. Therefore three back propagated neuro-load forecasters are derived and used to train data of the three clusters to achieve higher performances. When predicting electricity demand values, identifying the appropriate day type based on table 6 and feeding it into the correct neuro-load forecaster is shown to yield better results. If the considered time duration could be expanded, results can be improved and will be more accurate as there will be more observations for the sub categories. Further, instead of selecting the day type based on the extracted day types, one can automate the classification of a new data point to one of the three predefined clusters in order to increase the accuracy.

Table 6. Distribution of new day types across the three identified clusters

		Clusters based on the first difference series		
		1	2	3
		Row N %	Row N %	Row N %
Day Type	Monday	1.8	98.2	0.0
	Tuesday	97.2	1.9	0.9
	Wednesday	95.3	0.5	4.3
	Thursday	95.8	0.5	3.8
	Friday	96.7	0.0	3.3
	Saturday	8.0	0.9	91.1
	Sunday	8.6	0.0	91.4
	Poyaday I	1.8	0.0	98.2
	Poyaday II	100.0	0.0	0.0
	PBM Holiday (Except day following Wesak)	14.7	0.0	85.3
	PB Holiday	39.1	17.4	43.5
	Working day before a holiday	92.1	1.6	6.3
	Working day after a holiday I	0.0	100.0	0.0
	Working day after a holiday II	100.0	0.0	0.0
	Working day between a holiday and weekend	26.9	73.1	0.0
	Saturday after PB holiday	33.3	0.0	66.7
	Saturday after PBM holiday OR Poyaday	66.7	25.0	8.3
	Day following Wesak poya day	100.0	0.0	0.0

6. ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr. M.D.T. Attygalle, Dr. Liwan Liyanage Hansen and Ms. A. Karunaratne for the invaluable guidance, advice and the support given to do my research. Also I would like to thank the Vice Chancellor of University of Colombo, Dr. W.K. Hirimburegama, the Dean of the faculty, Prof. K.R.R. Mahanama for allowing me attend this conference. At last but not least to my loving husband, my little son, my father, my mother and my sister for their invaluable help and patience during the period I do my research.

REFERENCES

[1] Barzamini, R., Hajati, F., Gheisari, S., & Motamadinejad, M. B. (2012). Short Term Load Forecasting using Multi-layer Perception and Fuzzy Inference Syatems for Islamic Countries. Journal of Applied Sciences , pp40-47.

[2] Farahat, M. A., & Talaat, M. (2012). Short-Term Load Forecasting Using Curve Fitting Prediction Optimized by Genetic Algorithms. International Journal of Energy Engineering , pp23-28.

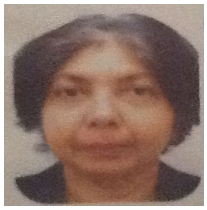
[3] Hernandez, L., Baladron, C., Aguiar, J. M., Carro, B., & Esguevillas, A. S. (2012). Classification and Clustering of Electricity Demand Patterns in Industrial Parks. Energies , pp5215-5227.

[4] Nagi, J., Yap, K. S., Tiong, S. K., & Ahmed, S. K. (2008). Electrical Power Load Forecasting using Hybrid Self-Organizing Maps and Support Vector Machines. International Power Engineering and Optimization Conference, (pp. 51-56). Selangor.

[5] Seetha, H., & Saravanan, R. (2007). Short Term Electricity Load Prediction Using Fuzzy BP . Journal of Computing and Information Technology , pp267-282.
[6] Soares, L. J., & Medeiros, M. C. (2008). Modeling and Forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. International Journal of Forecasting , pp630-644.
[7] The MathWorks, I. Statistics Toolbox.

Authors

Ms. K.A.D. Deshani is a Lecturer (Probationary) in the Department of Statistics, University of Colombo since March 2011. Before she was absorbed to the permanent cadre, she has been working in the same department as an Assistant Lecture, obtaining a B.Sc Special Degree in Statistics with Computer Science in 2008. She has a keen interest in developing computer systems to incorporate the dynamic nature to statistical interpretations. Her research interests are in the areas of Operational Research and Data Mining. Currently, she is a member of the research team of the project titled "Developing an Economical Strategy for the Future Electricity Generation Procedure in Sri Lanka"; which received the University of Colombo research grants 2011 which is carried out in collaboration with University of Western Sydney, Australia and Ceylon Electricity Board, Sri Lanka. In June 2012 she registered for a M.Phil under the said grant. With her interest in research, she was a contributed speaker at the International Statistics Conference on the publication titled "Analysis of Efficiency of a Multi-Queue against a Single Queue with Many Servers: A Study on Advertisement Counter Queues at a Leading Newspaper Company" and was published in the Proceedings of the International Statistics Conference 2011, Colombo Sri Lanka. In 2013 a research paper titled "A Study of the Dynamic Behaviour of Daily Load Curve for Short Term Predictions" was published in the proceedings of the International Symposium for Next Generation Infrastructure (ISNGI) Australia.



Dr Liwan Liyanage joined University of Western Sydney in the year 1989 with university level teaching experience at University of Colombo, University of Wollongong and King Saud University Riyadh, totalling 12 years. Qualifications: B. Sc (First Class), Graduate Diploma in Applied Statistics, Masters Degree in Theoretical Statistics and the Ph. D. in the area of Applied Probability gives her the breadth of coverage across the statistics disciplines. At UWS she has been instrumental in developing many degree programs in particular the integrated degree B. Maths and IT using data mining as the integrating tool. Senior lecturer (1995) and head of program of B. Maths & IT (1999). Her PhD was in random walk models, diffusion and related applications namely queuing theory and game theory. Thus her initial research was in applied probability, namely random walk models with difference equations, the master equation models with partial differential equations, and queuing models. This leads to differential equations representing diffusion and double diffusion. Her research, bridge the probabilistic models to the differential equation models of diffusion. Her passion to integrate disciplines and research methods have led her to the current research areas which include innovative work in "Operational Statistics" a new area developed in collaboration with UC Berkeley; Optimisation Techniques and Data Mining. Application areas include bio security, public health, climate change, electricity production and demand. From her 10 PhD/Masters students 8 had completed the research successfully. She has established ongoing national and international linkages and research collaborations and 30+ publications and a book chapter. Her paper on Operational Statistics was the 2nd most downloaded paper in April 2006 from Science Direct's TOP 25 articles.

Dr. Attygalle has been Head of the Department of Statistics from 2010 to present. As a Senior Lecturer attached to the Department of Statistics from 2006 to present, she has been routinely involved in Teaching, Research and many other administrative roles such as being the Coordinator of the MSc in Applied Statistics, BSc Special degree and Joint Special degree programmes conducted by the Department of Statistics. She holds professional memberships of the Sri Lanka Association for the Advancement of Science (SLASS) and the Institute of Applied Statistics -Sri Lanka.

Dr. Attygalle obtained her PhD in Statistics from the Lancaster University, UK, and a MSc in Statistics from the Warwick University, UK, in 2006 and 1996 respectively. Prior to this she completed a Diploma in Applied Statistics, from the University of Colombo in 1992. She obtained her first degree majoring in Statistics, Applied Mathematics and Pure Mathematics also from the University of Colombo, graduating with a first class in 1987. As professional qualifications she has obtained the Staff and Educational Development Association (SEDA)-UK accreditation as a teacher in higher education in 2005 and the Certificate in Teaching in Higher Education (CTHE) by the Staff Development Centre of the University of Colombo also in 2005. Her key research areas are Statistical Modelling, Model Diagnostics, Data Visualization and Sports Statistics. As a senior lecturer she has supervised many undergraduate and postgraduate research projects. Currently she is so-supervising two MPhil/PhD research students. She had been instrumental in developing industry links with many private and government organisations over the years and thus has carried out many consultancy projects and other training programs through the Department of Statistics. She had also won one of the University of Colombo research grants in 2011.

Mrs. A. Karunarathne is a former head of the Department of Statistics, and also had served as the former head of Department of Statistics and Computer Science. She had been in the service for more than 40 years and has been the key person to start the Special Degrees in Statistics and also initiate the Internship program in the Department. As a senior lecturer she is conducting lectures mainly in the field of Operational Research and Stochastic Processes.

She obtained a Diploma di. Sp.(Operational Research) from University of Rome and her first degree was B.Sc.(Mathematics) from the University of Colombo. She has contributed to the continuous development and transmittance of statistical knowledge through many diverse avenues, a key example of which is her involvement in the publication of a book on basics of statistics titled “Moolika Sankayanaya” written for University entrants and A/L Science Students in Sept 1997. As a senior lecturer she has supervised many undergraduate and postgraduate research projects.. Her key research areas are Stochastic Processes, Simulation, Queuing Models and Performance Modelling of Communication Networks.

INTENTIONAL BLANK

ASPECT-BASED OPINION EXTRACTION FROM CUSTOMER REVIEWS

Amani K Samha,Yuefeng Li and Jinglan Zhang

School of Electrical Engineering and Computer Science,
Queensland University of Technology, Brisbane, Australia
amani.samha@student.qut.edu.au
{y2.li, jinglan.zhang}@qut.edu.au

ABSTRACT

Text is the main method of communicating information in the digital age. Messages, blogs, news articles, reviews, and opinionated information abounds on the Internet. People commonly purchase products online and post their opinions about purchased items. This feedback is displayed publicly to assist others with their purchasing decisions, creating the need for a mechanism with which to extract and summarize useful information for enhancing the decision-making process. Our contribution is to improve the accuracy of extraction by combining different techniques from three major areas, namedData Mining, Natural Language Processing techniques and Ontologies. The proposed framework sequentially mines product's aspects and users' opinions, groups representative aspects by similarity, and generates an output summary. This paper focuses on the task of extracting product aspects and users' opinions by extracting all possible aspects and opinions from reviews using natural language, ontology, and frequent "tag"sets. The proposed framework, when compared with an existing baseline model, yielded promising results.

KEYWORDS

Data Mining, OpinionMining, Sentiment Analysis, Aspect Extraction, Customer Reviews

1. INTRODUCTION

The Internet contains vast amounts of textual information on people's expressed opinions, making the Internet an excellent source from which to gather data about a specific object within a specific domain. The ubiquity of customers' posted feedback has triggered the urgent need for systems that can automatically summarize documents. Searches for information about items available for purchase return enormous quantities of information, making it difficult to find useful data easily. Useful online information needs to be presented in a summarized form that includes the relevant data in easy-to-read and easy-to-understand format.

Reviews, forums, discussion groups, and blogs available on the Internet contain opinions and opinionated information. If extracted and summarized, those opinions could provide useful data for decision makers. The process of summarizing opinions relies primarily on identifying and extracting vital opinionated information from text. Efficiency of the process and quality of the resulting summary depends on the extraction of key information and exclusion of superfluous details. Both individuals and businesses seek opinion summaries to enhance their decision-making processes.

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 149–160, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4413

Feedback about purchased items can be objective and factual or subjective and opinionated. One customer's opinions may not fully represent the opinions of all customers, underscoring the importance of collecting and analysing opinions from many different opinion holders to evaluate the object under study. The need to understand customers' subjective feedback has made opinion extraction and summarization a hot subject in recent years. In opinion summarization, opinions are extracted, analysed, summarised, and then presented along with the corresponding opinionated information.

Researchers have studied various types of extraction and summarization, as well as methods to create and evaluate the final summary. This paper reviews recent work and covers some techniques on extracting and summarizing opinions. The primary focus is analysing customers' opinionated reviews, extracting opinionated aspects by applying the proposed framework to present extracted knowledge as" aspect-based opinion summary". The aim of this study is to achieve this goal by improving the accuracy of the aspect-based opinion summarization model to improve the quality of opinion summarization from customers' reviews. This paper documents development of a new technique to extract product aspects along with consumers' opinions about those products and aspects with the use of data mining techniques, natural language processing and ontologies. We begin with a discussion of some related work, followed by an explanation of the proposed framework, then the proposed extraction techniques, followed by experiment and evaluation, and finally conclusion with some recommendations for future work.

2. RELATED WORK

Opinion summarization from online customer reviews mainly consists of three tasks. First, aspects must be extracted. Then, associated opinion must be identified and oriented. Finally, sentence lists must be produced to form the final summary. The effectiveness of the final summary relies on aspect identification and extraction. Opinion is a perspective or a judgment formed about something; opinion is not necessarily based on fact or existing knowledge[1]. Conducting sentiment analysis is problematic [2, 3]because opinion is a quintuple of entity, aspect, orientation, opinion holder, and time[4].The entity is the item being studied (e.g., a product). The aspect can be feature, component, or function of the entity. While, orientation is the opinion provided about the entity and/or the aspect that was provided by the opinion holder at a specific time.

Summary is another concept of interest related to opinions; as explained in [5], a summary is "text that is produced from one or more texts, that conveys important information in the original text [6], which is no longer than half of the original text/s and usually significantly less than that. The Oxford Dictionary[1], defines summary as "a brief statement or account of the main points of something", and defines sentiment as "an exaggerated and self-indulgent feelings of tenderness, sadness, or nostalgia"[1].

Four broad categories of feedback for entities represent the types of words most frequently used: components, functions, features, and opinions[7, 8]. Entities for "camera" are demonstrated in Table 1. Some entities do not fit into any of the four established categories, so a fifth category, "other," is used to capture these terms.

Table 1. Entity Categories

Entity	Description
Components	Physical aspects, including the camera itself, LCD, viewfinder, battery
Functions	Capabilities, including movie playback, zoom, and autofocus
Features	Properties of components or functions, such as colour, speed, size, weight, and clarity
Opinions	Ideas and thought expressed by reviewers on product, features, components, or functions
Other	Other possible entities defined by the domain

To date, most methods have focused on extracting product aspect/features from online customer feedback and then summarizing the results, which is the first step to produce an opinionated aspect-based summary of the product under study. Hu and Liu [2, 3]presented a novel technique that performs extraction and summarization of customer reviews by using association rules based on an a priori algorithm. The system that Hu and Liu designed, extracted frequently used words representing aspects/features. In 2005, [9] proposed a modified version of the original system based on language pattern mining that identified explicit and implicit product aspect/features from positive and negative reviews.

Carenini et al. in [10]sought to improve the aspect extraction of prior designs using output from Hu and Liu's [2]model as input to their system to capture knowledge from customer reviews. The model worked by mapping the input to the user-defined taxonomy of the aspect hierarchy to eliminate redundancy and provide conceptual organization. Yi and Niblack in [11] developed a set of aspect extraction heuristics and selection algorithms to extract aspect from reviews. This model worked by extracting noun phrases, then selecting feature terms using likeness scores [12]. Popescu and Etzioniin [13] made more improvements to Hu and Liu's work[2, 3] by developing an unsupervised information system that extracted product aspects and opinions by mining reviews and removing frequently appearing nouns that are not aspects. The result of this improved system was increased precision but low recall compared to previous work.

Wu et al. in[14] proposed a novel approach to identify noun and verb phrases as aspect/features and opinion expressions, and then find the relationships between them. The method worked by extending traditional dependency parsing to the phrase level, which worked well in mining. Qiu et al. in [15]took a different approach by focusing on extraction of nouns and noun phrases, and then finding relationships between opinion words and target expressions based on dependency parsing. Both of these methods achieved normal recall performance and low precision but failed to extract infrequently cited aspects.

In [16], Qi and Chen proposed a discriminative model by using linear-chain conditional random files to mine opinions. Results of this model yielded improvements in recall and precision compared to other methods proposed by Turney[17] and Jin et al. [18]. Huang et al.[19]proposed aspect/feature extraction as a sequence label by implementing the discriminative learning model. This approach performed well, achieving an increase in both recall and precision.

3. PROPOSED FRAMEWORK

The proposed framework was designed to summarize customer reviews and produce "aspect - based opinion summary". To produce a representative summary, some essential information must be extracted. The framework is divided into four major tasks to use text files containing customer reviews as input and then perform the four tasks to produce the final output summary.

The first task is to mine entities (aspects and opinions) of the product under study and identify the associated opinion orientation of each aspect. The second task is to group aspects based on similarities. The third task is to select the most popular aspect sentences. The fourth task is to generate an opinionated summary that is based on product aspects. The architecture of the proposed framework is shown in
Figure 1.

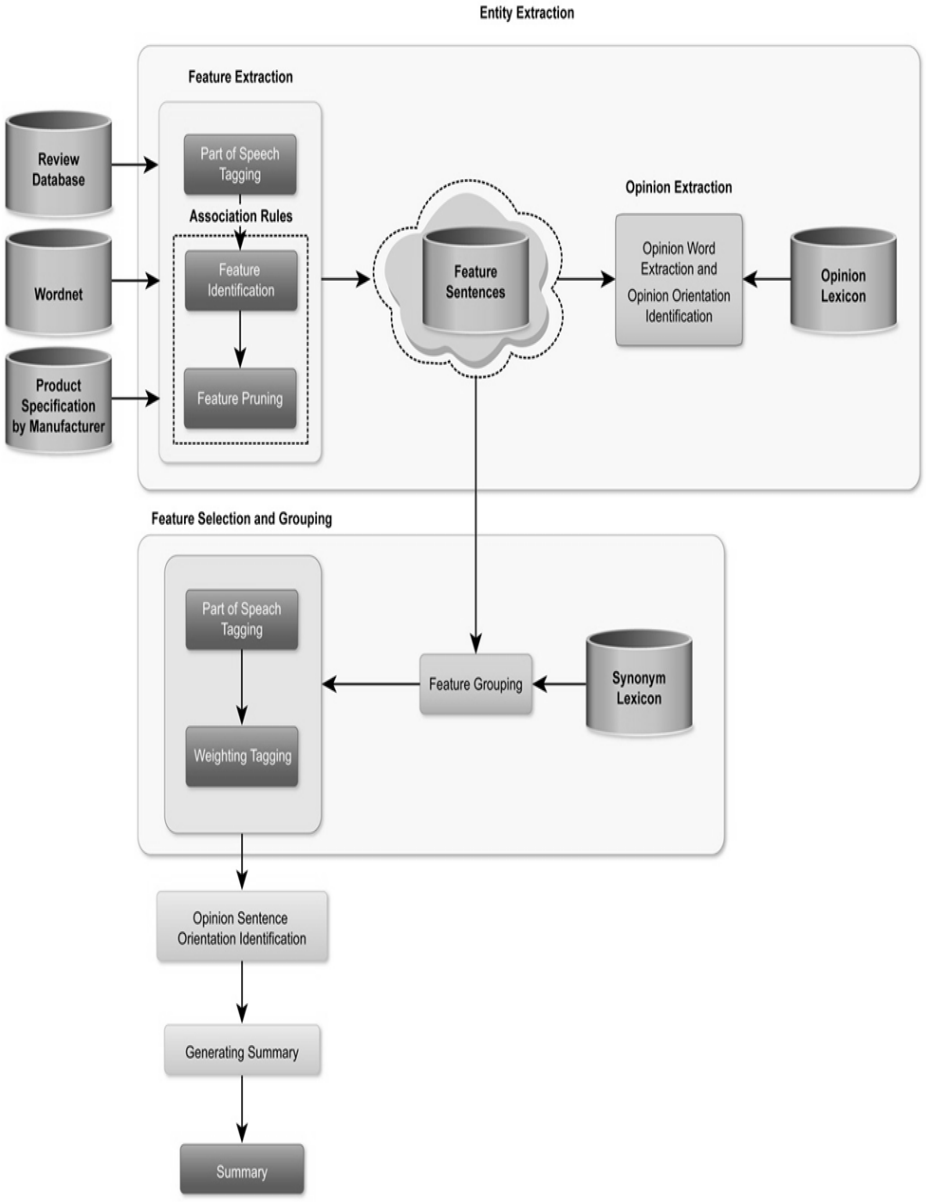


Figure 1. Proposed Framework

Although we touch on the four tasks, the focus of this paper is the proposed technique by which all possible aspects are extracted from customer reviews.

3.1. Entity Extraction

The first task of the proposed framework is "entity extraction". According to [7], entities include aspects/features, components, parts, functions, and opinions of the object being studied. For our work, entity extraction is handled as two extractions: product aspects extraction and opinions extraction. Furthermore, the extraction of aspects is decomposed into two-step process.

3.2. Aspect Grouping

Once entities have been extracted, they are grouped by based on synonyms. People may express their opinions about the same aspect using different words and/or phrases. To produce a useful summary, those different words about the same aspect must be grouped. Those words and/or phrases are domain synonyms—they share the same meaning and so must group them under the same aspect group. In a mobile phone domain, for instance, *"capacity"* and *"memory"* are two different expressions referring to the same aspect.

In this paper, aspect grouping is critical due to the numerous possible synonyms. The level of sufficiency is low for two reasons. First, although words may refer to the same aspects, some dictionaries do not consider words to be synonyms. Second, many synonyms are domain synonyms; they are likely to refer to the same aspect in one domain but not in another[20].

We aim to achieve aspect grouping using natural language possessing techniques, shared words and lexicon similarity. Some aspects may share words e.g., ("*battery*," "*battery life*," "*battery usage*," and "*battery power*"), all of which refer to the same aspect—"*battery*" [20]. Moreover, using lexicon similarity, we will match the extracted aspects to WordNet dictionary to obtain synonyms[21, 22].

3.3. Aspect Selection

After aspects have been grouped, the most representative aspect sentences must be selected to form the final opinionated summary. This step can be accomplished by analysing the strength of each opinionated sentence and then select sentences with the highest weight. The strength of all "adjectives, adverbs and verbs ", within the sentence, will determine the total weight of that sentence. Sentence importance is one of the most critical determinations of this proposed framework.

In this paper, we calculate the weights for all "adjectives, adverbs and verbs "for each the sentence. The calculation is done by adding up all weights for each" adjectives, adverbs and verbs "within the sentence, as presented in Table 2. For example, *"earpiece is very comfortable"*, the sentence has an "adjective = *comfortable*" and "adverb = *very*", therefore, the earned weight for this sentence is "2".

The weights are calculated based on the a method to score a combination of tags (adjective, verb, adverbs) to give weight to each aspect sentence, as indicated in Table 2 for adjectives and adverbs and Table 3 for verbs based on the approach proposed by [23] .

Table 2. Adjective and adverb weights

Tags	Description	Weight
JJ	Adjective	1
JJR	Comparative Adjective	2
JJS	Superlative Adjective	3
RB	Adverb	1
RBR	Comparative Adverb	2
RBS	Superlative Adverb	3

On other hand, verbs are treated differently from adjectives and adverbs. We used the categories proposed by [23]to weigh verbs, some categories are shown in Table 3.If the sentence contains a verb from positive categories, then "+1" will be added to the weight and if the verb is from negative categories then "-1" will be subscribed form the total weight. Based on final sentence's weights, the selection can be easily made. We will select sentences with the highest weight to be candidatures for the final summary.

Table 3. Verb weights

Verb category	Orientation	Verbs	Comments
Tell verbs	Positive	tell	Positively reinforce an opinion
Chitchat verbs	Positive	argue, chatter, gab	Positively reinforce opinion is being expressed
Advise verbs	Positive	advise, instruct	Positively reinforce an opinion
	Negative	admonish, caution, warn	Negatively reinforce the degree of certainty about a given opinion

3.4. Summary Generation

Summary generation is the final task of the process. It is based on the outcomes of the preceding tasks in which the extracted aspects and its corresponding opinion are selected and then weights are given to all sentences. The summary could be presented in various forms, such as diagram, text, or graph. Our expected output summary takes the form of pros and cons along with a horizontal histogram, where the pros indicate the set of positive product aspects/opinions and the cons represent the set of negative aspects/opinions. The horizontal histogram included as the percentage of positive opinions compared to negative opinions for all sentences.
Figure 2 is an example, of an aspect-based summary of "*MP3 player*".

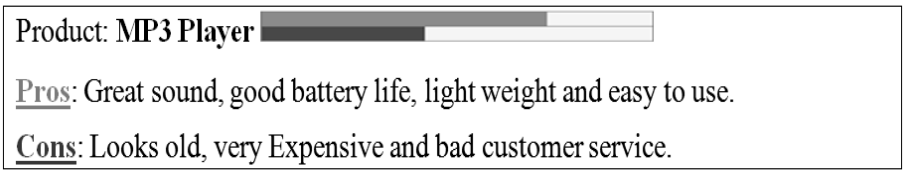


Figure 2.Aspect-based opinion summary

4. PROPOSED EXTRACTION TECHNIQUE

As illustrated in previous sections, system input is a list of customers' reviews of a specific product and the output is a summary of all reviews of that product. The initial tasks of this paper rely on part-of-speech (POS) tagging.

4.1. Part-of-Speech (POS) Tagging

To extract useful information such as aspects and opinions from reviews, the reviews must be parsed and parts of speech tagged accordingly. Part-of-speech (POS) tagging is the process of parsing each word of the sentence based on identifying linguistic tags. Table 4shows a list of linguistic POS tags. To illustrate the use of POS tagging, we offer the example of a customer's review of an iPhone5s. The original sentence is, *"I love my new Iphone5s, it is the best Smartphone ever, and it has a great camera that captures the best photos."* The tagged sentence is *"I/PRP love/VBP my/PRP$ new/JJ IPhone/NN 5s/NNS, /, it/PRP is/VBZ the/DT best/JJS smartphone/NN ever/RB, /, it/PRP has/VBZ a/DT great/JJ camera/NN that/WDT captures/VBZ the/DT best/JJS photos/NNS /"* where every word is tagged using the categories shown in Table 4.

Table 4.Part-of-speech (POS) tagging

Tag	Description	Tag	Description
JJ	Adjective	RBR	Comparative adverb
JJR	Comparative adjective	RBS	Superlative adverb
JJS	Superlative adjective	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
NN	Noun, singular or mass	VBG	Verb, gerund, or present participle
NNS/NNP	Noun, plural noun, singular	VBN	Verb, past participle
NNPS	Proper noun, plural	VBP	Verb, non-3rd-person singular/p
RB	Adverb	VBZ	Verb, 3rd-person singular present

Earlier research [2, 3] demonstrated that product aspects tend to be nouns or/and noun phrases and opinions tend to be adjectives or/and adjective phrases. In [23], sentiment analysis research showed that some combination of tags contribute to aspects and opinion extraction. Unlike these previous studies, the current research made more use of the sentence parsing process by considering more parts of the sentence to be aspects or/and opinions.

The proposed framework is designed to determine what people like and dislike about a given product. Identifying the aspects of this product is the first task, followed by finding the corresponding opinions. Understanding natural language is not easy, so the extraction process is not easy as well. The major difficulty is to understand the implicit meaning of a specific sentence. For example, *"using Iphone5 is a piece of cake,"* the phrase *"piece of cake"*means it is easy to use. However, there is no explicit word to show that hidden meaning. To solve such issues, semantic understanding is needed.

In this paper, we use OpenNLP, part of the SharpNLP Project package[24],which is a collection of natural language processing (NLP) tools that are written in C# programming language. For semantic understanding, we used a linguistic parser tool included as part of SharpNLP, OpenNLP, which parses each sentence of the review and yields the tags of each word (noun,

adjective, and so on). As Table 4shows all the POS tags taken from the Penn Treebank project POS tags [25].An additional tool, we used a WordNet database, SharpWordNet, to find synonyms in order to expand the aspect list. We use the produced output file from SharpWordNetto feed the proposed framework .

4.2. Product Aspects Extraction

Aspect extraction involves extracting aspects of the product being studied about which customers have expressed their opinions on. Aspects are usually nouns or/and noun phrases, for example,"*face recognition*", "*zoom*", and "*touch screen*" are aspects of the product "*camera*". To extract aspects, we must analyse all review sentences to know which POS items presented as aspects and which presented as opinions.

In natural language, people tend to write almost similar sentence structure. From here, we choose to use frequent sets based on its success in analysing and understanding customer purchasing behaviour. Mining frequent sets plays a great role in data mining, it aims to find interesting patterns form large amount of data. Frequent sets were introduced by [26] to analyse customer behaviour and how customers tend to purchase sets of items together. The main motivation to search frequent "tag"sets, came from the need to analyse how people tend to express their feelings in natural language. In other words, how people tend to write opinionated reviews.

To achieve the maximum number of possible aspects, we first build a list of aspects obtained from the product specifications and expand the list byword synonyms. Product specifications are aspects of the product provided by the manufacturer, while synonyms are derived from the WordNet dictionary [21]. We apply POS tagging technique to 260 sentences, then we analysis the tags based on manual observations. In order to determine how people tend to write their opinionated reviews. Then we apply opinion lexicon to match opinion words to which tags it expressed. From there, we look up for aspects by engaging the list of aspects and its synonyms.

The output is frequent sets, which consisted of frequent tags that define the product aspects, the opinion words and the relationship between those two tags. For instance, the tag of aspect appears first, therefore, the sequent of tags looks like *[NN][VBZ][RB][JJ]* which correspond to the sentence *"software is absolutely terrible"* . Figure 3 and Figure 4show tags that are more frequent, whereas Figure 5 shows how those tags are extracted.

- **[NN] [VBZ] [RB][JJ]** ***e.g.*** *"software is absolutely terrible"*
- **[NNS][VBP] [JJ]** ***e.g.*** *"pictures are razor-sharp"*
- **[NN][VBZ][RB][JJ]** ***e.g.*** *"earpiece is very comfortable"*
- **[NN] [VBZ] [JJ]** ***e.g.*** *"sound is wonderful"*
- **[NNS] [VBP] [RB]** ***e.g.*** *"transfers are fast"*
- **[VBZ][JJ]** ***e.g.*** *"looks nice"*

Figure 3. Frequent tags" Aspect appears first"

- **[JJ][NN] [IN] [NN]** ***e.g.*** *" superior piece of equipment"*
- **[JJ] [NN] [CC] [NN]** ***e.g.*** *"decent size and weight"*
- **[RB][JJ][TO][VB] [DT] [NN]** ***e.g.*** *"very confusing to start the program"*
- **[VBD] [NN]** ***e.g.*** *" improved interface"*
- **[JJ] [VBG]** ***e.g.*** *" great looking"*

Figure 4. Frequent tags" opinion appears first"

```
Algorithm AspectTagsExtraction ()
 //Input:      Sentences - List of sentences
             Dict - Feature Dictionary
             PSL - Positive Seed List
             NSL - Negative Seed List
 //Output:
             F1 - File Consisting of Possible features
             F2 - File Consisting: list of Feature & Opinion & sentence rows
2. for each sentence si ∈ Sentence do
3.     W = tokenize each word ∈si  /*Tokenized sentence */
4.     T   = tag each word   ∈ si      /*Tagged sentence */
5.     for each Wi ∈ si do
6.         if Wi ∈ Dict then
7.           apply_TwoRuleTag(si, PSL, NSL, Dict, W, T, index);  //index of the current
token in Wi
8.         else if Wi+1 ∈ Dict then
9.           apply_ThreeRuleTag(si, PSL, NSL, Dict, W, T, index);
10.        else if Wi+2 ∈ Dict then
11.          apply_FourRuleTag(si, PSL, NSL, W, T, index);
12.        endif
13.    endfor
14. endfor
```

Figure 5. Frequent tags extraction

4.3. Opinion Words Extraction

The second task of the extraction process is opinion extraction. This task involves extracting corresponding opinion words that customers used for every product aspects. Opinion words are usually adjectives that describe or express what customers think about product aspects. Usually, opinion words are located near aspects in the sentence. Some researches located opinion words as the closest adjective to the aspects[2, 3].Nevertheless, we first locate the opinions words in the sentence and from there we determine the corresponding aspects by searching the sentence backwards first for the closest aspect, if we did not find, then we search forwards.

In this paper, we use the opinion lexicon developed by Hu and Liu in [2, 3] to extract opinion words. It contains 6,800 positive and negative words in two different text files. If the word in our sentence matches the positive dictionary, the word is positive, and if a word matches the negative dictionary, then it is negative. Then, the weights for adjective are given based Table 2. Then we apply the frequent sets of tags to validate the relationship between the opinion word and the aspects. The extraction algorithm is shown in Figure 6.

```
Algorithm ApplyfrequentSets_ToTags ()   /* Aspect & opinion Extraction */
// Input:   PSL – Positive Seed List
            NSL – Negative Seed List
            W – Tokenized sentence
            T – Tagged sentence
            i – Current word/tag index
            AI – aspect index modifier
            OI – opinion index modifier
// output: aspect – extracted aspect
                     opinion – extracted opinion
2.  listOfTags1 = { " ", " ", " ", … } /* tags from predefined frequent sets */
3.  listOfTags2 = { " ", " ", " ", … }
4.  for each tag1 in listOfTags1 do
5.     for each tag2 in listOfTags2 do
6.        if Ti ∈ tag1 AND Ti + 1 ∈ tags2 then
7.           ifWi + OI ∈ PSL OR Wi + OI ∈ NSL
8.                aspect = Wi + AI
9.                opinion = Wi + OI
10.          endif
11.       endif
12.    endfor
13. endfor
```

Figure 6.Extraction algorithm.

5. EXPERIMENT AND EVALUATION

5.1. Data set

We conducted the experiment using Hu and Liu's dataset [2]consisting of annotated customer reviews of five different products:(Canon camera, DVD player, MP3 Player, Nikon and Nokia).These reviews, written by different customers , were collected from Amazom.com and Cnet.com and processed by Hu and Liu in [2]. The reviews contained 2,500 sentences. Each dataset consisted of more than 260 sentences found to be opinionated reviews written by 325 customers. The format of the datasets is unstructured text files. To evaluate the discovered aspects, a human tagger manually read all of the reviews and labelled aspects and associated opinions for each sentence. Before , we use the datasets, we pass the dataset to a pre-processing filter to remove all humane annotations and keep the original collected reviews.

5.2. Evaluation Criteria

To evaluate the performance of the proposed technique, we adopted three measurements named, precision, recall, and f-measure, and then we compared these measures to the baseline model proposed by Hu and Liu [2].The evaluation involved two perspectives: the effectiveness of aspect extraction and opinions extraction processes.

6. RESEARCH RESULTS

Having completed the aspect and opinion extraction, we reviewed our results. As shown in Table 5, our framework yielded improved precision and maintain the same recall compared with the novel work proposed by HU & Liu in [2].

Table 5shows the average precision and recall of the five products reviews named (Canon camera, DVD player, MP3 Player, Nikon and Nokia), along with the calculated f-measure of precision and recall. The precision reflects the ration of accuracy of classified aspects and

opinions to the number of all reviews, while recall reflects the ration of completeness of all reviews classified correctly.

Table 5.Comparison of proposed technique and baseline model

Average Precision		
	Aspect extraction	Opinion extraction
Baseline [2]	0.7	0.64
Proposed technique	0.99	0.56
Average Recall		
	Aspect extraction	Opinion extraction
Baseline [2]	0.79	0.69
Proposed technique	0.64	0.61
F-measure		
	Aspect extraction	Opinion extraction
f-measure for Baseline [2] and Proposed technique	0.74	0.65
	0.77	0.60

From previous results, we conducted t-*tests* to quantify the improvement of precision and recall for the extraction processes. The value of the t-*test* for precision for aspect extraction is ***"0.0001"*** and for recall for aspect extraction is ***"0.0172"*** which considered being extremely statistically significant. The value of the t-*test* for precision for opinion extraction is ***"0.0851"*** and for recall for opinion extraction is ***"0.0941"***which performs normally compared to the baselines model and leave us with a room to improve the opinion extraction.

7. CONCLUSION AND FUTURE WORK

In this paper, we proposed framework to produce an opinionated summary from customer reviews. The main achievement involved the task of aspect and opinion extraction. The extraction was based on data mining, natural language processing and ontology techniques. The main objective of this study is to provide "aspect-based opinionated summary" from customer reviews of online sold products. Our experimental results showed great promise for the technique. At this stage, we achieved very high precision and a normal recall performance compared to the baseline model in extracting aspect and opinion .In future work, we plan to improve and enhance our technique to achieve higher results.

REFERENCES

[1] J. Scott and G. Marshall, Oxford dictionary of sociology: Oxford University Press, 2009.
[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.
[3] M. Hu and B. Liu, "Mining opinion features in customer reviews," in AAAI, 2004, pp. 755-760.
[4] B. Liu, Web data mining: exploring hyperlinks, contents, and usage data: Springer Verlag, 2007.
[5] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," Computational linguistics, vol. 28, pp. 399-408, 2002.
[6] G. Somprasertsri and P. Lalitrojwong, "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization," J. UCS, vol. 16, pp. 938-955, 2010.
[7] S. Banitaan, S. Salem, W. Jin, and I. Aljarah, "A formal study of classification techniques on entity discovery and their application to opinion mining," 2010, pp. 29-36.
[8] H. Binali, V. Potdar, and C. Wu, "A state of the art opinion mining and its application domains," in Industrial Technology, 2009. ICIT 2009. IEEE International Conference on, 2009, pp. 1-6.

[9] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in Proceedings of the 14th international conference on World Wide Web, 2005, pp. 342-351.
[10] G. Carenini, R. T. Ng, and E. Zwart, "Extracting knowledge from evaluative text," in Proceedings of the 3rd international conference on Knowledge capture, 2005, pp. 11-18.
[11] J. Yi and W. Niblack, "Sentiment mining in WebFountain," in Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, 2005, pp. 1073-1083.
[12] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," Computational linguistics, vol. 19, pp. 61-74, 1993.
[13] A. M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," 2005, pp. 339-346.
[14] Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, 2009, pp. 1533-1541.
[15] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Computational linguistics, vol. 37, pp. 9-27, 2011.
[16] L. Qi and L. Chen, "A linear-chain crf-based learning approach for web opinion mining," in Web Information Systems Engineering–WISE 2010, ed: Springer, 2010, pp. 128-141.
[17] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," 2002, pp. 417-424.
[18] W. Jin, H. H. Ho, and R. K. Srihari, "OpinionMiner: a novel machine learning system for web opinion mining and extraction," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 1195-1204.
[19] S. Huang, X. Liu, X. Peng, and Z. Niu, "Fine-grained product features extraction and categorization in reviews opinion mining," in Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, 2012, pp. 680-686.
[20] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Clustering product features for opinion mining," in Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp. 347-354.
[21] C. Fellbaum, WordNet: An electronic lexical database Springer, 2010.
[22] J. J. Jiang and D. W. Conrath, "Semantic simil
taglaba(n)11.962(e)-18.349ion_comods(i)-4.5465I(n)11.965(s)3.70931(t)-4.5408(i)-4.5465(s)3.709319(x)11.962(i)-4.546108(d)11.962
53(i)-4.548070.87954( )-340.78iimt 27824j /R9 9.35(w)]f 0.44(2 -4..11 105(5.84 63 Tm ( )Tj /R9 9.36096 Tf 0.861488 -4..11 105[976(y)24.806k(i)-17.3911)-0.87

QUERY OPTIMIZATION IN OODBMS: IDENTIFYING SUBQUERY FOR COMPLEX QUERY MANAGEMENT

Sheetal S. Dhande[1] & Dr. G. R. Bamnote[2]

[1]Department of Computer Engineering,
Sipna's College of Engg & Tech,Amravati, Maharashtra, India,
[2] Prof & HOD Computer Science &Engineering Prof. Ram Meghe.
Institute of Research , Badnera Amravati, Maharashtra, India
[1]sheetaldhandedandge@gmail.com, [2]grbamnote@rediffmail.com

ABSTRACT

This paper is based on relatively newer approach for query optimization in object databases, which uses query decomposition and cached query results to improve execution a query. Issues that are focused here is fast retrieval and high reuse of cached queries, Decompose Query into Sub query, Decomposition of complex queries into smaller for fast retrieval of result.

Here we try to address another open area of query caching like handling wider queries. By using some parts of cached results helpful for answering other queries (wider Queries) and combining many cached queries while producing the result.

Multiple experiments were performed to prove the productivity of this newer way of optimizing a query. The limitation of this technique is that it's useful especially in scenarios where data manipulation rate is very low as compared to data retrieval rate.

KEYWORDS

Query Caching, Query Decomposition, Query Optimization, Stack Based Approach (SBA), Stack-Based Query Language (SBQL), Object Databases.

1. INTRODUCTION

Object-Oriented database technology is a marriage of object-oriented programming and database technologies. How these programming and database concepts have come together to provide what we now call object-oriented databases. Perhaps the most significant characteristic of object-oriented database technology is that it combines object-oriented programming with database technology to provide an integrated application development system. There are many advantages, including the definition of operations with the definition of data. First, the defined operations apply ubiquitously and are not dependent on the particular database application running at the moment. Second, the data types can be extended to support complex data such as multi-media by defining new object classes that have operations to support the new kind of information.

Query processing and its optimization are the two most popular areas of research in the database community. Query processing is the sequence of actions that takes as input a query formulated in

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 161–177, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4414

the user language and delivers as result the data asked for [1]. Query processing involves query transformation and query execution. Query transformation is the mapping of queries and query results back and forth through the different levels of the DBMS. Query execution is the actual data retrieval according to some access plan, i.e. a sequence of operations in the physical access language [2].

DBMSs support high performance of query processing by using special data structures (e. g. indexes, stacks) or by performing operations like I/O operation simultaneously. However, the major potential of performance improvement can be achieved by special preprocessing of query before execution .This preprocessing referred to as query optimization, is performed by special module called query optimizer.[3]

The task of query optimization ideally is to find the best execution plan, i.e. the execution plan that costs the least, according to some performance measure. Usually, one has to accept just feasible execution plans, because the number of semantically equivalent plans is too large to allow for enumerative search [4].

In various types of Database Systems (Relational as well as Object-Oriented), many techniques for query optimization are available [5]. Few of them are Pipelining, Parallel Execution, Partitioning, Indexes, Materialized Views, and Hints etc [6] [7].

One technique which has not been convincingly implemented is Query Caching [8].
Query Caching will provide optimum performance. Instead of spending time re-evaluating the query, the database can directly fetch the results from already stored cache. The most obvious benefit of Query Caching can be seen in systems where Data Retrieval rate is very high when compared to Data Manipulation. Hence database i.e. data store get modified after the long periodic intervals. During these intervals if a particular [25]

2. RELATED WORK

The task of a Database Management System (DBMS) is to safely store usually large amounts of consistent data and to provide easy and fast access to these data, either for retrieval or update purposes. The data model of a DBMS provides possible structure of the data so that it provides easy access to the user. The implementation of such a high-level query language requires an enormous effort; it is the task of the query optimizer to ensure fast access to the data stored in the database [1].

A DBMS is a complex piece of software that consists of many layers. The three main layers distinguished are the user interface, an intermediate layer that is called the logical algebra, and the bottom layer called the physical algebra, which provides mechanisms to actually access the data stored in files.

[5][9]In query processing, the user query is first mapped into a logical algebra expression, and this expression in turn is mapped into an expression of the physical algebra. Traditionally, query optimization consists of two phases:

- Logical optimization, which is the rewriting of a logical algebra expression into one that (hopefully), can be evaluated with less cost.
- Cost-based optimization, which concerns the mapping of logical algebra expressions into expressions of the physical algebra.

Object identity can be employed to speed up join processing, the presence of inheritance hierarchies and path expressions allows to design new index structures.

New approach present in [8]for coupling OODBMSs (Object Oriented Database Management Systems) and IRSs (Information Retrieval Systems) that provides enhanced flexibility and functionality ,This approach allows to decide freely to which document collections, that are used as retrieval context, document objects belong, which text contents they provide for retrieval and how they derive their associated retrieval values, either directly from the retrieval machine or from the values of related objects

Seen the success [8][9]that the query optimization knew in the relational model, in OODBMS data are represented in the basis as of objects. Associations are implemented by the direct ties via object identifying that permit a fast navigational access between the different objects.

[11]The usual cost measure especially for large databases is the total no of disk access (since storage access is most costly operation) such number is determined among others by the size of page which is an atomic unit of data that can be processed by a system (the most popular size of page is 4-8 KB).For smaller databases, where most of the data involved in a query can be completely stored in memory, the emphasis is usually on minimizing computation cost.

[12]While specifying in the class diagram a mono-valuated or multi-valuated tie toward another class, queries can follow this ties in order to attend the searched objects. Specific features of object oriented data models offer many additional opportunities for optimization.

Cost-based optimization [25] [26] is guided by specific database characteristics such as table size (i.e. cardinality and tuple width), the presence of indices, etc. Usually, user languages are high-level, declarative languages allowing stating what data should be retrieved, not how to retrieve them. For each user query, many different execution plans exist, each having its own associated costs. The task of query optimization ideally is to find the best execution plan, i.e. the execution plan that costs the least, according to some performance measure.

From the conceptual point of [13][14] view transparency is the most essential property of a cached query. Caching of query results yields relatively most improvement in a query evaluation performance, i.e. significantly decreases the time of anticipation for a response from database management system. Additional advantage of the query caching method is that reuse time of cached results is independent of query type, its complexity and current database state. The optimization method needs some time for storing in the cache queries, their results together with proper structures for maintenance purposes.

[15] [16]. Database operations typically involve obtaining a database root from the OODBMS which is usually a data structure like a graph, vector, hash table, or set and traversing it to obtain objects to create, update or delete from the database. When a client requests an object from the database, the object is transferred from the database into the application's cache where it can be used either as a transient value that is disconnected from its representation in the database (updates to the cached object do not affect the object in the database) or it can be used as a mirror of the version in the database in that updates to the object are reflected in the database and changes to object in the database require that the object is refetched from the OODBMS.

[17] Accessing secondary storage is much more expensive than accessing main memory , therefore some data (partial query result) are stored in special main memory buffers.

[18]sometime function can be very expensive to execute, it is sometimes advantageous to cache its result in case it is invoked multiple times with the same arguments. In such cases function input and output are stored in a special data structure. If the function is invoked afterwards for same arguments, it is enough to find the appropriate value in the structure , the subprogram does not have to be executed again . there are three main methods to implements function caching.

Memorization: computation result are stored in main memory hash table.

Sorting: it is useful when an expensive function is to be executed for the value of relation column; the input relation is stored on the input column and cache of only the last input/output pair for the function is maintained

Hybrid cache: the basic idea is to be memorization , but to manage the input stream so that main memory hash table never exceeds a maximum size; this is accomplished by staging tuples with previously unseen input value to disk and rescanning them later

Query is calculated only once i.e. for the first time and 999 times the stored result is reused [15]. Data Manipulation can invalidate the cache results because the inserted /modified /deleted data can bring the difference between the cached results and the actual results. Hence regeneration of the cached results will be required to restore the results back again to the useful state .

Oracle 11g database system offers this mechanism for SQL and PLSQL. Mysql database also implemented query chaining where only full selected query texts together with the corresponding result stored in cache. LINQ, Microsoft .NET environment also have this kind of facility [5][19].

Our research is focused on how cached queries transparent mechanism can be used in query optimization, assuming no changes to syntax, semantics and pragmatics of query language itself [16]. But there is not any result caching solution implemented in current commercial and non – commercial object oriented database system [20].

The functionality of a currently development database programming methodology called ODRA (Object Database for Rapid Application development) which works fully on the object oriented principles. The database programming language is called SBQL (Stack based query language).There are different optimization techniques, which are available in ODRA like [21]

- Query Optimization through Cached Queries for Object- Oriented Query Language SBQL.
- Query Optimization by Result Caching in the Stack-Based Approach.
- Query Optimization by Rewriting Compound Weakly Dependent Subqueries.

The most important concept of this work, Query caching have been implemented by constructing a prototype [22] which suggested in future scope (methods for reusing of some parts of cached results or results of many cached queries combined together.) in Query Optimization through Cached Queries for Object- Oriented Query Language SBQL

The experimentation for optimization of query will be done based on queries written in Stack based query language (SBQL) syntax which is designed and implemented using a stack-based architecture (SBA) framework [23][24].The performance gains will be measured by comparing the results against the performance of the identical queries written and executed in Prototype with cached and without cached queries.

This paper organized as section 1 Introduction. Section 2 Related work about Query optimization in object database. Sections 3 stack base Approach and section 4 describes description of optimization strategies-Query caching and further reuse the result in cached by earlier execution. Section 5 shows experimental findings and section 6 Discuss Experimental Result of cached semantically mapped queries and concludes.

3. STACK-BASED APPROACH

Stack-Based Query Language (SBQL) is useful for the design and implementation of wide range of database models [23]. SBQL is developed according to the Stack-Based Architecture (SBA), a conceptual framework for developing object-oriented query and programming languages [24]. In SBQL the data is stored in the form of persistent objects and the collection of data objects is called as Object Store. Hence adding, deleting or updating information in Object-oriented Databases is nothing but adding, deleting or updating the objects. Objects may contain other objects (aggregation) or references to other objects. Hence the Object-Oriented Modeling concepts of complex objects, associations between objects, classes, types, methods, inheritance, dynamic roles, encapsulation, polymorphism, semi-structured data and other features are employed in the creation of Object Store Models, a representation of the database in Object Databases.

3.1 QUERY LANGUAGE (SBQL)

SBQL permits the use of all well-known query operators such as selection, projection, navigation, path expressions, join, quantifiers, etc. SBQL has special as and group as alias operators, apart from binary operators [either of algebraic or non-algebraic type.

In the evaluation of SBQL queries two stack are in use namely ENVS (Environmental Stack) and QRES (Result Stack). In the processing of algebraic operators ENVS is not required to be used [24]. The examples of algebraic operators are Operators for arithmetic and string comparisons, set-oriented operators, aggregate functions, auxiliary naming operators, Boolean operators, etc. In contrast, processing non-algebraic operators involves operations on ENVS. The examples of non-algebraic operators are selection (where), projection/navigation and path expressions (dot), dependent join (join) etc.

3.1.1 Distinction of SBQL Queries

In contrast to SQL and OQL, SBQL queries can be easily decomposed into subqueries, down to atomic ones, connected by unary or binary operators. This property simplifies implementation. In query optimization decomposed atomic queries along with query caching plays an important role.

3.2 Object Database Models

SBA assumes a family of object store models which are enumerated AS0, AS1, AS2 and AS31. The simplest is AS0, which covers relational, nested-relational and XML-oriented databases. AS0 assumes hierarchical objects with no limitations concerning the nesting of objects and collections. AS0 also covers pointer links (relationships) between objects.

4. QUERY OPTIMIZATION ARCHITECTURE

Trasform query into syntatic tree

Static evaluator and type checker , whether query syntatically correct , validate table name , coloum name,operator , object name , method name , function name involed in query.

According to rules of normalization and send query to query optimizer (which is based on chche query optimizer)

Interpret query , find result.

Find and dispaly query result.

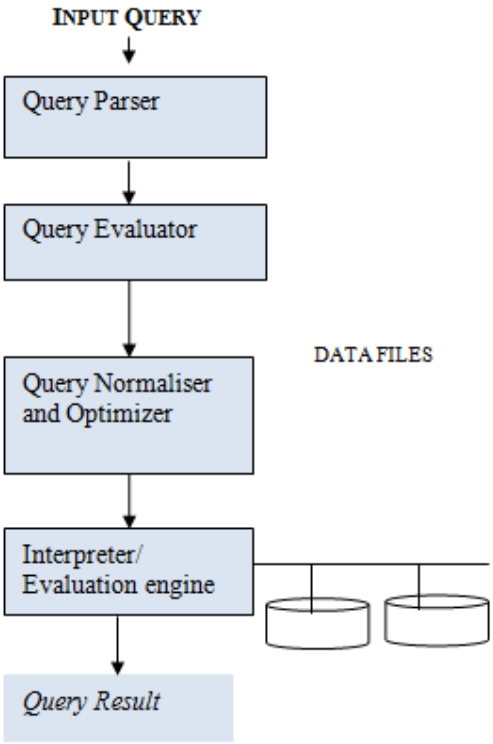


Figure 1: Praposed Architecture of Query Optimization .

4.1 Query Optimization

The scenario of the optimization using cached queries in query evaluation environment for SBA is as follows [10].

1) A user submits a query to a client-side user interface.

2) The user interface system passes it to the parser. The parser receives it and transforms into a syntactic tree

3) The query syntax tree is then received by static evaluator and type checker. It checks whether the query is syntactically correct or not. If not, it will report the errors. It also validates the table names, column names, operators, procedure names, function parameters involved query. Hence it will check the query's semantics. For this purpose, it will use the Metabase present on the server side. Metabase is a part of the database system which contains the Meta information related with the data in the various objects. This is static evaluation of the various nodes in the query syntax tree [12].

4) This type checked and statically evaluated query tree is then sent to the query normaliser which reconstructs the query according to the rules of normalization. This normalised query is then send to the query optimizer. All these components query parser, query type checker; query normaliser, query optimizer and query interpreter are employed on the client side database system.

5) The query optimizer rewrites the received normalized query using particular strategies like query decomposition. Each decomposed part of the complex query is send to the server [15]. Server checks whether the received sub query is already cached or not. If sub query is present in cache, the Unique Identification number of the entry in cache which corresponds to the result of the given sub query is dispatched to the query optimizer. Optimizer replaces (rewrites) the sub query tree of the total query tree by a node containing that unique identification number [16]. This UIN will be used by query interpreter to directly fetch the result from the server. Hence all the parts of the query whose results are already stored in cache will get replaced by their respective UIN. Due to this all sub queries which get replaced by corresponding UIN, their results will be brought from the cache & hence their re-evaluation will be avoided. This rewriting will generate the best evaluation plan which promises to give the best performance & having a least cost in terms of time and storage.

6) The optimized query evaluation plan is then sent to query interpreter.

7) The plan is executed by the query interpreter.

4.2 CLASS DIAGRAM OF EXAMPLE DATABASE

To implement this prototype we have design Match database. The additional information included is as follows

- Information about officials involved in a test match will be included in the database. These people include umpire, TV umpire, match referee and reserve umpire.
- To illustrate inheritance, here we created class *person.* All people participating in test match will inherit properties from object *person*.

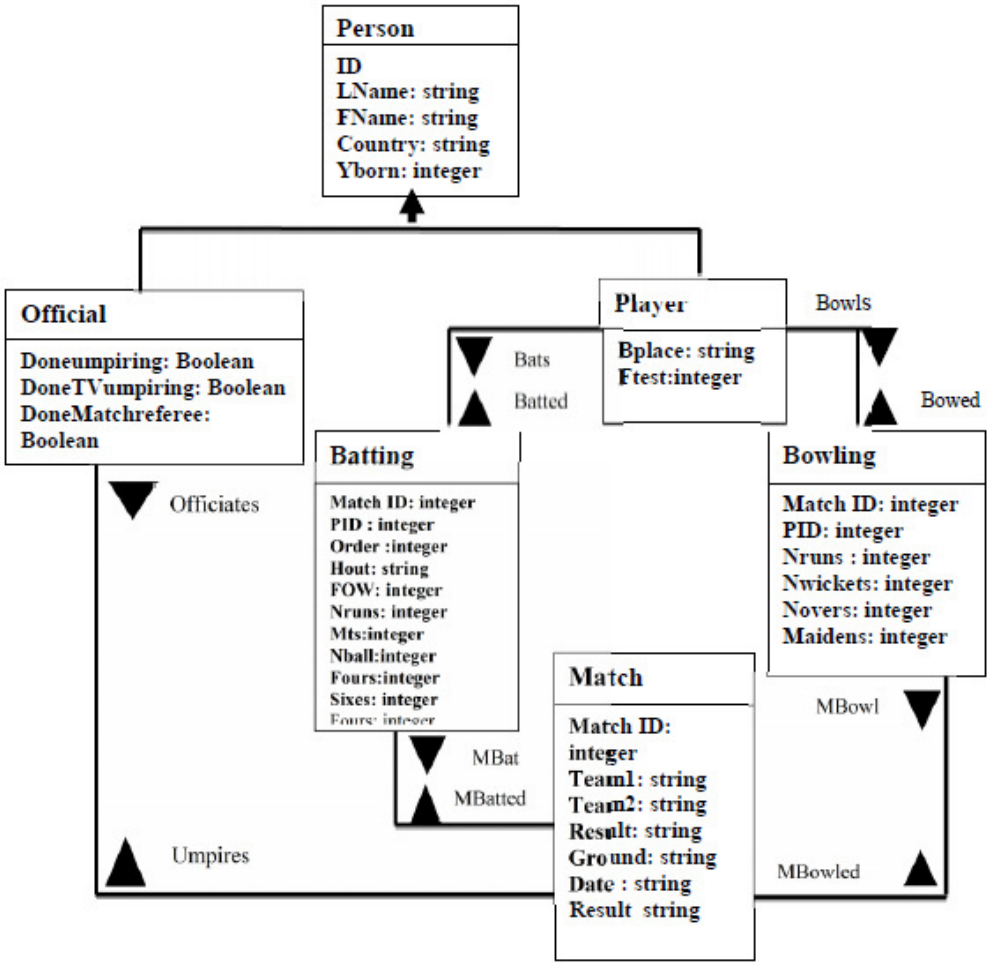


Figure 2: Class Diagram of Match Database

4.3. Query Caching

Once the optimised evaluation plan is executed successfully, the query is cached on the server side in pair <query, result>. Following that the calculated result of the query is send to the client user who has submitted the query.

When the semantically equivalent query (written in the same or different way) is submitted by the same or other user, after parsing, type checking and normalization of the query, optimizer sends the query to the server side query cache manager. Query cache manager searches for the query in the query cache registry and if found there will return the unique identification number (UIN) of the corresponding result to the client, thus avoiding the recalculation of previously stored result. Using this UIN, query interpreter (on the client system) can fetch the stored result of the query directly from the server. If the query is not found in query cache registry, query cache manager will send a message to an optimizer (on the client side database system) indicating that a query is not cached and hence its result needs to be calculated. Optimizer then does not rewrite the query i.e. does not reconstruct a parse tree. That part of the query will be then calculated by the query

interpreter at runtime using runtime ENVS (Environmental Stack) and runtime QRES (Result Stack) [22] [24].

Description of few components on server side:

Query Cache Manager – This is a program running in the server, its job is to check the Query Cache Registry and figure out if the query is cached. The Query Optimizer with pass normalized query (or normalised inner sub-query) to the Query Cache Manager.

Query Cache Registry – This contains all the cached queries along with the results.

4.4. Query Normalization

To prevent from placing in the cache, queries with different textual forms but the same semantic meaning (& hence also will generate the same result), several query text normalization methods will be used. Hence if a query is already stored in the cached with its result, all semantically equivalent queries will make reuse of the stored result, as all those queries will be mapped with the already stored query (due to normalization) [9].

Examples of few techniques useful in the process of normalization are:

a) Ordering of operands

Sum and multiply operations are put before subtractions or divisions [23], i.e. an arithmetic
expression is transformed as follows:
p / q / r * s / t
is normalized to:
p * s / q / r / t

b) Unification of Auxiliary Names

Auxiliary name used here for *as* or *group as* operators are unified. , it is not root of syntactic tree, in other case , it ii memorized query result, evaluated earlier by static evaluator.

Query Q*:*

((fname where country = "india") as f) join (f.batting as h)) . f . Ftest, h.MatchId)

Is normalized to

(((fname where country = "india") as cache _aux1) join (cache _aux1.batting as cache_aux2)) . cache _aux1. Ftest, cache_aux2.MatchId)

c) Ordering based on column names (in the order in which they appear in the object description).

d) Column names should be maintained to the left side of each operator .

4.5 Query Decomposition and Rewriting

After normalization phase query is decomposed, if possible, into one or many simpler candidate sub queries. Query decomposition is a useful mechanism to speed up evaluating a greater number of new queries. If we materialize a small independent subquery instead of a whole complex query, then the probability of reusing of its results is risen. In addition, a simple semantic of the decomposed query reduces the costs of its updating. Each isolated subquery and finally a whole query is independently analyzed in context of the set

of cached queries defined in the query cache registry and if it hasn't yet cached, it becomes a new candidate for caching.

Decomposition of queries in our proposed approached based on operators in SBQL , that are algebraic and non algebraic in categories.

Nature of operator decided query is dependent and independent, if query join with algebraic operator , always consider independent subqueries . if query involves a non algebraic operator , then subquery evaluated in the context determined by another subquery involved query.

4.5.1 Operator that identifies subqueries

In SBQL queries are combined by operators , which is algebraic and non algebraic , the majority of operators in sbql are algebraic . they include numerical comparisons , numerical operators , string comparisons, Boolean and, or, not , aggregate function , set ,bag sequence operators and comparisons , the Cartesian product (denoted by comma), etc[24].

Algebraic Operator

The definitions of algebraic operator are follows

Let q1 Δ q2 be a query formed of two binary algebraic operator Δ, evaluation of this operation is as follows

Algo *evaluatequery(query)*
Begin
……
……
resultarray : array
*Case query is q1 Δ q2 (* Δ is algebraic operator)*
Begin
Evaluate query q1 and push the result into result stack
Evaluate query q2 and push the result into result stack

Read the result of q2
Save the result of q2 in resultarray
Read the result q1
Save the result of q1 in resultarray+1
Apply Δ (temp Δ temp+1) to the result of both the subqueries and save the final result in result stack
End.

And another useful algebraic operators is defining of an auxiliary name n (for the result of query q). it assigns the name n to each row of the result table / result array returned by q
Following are semantics

Match
Return the result table/ result array
<i1> , <i2>

Then the query

Match as *M*

Return the result table/ result array

{<M(i1)>, <M(i2)>}

Group as

This names entire result of query. The semantics of this operator is as follows: if q returns a resulttable /result array t , then query

q group as N

returns a result of N(t) .

Non algebraic operator

The evaluation of non algebraic operators (a selection , a dot, a dependent join, quantifiers, etc) is more complicated. In algebraic operator, subqueries q1 and q2 occurring in query q1 Δ q2 are evaluated independently, then result make up the final query result. In non algebraic operator if query q1 and query q2 involves non algebraic operator θ, then q2 is evaluated in the context determined by q1. This is the reason this operator are referred as non algebraic.

The SBA semantics is uniform basis for the definition of several non algebraic operators of OQL like language

Selection: q1 where q2 , where q1 is any query and q2 is Boolean valued query . If q2 returned TRUE for the row r returned by q1, then r is the element of final result table; otherwise it is skipped.

Navigation, Projection, path expression: q1.q2 the final result table is the union of tables returned by q2 for each row r returned by q1. This construct covers generalized path expressions, e.g. q1.q2.q3.q4 is understood as ((q1.q2).q3).q4

Dependent join , navigational join : q1σ q2 a partial result for particular r returned by q1 is a table obtained by a concatenation of row r with each row returned by q2 for this r. the final result is the union of partial results.

Quantifiers : q1 (q2) and q1(q2), where q1 is any query , and q2 is Boolean –valued query for the final result is false for at least one row r returned by q1; otherwise the final result true . for actual definition is applied., if q2 returns FLASE

4.5.2 Distributive Property for Identifying sub queries

Here considering non algebraic operators are distributive. The properties is defined below

Definition

A non algebraic operator θ is distributive, if any syntactically correct queries q1, q2, q3 (q1, q2 are union compatible) and for any database holds following
(q1 union q2) θ q3

Is semantically equivalent to

(q1 θ q3) union (q2 θ q3)

In other world, a non algebraic operator θ is distributive , if for any query q1θ q2

Its result can be calculated as union of all result of r θ q2, where r is a row of the table returned by q1.

That is, the final result is always a union of partial results taken for all rows returned by q1.

Proposition I

Operator where, dot, and independent join are distributive
Union of partial result obtained for each row returned by the sub query occurring on the left side of such an operator. Quantifiers are not distributive. There are also others non – algebraic operators that are not distributive (e.g. grouping, ordering) , not consider in this paper.

Proposition II

The following associativity properties hold

1. (q1.q2).q3 is equivalent to q1.(q2.q3)
2. (q1 ⋈ q2) ⋈ q3 is equivalent to q1 ⋈ (q2 ⋈ q3)
3. (q1 ⋈ q2) . q3 is equivalent to q1.(q2.q3)
4. (q1 ⋈ q2) where q3 is equivalent to q1 ⋈ (q2 where q3)

Finding out subqueries

The concept of handing independent subqueries is investigate [] . in our approach we are finding subqueries for handling wider and complex queries.(we called it as complex query , because it combines different queries with non algebraic operator). Handling subqueries with context of sub query result.

Query Q1: Find MatchID of all matches in database in which Tendulkar batted.

This query need to use two classes as the player name is given only in the class Player and batting performance is given only in the lass batting . we will use a subquery to obtain the PID of Tendulkar and use that information to obtain matched of matches that he has played in.

This query is simple use of a subquery. The subquery only return a single constant which is compared with another value in the WHERE clause. If the comparison's result is true then the

row in the outside query is selected otherwise it is not. In a query like this in which the subquery returns only value, PID IN could be replaced by PID = but if are not certain that the result of the subquery will be constant, then it is best to use the IN operator.

Batting where PID IN((player where Lname= "Tendulkar").playerID).matchID

In this query, we find subquery

(player where Lname= "Tendulkar").playerID group as P

And transform whole query to following form
((((player where Lname= "Tendulkar").playerID) group as P). (Batting where PID IN .P).matchID)
We decompose query by identifying subqueries , if subquery is matched and proposed as cached query uniquely identified by its reference id ,(reference id , which we created to identify cache result stored in cache to identify query)

Batting where PID IN($cache(reference_id#1).matchID)

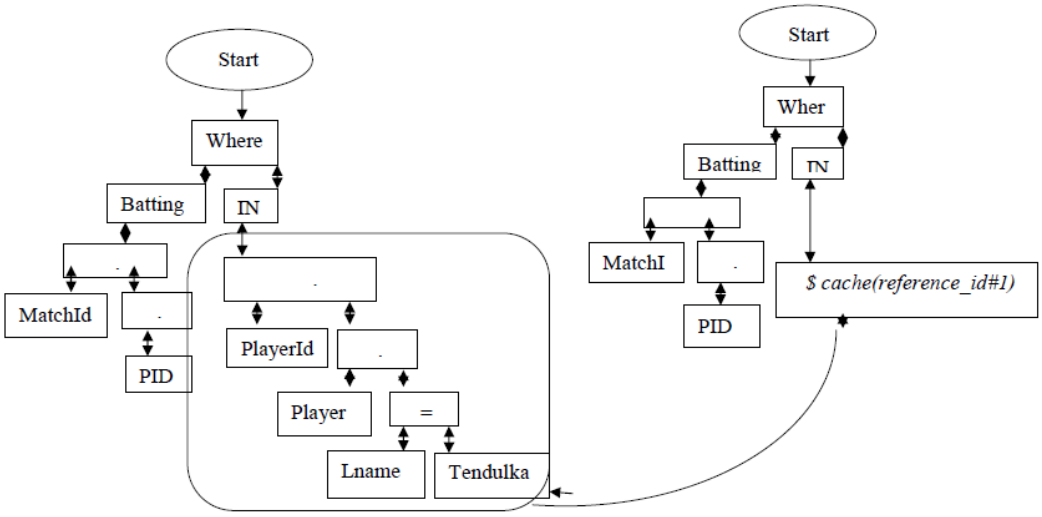


Figure 3: Sample Query Optimization

Using aggregate Function

Query Q2: Find maximum score of Tendulkar and MatchId of all matches in database in which Tendulkar batted.

This query need to use two classes as the player name is given only in the class Player and batting performance is given only in the class batting . it is same query as above but here we need to find Maximum score of Tendulkar , here we write query by using aggregate function *Max* on attribute Nruns .we will use a subquery to obtain the PID of Tendulkar and use that information to obtain matched of matches that he has played in.

This query is simple use of a subquery. The subquery only return a single constant which is compared with another value in the WHERE clause. If the comparison's result is true then the

row in the outside query is selected otherwise it is not. In a query like this in which the subquery returns only value, PID IN could be replaced by PID = but if are not certain that the result of the subquery will be constant, then it is best to use the IN operator.

Batting where PID IN((player where Lname= "Tendulkar").playerID).matchID,Max(Nruns)

In this query , we find subquery

Qsub1: (player where Lname= "Tendulkar").playerID group as P
and
Qsub2: max(Bats.Batting.Nruns) group as q

And transform whole query to following form

(((player where Lname= "Tendulkar").playerID) group as P).((Batting where PID IN .P). matched , batted.player.q)

We decompose query by identifying subqueries , if subquery is matched and proposed as cached query uniquely identified by its reference id ,(reference id , which we created to identify cache result stored in cache to identify query)

Batting where PID IN
($cache(reference_id#1).matchID,(batted .player($cache(reference_id#2))

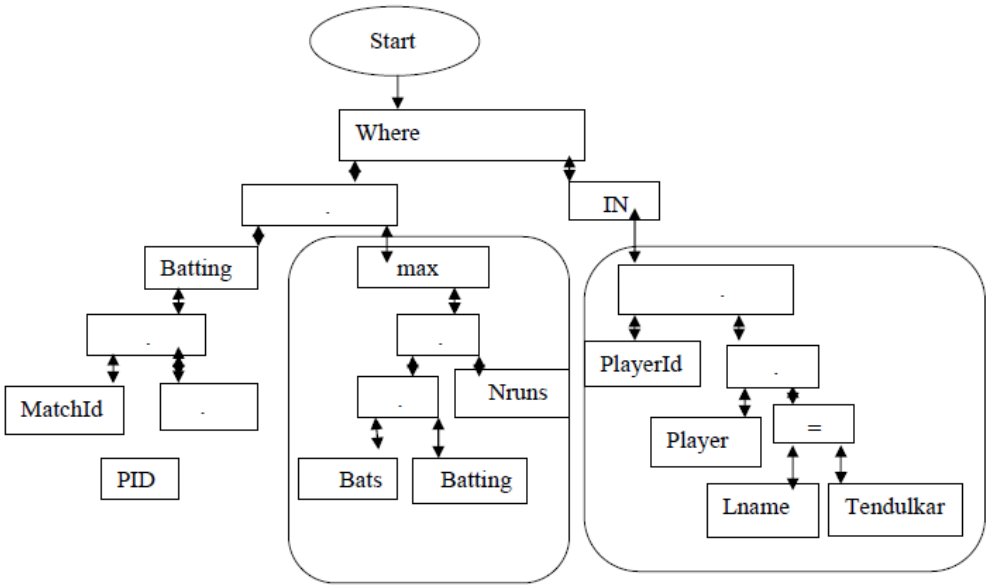


Figure 4 : Sample of Query Optimization , Handling complex queries through Cache Reference.

Subquery is matched and proposed as cached query uniquely identified by its reference id ,(reference id , which we created to identify cache result stored in cache to identify query)

($cache(reference_id#1).matchID,(batted .player($cache(reference_id#2))

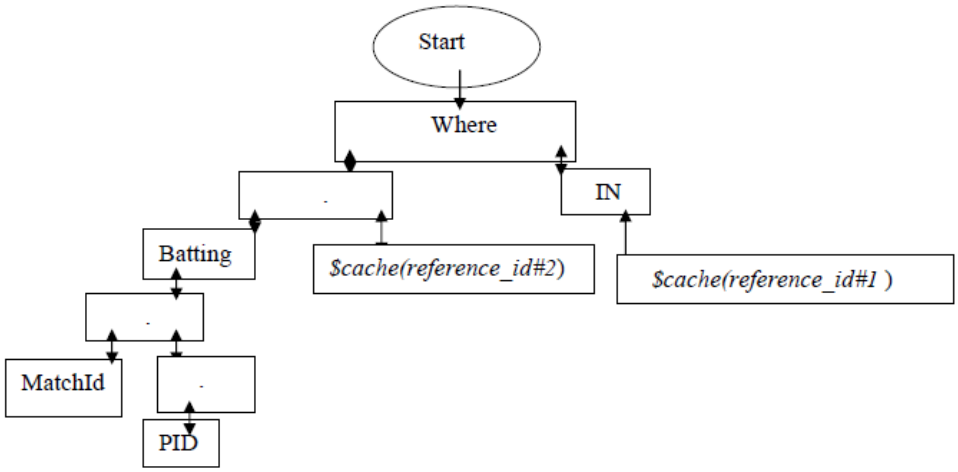


Figure 5: Sample Query Optimization through Reference Cache.

5. EXPERIMENTAL RESULT

We have the performance of the optimizer by calculating response times 100 subsequent results using set of queries retrieving data from database containing over 15000 objects. We have compare data with three optimization strategies with cache, without cache, with Db40, and in many cases especially complex queries, wider queries response time were 100 times faster. Table1 shows result in term of computation time (in micro second) of three different approaches of three different complex queries (considering complex queries , which contain many join operation , aggregate functions and many subqueries)

Approach	Query One	Query Two	Query Three
DB4o [no optimization]	150087	144779	117553
Prototype [no caching]	375127	391564	345777
Prototype[with caching]	21000	19845	20221
Note: Numbers indicates time taken to calculate the result of the query in microseconds			

6. CONCLUSION AND FUTURE SCOPE

Based on the experimental results we can state that Decomposition and Caching techniques in Object Oriented Queries have resulted in approx around hundred percent increases in performance and query output.

We have presented a new approach of query optimization, where query execution done using caching of result of previously answered queries; here we try to address future scope of [17]. Issue that we are try to handle is recognizing some parts of cached result helpful for answering other queries and combining many cached queries while producing a result of one wider query.
The work on cached queries is continued. There are many future extensions of this work, some are extending caching solutions for distributed environment .

REFERENCES

[1] Narayanan K.R.S. and Jayanthi T., "Overview of Object Oriented Databases", Conference on Recent Advances in Information Technology ,READIT - 2005

[2] H. Mistry, P. Roy, S. Sudarshan, and K. Ramamritham, "Materialized view selection and maintenance using multi-query optimization," in Proc. of ACM SIGMOD, pp. 307–318, 2001.

[3] Hanna Kozankiewicz, Krzysztof Stencel, Kazimierz Subieta "Distributed Query Optimization in the Stack-Based Approach" Springer Journal Lecture Notes in Computer Science, Volume 3726,pages 904-909, Springer-Verlag Berlin Heidelberg 2005.

[4] Roberto V. Zicari, Editor ODBMS.ORG "A New Renaissance for ODBMSs" ICOODB 09 conference on July 2, 2009, Zurich

[5] Silberchatz ,Korth ,Sudharshan "Database System Concepts" 4th edition Mc-Graw-Hill, ISBN 0-07-120413-X chapter 8, Object Oriented Database pages 307-333, chapter 13& 14 Query processing, Query Optimization,2002.

[6] Antonio Albano, Giorgio Ghelli, and Renzo Orsini "Programming Language of Object Database" in Very Large Data Bases VLDB Journal, Volume 4, 1995, Pages 403-444.

[7] Belal Zaqaibeh and Essam Al Daoud, "The Constraints of Object-Oriented Databases" International Journal of Open Problems in Computer Science and Mathematics, IJOPCM, Volume 1, No. 1, June 2008.

[8] Yannis E. Ioannidis, "Query Optimization" in International Journal on Very Large Data Bases VLDB Journal ,Volume 6, Issue No 2, May 1997 Springer- Verlag New York, USA.

[9] Christian Rich, Marc H. Scholl "Query Optimization in OODBMS" in Proceeding of The German Database Conference BTW., Springer, ISBN 3-540-56487-X March 1993.

[10] S. S. Dhande,G. R. Bamnote "Query Optimization in OODBMS: Decomposition of Query and cached for Wider Query Management" in International Conference on Research and Scientific Innovation ICRSI -2013, by Research and Scientific Innovation Society Australia | New Zealand | India , International Journal of Latest Technology in Engineering Management &Applied Science A Unit of International Standards Publication ISSN 2278 – 2540, Volume III, Issue I, January 2014

[11] Minyar Sassi and Amel Grissa-Touzi "Contribution to the Query Optimization in the Object-Oriented Databases" in Journal of World Academy of Science, Engineering and Technology, WASTE Issue No. 11, 2005.

[12] Bleja, M. Stencel, K. Subieta, K., Fac. "Optimization of Object-Oriented Queries Addressing Large and Small Collections" " in International Multiconference Computer Science and Information Technology, IMCSIT 09. ISBN: 978-1-4244-5314-6, 2009.

[13] Agathoniki Trigoni "Semantic optimization of OQL queries" Technical Reports published by the University of Cambridge ISSN 1476-2986, October 2002.

[14] Piotr Cybula and Kazimierz Subieta "Query Optimization Through Cached Queries for Object-Oriented Query Language SBQL" Springer Journal Lecture Notes in Computer Science, volume 5901/2010, pp.308-320, 2010.

[15] S. S. Dhande,G. R. Bamnote "Query Optimization in Object oriented Database through detecting independent sub queries" in International Journal of Advanced Research in Computer science and software Engineering. (IJARCSS), ISSN: 2277-128X, VOL-2,ISSUE-2 , FEB 2012.

[16] S. S. Dhande,G. R. Bamnote "Query Optimization of OODBMS: Semantic Matching of Cached Queries using Normalization" in International Conference on "Emerging Research in Computing, Information, Communication and Applications"ERCICA-2013, Elsevier Publication DBLP,ISBN:978-93-5107-102-0, Aug 2013.

[17] K. Bratbergsengen, K. Novag " improved and optimized partitioning techniques in database query processing" , springer LNCS 1271, PP 69-83, 1997.

[18] J.M Hellerstein , J. F. Naughton , " Query execution technique for caching expensive methods" , proceeding of SIGMOD , PP 423-434, 1996.

[19] "Next-Generation Object Database Standardization" Date: 27-September-2007 Object Database Technology Working Group White Paper. in International Multiconference Computer Science and Information Technology, IMCSIT 09. ISBN: 978-1-4244-5314-6, 2009.

[20] "On Oracle Database 11g" Oracle Magazine, VOL XXI, No 5 , 2007

[21] Mrs. Laika satish and Dr. Swami Halawani , "A fusion algorithm for joins based on collections in Odra (Object Database for Rapid Application development)." In IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 2, July 2011 ISSN (Online): 1694-0814, PP 289-292

[22] Piotr Cybula, Kazimierz Subieta "Decomposition of SBQL Queries for Optimal Result Caching" Proceedings of the Federated Conference on Computer Science and Information Systems, ISBN 978-83-60810-22-4, 2011 pp. 841–848

[23] K. subieta, C. Beeri, F. Matthes, and J. W. Schmidt, "A Stack Based Approach to query languages", in Proc. of IInd Springer Workshops in Computing, 1995.

[24] K.. Subieta, "Stack-Based Approach (SBA) and Stack-Based Query Language (SBQL)" http://www.sbql.pl/overview/, 2008.

[25] M. Tamer, Jose A. Blakeley "Query Processing in Object Oriented Database System" in Proceeding of ACM SIGSOFT Software Engineering Notes ,volume 35 , Issue No. 6, ISBN:0-201-59098-0, November 2010.

[26] David Dueck, Yiwen Jiang, and Archana Sawhney . "Storage Management for Object-Oriented Database Management Systems: A Comparative Survey"

.

INTENTIONAL BLANK

AN ANTI-CLONE ATTACK KEY MANAGEMENT SCHEME FOR WIRELESS SENSOR NETWORKS

Heshem A. El Zouka

Department of Computer Engineering, College of Engineering and Technology
Arab Academy for Science & Technology and Maritime Transport,
Alexandria, Egypt
helzouka@aast.edu

ABSTRACT

Wireless Sensor Networks (WSNs) are subject to various kinds of attacks such as replaying of messages, battery exhausting, and nodes compromising. While most of these attacks can be dealt with through cryptographic security protocols provided by key management schemes, there are always a few that manage to really cause problems. One such attack that is most common and significant in WSNs is cloning attack. In clone attack, the intruder tries to capture and compromise some nodes and inject them into several locations throughout the network in order to conduct other types of attacks. Moreover, if this attack is not detected early, then these replicated injected nodes will consume a large amount of the network resources. In this paper, we analyze several key management schemes that can be used for checking integrity and preventing cloning attacks. After analyzing the problems associated with these schemes, we propose a model that allows us to distinguish between legitimate nodes and cloned nodes in such sensor networks.

KEYWORDS

Node replication; Network security; Energy efficient; clone attacks ; Key management schemes.

1. INTRODUCTION

WSNs are increasingly becoming the networks of choice in many areas, including military, industrial, environmental, and medical applications. Sensor nodes are gaining interest due to their low-cost and their low power consumptions. A WSN consist of a set of sensor nodes that are distributed over a large geographic area in order to cooperatively pass the sensed data. It is expected to operate cooperatively over a long time with minimal power consumptions. Sensor nodes themselves consist of sensing, data processing, coordinating circuits and communicating components.

Therefore, the design of secure and survivable node is one of the most vital issues in designing energy-efficient protocols for wireless sensor network where the energy, memory and

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 179–190, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4415

computational power of sensor nodes are limited. In this paper, some of the challenges facing the key management schemes in WSNs are discussed in attempting to evaluate them and propose a based security solution against cloning attacks, and hence securing the communication channel.

Furthermore, utilizing the existing security protocols in wireless sensor networks has led us to propose a secure framework which incorporates Kerberos authentication protocol [1] in a way that reduce the communication overhead especially over low bandwidth networks.

The rest of this paper is organized as follows: In the following sections different types of cloning attacks will be reviewed with respect to the existing pairwise key setup schemes and their vulnerabilities. Then, the case with which sensor nodes can be compromised using regular off the shelf technology and readily available free software will be demonstrated, thereby examining the vulnerability of the existing key pre-distribution schemes. Following that, additional issues associated with cloning attacks focusing on preventive techniques rather than detective approaches will be described, for example, several possible approaches are suggested to improve the effectiveness of key management in WSNs and to avoid the problem of cloning. Finally, the last section gathers everything together; the implementation discussed along with all the simulation results obtained and a comparison of the results is presented.

2. RELATED WORK

Several possible approaches are proposed in the literature to improve the security, authentication protocols, and key management schemes in WSNs. Indeed, most existing key management schemes in sensor networks are designed to establish a pairwise key among the nodes, no matter whether these nodes communicate with each other or not, and this cause the network to suffer from many attacks and vulnerabilities [2].

These vulnerabilities allow remote attackers to sniff the network, easily create clones in the compromised nodes and inject them in several locations on the network trying to launch other types of attacks. In fact, the simplicity and low-cost of these sensor nodes can make cloning attacks more likely, especially during the maintenance phase, where some of the network nodes are replaced with new ones to prolong the battery's lifetime.

Recently, several solutions have been presented to defend a WSN against these attacks. Most of these solutions have been proposed based on the use of strong cryptographic techniques and robust key management schemes that control access among sensor nodes [ ].

To control access and secure the communication channels between nodes, each of the proposed schemes try to establish a symmetric key between every pair of neighboring nodes. The use of strong symmetric cryptography system, however, requires a robust key management scheme to handle, distribute and when needed, revoke and refresh the symmetric shared keys used for securing the communications between nodes. These established keys are often used to ensure the integrity of the overall traffic exchanged between the network nodes.

However, the establishment of pairwise keys between communicating neighbor nodes is a challenging problem due to the dense deployment and randomness nature of sensor networks. Hence, in most key management schemes, the problem of joining new node and discovering its direct neighbors in order to establish a proper pairwise keys, may remain a difficult task since the

nodes are randomly scattered across large geographical area, causing non uniform distribution of the nodes. Yet, there are many other issues that affect the design of robust and secure key management schemes. For example, the design of energy efficient protocols pushed researches to develop lightweight authentication protocols that can be used to validate the legitimate nodes in WSNs [4]. Many of these proposed protocols were presented, but none of which employs asymmetric cryptography schemes due to the limited resources of the sensor nodes.

Moreover, the lack of hardware memory protection may allow the attackers to extract sensitive information from the physical memory of the nodes. Even with well hardware protection, nodes in W|SNs are prone to failure due to hardware malfunction caused by their dense deployment of sensor nodes, thereby exposing the information stored in nodes [5]. All of these vulnerabilities may allow the attacker to reproduce new clones and inject them in several locations of the network. These clones can be easily project themselves as legitimate nodes to the network and explore other types of attacks [6]. Therefore, the detection of clone attacks is another major challenge in securing wireless networks, and will be discussed further in the following sections. Analysis of current master key based schemes in WSNs

3. CLONNING ATTACK AND KEY MANAGEMENT

To minimize the impact of cloning attacks in WSNs, a variety of key management schemes have been proposed over the past few years. These schemes can be classified into three main categories: (1) Time based schemes; (2) Geographic location based schemes; and (3) Third party based schemes. These three basic schemes are analyzed for their defense against cloning attacks, where, for example, the sensor nodes are subject to physical compromise that is hard to defend against. However, in order to analyze these schemes, it is useful to consider some assumptions which permit us to generalize the protection scope against cloning attacks. First, it is assumed that all nodes' locations are fixed and there are no mobile nodes. Secondly, all sensor nodes are deployed in a two dimensional area and each node has the knowledge of its own position and its own ID. Thirdly, it is assumed that there is a time limit T_{min} to compromise the node, and the attacker can successfully compromise the node within that time limit and obtain all the stored keys. Finally, it is assumed that every node has a setup time T_{set}, where T_{set} is the maximum time a newly deployed node needed to discover its immediate neighbors in order to establish a trusted pairwise keys with them.

Meanwhile, the base station (BS) maintains the record of IDs, master key, and positions of all sensor nodes. All the data mentioned above can be acquired during either the initial deployment of the sensor nodes or during maintenance phases of WSN.

A. Time based schemes

In this key management schemes, a master key (K_m) is preloaded into each sensor node. A sensor node uses this key to set up a pairwise key with each of its neighbors. After completion the key setup phase, each node erases the key K_m from its memory. Localized Encryption and Authentication Protocol (LEAP) is one of the most popular example of this schemes [7]. In LEAP, every node is preloaded with a master key K_m (sometimes called the primary key) under the assumption that this master key will be removed when the network is deployed.

In a network of N nodes, each node is assigned with an ID from 0 to N-1, where a node with IDu and its key Km can establish a secure one way hash function Ku = *f* Km (IDu). Then, in the neighbor discovery stage, node u broadcasts a message containing its identity IDu and set a timer, which will be triggered when the elapsed time of neighbor discovery is greater than Tmin. The response message from a neighbor node v contains its identity and message authentication code (MAC) will be used later for verifying node v's identity. In general, the following example shows how the conversation is established to generate a pairwise key between any two adjacent nodes:

$u \rightarrow$ □ ; Broadcast to all neighbors (1)
$v \rightarrow u : v \mid MAC(K_v , u|v)$; Response message (2)
$K_{u,v,\_} = f(f(k_m , v) , u)$; Computed pairwise key (3)

Therefore, by exchanging ID numbers, each node can set up a shared key with its neighbor nodes. Once T_{min} is expired, every node, such as node v, will erase the master key K_m from its memory, while keeping its own individual key (K_v). However, in case of a cloning attack, a number of security breaches can be introduced in this keying scheme. Most important, if the initial master key becomes known to the attacker at any time less than T_{min}, then the attacker can easily forge any pairwise key between two adjacent nodes. In this case, the attacker will not only be able to compromise all previously established pairwise keys in the network, but will also be able to compromise all future pairwise keys. Moreover, even if the master key is not compromised, the attacker can inject any number of malicious nodes during the maintenance operation phase of the network. In case of hardware failure of node components, the node keeps the initial master key in memory without erasing it and hence the key will be captured easily. The chance of hardware failure is more likely to increase if a deployment method uses an airplane to deploy sensor nodes.

To overcome these vulnerabilities in the basic LEAP scheme, S. Zhu et al. further proposed the extended scheme to LEAP, which was named as LEAP++ [8]. In this scheme, authors assume that the attacker is capable of recovering K_m before T_{est}. They propose a solution to this problem by having time slots for the distributed keys. Therefore, every master key is only valid for a certain time slot T, and every new joining node in the network is preloaded with a master key and a set of individual keys for all other time periods t , where $t > T$. In this scenario, if the master key is compromised, the attacker can only know the pairwise keys setup within that time T, and the pairwise keys setup in other time periods are still secure.

However, this solution introduces other potential problems, which make LEAP++ less attractive in terms of timing, control, and process. For example, one key question is how to calculate the length of time slot. If the length of time slot is too long and there are many nodes required to set up keys during this time, the approach is not relatively new compared to the LEAP protocol. On the other hand, by reducing the length of time slot, then the number of compromised pairwise keys will be also reduced. Clearly that shorter time duration will also increase the difficulty of management and deployment.

Another problem with this approach is that it does not offer support for backward authentication. So, encrypted data recorded earlier can be easily decrypted including key exchanging data between neighboring nodes. Therefore, the vulnerability of cloning attacks remains high due to the lack of backward authentication between nodes. Additionally, the attacker can add malicious nodes to the network if he in possession of the initial master key [9]. The open broadcast nature of radio communications also makes it possible for any faulty node to be impersonated without knowing it, and hence revealing the stored keys [10].

B. Geographic Location Based Schemes

In localization based schemes, each sensor node knows the coordinates of its location using either global positioning system (GPS) or any other localized methods.

For example, in case of deterministic deployment, the position of the node is calculated according to its relative distance to neighbors, and in which any pair of nodes comes under transmission range of the WSNs are considered neighbor nodes.

Generally, all localization schemes are based on Eschenauer and Gligor's random key pre-distribution (RKP) [11]. RKP scheme is a probabilistic key management scheme where each node is preloaded with a number of keys that are randomly selected from a large key pool. Neighboring nodes use these preloaded keys to set up their pairwise keys. All communication will then use this pairwise keys to authenticate and verify the integrity of the exchanged messages. In addition, based on the location of nodes, the confidentiality is maintained by assigning an index for each key, and the index of keys is exchanged between nodes and their neighbors to determine their shared pairwise keys. Therefore, information about the position of the node can be used to ensure confidentiality between neighbor nodes and hence preventing cloning attacks.

However, compromising one node will reveal its keys and any established pairwise keys, although the attacker cannot inject malicious nodes elsewhere into the network. This is because of the location of the nodes which were deployed on predefined regions of the network. Another problem with the location based schemes is that they consume more memory than other key management schemes of WSNs since each node needs to store the coordinates of its neighbors, and the relative amount of memory in WSN is very limited. However, in such schemes, the energy consumption will be balanced among all the sensor nodes and hence the network lifetime can last longer.

C. Third parity based schemes

These types of schemes depend on a trusted third party (e.g. the base station) or a server that acts as a key distribution center (KDC) where a pairwise key is generated upon request of any two sensor nodes in the SN wishing to communicate. The KDC normally sends this key in encrypted form to the communicating nodes. An example of this scheme is Kerberos, which was built on the Needham- Schroeder protocol. Kerberos was originally designed to enable two parties to exchange secret information across an otherwise open network [12,13,14].

In this key management scheme, each sensor node of the network shares a different secret key with the KDC, which enables the nodes to verify the received message originated from the base station. The Kerberos server itself provides a centralized server whose function is to validate sensor nodes by providing them with ticket to grant request to the base station. Actually, both uthentication server and a ticket granting server, the main two components of Kerberos, work together as a trusted third party (TTP), and the authentication server knows all the nodes' passwords and stores them in a centralized place.

Actually, both authentication server and a ticket granting server, the main two components of Kerberos, work together as a trusted third party (TTP), and the authentication server knows all the nodes' passwords and stores them in a centralized place.

On the other hand, the purpose of the ticket granting server is to certify to the server/Base station in the network and to ensure that a node is really what it claims to be. In this way, both the authentication and authorization servers are used to authenticate node to each other in WSN.

Figure 1 describes how the node and the base station are jointly configured to verify each other's identity via the Kerberos server. In this flat connection protocol, the Kerberos key exchange mechanism specifies three exchanges: the Kerberos authentication exchange, the key granting service exchange and base station to node service exchange. In this way the connection is established between the nodes and the servers to enable them to exchange the keys and certificates. However, the deficiency with these protocols is that they use what is known as "hierarchical authentication protocols" where each sensor node in network has only one authentication provider, which is Kerberos in this case. When the network density is high, all the sensor nodes have to wait for a long time to be authenticated and establish a semi SSL connection with the base station. From energy consumption perspective, most amount of energy is consumed in such authentication and authorization process. In order to avoid energy consumption and unnecessary traffic, which in turn may increase the average delay and cause a cloning attack, an alternative practical approach that uses the envelope model is presented and described in section 3 of this paper, but with some changes.

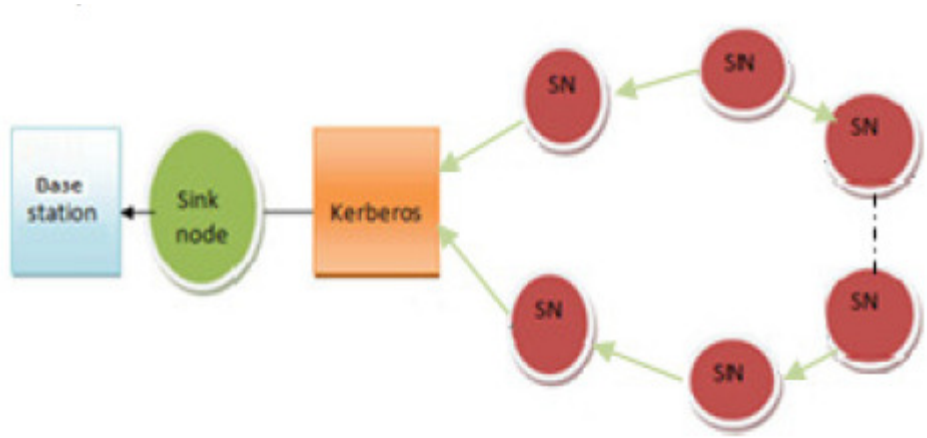


Figure 1. Flat Connection Model

In this model, the network is divided into clusters and a set of Kerberos controllers as shown in Figure 2. Each controller works as an authentication authority and a key management for one cluster in the control group of the WSN. On the other hand, all the nodes inside each cluster will communicate with the CH node using AES encryption Algorithm.

The CHs themselves will authenticate and communicate securely with each other using Kerberos. The effectiveness of this model is that it distributes the keys among the upper hierarchy of CHs using Kerberos authentication, and strong symmetric cryptosystem among cluster nodes, making it impossible for cloning attacks to take place. Even if the attacker succeeded to compromise one cluster, the other clusters are still protected.

The proposed Hierarchical model uses multiple Kerberos controller as apposed to the Flat model.

Clearly, because of the constraints imposed on WSNs, such as energy limitation, the cost of having many Kerberos controllers tend to be quite complex and usually defy analytical methods that have been proved to be fairly effective for Flat connection model.

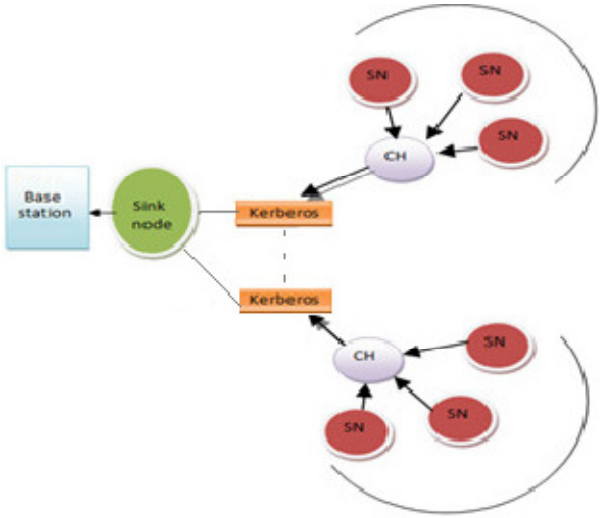


Figure 2. Hierarchical Connection Model

Another advantage of Hierarchical model over Flat is the minmized overhead as there is no common master key shared between the nodes across the network to help each node realize its closest neighbor. Also the Flat model creates a single point of failure acting as a bottleneck in the whole network. In addition, the model uses AES-128 encryption in the communication between nodes of the same cluster which offered faster computation, thus minimizing energy dissipation in these cluster nodes.

However, because of many constraints imposed in modeling Hierarchical networks, such as the dependency measures of multiple Kerberos controllers, modeling of such networks tend to be quite complex. Furthermore, few controllers have come into existence, for there are still many research experiments that need to be considered.

4. DISCUSSION AND SIMULATION RESULTS

In this paper the cloning attack problem and its impact with respect to three categories of key management schemes were presented. In time based scheme, the master key in basic LEAP protocol is used to calculate all of its neighbor pairwise keys. We noticed that the node can be compromised by reproducing clones which will allow the intruder to infiltrate the sensor network, and then other types of attacks can be conducted. Therefore, the first type of key management scheme exploits seriously degrades the resilience of such schemes.

To overcome the vulnerabilities in basic LEAP protocol, we showed how LEAP++ used a time slots for the distribution of the pairwise keys. In this protocol, every master key is valid for certain time slot T, and every new joining node is preloaded with a master key. However, we found that LEAP++ did not offer advantages compared to LEAP in terms of timing, control, and process. Besides, it is not easy to calculate the length of time slot. The analysis also showed that

the vulnerability in time based schemes remains high due to the lack of backward authentication between nodes, which make these schemes vulnerable to the cloning attacks.

We analyzed the localization based schemes, and found several constraints and limitations which can limit the use of such schemes. We defined the problem of localization systems as estimating the position or coordinated of sensor nodes. In localization schemes, nodes can be equipped with a GPS system, but this is a costly solution in terms of memory and power consumption. We also found that most of the deterministic deployment algorithms were not aware of range measurement inaccuracy or had not considered the scaling problems in designing their localization algorithms. However, one of the benefits of using localization ased schemes is their ability to store all the information needed to determine the position of the nodes which can assist in strengthening the process of key establishment and hence, in preventing cloning attack.

Then, we examined the schemes which involve the base station in the process of key management. We presented the strengths and weaknesses and what are the possible attacks to these management schemes in general. In these schemes, the base station plays a central role in generating the pairwise keys and authenticating the nodes. Two authentication schemes were discussed, one is Flat connection model and the other is Hierarchical connection model.

In Flat model, the connection is established between nodes and servers in a manner that is secure and efficient in terms of authenticity. However, the performance of these schemes degrades significantly when the number of sensor nodes increases. Clearly, a network that has only one authentication provider will cause considerable routing overheads and longer authentication time.

In Hierarchical model, the cluster heads are selected according to their battery life time and in a way similar to [15, 16]. In this scheme sensor nodes play the roles of cluster heads periodically. Whenever a cluster head is elected in a cluster, the CH broadcasts a message to other member in the cluster that it becomes a cluster head.

We evaluated the performance of Hierarchical compared with Flat structure in detail including energy consumption and battery life time. We used OMNET [17] as a simulator to analyze the performance of Flat and Hierarchical.

The basic assumptions used in performance analysis assumes that different energy consumption values would be generated according the key management process performed by nodes and servers, making a distinction between the distance among sensor nodes and the authentication servers. The network size was simulated as a square area of 100 x 100 m3, and the performance of algorithms was analyzed with respect to the lifetime of the network.

Table 1 Network variables

Items	Value
Sensing area (m^2)	100 x 100 m
Number of nodes	100, 500
Initial energy (J)	10 J
Tx energy	50 nJ/bit/m^3
Rx energy	50 nJ/bit/m^3
Packet size (bytes)	32 bytes

On the other hand, the amount of consumed energy was measured by considering the energy consumption required for the replacement of cluster heads and the broadcasting messages between all nodes and their servers. The model is implemented based on the assumptions listed in table 1. As shown in the table, 100 sensor nodes were randomly deployed over an area of 100x100 m3 to be used in the simulation, and then we increased the number of sensor nodes to be 500 distributed over the same area.

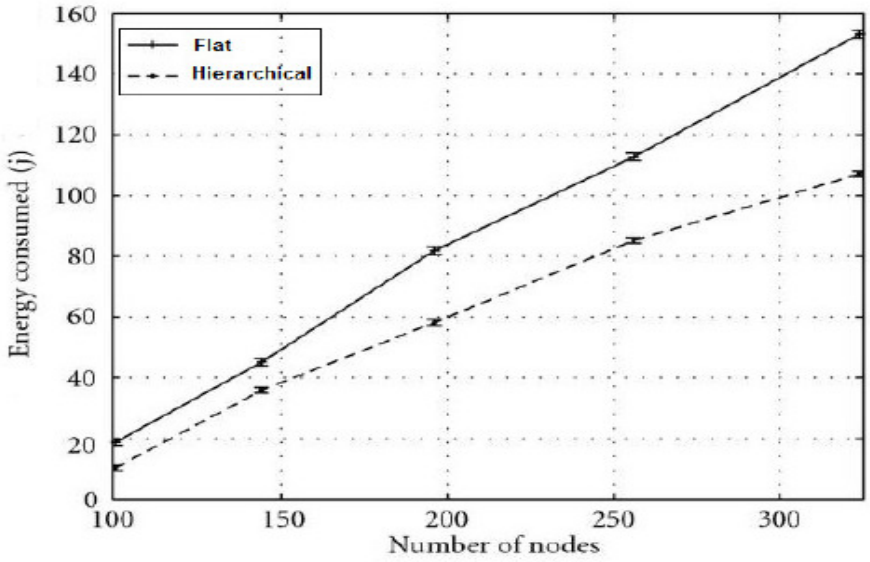


Figure 3. Energy Consumption

All nodes are assumed to have fixed locations and no mobility feature. All nodes are homogeneous and have the same initial energy of 10 J. The energy required by the radio to run the transmitter or receiver circuitry = 50 nJ/bit/m3. For modeling the Kerberos authentication server, we applied a four byte SHA-1 algorithm such that an intruder has to generate 231 packets on average and the sensor nodes would be dead. The compressed data packet size in bytes = 16. We plotted the average of 100 simulate experiments, and the compare results are shown in Figure 3 and Figure 4.

As illustrated in Figure 3, we can observe that Hierarchical is more energy efficient than Flat. Based on these results, we noticed that more than 75% of the sensor nodes in the Hierarchical model preserved their energy as the energy is consumed mostly around the cluster heads. On the

other hand, the Flat model introduces more energy consumption due to the longer paths to Kerberos and consequently higher end-to-end packet transmission time. Therefore, based on these results, we conclude that Hierarchical model is better than Flat in terms of balancing the energy consumption in wireless sensor networks.

As illustrated in Figure 4, the average throughput measured over the Hierarchical model tends to be higherthan the Flat model due the aggregation of all packets at the CHs. Clearly, the flat model offers a higher end-to-end delay as the data travels a long distance before it reaches the BS/Kerberos controllers.

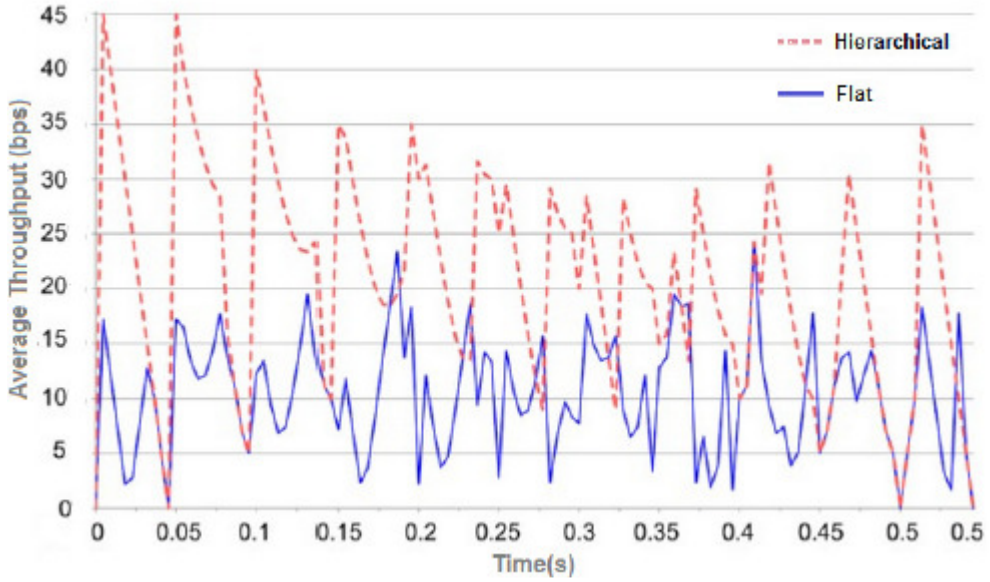


Figure 4. Communication Overhead

On the other hand, the Hierarchical model offers a higher throughput, faster key management scheme, and lower authentication delay than the Flat model. Therefore, we can finally conclude that the Hierarchical model has achieved better simulation results than the traditional Flat model in terms of energy, throughput performance, and network life time.

5. CONCLUSION AND FUTURE WORK

In this paper, the challenges and the approaches for the security and routing protocols of WSNs were surveyed. Then, a framework that secures the communications between the wireless nodes was proposed. In the first experiment, a Hierarchical model that uses Kerberos controller along with a cluster head in a hybrid manner to preserve the energy and increase the life time of WSN was implemented. In the second experiment, the process of employing the base station to enhance the authentication protocol of the sensing nodes was examined. To improve the performance of the Flat model, the proposed Hierarchical architecture is implemented using two security layers, one for establishing authenticity and one for generic trust that authenticates the distributed Cluster Heads. The existing key management schemes were surveyed, and based on their response, a

Hierarchical model that uses multiple Kerberos controllers to improve the effectiveness of key management in WSNs was proposed.

The analysis showed that the proposed Hierarchical model provides a significant increase in the life of the entire network as more than 75% of the nodes reserved their energy while the consumption is limited to the CHs. As for evaluating the effectiveness of employing a strong authentication technique, the analysis showed that the distributed Kerberos controllers experienced fewer losses by sending fewer instructions per packet and the resulting compressed data rate was improved.

In the future, the scale of the network will be increased and more than one base station will be examined, also we plan to make our protocols aware of data freshness by adding time stamp to the authenticated packet. Additionally, we plan to study the performance of our model on different motes and build a comparison over different architectures.

REFERENCES

[1] J. G. Steiner, C. Neuman, and J. I. Schiller, ``Kerberos, an Authentication Service for Open Network Systems'," USENIX Association Conferences Proceedings, February 1988, pp. 191-202.

[2] D. Manivannan and P. Neelamegam, "WSN: Key Issues in Key Management Scheme – A review," Research Journal of Applied Science, Engineering and Technology, vol. 4, 2012, pp. 3188-3200.

[3] S. Othman, A. Trad, and H youssef, "Performance Evaluation of Encryption Algorithm for Wireless Sensor Networks," International Conference on Information Technology and e-Service (ICITeS), March 2012, pp. 23-35.

[4] O. D. Mohatar, A. F. Sabater, and J. M. Sierra, "A lightweight Authentication Scheme for Wireless Sensor Networks," Ad Hoc Networks, vol. 9, no. 5, 2010, pp. 727-735.

[5] C. Sreedhar, S. Vema, and P. Kasiviswanath, "A Survey on Security issues in Wireless ad hoc Routing Protocols, International Journal 2(2), 2010, pp. 242-232.

[6] A. Pandey and R. Tripathi, "A Survey on Wireless Sensor Networks Security," International Journal of Computer Applications, vol.3, no.2, June 2010, pp. 8887 – 8975.

[7] S. Zhu, S. Setia, and S. Jajodia, " LEAP: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks," ," in Proceedings of CCS'03, The 10th ACM Conference on Computer and Communications Security, Washington D.C, USA, October 2003, pp. 27-31.

[8] S. Zhu, S. Setia, and S. Jajodia, "LEAP+: Efficient Security Mechanisms for Large Scale Distributed Sensor Networks," ACM Transactions on Sensor Networks vol. 2, 2006, pp. 500 – 528

[9] J. Jang, T. Kwon, and J. Song, "A time-based key management protocol for wireless sensor networks," in Proceedings of ISPEC07, Information Security Practice and Experience, 2007, pp. 314–328.

[10] B. Tian, S. Han, L. Liu, S. Khadem, and S. Parvin, "Towards Enhanced Key Management in Multi-phase ZigBee Network Architecture," , in Proceedings of Computer Communication, vol.35, no.5, pp. 579-588.

[11] Laurent Eschenauer and Virgil D. Gligor, "A Key Management Scheme for Distributed Sensor Networks," in Proceedings of the 9th ACM Conference on Computer and Communication security, November 2002, pp. 41-47.

[12] R. Needham and M. Schroeder, "Using Encryption for Authentication in Large Networks of Computers," Communications of the ACM, vol. 21, no. 12, December 1978, pp.993-999.

[13] B. Clifford and Theodore Ts'o, " Kerberos: An Authentication Service for Computer Networks," From IEEE Communications Magazine, vol. 32, no. 9, September 1994, pp. 33-38

[14] C. Chang, D. J. Nagel, and S. Muftic, "Assessment of Energy Consumption in Wireless Sensor Networks: A Case Study for Security Algorithms," In 4th IEEE International Conference on Mobile Ad Hoc and Sensor Systems (IEEE MASS 2007), Pisa, Italy, October 2007, pp. 1-6.

[15] S. Mostafa, H. El Zouka, and M. Abouelnasr, "Hybrid Encryption Secure Routing Protocols for Wireless Sensor Networks," Proceeding of the ISCA, First International Conference on Sensor Networks and Applications (SNA), San Francisco, November 2009, pp. 109-114

[16] H. El Zouka, "Challenges in Securing Wireless Sensor Networks," in Proceedings of SENSORCOMM 2013, The Seventh International Conference on Sensor Technologies and Applications, Barcelona, Spain, August 2013, pp. 145-150.

[17] The OMNeT++ Simulator. http://www.omnetpp.org [Retrieved on January, 2014].

ALGORITHMS FOR PACKET ROUTING IN SWITCHING NETWORKS WITH RECONFIGURATION OVERHEAD

Timotheos Aslanidis and Marios-Evangelos Kogias

National Technical University of Athens, Athens, Greece

ABSTRACT

Given a set of messages to be transmitted in packages from a set of sending stations to a set of receiving stations, we are required to schedule the packages so as to achieve the minimum possible time from the moment the 1st transmission initiates to the concluding of the last. Preempting packets in order to reroute message remains, as part of some other packet to be transmitted at a later time would be a great means to achieve our goal, if not for the fact that each preemption will come with a reconfiguration cost that will delay our entire effort. The problem has been extensively studied in the past and various algorithms have been proposed to handle many variations of the problem. In this paper we propose an improved algorithm that we call the Split-Graph Algorithm (SGA). To establish its efficiency we compare it, to two of the algorithms developed in the past. These two are the best presented in bibliography so far, one in terms of approximation ratio and one in terms of experimental results.

KEYWORDS

switching networks, packet, routing, scheduling, reconfiguration cost, approximation

1. INTRODUCTION

As the need for communication and dissemination of information increases in modern technology based societies, so does the need for faster and more efficient networks and routing of packages between stations. In this context switching networks and the transmission of large packets of data between them, has become an issue of major importance and as information loads continue to increase rapidly, it is expected that the need for well scheduled data transfers to decrease time and resource usage, will keep on being an often addressed subject for many scientists and engineers.

In this manuscript and in the context of message scheduling and transmitting through switching networks we consider the Preemptive Bipartite Scheduling problem (encountered as PBS in bibliography). Given a set of n transmitting stations and a set of m receiving stations, we are required to send across messages, each initiated from a specific transmitter, to reach an also prespecified receiver station. The duration of each message is also predetermined for all messages to be transmitted. Restrictions in the systems considered, are that no transmitter may transmit data towards more than one receiver at any time, nor may a receiver receive, more than one message at a time. Messages are sent in packages and to enhance transmission speed we are allowed to preempt any package and continue transmission of any part of that package at a later time. Unfortunately, since the system has to reconfigure after each preemption, any interruption of

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 191–196, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4416

packets transmission will come with a time cost. Information on the data that was not sent has to be saved and a new setup has to be initiated for the next packet to start transmitting. Consequently prior to sending any of the packages there will be a setup overhead. We consider this overhead to be constant for all transmission initiations. In this paper we aim to minimize the duration of the aforementioned process.

2. PREVIOUS RESULTS

PBS is known to be NP-Complete [9] and proved to be 4/3-ε inapproximable for any ε>0, unless P=NP in [4].

As PBS algorithms can be implemented in various applications, many polynomial time algorithms have been designed to produce solutions close to the optimal, found in [1], [9], [12], [8]. The best guaranteed approximation ratio so far is $2-\frac{1}{d+1}$, where d is the reconfiguration cost, and is found in [1]. Experiments on the performance of various algorithms are presented in [4] and [5].

The problem can be solved in polynomial time if we consider a zero setup cost or if we only want to minimize the number of switchings [9]. Another variation of the problem for which the optimal schedule can be calculated in polynomial time is presented in [1].

For the purposes of this paper we consider 2 algorithms published in the past:

- A-PBS(d+1), found in [1], which so far is the one with the lowest approximation ratio, and
- A1, found in [5], which according to past experiments yields the best experimental results.

To compute each packet to be transmitted, A-PBS(d+1) rounds up the time of each message to the closest multiple of d+1 and calculates the packet reducing the workload of each station to the minimum multiple of d+1.

On the other hand A1 computes an arbitrary packet with a maximum number of messages and decides how to preempt by calculating a lower bound to the remaining transmissions cost to be the minimum possible.

3. GRAPH REPRESENTATION AND NOTATIONS

Our data representation will be through a bipartite graph G(V,U,E). V will be the set of transmitters, U the set of receivers while E, the set of edges, will correspond to the messages that have to be transmitted from V to U. A weight (or cost) c(v,u), will be assigned to each of the edges e=(v,u), to denote the time required to transmit the message from node v to node u. Edge weights are considered to be non-negative integers.

Furthermore the following notation will be used: $\Delta=\Delta(G)=\max\{\max_{v\in V}(\deg(v)), \max_{u\in U}(\deg(u))\}$, that is, Δ will denote the maximum number of messages that need to be either sent or received from or to any of the stations.

The function t: $V \cup U \rightarrow Z^*_+$ will denote the total workload of any station, namely $t(v)=\sum_{u \in U} c(v,u)$ for any $v \in V$ or $t(u)=\sum_{v \in V} c(v,u)$ for any $u \in U$.

$W=W(G)=\max\{ \max_{v \in V}(t(v)), \max_{u \in U}(t(u)) \}$, that is W will denote the maximum transmission time of the messages either sent to or received from any station.

$d \in Z^*_+$ will denote the overhead to start the next transmission.

4. A HEURISTIC WITH IMPROVED RESULTS

For the purposes of our algorithm the initial graph is split in two parts. G_M comprises edges of weight at least d and G_m contains all edges of weight less than d. Our main concern for G_M is to keep reducing the workload for each of the stations, achieving the minimum transmission time possible, whereas in the case of G_m, where edge weights are small in comparison to d, we aim in minimizing the number of switchings. The intuition in designing this algorithm is that for messages of long duration, priority on how to schedule has the message duration rather than the number of preemptions, whilst for messages of shortest duration prioritized is the minimization of the number of preemptions. In particular:

The Split-Graph Algorithm (SGA)

Step 1: Split the initial graph G(V,U,E) in two bipartite graphs $G_m(V_m,U_m,E_m)$ and $G_M(V_M,U_M,E_M)$, where $V_m=V_M=V$, $U_m=U_M=U$ and E_m contains all edges of weight less than d, E_M contains all edges of weight d or more. Clearly in this initiation step $E=E_m \cup E_M$ and $E_m \cap E_M=\varnothing$.
Step 2: Use subroutine1 to find a maximal matching M, in G_M.
Step 3: Use subroutine 2 to calculate the weight of the matching to be removed. Remove the corresponding parts of the edges.
Step 4: Add edges to M, from E_m to maximize |M| and remove them from E_m.
Step 5: Move edges of weight less than d, from the graph induced by step 3 to E_m.
Step 6: Repeat steps 2 to 5 until all edges initially in E_M have been completely removed.
Step 7: Use subroutine 3 to calculate Δ_m maximum matchings in G_m, where Δ_m is the degree of G_m.
Step 8: Schedule the messages as calculated in steps 2, 3 and 7.

Subroutine 1:

Step 1: $M=\varnothing$ (Initialization of the matching).
Step 2: For each node $w \in V_M \cup U_M$ calculate t(w).
Step 3: Sort all nodes $w \in V_M \cup U_M$ in decreasing order of t(w). Let L be the induced list of nodes.
Step 4: Let w_0 be the 1st node to appear in L. Run sequential search in L to find the 1st neighbor of w_0 appearing in L. Denote that neighbor by w_1.
Step 5: $M \leftarrow M \cup \{w_0, w_1\}$.
Step 6: Remove w_0, w_1 from L.
Step 7: Repeat steps 2, to 6 until M becomes maximal.

Subroutine 2:

Step 1: For each edge e=(v,u) of the matching M, with corresponding weight c(e) calculate what the value W(G´) of the induced graph G´(V´,U´,E´) would be if all edge weights in the matching were to be reduced by c(e). In this case edges of cost less than c(e) would be completely removed. Set

$$r(e)=\begin{cases} c(e) & ,\text{if} \quad W(G')=W(G)-c(e) \\ 0, & \text{otherwise} \end{cases}$$

Step 2: Calculate $R=\max\{r(e) \mid e\in M\}$

Step 3: For each edge in M set its new weight $c(e)=\begin{cases} c(e)-R, & \text{if} \quad c(e)>R \\ 0, & \text{otherwise} \end{cases}$.

Subroutine 3:

Step 1: Add nodes and edges to make G_m a regular graph of degree Δ_m. New edges will be of zero weight. In a regular graph, all nodes will be of the same degree.
Step 2: Calculate a maximum matching M_m in G_m and remove all edges of M_m from G_m. G_m's degree will now be reduced by 1.
Step 3: Repeat step 2 until $G_m=\varnothing$.

5. EXPERIMENTAL RESULTS AND COMPARISON

Figure 1 represents each algorithms' performance in terms of approximation ratio. 1000 test cases have been ran for a 15 transmitters-15 receivers system for values of setup cost varying from 1 to 200 and message durations varying from 1 to 50. SGA performs significantly better than both A1 and A-PBS(d+1) and as the overhead increases it shows an increasingly improved performance. It is important to mention that in practice, as information loads exponentially increase, the number of stations and communication tasks increases and so does the setup cost. That is in fact the most encountered situation nowadays.

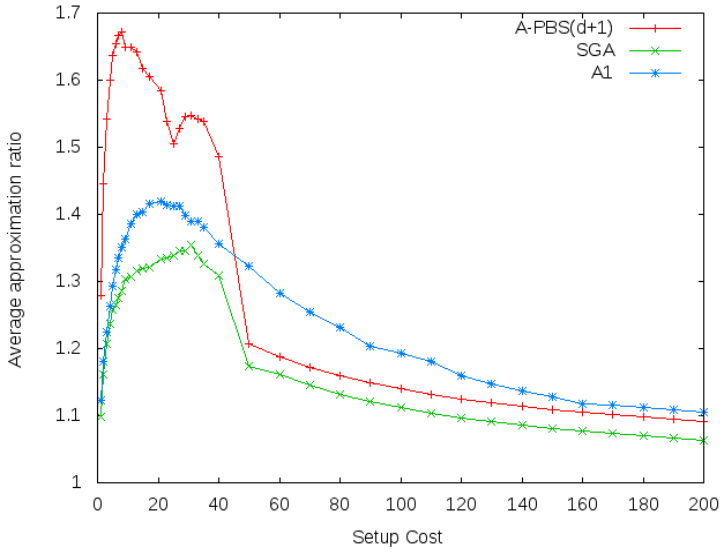


Figure 1. Average approximation ratio comparison

Figure 2 presents the worst performance that each algorithm had depending on the setup cost, in terms of approximation ratio again. SGA is found to be once again a lot more efficient. Furthermore, SGA in most cases has a worst case really close to the average showing that its performance does not fluctuate much, making it an all cases a reliable tool for this type of scheduling.

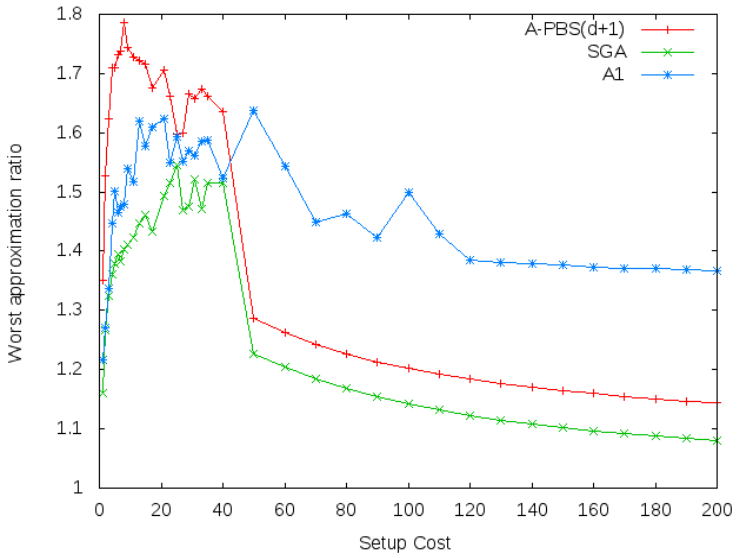


Figure 2. Worst approximation ratio comparison

In terms of running time, SGA also appears to be a lot more efficient, as experiments have shown that A1 and A-PBS(d+1) are by up to 500% slower. This is mainly because SGA is based on an entire different approach on how to schedule the messages, using subroutines that in general are a lot faster than those used in previous papers.

6. DISCUSSION AND FUTURE WORK

Our newly presented algorithm (SGA), has proven to produce much more efficient routings both in terms of hardware usage as well as time span. Therefore, we believe that the idea of splitting the initial graph in parts can be further researched and depending on the magnitude of the edges as well as the setup cost, the Split-Graph Algorithm's efficiency can be further improved. An approximation ratio for SGA could be established to be less than 2. Exploiting to a greater extend algorithms that provide optimal solutions for special instances of the problem might also yield interesting new approximation algorithms. Finally, lifting limitations of the problem or introducing new ones could help in developing new classes of graphs for which polynomial algorithms might provide an optimal schedule. Such algorithms could be the tools to designing new and improved approximation algorithms.

REFERENCES

[1] F. Afrati, T. Aslanidis, E. Bampis, I. Milis, Scheduling in Switching Networks with Set-up Delays. Journal of Combinatorial Optimization, vol. 9, issue 1, p.49-57, Feb 2005.

[2] G. Bongiovanni, D. Coppersmith and C. K.Wong, An optimal time slot assignment for an SS/TDMA system with variable number of transponders, IEEE Trans. Commun. vol. 29, p. 721-726, 1981.

[3] J. Cohen, E. Jeannot, N. Padoy and F. Wagner, Messages Scheduling for Parallel Data Redistribution between Clusters, IEEE Transactions on Parallel and Distributed Systems, vol. 17, Number 10, p. 1163, 2006.
[4] J. Cohen, E. Jeannot, N. Padoy, Parallel Data Redistribution Over a Backbone, Technical Report RR-4725, INRIA-Lorraine, February 2003.
[5] P. Crescenzi, X. Deng, C. H. Papadimitriou, On approximating a scheduling problem, Journal of Combinatorial Optimization, vol. 5, p. 287-297, 2001.
[6] R. L. Cruz and S. Al-Harthi, Packet Scheduling with Switch Configuration Delays, in Proc. 39th Annu. Allerton Conf. Commun., Contr., Comput., 2001.
[7] S. Even, A. Itai, A. Shamir, On the complexity of timetable and multicommodity flow problems SIAM J. Comput., vol. 5, p. 691-703, 1976.
[8] I. S. Gopal, G. Bongiovanni, M. A. Bonucelli, D. T. Tang, C. K. Wong, An optimal switching algorithm for multibeam satellite systems with variable bandwidth beams, IEEE Trans. Commun. vol. 30, p. 2475-2481, Nov. 1982.
[9] I. S. Gopal, C. K. Wong Minimizing the number of switchings in an SS/TDMA system IEEE Trans. Commun. vol. 33, p. 497-501, 1985.
[10] T. Inukai, An efficient SS/TDMA time slot assignment algorithm IEEE Trans. Commun. vol 27, p. 1449-1455, Oct. 1979.
[11] E. Jeannot and F. Wagner, Two fast and efficient message scheduling algorithms for data redistribution over a backbone, 18th International Parallel and Distributed Processing Symposium, 2004.
[12] A. Kesselman and K. Kogan, Nonpreemptive Scheduling of Optical Switches, IEEE Transactions in Communications, vol. 55, number 6, p. 1212, 2007.
[13] K. S. Natarajan and S. B. Calo, Time slot assignment in an SS/TDMA system with minimum switchings IBM Res. Rep. 1981.
[14] B. Towles and W. J. Dally, Guaranteed Scheduling of Switches with Configuration Overhead, in Proc. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM '02. pp. 342-351, June 2002.

Authors

Timotheos Aslanidis was born in Athens, Greece in 1974. He received his Mathematics degree from the University of Athens in 1997 and a master's degree in computer science in 2001. He is currently a phd candidate at the National and Technical University of Athens in the department of electrical and computer engineering. His research interests comprise but are not limited to computer theory, number theory, network algorithms and data mining algorithms.

Marios-Evangelos Kogias was born in Athens in 1991. Currently, he is an undergraduate student at the National Technical University of Athens pursuing a diploma in Electrical and Computer Engineering. He is a passionate coder and interested mainly in operating systems, computer networks and algorithms.

COIFLET-BASED FUZZY-CLASSIFIER FOR DEFECT DETECTION IN INDUSTRIAL LNG/LPG TANKS

Uvais Qidwai and Mohamed Shakir

KINDI Research Center, Department of Computer Science & Engineering,
Qatar University, Doha, Qatar
uqidwai@qu.edu.qa, shakir@qu.edu.qa

ABSTRACT

This paper describes a classification method for raw sensor data using a Fuzzy Inference System to detect the defects in large LNG tanks. The data is obtained from a Magnetic Flux Leakage (MFL) sensing system which is usually used in the industry to located defects in metallic surfaces, such as tank floors. A robotic inspection system has been developed in conjunction with the presented work which performs the same inspection tasks at much lower temperatures than human operators would thus reducing the shutdown time significantly which is typically of the order of 15-20 million Dollars per day. The main challenge was to come up with an algorithm that can map the human heuristics used by the MFL inspectors in field to locate the defects into an automated system and yet keep the algorithm simple enough to be deployed in near real-time applications. Unlike the human operation of the MFL equipment, the proposed technique is not very sensitive to the sensor distance from the test surface and the calibration requirements are also very minimal which are usually a big impediment in speedy inspections of the floor by human operator. The use of wavelet decomposition with Coiflet waves has been utilized here for deconvolving the essential features of the signal before calculating the classification features. This wavelet was selected to its canny resemblance with the actual MFL signals that makes these wavelets very natural basis function for decomposition..

KEYWORDS

Nondestructive Testing (NDT), Fuzzy Inference System, Defect Detection, Classification, filters, Wavelet based Deconvolution, Coiflet Transform.

1. INTRODUCTION

Liquefied Natural-Petroleum Gas (LNG), Liquefied Petroleum Gas (LPG) and Field/Plant Condensate are some of the main products of oil and gas sector in the Middle East and phenomenal revenue is associated with their trade. Both LNG and LPG are usually stored at very low temperatures which pose a major issue in maintaining the tank since the tank must be shut down for a period of 10-14 days before the tanks can be safe for human ingress. The maintenance cost and time for such an operation can be significantly reduced if the shutdown period can be decreased. One of the ways that this can be achieved is by deploying a specially designed robot, capable of working at very low temperatures and in hazardous conditions, thus, replacing the need for risking human personnel during initial stages of inspection and at the same time reducing the shutdown cost. Such robotic system is capable of inspecting the floor which sits on a solid

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 197–207, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4417

concrete flooring and is virtually impossible to be accessed from outside [1, 2]. The presented technique is in conjunction with the design of one such robot and its application in real industrial environment.

Robotics is an extension of human abilities and can create a safety environment for work where needed. Robots can, through adaptive design, perform tele-operation and act as smart tools, handle an assortment of tasks while communicating and interacting with workers located on-site or remotely. This supports sustainable production at the cutting-edge of efficiency and creativity. Remotely Operated Vehicles (ROVs) have become an important tool in many industrial and environmental applications. On one hand, such robots reduce the hazardous exposures of the human personnel by operating in such dangerous or harsh environments, and at the same time they increase the level of accuracy in measurements and inspections by not being limited with the human weaknesses and subjectiveness. The developed robotic system is exactly based around this idea. It can operate at very low temperature where human exposure will be too dangerous as well as too expensive. At the same time, the intelligent Fuzzy classification algorithm accumulates the experiences from several experts from the industry and is not limited by human weaknesses. By using this intelligent system in the robotic platform, the tank floor can be inspected remotely as well as autonomously with more accurate detection.

Large chemical and petroleum product storage tanks are mostly made of steel plates that are welded together to form the structure. The material and the welds are inspected for manufacturing defects when constructed but must also be periodically inspected throughout their service life for signs of damage. The carbon steel is prone to attack by corrosion and in some circumstances cracks can form over time. NDT personnel use visual, X-ray, ultrasonic and other inspection methods to search for flaws and service-induced damage.

2. EXPERIMENTAL SETUP

The MFL inspection system used in this work is quite standard in the industry for this application. The detection the defects in the floor of the tank are sensed by used MFL sensor using the theory of Hall Effect sensing. Magnetic Flux Leakage (MFL) is a phenomenon in which the magnetic flux escapes out of the normal magnetic path due to disruptions in the path or broken paths. After ferromagnetic materials are magnetized, the leakage magnetic field can be created above its surface because of its internal defects, the method to measure leakage magnetic field by sensors is called magnetic flux leakage testing (MFL) [3-10]. Figure 1 shows the robot used in this work.

(a)

(b)

Figure 1. The robotic system used in this work for data collection. (a) The inspection robot on a test plate, and (b) the MFL sensor fixed under the robot's body.

The MFL Data is an exciting source for a variety of signal processing techniques and researchers have been quite busy in applying all sorts of conventional and non-conventional approaches in identifying the defects from the data. Techniques like Wavelet Transforms, Genetic Algorithms, Particle Swarm Optimization, and other well-known filtering techniques have been applied on such data [11-15]. Each technique has its own merit in terms of the application. However, all of these do have a large computational time requirements. In this work, the main focus was on developing a technique that could be utilized in near-real-time scenarios for active inspections using the robotic platform.

3. ALGORITHM

As the robot moves across the testing area, the MFL data is collected, standard signal processing techniques are applied on it. Primarily this corresponds to normalization and low-pass filtering for noise removal. Based on the shapes of the MFL data waveforms, following strategy was adapted:

1. Coiflet wavelet [16, 17] was selected based on its similarity in shape with the MFL waveform. Coiflets are discrete wavelets designed by Ingrid Daubechies, at the request of Ronald Coifman, to have scaling functions with vanishing moments.

2. Comparing with the actual MFL waveform, this wavelet must be smoothed and skewed. This was done using convolution with a Gaussian wavelet. The one dimensional Gaussian wavelet has an impulse response given by:

$$g(x) = \sqrt{\frac{a}{\pi}}\, e^{-ax^2} \tag{1}$$

Where x is the distance from the origin in the horizontal axis and σ is the standard deviation of the Gaussian distribution. A typical coiflet generated as part of this work is shown in Figure 2 along with the actual Coiflet mother wavelet.

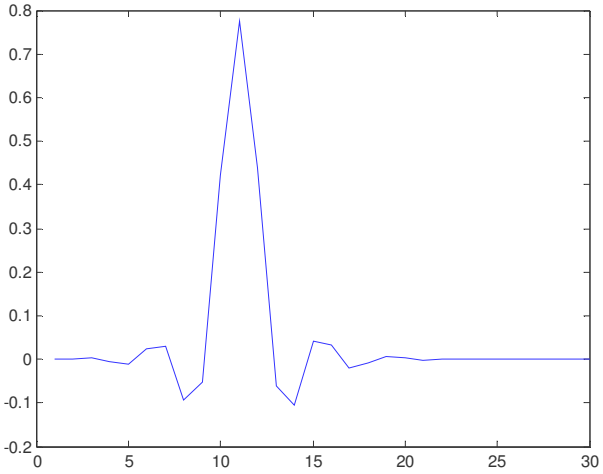


Figure 2. Coiflet mother wavelet used in this work for decomposing the MFL signals into low and high frequency components.

By using Coiflet-based deconvolution , removing all the noise and base signal artifact from the raw MFL signal is achived. This is done by:

- Decomposing the raw signal into its Approximation and Detail components [CA and CD].
- Then the CA is decomposed again into the respective components [CAA and CDD].
- These two decompositions have been found to be sufficient in removing all the unwanted sensor noise, metallic approximations, and quantization artifacts from the raw signal.
- Then, an inverse transform is applied in order to regenerate the original signal by combining the two approximation components (CAA and CA) with corresponding detail components reset to zero.

This results in a cleaner signal which can be better analyzed for the subsequent classification step. Figure 3 shows the result of deconvolution on the raw MFL signal.

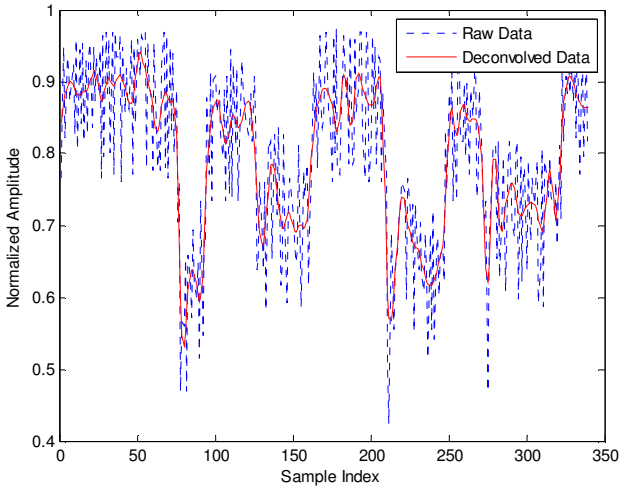


Figure 3. Deconvlution result on the MFL signal. The dotted blue line shows the Raw MFL data and the solid red line shows the deconvolved signal.

Figure 4 represents the complete Algorithm, including the subsequent steps of Feature Calculation and Fuzzy Inference System, which are further elaborated in the following sections.

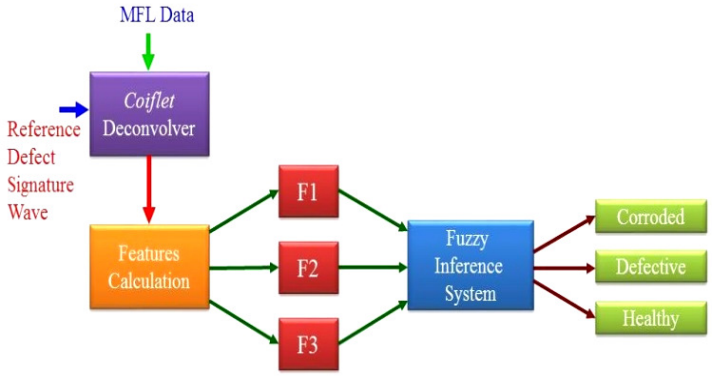


Figure 4. Main signal processing algorithm applied on the raw MFL data.

The algorithm is applied on a finite length of the stored or a window of on-the-fly data from the robot's sensors and correlates it to the points on the floor map of the tank. For each window of samples taken for analysis, the algorithm results into the classified areas of healthy, corroded, and defective states. For the possible defective states, recommendation is made for the maintenance staff to look more carefully into the possibility of having a further classification. The first 100 Coiflet coefficients, the feature set is also stored in a simple database in order to be used later on for more detailed classification.

4. THE FUZZY INFERENCE SYSTEM

The actual understanding of the waveforms is obtained by the program through a set of measurable values, called Features. In order to select a meaningful feature set, human heuristics were heavily utilized in this work. By quantifying the human perception according to the general understanding of the information present in the measurements by the NDT expert, a better classification is obtained that will assist the technicians to make a better decision concerning maintenance, material selection and replacement. Fuzzy Inference System is probably the most suitable tool for this purpose. While it is a wonderful tool to map the imprecise or heuristically oriented information into quantitative data, it also maps the human decision making capabilities using a subjective Rule base. This may raise some issues related to the accuracy since it is based on less precise information and the usefulness of the system on the whole may be questioned. Actually, when limitations of NDT personnel in the field are considered, these are considered subjectively. This means that the accuracy of the NDT process is limited by expertise of a single individual. This is supplemented further with the classical limitations associated with human operators such as tiredness, physical and mental uprightness, and proper understanding of the problem. In order to automate this process, the heuristical perceptions of the human operators must be quantified into some measureable parameters (or features). The Features selected for the work presented in this work are, therefore, based on translating the human understanding into the mathematical quantities. In the real practice, the NDT expert will use the MFL scanning system and drag it over the test surface. An audio signal (beep) is the target outcome from these experiments. The degree amplitude and durations of these beeps are directly related to the depth and width of the defect under the sensor head, respectively. Once a signal is heard, the place is marked with spray-paint for more careful follow-up inspections. These characteristics of the testing procedure are captured in this work by using two sources:

1. Human heuristics from NDT experience, and
2. Deconvolved signals' statistical features.

In order to understand the mechanism of how a human expert would classify the metal health using MFL data, several experts were consulted with including experts from the top manufacturer of the MFL sensing equipment used in the paper, Silver Wing UK. Based on the collective experiences of all of these experts, the codes defined in the MFL based inspections, and the way the MFL sensors are used by the human operators, following heuristical features can be deduced as indicators of various health parameters of the sample:

- Big change in the MFL signal representing a deep flaw.
- Long voids (amplitude changes) in the MFL signal represents wider surface defects.
- Grass like; low-amplitude signals represent healthy metal.
- Corrosion is defined by several modes of the grass signals that fall between the deep defects and the healthy metal.

Based on the human heuristical parameters, following features were devised to be calculated from the raw and deconvolved signals:
F1. AWR (Amplitude to Width ratio) of the big change in the MFL signal (deconvolved signal only)
F2. GE (Grass Energy) is the sum squared values of the grass signal below 10% of the maximum value in the MFL signal (Raw signal only)

F3. BE (Bush Energy) is the sum squared values of the grass-like signal between 10% and 30% of the maximum value in the MFL signal (Deconvolved signal only).

Using these features, following Membership functions were made to map the actual values from the data into the Fuzzy space (fuzzification of data). The overall Fuzzy classification system is shown in Figure 5.

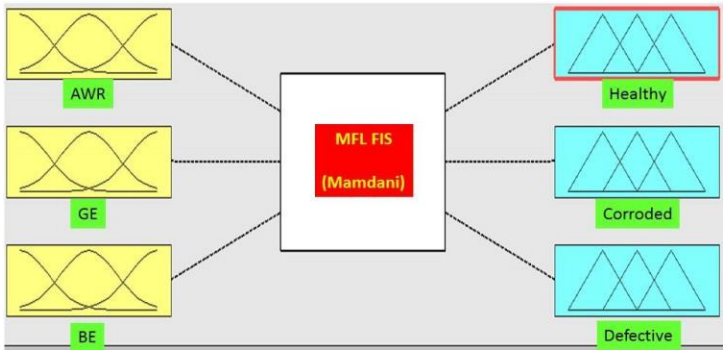


Figure 5. Fuzzy Inference System for classifying the input features (AWR, GE, and BE) into one of the output classes (Healthy, Corroded, Defective).

The input features are calculated on the fly as the data is obtained and are then fuzzified by the input membership functions as shown in Figure 6.

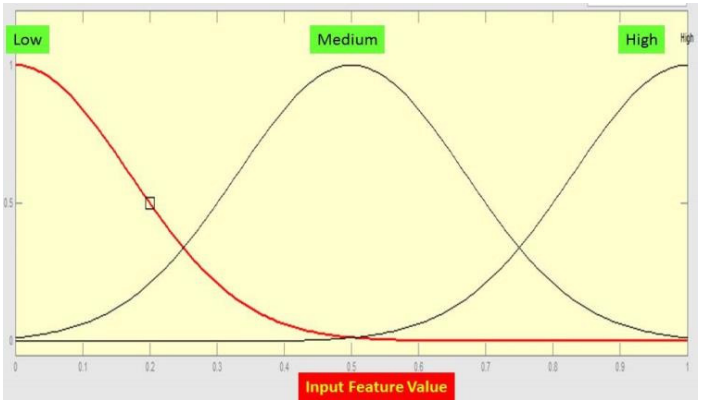


Figure 6. Input feature fuzzification memberships. All the three inputs are fuzzified in the same degrees.

The output classes are also kept quite similar to each other and are shown in Figure 7. All the three classes get some score after each classification is done. The largest value is considered as the ultimate class.

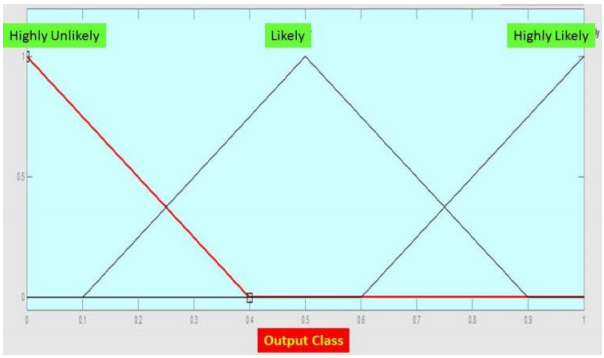


Figure 7. Output classes. All the three outputs have same membership degrees.

The actual mapping from the input to output is done through a set of rules:

1. If (AWR is High) and (GE is Low) and (BE is Low) then (Healthy is Highly-Unlikely)(Corroded is Highly-Unlikely)(Defective is Highly-Likely)
2. If (AWR is Low) and (GE is Low) and (BE is High) then (Healthy is Highly-Unlikely)(Corroded is Highly-Likely)(Defective is Highly-Unlikely)
3. If (AWR is Low) and (GE is Low) and (BE is Low) then (Healthy is Highly-Likely)(Corroded is Highly-Unlikely)(Defective is Highly-Unlikely)
4. If (AWR is Low) and (GE is Medium) and (BE is Low) then (Healthy is Highly-Likely)(Corroded is Highly-Unlikely)(Defective is Highly-Unlikely)
5. If (AWR is Low) and (GE is Medium) and (BE is Medium) then (Healthy is Highly-Likely)(Corroded is Highly-Unlikely)(Defective is Highly-Unlikely)

These rules are translated into decision surfaces which can be used on the fly to classify the health of the sample under study. Figure 8 shows these classes.

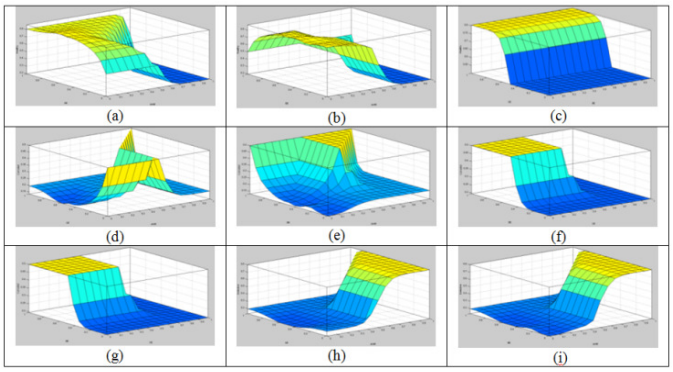


Figure 8. Various decision surfaces. (a) through (c) for Healthy class between the parameters AWR-GE, AWR-BE, and GE-BE. (d) through (f) for Corroded class between the parameters AWR-GE, AWR-BE, and GE-BE. (g) through (i) for Defective class between the parameters AWR-GE, AWR-BE, and GE-BE.

5. TESTING AND RESULTS

The algorithm was tested by moving the robot on a 3.5 m diameter iron plat with a horizontal weld defect in the upper half of the plate. Figure 9 shows the MFL Sensor data received without the location constraints.

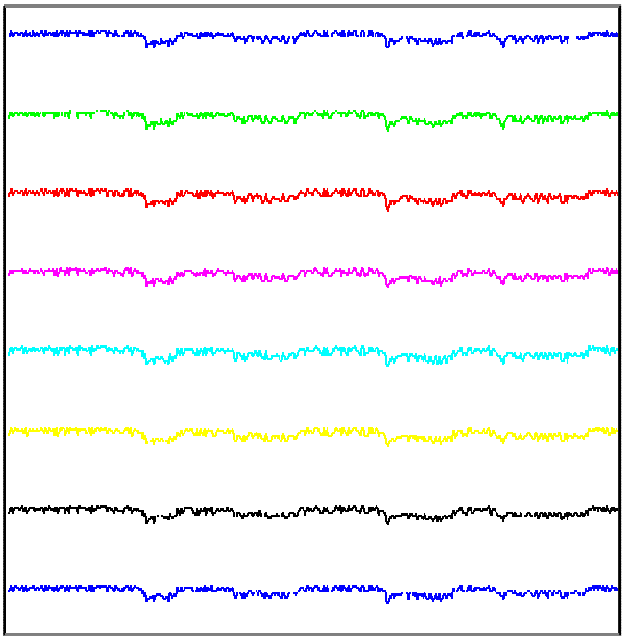


Figure 9. Sensor data from 8 out of 16 sensors in the MFL sensor head.

Once the calculations are done on the data based on the algorithm discussed above, the location is also decided using the laser ranging mechanism of the robot. Figure 10 shows the resulting floor map for this test run. The area shown in the figure is 30x15 cm2 and clearly shows the three classes in three color levels. The dark brown color represents the weld, the bluish tint represents the corrosion and the orange color represents the healthy metal. Rest of the sample area is kept at dark blue color to indicate the area that was not scanned.

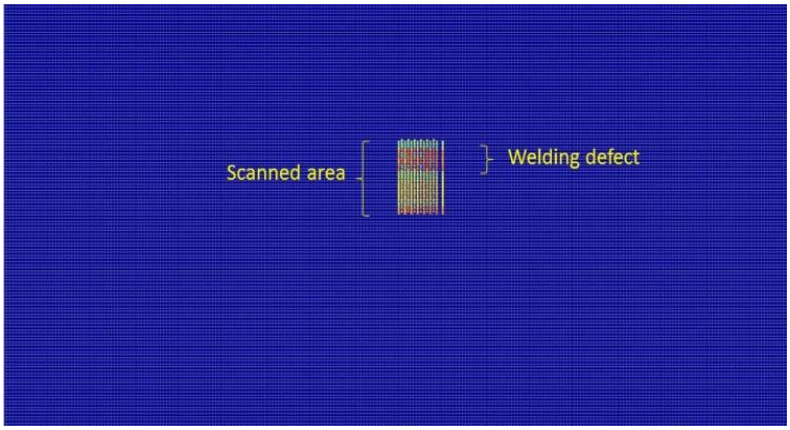


Figure 10. Floor map with Coiflet deconvolution and Fuzzy classification applied.

6. CONCLUSION

The Robotic Inspection System developed in this work has shown some very promising outcomes in the testing phase and is expected to improve in terms of its performance and intelligence as more tests will be performed on more realistic defects. The basic sensing mechanism can be used to test the MFL sensing capabilities on a number of test samples and more realistic database can be developed. More realistic samples will provide more insight into how to detect a specific type of defect.

MFL based inspection techniques are quite common in the oil and gas industry. The human dependence in the process limits the accuracy due to the way the MFL data is perceived by the operator of the inspection equipment. The Robotic inspection system presented here is quite superior to these limitations but requires a lot of signal processing to evaluate the signals obtained from the sensors on the robot. The presented technique has shown very promising results with real MFL data and is being tested on industrial targets to build more confidence and accuracy in the inspection procedure.

While several algorithms were tested, the presented algorithms in this report, by far, are the best we have for the purpose of deconvolution and classification. The Fuzzy classifier will enhance with time as newer samples will provide newer features which will make classification more reliable. We really hope that this robotic system will flourish further into an industrial platform that can add to the value of the prime industry in Qatar, the oil & gas sector.

The raw data waveforms are clearly indicators of the complexity of the data type involved and the need for a better proximity decision related to the inspection. As can be seen from the presented results, the trends are preserved quite well but the unwanted artifacts are significantly reduced, if not removed all together using the deconvolution method.

ACKNOWLEDGEMENTS

This publication was made possible by a grant from Qatar National Research Fund under its National Priority Research Program, for project NPRP 08-276-2-091. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of Qatar National Research Fund.

REFERENCES

[1] The art of Tank gauging, Technical notes, ENRAF Netherlands, 2006.
[2] http://www.ndted.org/AboutNDT/SelectedApplications/TankInspection/TankInspection.htm.
[3] Dury, J.C., and Marino, A., "A Comparison of the Magnetic Flux Leakage and Ultrasonic Methods in the detection and measurement of corrosion pitting in ferrous plate and pipe"www.ndt.net/article/wcndt00/papers/idn701/idn701.htm
[4] Song, Q., "Interacting effects of clustering defects on MFL signals using FEA",
[5] BINDT Insight, Vol 55, No. 10, 558-560, October 2013.
[6] The MFL Compendium: Articles on Magnetic Flux Leakage, ISBN: 978-1-57117-210-5, 2010, American Society of Nondestructive Testing (ASNT).
[7] Kim, D. K., et. Al., "Development of MFL system for in-pipe robot for unpiggable natural gas pipelines", 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), 2013, pp. 51-54.
[8] Kandoori, M. R., Shirani, F., Araabi, B.N. , Ahmadabadi, M.N. , and Bassiri, M.M., "Defect detection and width estimation in natural gas pipelines using MFL signals", 9th IEEE Asian Control Conference (ASCC), 2013, pp. 1-6.
[9] Kikuchi, H., Sato, K., Shimizu, I., Kamada, Y., and Kobayashi, S., "Feasibility Study of Application of MFL to Monitoring of Wall Thinning Under Reinforcing Plates in Nuclear Power Plants", IEEE Transactions on Magnetics, Vol. 47, No. 10, 2011, pp. 3963-3966.
[10] Sun, L. Y., Li, Y. B., Sun, L. B., Li, L. G., "Comparison of Magnetic Flux Leakage (MFL) and Acoustic Emission (AE) techniques in corrosion inspection for pressure pipelines", 31st Chinese Control Conference (CCC), 2012, pp. 5375-5378.
[11] Han, W., and Que, P., "A modified wavelet transform domain adaptive FIR filtering algorithm for removing the SPN contained in the MFL data", IEEE International Conference on Industrial Technology, 2005, pp. 152-157.
[12] Han, W., Xu, J., and Tian, G., "MFL inspection defect reconstruction based on self-learning PSO", Far East Forum on Nondestructive Evaluation/Testing: New Technology & Application (FENDT), 2013, pp. 50-54.
[13] Han, W., and Que, P., "Defect reconstruction from MFL signals using improved genetic local search algorithm", IEEE International Conference on Industrial Technology, ICIT 2005, pp. 1438-1443.
[14] Lee, J. Y., Afzal, M., Udpa, L., Udpa, S., and Massopust, P., "Hierarchical rule based classification of MFL signals obtained from natural gas pipeline inspection", Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000, Vol. 5, pp. 71-78.
[15] Melikhov, Y., Lee, S.J., Jiles, D.C., Lopez, R., and Brasche, L., "Analytical approach for fast computation of magnetic flux leakage due to surface defects", Digests of the IEEE International Magnetics Conference, 2005, (INTERMAG Asia 2005), pp. 1165-1166.
[16] Wei, D., Coiflet type Wavelets: Theory, Design, and Applications, Ph.D. dissertation, University of Texas at Austin, 1998.
[17] MATLAB's link for the Coiflet Wavelet: http://www.mathworks.com/help/wavelet/ug/wavelet-families-additional-discussion.html.

Authors

Uvais Qidwai received his Ph.D(EE). from the University of Massachusetts–Dartmouth USA in 2001, MS(EE) in 1997 from KFUPM Saudi Arabia, and BS(EE) in 1994 from NED University of Engineering & Technology, Karachi, Pakistan. He taught in the Electrical Engineering and Computer Science Department, Tulane University, in New Orleans as Assistant Professor, and was in-charge of the Robotics lab as well as a research member of Missile Defence Centre, during June 2001 to June 2005. He joined the Computer Science and Engineering Department, Qatar University, in FALL of 2005 where he is currently working as Associate Professor. Dr. Qidwai's present research interests include Robotics, Signal and Image Processing, Expert Systems design for Industrial Applications, and Intelligent Algorithms for medical informatics. He has participated in several government- and industry-funded projects in the United States, Saudi Arabia, Qatar, UAE, Singapore, Malaysia, and Pakistan, and has published over 100 papers in reputable journals and conference proceedings.

Mohamed Shakir received the B.Tech degree in Applied Electronics and Instrumentation from MES College of Engineering, Calicut University, India, in 2005 and the M.S. degree in Electrical Engineering from Washington International University, USA in 2011. He is currently pursuing the Ph.D. degree in Electrical Engineering at Univeristi Teknologi Petronas, Malaysia. From 2011 to 2014, he was a Research Assistant with the Computer Engineering Department, Qatar University, Qatar. His research interest is in Robotics.

INTENTIONAL BLANK

EMBED SYSTEM FOR ROBOTIC ARM WITH 3 DEGREE OF FREEDOM CONTROLLER USING COMPUTATIONAL VISION ON REAL-TIME

Luiz Cortinhas[1], Patrick Monteiro[1], Amir Zahlan[1], Gabriel Vianna[1] and Marcio Moscoso[2]

[1]Instituto de Ensinos Superiores da Amazonia, Belém, Pará
iesam@iesam.com.br
[2]Instituto Federal do Pará, Belém, Pará
ifpa@ifpa.edu.br

ABSTRACT

This Paper deals with robotic arm embed controller system, with distributed system based on protocol communication between one server supporting multiple points and mobile applications trough sockets .The proposed system utilizes hand with glove gesture in three-dimensional recognition using fuzzy implementation to set x,y,z coordinates. This approach present all implementation over: two raspberry PI arm based computer running client program, x64 PC running server program, and one robot arm controlled by ATmega328p based board.

KEYWORDS

Robot, Arm, Embed, System, Sockets, MultiPoint, Hand, Recognition, Webcam, Raspberry, High Pass Filter.

1. INTRODUCTION

Gesture Recognition is an important, yet difficult task on arm-based embeds systems [1].
It is a versatile and intuitive way to approach the more natural form to human-machine interaction just need glove with five light emissor diode at fingertips, tracking the movements over filtered images sequences captured by webcam and recovering data to 3D structure on real time. At same time arm-based system is a low-cost and newest trend to approach mobile world, these reasons make Raspberry PI the best choice. This presented is designed to JavaSE version 7 solution because this code language is perfect to minimize code creations for different architectures[4] and compiled programs owing to virtual machine developed to all specified devices architecture: X86, X64 and ARM, all these running a generic Linux SO.

2. THE GLOVE

To avoid complex implementations of Image Processing and obtain major precision was built a simple glove fig.1 on fingertips located LEDs with a wavelength 850 nm (infra-red) and 5mm of diameter. The glove has a 3,3v coin battery that energizes the five leds through different five 220 ohm resistors. This implementation is very cheap and capable to make more smooth detected movements.

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 209–216, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4418

Figure 1. The Glove

3. THE HARDWARE

Except by the glove described above, there are two hardwares: Minimum Computer Requirements and Robot Arms.

A. Minimum Computer Requirements

Considering the propose to support cheap computers and embed systems, the minimum requirements is based on raspberry PI requirements: processor at 700 Mhz, 384 Mb de RAM, 128 Mb for Video Memory and at least 1Mbps connection.

B. Robotic Arm

The robotic arm on fig.8 is named Mark II build in aluminium on thickness of 1mm in four body parts:

- Base: Built on box on (16 cm x 16 cm x 11 cm) format content power supply AC/DC converter with 12v at 5A and 5v at 1A outputs, 12 v is used to energize: four servo motor model MG995 – TowerPro each projected to load 15kg/cm on maximum and one, the 5 v Supply is used to energize atmega328P responsible to convert input data in degrees to coordinates movements, supporting two servo motors the base weight proximate at 3kg.
- Lower Arm: starting from base joint measuring 40 cm in length and 4cm in width, supports 1 servo motor on upper end, this element weight 180 g measured.
- Higher Arm: Starting from Lower Arm joint above described measuring 20 cm in length and 2,7cm in width, support one servo motor on upper end, weight 110 g measured.
- Claw Grip: For purposes especial to minimal weight and force to grip function, the claw use a little servo motor model TowerPro MicroServo 9 g – SG90 measuring 9 g and concentrate force 1kg/cm in maximum load, generate a higher grip force, measured all weight 20 g.

4. COMPUTATIONAL VISION

This section is focused on motion analysis of hand with glove presented revealed that gesture can be characterized based on four different aspects: shape, motion, position and orientation. All gesture recognition approaches try to approach the problem by concentrating on one or more of above aspects according in [1], on this case will go use position and orientation. To solve problem

of low hardware specification adopted: this method of recognition is easiest using just one hand whit five dots to interpretation.

The motion analysis start at first captured sequence images on second step the image processing is applied on source the High Pass algorithm developed show only higher value pixels, works scaling between black to white on black pixel equal 0 and white pixel equal 255, this method is developed to dynamical approach to solve problem of different conditions of illumination every searching minor and major pixel and set interval minimum and maximum range on scale, the High Pass algorithm filter for pixel value between 240 and 255 finally the image is converted to binary format fig.2, on next step is applied Hough Transform[2] searching for circular shape [5]-[6] obtain only position of five most approximate diameter shapes calculate center of white shapes shown each as blue circles to calculate the hand's center fig.4 generating coordinate to be displayed on red dot with yellow trace, it all is shown in fig.3,if this is a first capture this hand's center coordinates are saved and adjusted to zero on next image all processes above will happen and newest "hand's center" coordinates (x,y and z) and last "hand's center" calculate the difference on each resulting on pixel moved parsing to Fuzzy Logic, except Z, that will define the real movement to be sent to robotics arm on angular form ranged (0º-180º).

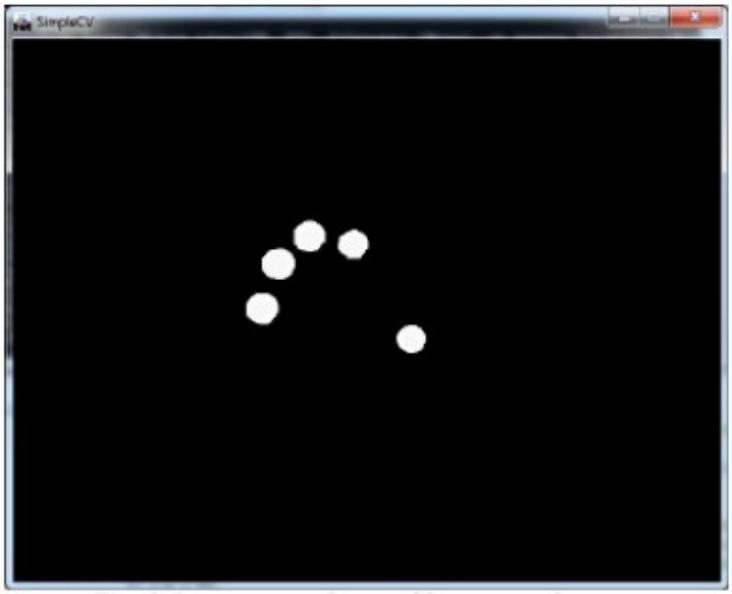


Figure 2. Result of High Pass Filter Algorithm applied on all pixels

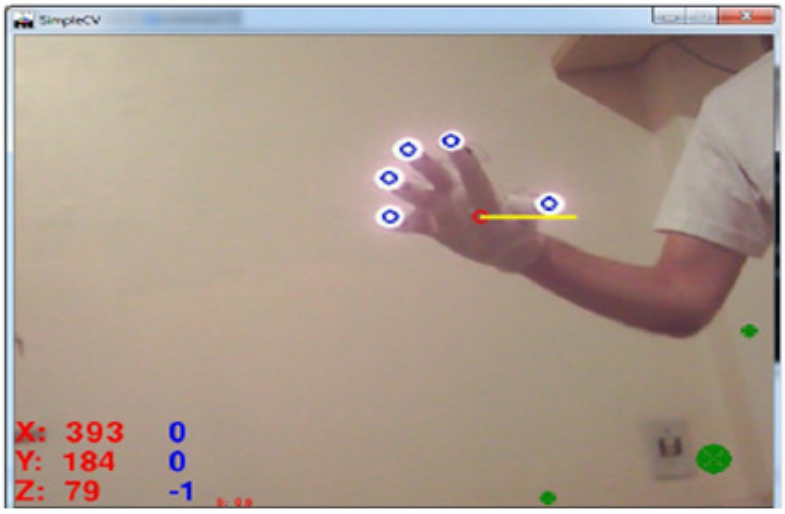


Figure 3. Result of Analysis and Image Processing

$$xCenter = \frac{\sqrt{((max(x) - min(x))^2)}}{2},$$
$$yCenter = \frac{\sqrt{((max(y) - min(y))^2)}}{2},$$
$$zCenter = (\frac{\sum x - xCenter}{5}) - (\frac{\sum y - yCenter}{5})$$

Figure 4. Formulas to calculate hand's center point.

5. FUZZY LOGIC

The fuzzy is a multivalued logic responsible to transform not exact information on acceptable output supporting stochastic outputs based on rules, this way approximate human to machine [3] this start with the receipt of the amount of displacement and the direction of movement by the center point of the hand applying the rules demonstrated on fig.5 for movement on X hand axis and fig.6 for Y hand axis, each of these rules on axis x on fig.5 represent the number of pixels displaced taking as reference last position of hand center point, and y axis on fig.5 represents the degree of relevance calculed. For these entrances fuzzy generate one output based on result of each entrance transformed by respective rule converted to moviment signal on degrees, see output rule on fig.7 to arm robotics axis: X and Y, verify this important output necessary to eliminate the problem of inaccurate movements of the human hand according in [8] the Device's Human Resolution (DHR) more precisely on Static Stability is the lower precision indicator especially in free space and without support. The fuzzy resultant output is sent to robotic arm on format: "(x moviment,y moviment,z moviment)" each on their degree information result

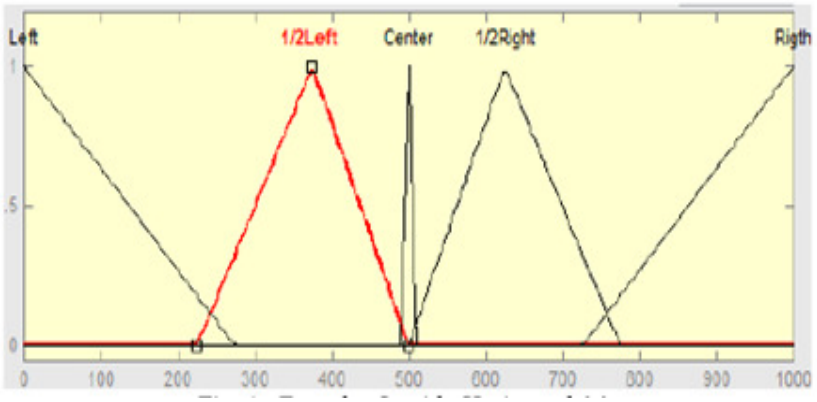


Figure 5. Fuzzy rule Input for x axis

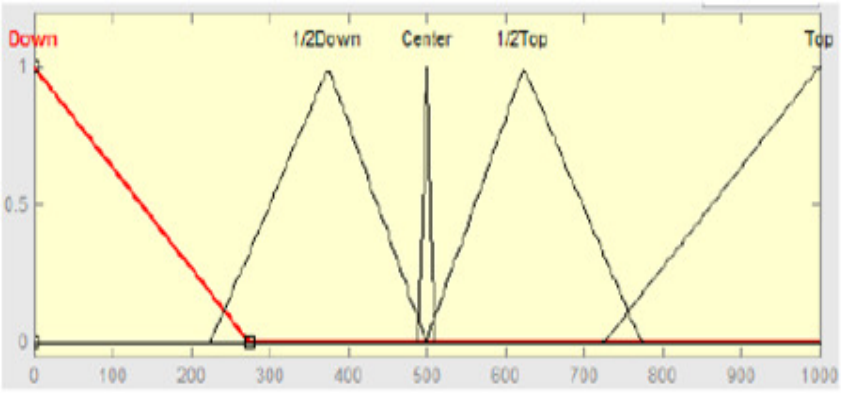


Figure 6. Fuzzy rule Input for y axis.

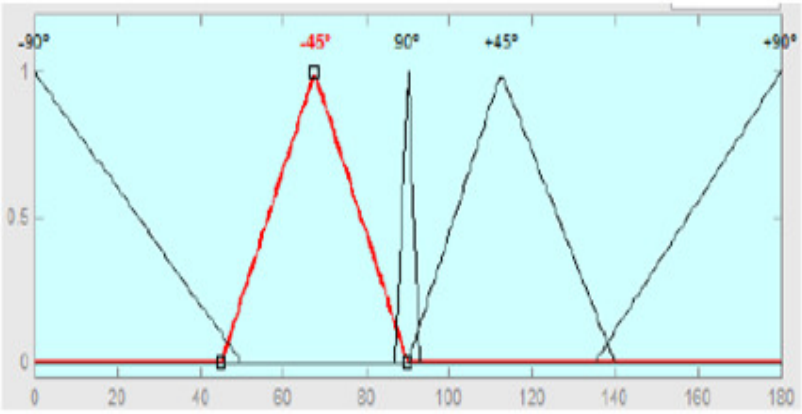


Figure 7. Fuzzy Logic rule Output on degrees

Figure 8. Robotics Arm Mark II (Stable Version)

6. CONCLUSIONS

This presented work all objectives are presented and built a possible embed system using processing image on acceptable scale interacting with hardware robotics arm developed to symbolize an industrial robotic arm and show that the relationship man - machine can become more human and get better results from movement with little investment, besides the mobility applied allowing almost any machine can be included in the system. The major process included on this process is presented in fig.9 below, this project reaches its goals and demonstrates how ARM architecture is being inserted how low-cost machine.

The robotic arm demonstrate stability and security same in poor conditions of hand stability, because the Fuzzy Model works on stability controller in 2D (X,Y) the Z axis is don't demonstrate instability because this is calculated using means minimizing little movements[7].

On tests Raspberry Pi with client running only receives Camera Images only applying Movement Analysis, Fuzzy and Send Data to hardware by way of Serial UART (at 9600bps), this approach reached goal with successful movements fast and accurate.

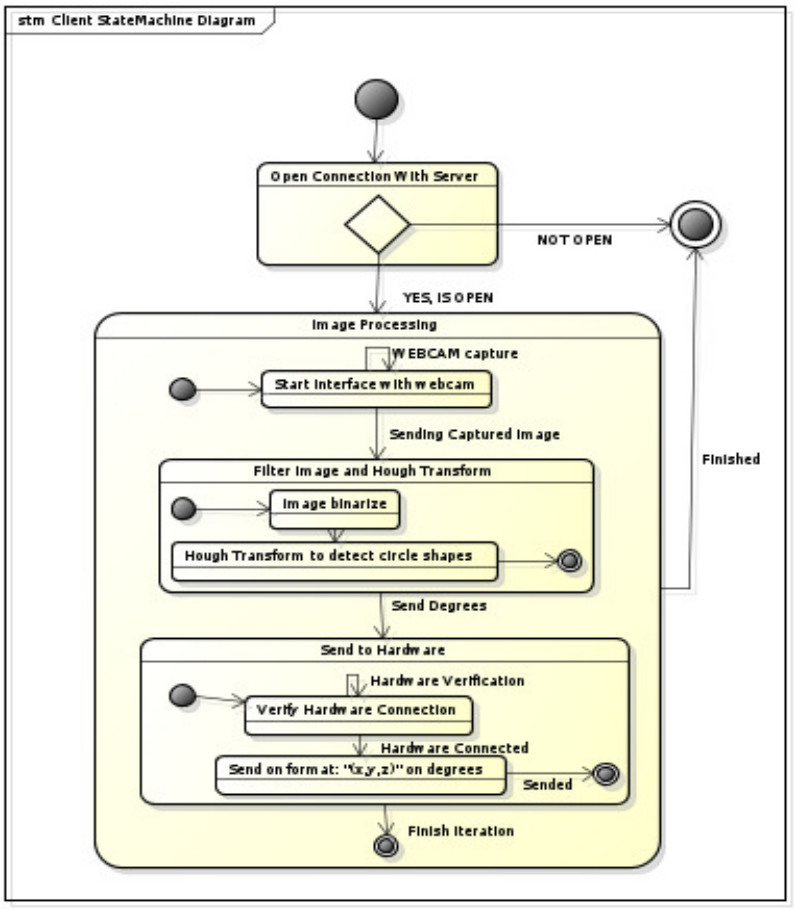


Figure 9. State Machine Diagram Overall Client Program

ACKNOWLEDGEMENTS

Thanks everyone include my literary agent!

REFERENCES

[1] J. Laura Dipietro, Angelo M. Sabatini, Paolo Dario. 2008. "A Survey of Glove-Based Systems and Their Applications". IEEE TRANSACRIONS ON SYSTEMS, MAN, AND CYBERNETICS - PART C: APPLICATIONS AND REVIEWS. VOL. 38 NO. 4 JULY 2008: 461-482.

[2] V. F. Leavers "Survey: Which Hough transform?",CVGIP: Image Understanding, vol. 58, no. 2, pp.250 -264 1993

[3] Tang Wen, Hu Jianbin and Chen Zhong, "Research On a Fuzzy Logic-based Subjective Trust ManaC. Canudas De Wit, B. Siciliano and G. Bastin, Theory of robot control, London: Spring-Verlag, 1996. gement Model", 2005,pp.1654-1659.

[4] W. Binder and J. Hulaas. Java bytecode transformations for efficient, portable CPU accounting. In First Workshop on Bytecode Semantics, Verification, Analysis and Transformation (BYTECODE 2005), volume 141 of ENTCS, pages 53-73, Edinburgh, Scotland, April 2005.

[5] T. J. Atherton and D. J. Kerbyson, "Size invariant circle detection," Image and Vision Computing, vol. 17, no. 11, pp. 795-803, 1999.

[6] D. Kerbyson and T. Atherton, "Circle detection using hough transform filters," in Image Processing and its Applications, 1995., Fifth International Conference on, 1995, pp. 370-374.

[7] C. Canudas De Wit, B. Siciliano and G. Bastin, Theory of robot control, London: Spring-Verlag, 1996.
[8] François Bérard,Guangyu Wang and Jeremy R. Cooperstock,On the Limits of the Human Motor Control Precision: the Search for a Device's Human Resolution, University of Grenoble – INP France.

Authors

Luiz Cortinhas Ferreira Neto
Short Biography
Graduating on Computer Engineering – IESAM since 2010
Member of LINC – Laboratory of Inteligence Computational / UFPA

Gabriel Vianna Soares Rocha
Graduating on Computer Engineering – IESAM since 2010
Member of LINC – Laboratory of Inteligence Computational

Amir Samer Zahlan
Graduating on Computer Engineering – IESAM since 2010

Patrick Monteiro
Graduating on Computer Engineering – IESAM since 2010

Marcio Nazareno de Araujo Moscoso
Master on Eletrical Engineer on UFPB, since 200
Graduated on Eletrical Engineer on UFPA, since 1998

ROBUST COLOUR IMAGE WATERMARKING SCHEME BASED ON FEATURE POINTS AND IMAGE NORMALIZATION IN DCT DOMAIN

Ibrahim Alsonosi Nasir

Department of Electronic and Computer Engineering,
Sebha University, Sebha, Libya
Ibrn103@yahoo.com

ABSTRACT

Geometric attacks can desynchronize the location of the watermark and hence cause incorrect watermark detection. This paper presents a robust colour image watermarking scheme based on visually significant feature points and image normalization technique. The feature points are used as synchronization marks between watermark embedding and detection. The watermark is embedded into the non overlapped normalized circular regions in the luminance component or the blue component of a color image. The embedding of the watermark is carried out by modifying the DCT coefficients values in selected blocks. The original unmarked image is not required for watermark extraction Experimental results show that the proposed scheme successfully makes the watermark perceptually invisible as well as robust to common signal processing and geometric attacks.

KEYWORDS

Watermarking, DCT domain, image normalization, feature points.

1. INTRODUCTION

Visual data such as image and video can be easily copied, altered and distributed over the internet without any loss in quality. Therefore, the protection of the ownership of multimedia data has become a very challenging issue. Watermarking is the process of embedding hidden information called a watermark into the digital media, such that the watermark is imperceptible, robust and difficult to remove or alter [1]. In recent years, attacks against image watermarking systems have become more complicated [2]. In general, these attacks can be classified into two broad categories: signal processing and geometric attacks. While signal processing attacks reduce the watermark energy, geometric attacks can induce synchronization errors between the encoder and the decoder of the watermark. As a result, the decoder is no longer able to detect the watermark. Robustness to geometric attacks is still challenging in the image watermarking community. Most existing watermarking algorithms focus mainly on embedding watermarks into grey-scale images in spatial or frequency domain. The extension to colour images is usually accomplished by marking the image luminance component or by processing each colour channel separately [3, 4]. Kutter et al. [5] suggested embedding the watermark in the blue channel, because the human eye is less sensitive to changes in this band. Lian et al. [6] suggested that the watermark should be

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 217–227, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4419

embedded into the green component. This is because the loss of energy of the blue and red components is higher than the green component when the watermarked image is attacked by JPEG compression. However, the human eye is more sensitive to changes in the green band. Barni et al. [7] introduced another colour image watermarking method based on the cross-correlation of RGB channels. However, it has relative high computing costs and low processing speed since the full-frame DCT is used for three colour channels. Kutter at al. [8] investigated watermarking of luminance and blue-channels using a perceptual model, which takes into account the sensitivity and the masking behaviour of the HVS. Nasir et al. [9] suggested embedding the watermark into luminance component of the color image and use image normalization technique to reduce the effect of synchronization errors. However, this method cannot resist cropping attacks. Several grey-scale image watermarking methods have been developed to overcome of synchronization errors caused by geometric attacks. These methods can be roughly classified into template-based, invariant transform domain–based, moment-based, histogram-based, and feature extraction-based methods. The template-based watermarking methods are based on embedding a template in addition to the watermark to assist the watermark synchronization in the detection process. This may be achieved using a structured template embedded in the DFT domain to estimate transformation factor to resynchronize the image [10-12]. In [13-14], watermarks are embedded in affine-invariant domains such as the Fourier-Mellin transform or log-polar domain to achieve robustness against affine transforms. In [15, 16], the watermark is embedded in an affine-invariant domain by using generalized random transform and Zernike moment, respectively. However, watermarking methods involving invariant domains are difficult to implement due to the log-polar mapping [17]. Based on the fact, that the histogram is independent of the position of the pixels, the authors in [18, 19] presented histogram-based watermarking approaches. However, these approaches suffer from robustness limitations under histogram enhancement and equalization attacks. To overcome the issue of synchronization, feature points are used as reference points for both watermark embedding and detection. In [20], Mexican hat wavelet method is used to extract feature points. In [21, 22], the Harris detector is used to extract the feature points. However, Mexican hat wavelet or Harris detector are sensitive to image modification. In [23], the end-stopped wavelets feature detector is used to extract feature points. To resist image geometric attacks and to eliminate synchronization errors between the watermark embedding and the detection, this paper presents a robust color image watermarking scheme, which combines the advantages of feature points extraction and image normalization and investigates watermarking of luminance and blue-channels by modifying the DCT coefficients values in selected blocks.

The rest of this paper is structured as follows. Section 2 describes the proposed watermarking scheme and section 3 presents experimental results. Conclusions are drawn in section 4.

2. THE PROPOSED WATERMARKING SCHEME

The block diagram shown in Fig. 1 provides an overview of the proposed watermarking scheme. First, the Luminance (Y) component in YIQ (Luminance, Hue, and Saturation) or the blue component in RGB (Red, Green, and Blue) color models is obtained from the original image for embedding the watermark; second, wavelet based feature detector is utilized to extract steady feature points from the B or Y component of the original image; then the circular regions are normalized by image normalization process. To enhance the robustness, the watermark bits are embedded into all circular images. Finally, the watermarked image is reconstructed. During the detection process, we claim the existence of the watermark if one copy of the embedded watermark is correctly detected in one embedding circular region.

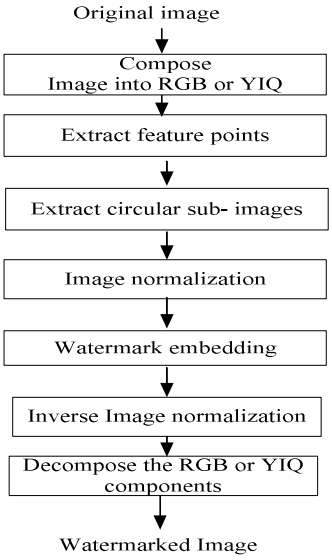


Figure 1. Watermark embedding scheme

2.1. Feature Extraction Detector

Monga et al, [24] proposed an iterative feature detector to extract significant geometry preserving feature points. The detector determine the feature points by computing a wavelet transform based on an end-stopped wavelets obtained by applying the first-derivative of Gaussian (FDoG) operator to the Morlet wavelet. Monga et al, [24] evaluate the performance of this detector with three commonly used detectors that are Harris corner detector, the maximally stable extremaly region (MSER) detector and Hessian Affine and conclude the feature detector based on end-stopped wavelets is the most robust. Therefore, in the present scheme, this detector has been adopted to extract the feature points. The feature detection process can be divided into the following steps:

(i) For each image location, the wavelet transform is computed as given in reference [24]
(ii) The significant features points are identified by looking for local maxima of the magnitude of the wavelet coefficients in a pre-selected neighbourhood.
(iii) A threshold is applied to eliminate spurious local maxima in featureless regions of the image.

To determine the regions for each determined feature point for embedding the watermark, a search is carried out within a circular neighbouring region whose radius is set to be R. If the detector response at the centre of the region achieves local maximum, the feature point is selected. Otherwise, it is discarded. To obtain non-overlapping regions, the most stable feature points are first selected. Then, any feature points whose corresponding region overlaps with the selected feature points are excluded.

2.2. Image Normalization

Synchronization errors between the embedding and the detection of the watermark may be introduced by geometric attacks such as rotation, shearing and translation and although the watermark is still present in the watermarked image, it can no longer be detected. Image

normalization techniques developed for pattern recognition [25] can be used to overcome this problem as suggested in [26]. In the proposed scheme, an image normalization technique is performed on extracted circular images.

2.3. Watermarking Embedding Process

We assume that the watermark of length N_w is a binary and denoted by $W = \{w_i , i = 1,.....,N_w , w_i \in (0,1)\}$, which is a key-based PN sequence. The private key is shared with the detector to make decision whether a given watermark is present or not. The watermark is embedded into DCT coefficients of M×M block. The proposed watermark embedding process is described as follows.

- The Luminance (Y) component or the blue component of the original image is selected to embed the watermark.
- The feature detector based on end-stopped wavelets is applied to the image to determine the feature points as described in section 2. These feature points are used for the reference centers of circular subimages for watermark embedding and detection.
- For each determined feature points, search within a circular neighbouring region, whose radius is set to be R to extract non overlapped circular images for embedding the watermark.
- The normalization process is applied to each extracted circular image.
- The normalized circular image can not be transferred directly into frequency domain. Therefore zero-padding operation could be performed on the normalized circular image. In the proposed method, a subimage is extracted from the normalized circular image because zero-padding operation will introduce error after applying the inverse DCT transform method.
- The discrete cosine transform (DCT) is applied to a selected 8×8 blocks of the sub-images.
- To achieve robustness against common signal processing attacks, the low frequency coefficient of the selected DCT block is used to embed the watermark. In the proposed scheme, the DC coefficients are kept unmodified and the first four AC coefficients in zigzag order are selected to embed the watermark. In order to reduce the visual degradation on the watermarked image, the number of AC coefficients for embedding a watermark bit in each selected DCT blocks is set to 4. This is because using more coefficients for embedding a watermark bit will cause more distortions of the watermarked image.

  The watermark embedding process is carried out by quantizing the absolute value of the second largest DCT coefficients in the selected DCT blocks to the nearest values M_0 or M_1 as shown in Figure 2 by dashed vertical lines. The watermark embedding algorithm can be described as follows:
  Firstly, the length of embedding intervals for bit 0 and bit 1 is defined as given in (1)

$$L_0 = L_1 = \frac{|AC_1|}{L} \qquad (1)$$

  where L_0 and L_1 are the length of embedding intervals for bit 0 and bit 1, respectively. L represents the number of embedding intervals and $|AC_1|$ is the absolute value of the largest DCT coefficients selected from the first four AC coefficients in zigzag order. Secondly, to embed watermark bit 0 or bit 1, the absolute value of the second largest DCT

coefficient $|AC_2|$ is quantized to the nearest M_0 to embed '0' or to the nearest M_1 to embed '1' as follows:

$$AC_2^* = \begin{cases} M_0 & \text{if } w = 0 \\ M_1 & \text{otherwise} \end{cases} \quad (2)$$

Where AC_2^* is the watermarked coefficient, M_0 and M_1 are the middle values of the quantization level '0' and level '1', respectively. The $|AC_3|$ and $|AC_4|$ coefficients are only quantized to the value AC_2^* if they are greater than the watermarked coefficient AC_2^*. The signs of the watermarked coefficients are determined as given in (3)

$$AC_2^* = \begin{cases} -AC_2^* & \text{if } AC_2 < 0 \\ AC_2^* & \text{otherwise} \end{cases} \quad (3)$$

The watermarked subimages are obtained by applying the IDCT transform. Finally, the inverse normalized process is applied to each watermarked circular image and the watermarked image is reconstructed.

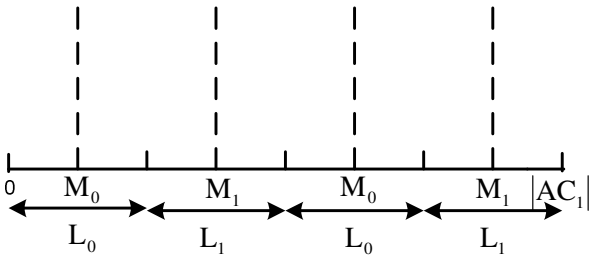


Figure 2 Quantization process for watermarking
Watermark Extraction process

2.4. Watermark Extraction Process

The proposed watermark extraction process is performed without use of the original image. In the extraction process, the first six steps are similar to that used in the watermark embedding process. The watermark bit is extracted as given in (4)

$$W_i^* = \begin{cases} 0 & \text{if } |AC_2^*| \in L_0 \\ 1 & \text{if } |AC_2^*| \in L_1 \end{cases} \quad (4)$$

where $|AC_2^*|$ is the absolute values of the second largest DCT coefficients of the first four AC coefficients in the selected DCT blocks of size 8×8. The AC coefficients are selected in zigzag order, W_i^* is the extracted watermark bit and L_0 and L_1 are the embedding intervals for bits 0 and 1, respectively. The extracted watermark is then compared with the original embedded watermark to decide a success detect. The normalized (NC) given in [22] is used to evaluate the similarities between the original and the extracted watermarks.

Main-body text is to written in fully (left and right) justified 11 pt. Times New Roman font with a 6pt. (paragraph) line spacing following the last line of each paragraph, but a 12pt. (paragraph) line spacing following the last paragraph. Do not indent paragraphs.

3. EXPERIMENTAL RESULTS

The watermark imperceptibility and robustness are evaluated by using 10 different colour images of size 512×512 including Lena, Peppers, Baboon, Lake, etc. In the experiments, a pseudorandom sequence of size 16-bits is used as a watermark and the radius of each circular image is 71.

3.1 Watermark Imperceptibility

The distortion of an image depends on the watermark length, the number of quantization levels for embedding the watermark, the number of extracted sub-images and the number of AC coefficients for embedding a watermark bit in each 8×8 DCT block. The larger the number of AC coefficients used for embedding, the more significant the distortion. Also the more the quantization levels (L) for embedding watermark bits, the smaller the distortion. In the other words, increasing the number of quantization levels leads to a small change in the AC coefficients. Hence there is a trade off between robustness and imperceptivity. The Peak Signal to Noise Ratio (PSNR) is adopted to evaluate the perceptual distortion of the proposed scheme. The PSNR values for ten watermarked images are between 39 and 53 db. These values are all grater than 30 db, which is the empirically tested threshold value for the image without any perceivable degradation [21]. Taking Lena, Peppers, as an example, the watermarked images and circular feature regions from Y component are shown in Figure 3.

Table I. PSNR between watermarked image and the original image (db)

Image	RGB model	YIQ model
Lena	52.63	44.93
Peppers	51.69	43.08
Baboon	44.57	39.19

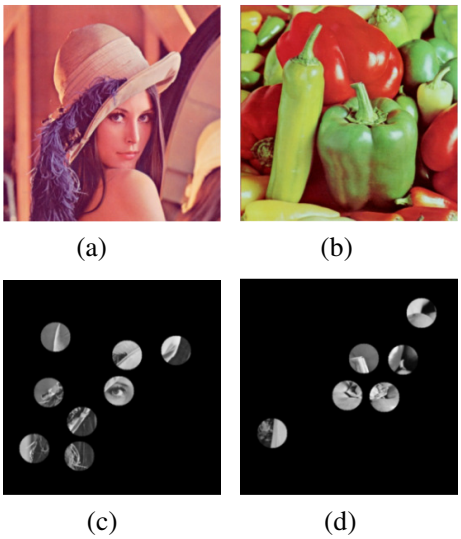
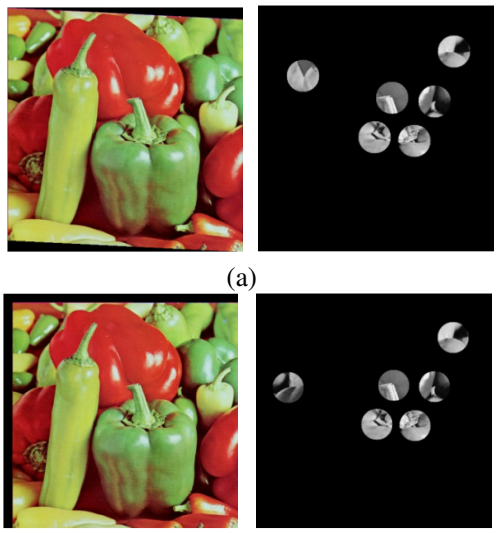


Figure 3. (a) and (b) Watermarked Lena and Peppers images; (c) and (d) circular feature regions

3.2 Watermark Robustness

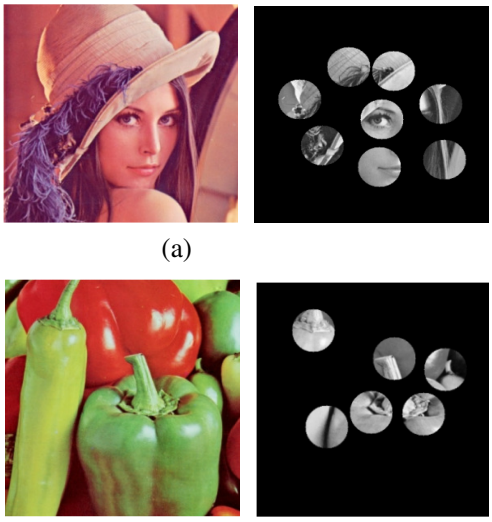
To evaluate the robustness of the proposed watermarking scheme, various common signal processing and geometric attacks were applied to the watermarked images. These attacks include JPEG-lossy compression, median filtering, low-pass filtering, Gaussian filtering, and cropping, shearing, rotation, row and column removal attacks. As an example, results for some geometric attacks are shown in Figure 4 and Figure 5.



(a)

(b)

Figure 4. Results of geometric attacks; (a) shearing x-0%, y-5%, (b) Translation-x-20 and y-20.



(a)

(b)

(c)

Figure 5. Results of cropping attacks; (a) cropping Lena 25 % off , (b) cropping Peppers 25 % off, (c) centred cropping 10%.

Table 2 and Table 3 summarize experimental results by applying common signal processing attacks on Lena, Peppers and Baboon images watermarked in RGB model and YIQ model. For JPEG lossy compression attacks, the quality factor varied from 30% (high compression) to 100%. As can be seen from Table 1, the embedded watermark in Y component can be correctly extracted even under JPEG compression with a quality factor as low as 30%. As shown, better performance is achieved when the watermark is embedded in Y component than the B component. This robustness is achieved by embedding the watermark into the low frequency coefficients of the DCT, which are less affected by JPEG compression attacks. For filtering attacks, the watermarked images were subjected to median, low pass and Gaussian filtering. As shown in Table 2, more robustness to these attacks is achieved when the watermark is embedded in Y component in YIQ model.

Table 3 shows that better robustness is achieved when the watermarked in embedded in Y component. As can be seen, the watermark can be correctly detected when the watermarked image attacked by geometric attacks. The proposed scheme overcomes the synchronization problem caused by geometric attacks by combining the advantages of using image normalizing and geometrically invariant feature points. Robustness against cropping attacks is achieved because the normalization process is applied into sub-images rather than the entire image.

The performance of the proposed watermarking scheme in YIQ model is better than RGB model due to the following factors:

(i) Loss of energy of the blue component is high when the watermarked image is attacked by JPEG compression or low pass- filtering attacks [6].

(ii) The blue component of an image in RGB model is more sensitivity to rotation because such a geometric transformation is based on interpolation which is a low-pass local filtering that affects the high frequency content. Consequently, the watermark is less robust to the rotation attack when is embedded in this component.

The more distortion on the blue component, the less accurate normalization angle can be used at extraction.

Table 2 Watermark detection results for signal processing attacks (detection rates)

Attacks	RGB Model			YIQ Model		
	Lena	Peppers	Baboon	Lena	Peppers	Baboon
Jpeg 100%	2/8	1/6	5/11	4/8	4/6	3/11
Jpeg 80%	1/8	1/6	2/11	4/8	4/6	3/11
Jpeg 60%	1/8	1/6	3/11	5/8	4/6	2/11
Jpeg 50%	1/8	0/6	2/11	3/8	4/6	2/11
Jpeg 30%	1/8	1/6	1/11	3/8	4/6	1/11
Median filtering 3×3	1/8	0/6	0/11	3/8	3/6	3/11
Low-pass filtering3×3	0/8	1/6	0/11	3/8	2/6	3/11
Gaussian filtering 3×3	0/8	1/6	1/11	2/8	2/6	4/11

Table 3 the watermark detection results for geometric attacks (detection rates)

Attacks	RGB Model			YIQ Model		
	Lena	Peppers	Baboon	Lena	Peppers	Baboon
Rotation 1°	1/8	0/6	2/11	3/8	3/6	4/11
Rotation 5°	2/8	2/6	2/11	3/8	2/6	2/11
Shearing x-1%,y-1%	1/8	1/6	3/11	3/8	2/6	3/11
Shearing x-0%,y-5%	1/8	1/6	2/11	2/8	3/6	3/11
Shearing x-5%,y-5%	1/8	1/6	2/11	1/8	3/6	3/11
Translation-x-5 y-5	1/8	1/6	3/11	5/8	4/6	2/11
Translation-x-10 y-10	1/8	1/6	3/11	5/8	4/6	2/11
Translation-x-20 y-20	1/8	1/6	2/11	5/8	4/6	2/11
Centered cropping 5%	1/8	3/6	6/11	3/8	3/6	1/11
Centered cropping 10%	1/8	2/6	4/11	1/8	3/6	2/11
Centered cropping 20%	1/8	1/6	1/11	2/8	1/6	1/11
Cropping 25% off	1/8	2/6	1/11	3/8	2/6	1/11
Remove 1Row & 5 Col	1/8	1/6	2/11	4/8	4/6	2/11
Remove 5Row & 17 Col	1/8	1/6	1/11	3/8	3/6	2/11
Remove 17Row & 5 Col	0/8	1/6	1/11	4/8	2/6	2/11

4. CONCLUSIONS

This paper presents a robust colour image watermarking scheme, which is designed to be robust against both signal processing and geometric attacks. In order to eliminate synchronization errors between the watermarks embedding and the detection, perceptually significant feature points and image normalization technique were used. The reference image is not required at the detector. The watermark is embedded into the image luminance in YIQ model or in the blue channel in RGB model by modifying the DCT coefficients values in selected blocks.

Experimental results show that the proposed scheme succeeds in making the watermark perceptually invisible and also robust against various signal processing and geometric attacks. Further research is to improve the results by using the color components of YIQ model to determine the feature points.

REFERENCES

[1] L. M. Marvel, C. G. Boncelet, Jr., and C. T. Retter, "Spread spectrum image steganography," IEEE Transactions on Image Processing, vol. 8 (8), pp. 1075-1083, 1999.

[2] M. Barni, I.J. Cox, T. Kalker, Digital watermarking, 4th International Workshop on Digital Watermarking, Siena, Italy, Lecture Notes in Computer Science 3710, Springer 2005.

[3] K. I. Hashida and S. A., "A method of embedding robust watermarks into digital color images," IEICE Transactions Fundamentals, vol. E81-A(10), pp. 2133-2137, 1998.

[4] N. Nikolaidis and I. Pitas, "Robust image watermarking in spatial domain," Signal Processing, vol. 66(3), pp. 385-403, 1998.

[5] M. Kutter, F. Jordan, and F. Bossen, "Digital watermarking of color images using amplitude modulation," Journal of Electronic Imaging, vol. 7, pp. 326-332, 1998.

[6] L. Lian-Shan, L. Ren-Hou, and G. Qi, "A new watermarking method based on DWT green component of color image," in International Conference on Machine Learning and Cybernetics, vol. 6, 2004, pp. 3949-3954

[7] M. Barni, F. Bartolini, and A. Piva, "Multichannel watermarking of color images," IEEE Transactions on Circuits and Systems for Video Technology, vol. 12(3), pp. 142-156, 2002.

[8] M. Kutter and S. Winkler, "A vision-based masking model for spread-spectrum image watermarking," IEEE Transactions on Image Processing, vol. 11(1), pp. 16-25, 2002.

[9] I. Nasir and A. Abdurrman, "A Robust Color Image Watermarking Scheme Based on Image Normalization, " World Congress on Engineering, pp. 2238-2242, 2013.

[10] S. Pereira and T. Pun, "Robust template matching for affine resistant image watermarks," IEEE Trans. on Image Processing, vol. 9(6), pp. 1123-1129, 2000.

[11] X. Kang, J. Huang, Y. Q. Shi, and Y. Lin, "A DWT-DFT composite watermarking scheme robust to both affine transform and JPEG compression," IEEE Trans. on Circuits and Systems for Video Technology, vol. 13(8), pp. 776-786, 2003.

[12] J. L. Dugelay, S. Roche, C. Rey, and G. Doerr, "Still-image watermarking robust to local geometric distortions," IEEE Trans. on Image Processing, vol. 15(9), pp. 2831-2842, 2006.

[13] C.Y. Lin, M. Wu, J.A. Bloom, I.J. Cox, M. L. Miller, and Y. M. Lui, "Rotation, scale, and translation resilient watermarking for images," IEEE Trans. on Image Processing, vol. 10(5), pp. 767-782, 2001.

[14] D. Zheng, J. Zhao, and A. E. Saddik, "RST-invariant digital correlation," IEEE Trans. on Circuits and Systems for Video Technology, vol. 13(8), pp. 753-765, 2003.

[15] X. Kang, J. Huang, et al., A DWT-DFT composite watermarking scheme robust to both affine transform and JPEG compression, IEEE Trans. on Circuits and Systems for video Technology, vol. 13, no. 8, pp. 776-786, 2003.

[16] D. Simitopoulos, D.E. Koutsonanos, Robust image watermarking based on generalized random transformations, IEEE Trans. On Circuit and Systems for Video Technology, vol. 13, no. 8, pp. 732-745, 2003.

[17] C.Y. Lin, M. Wu, J.A. Bloom, I.J. Cox, M. L. Miller, and Y. M. Lui, "Rotation, scale, and translation resilient watermarking for images," IEEE Trans. on Image Processing, vol. 10(5), pp. 767-782, 2001.

[18] S. Roy and E.C. Chang, Watermarking color histogram, in proc. Int. Conf. Image Process, pp.2191-2194, 2004.

[19] S. Lee, Y. Suh, and Y. Ho, Lossless data hiding based on histogram modification of different images, in Proc. Pacific-Rim Conf. Multimedia, vol3, pp. 340-347, 2004.

[20] S. Xiang, H. Joong, and J. Huang, Invariant Image Watermarking based on statistical features in low-frequency domain, IEEE Trans. on Circuit and Systems for video Technology, vol. 18, no. 6, pp. 777-789, 2008.

[21] X. Qi, J. Qi, A robust content-based digital image watermarking scheme, Signal processing, vol. 87, pp. 1264-1280, 2007.

[22] L. Li, and B. Guo, “Localized image watermarking in spatial domain resistant to geometric attacks", Int. Journal of Elec. And Comm., vol. 63, pp. 123-131, 2009.

[23] I. Nasir, F. Khelifi, J. Jiang, and S. Ipson, "Robust image watermarking via geometrically invariant feature points and image normalisation", Image Processing, IET, vol.6, no.4, pp.354-363, 2012.

[24] V. Monga and B. Evans, Percetual image hashing via feature points: performance evaluation and tradeoffs, IEEE Trans. on Image processing, vol. 15, no. 11, pp. 3453-3466, 2006.

[25] M. Alghoniemy, and A. H. Tewfik, "Geometric invariant in image watermarking", IEEE Trans. on Image Processing, vol. 13, no. 2, pp. 145-153, 2004.

[26] S. C. Pei and C. N. Lin, "Image normalization for pattern recognition", Image Vision. Computing, vol. 13, no. 10, pp. 711-723, 1995.

Authors

Ibrahim Nasir received B.Eng. from Sebha University, Libya in 1994, M.Sc. degree from Heriot-Watt University, Edinburgh, UK in 2005, and PhD from the University of Bradford, UK in 2010. In 2011, He joined the Department of Electronic and Computer Engineering, University of Sebha, Libya. His interest research topics are Image processing, Mobile robotic and Embedded system.

INTENTIONAL BLANK

INDUCTIVE LOGIC PROGRAMMING FOR INDUSTRIAL CONTROL APPLICATIONS

Samiya Bouarroudj[1,2] and Zizette Boufaida[2]

[1]Department of Mathematics and Computer Science,
High School ENSET, Skikda, Algeria
[2]LIRE Laboratory, Constantine 2 University, Constantine, Algeria
bouarroudjsamia@yahoo.fr[1]
zboufaida@gmail.com[2]

ABSTRACT

Advanced Monitoring Systems of the processes constitute a higher level to the systems of control and use specific techniques and methods. An important part of the task of supervision focuses on the detection and the diagnosis of various situations of faults which can affect the process. Methods of fault detection and diagnosis (FDD) are different from the type of knowledge about the process that they require. They can be classified as data-driven, analytical, or knowledge-based approach. A collaborative FDD approach that combines the strengths of various heterogeneous FDD methods is able to maximize diagnostic performance. The new generation of knowledge-based systems or decision support systems needs to tap into knowledge that is both very broad, but specific to a domain, combining learning, structured representations of domain knowledge such as ontologies and reasoning tools. In this paper, we present a decision-aid tool in case of malfunction of high power industrial steam boiler. For this purpose an ontology was developed and considered as a prior conceptual knowledge in Inductive Logic Programming (ILP) for inducing diagnosis rules. The next step of the process concerns the inclusion of rules acquired by induction in the knowledge base as well as their exploitation for reasoning.

KEYWORDS

Inductive Logic Programming (ILP); SHIQ+log; Hybrid Reasoning; Semantic Web Technologies; Control System; Knowledge Management.

1. INTRODUCTION

The increasing demand in quality, safety, availability and cost optimization of industrial processes requires the use of Advanced Supervision Systems. Traditionally, this aspect of the supervision was under the responsibility of the human operators, possibly assisted by a set of sensors and detectors. Nevertheless this apparatus was set up to control the process and not for the detection and the diagnosis of faults.

Consequently, the development of new approaches is essential for a robust monitoring. In this context, many approaches are developed for fault detection and diagnosis (Cf. § 2.3). They include data-driven, analytical, and knowledge based approaches [1]. Methods of faults detection and diagnosis mentioned above have their strengths and weaknesses. Thus the combination of complementary methods is an effective way to achieve high performance.

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 229–241, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4420

Since supervision models are dependent on disparate information drawn from distributed sources, shared semantics based on a common ontology offers a way to develop these linkages. We address this critical need in this paper. Ontologies are a suitable formal representation able to convey this complex knowledge, but their use in learning algorithms is still a research issue.

Inference rules may be crafted by the domain expert as part of the ontology design, or automatically learned by machine learning techniques. We focus on this latter case as a generic component to easily adapt them to new domains. However, as opposed to previous approaches, learning takes place in the ontology language to produce deductive diagnosis rules which is possible with inductive logic programming (ILP).

We propose also a framework allowing the cohabitation of rules acquired by induction and the ontology as well as their exploitation for reasoning and researches.

The use of large steam boilers is quite common in industry due to their advantageous features [2][3]. However, such facilities are subject to several operating failures that could expose the system structural integrity to serious hazard and huge economic and human life losses. Early detection of such faults under operation is of great importance: it helps in reducing possible damage to equipments and productivity loss caused by (otherwise) unscheduled boiler shut-down, and also ensures safety operation of the systems.

The paper is organized as follows: initially, a discussion of current methods of fault diagnosis is presented. So, the proposed fault diagnosis system is developed. After discussing system structure, the main steps of the methodology designed are described in detail.

2. APPLICATION FIELD

In order to make natural gas practical and commercially viable to transport from one country to another, its volume has to be greatly reduced. To obtain maximum volume reduction, the gas has to be liquefied (condensed) by refrigeration to less than -161 °C. This process also requires very strict safety measures and precautions during all liquefaction stages, due to the flammable nature of the gas involved.

The LNG (Liquefied Natural Gas) plant (GL-1K complex) is located at 5km east side of Skikda, Algeria. It has an area of 92 hectars and has been in production since the early 1970s. Gas is sourced from the Hassi R'mel fields, which also supply Arzew plants.

The plant, which is owned and operated by Sonatrach -owned oil and Gas Company –, had grown to six trains by the 1990s with the last of these commissioned in 1981. An LNG train is a liquefied natural gas plant's liquefaction and purification facility. Each LNG plant consists of one or more trains to compress natural gas into liquefied natural gas. A typical train consists of a compression area, propane condenser area, methane, and ethane areas.

All the trains received upgrades in the 1990s to bring them up to required specifications and the plant was capable of producing 7.68 million tons of LNG per year.

2.1. Steam boiler description

The industrial steam boiler, an ABB ALSTOM type, installed in the complex of natural gas liquefaction, generates a nominal steam capacity of 374 tons/h at superheated steam conditions of 73 bars and 487 ◦C. It is composed of three main parts: the main feedwater line, the steam

generator and the main superheated steam line. Figure 1 shows a schematic representation of the steam generator.

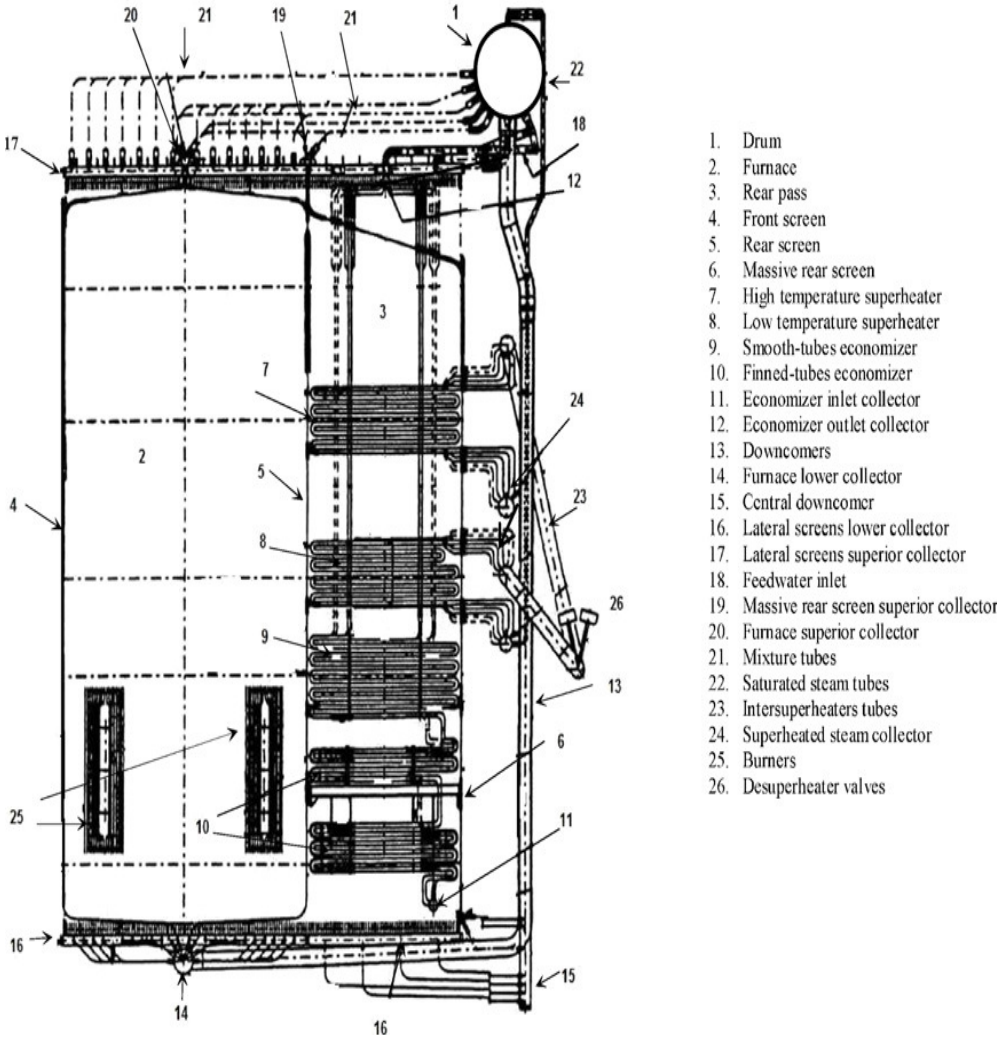


Figure 1. Cross-section view of steam generator

The integrated control and monitoring systems include various safety systems designed to prevent postulated damage during normal and abnormal transients. The top of the steam drum is equipped with three safety valves that control system pressure. Two other safety valves are installed on the main superheated steam line and there are four isolation valves and four flapper valves in different locations of the steam boiler facility.

2.2. Problem of supervision

In most chemical plants, distributed control systems (DCS) are used to simultaneously control thousands of process variables such as temperature and pressure. The main activities of plant operators in this control are to supervise these highly automated systems; fault detection and diagnosis, planning countermeasures; compensating and correcting abnormal situations. Operators are faced with these complex decision makings in managing abnormal situations.

Monitoring manufacturing plants is a complex task. We are interested, in one hand, in monitoring and decision-aiding in case of malfunction of the production facility. In the other hand, when a malfunction occurs on the plant, this leads to some higher level problems. In fact, in such a case, we do not have any formal model of the plant, of the process and of the malfunction that occurred. For such high-risk production, we are interested in the more efficient way to handle a malfunction at the moment is to stop the production, which entails a financial impact. Our system allows extracting knowledge about the production facility during malfunction situations. Later, this knowledge may be used during a decision-aid step.

2.3. Overview of existing approaches of fault detection and diagnosis

For developing decision support systems in process control engineering, three solution approaches of data driven, analytical, and knowledge-based have been identified as follows:

2.3.1. Data-driven Approach

These approaches have been increasingly adopted for feature extraction from historical databases developed from process operations. The most popular data-driven process monitoring approaches include principal component analysis (PCA) [4], Fisher discriminant analysis and partial least-squares analysis (PLS).

2.3.2. Analytical Approach

The analytical approach generally involves detailed mathematical models which use some measured input u and output y, and generates features such as residuals r, estimation parameter p, and estimation state x. Then, based on these values, fault detection and diagnosis can be performed by comparing the observed feature values with those associated with normal operating conditions either directly or after some transformations. Analytical methods can be categorized into the two common methods of parameter estimation and observer-based method [5].

2.3.3. Knowledge-based Approach

The artificial intelligence technologies which are associated with knowledge-based approaches and adopted for monitoring, control, and diagnosis in the industrial process include expert systems, fuzzy logic,...[6]

2.3.4. Integrated approaches for monitoring, diagnosis and control of industrial processes

The approaches described above are often combined in existing systems. Due to growing complexity of current systems, the integration of these approaches into an intelligent system requires a framework which coordinates communication among the different solution modules. Advanced Monitoring Systems of processes have been recognized by academia and industry as a vital research area, which many research programs and industrial projects were initiated to investigate. The most relevant ones are presented in the following.

CHEM-DSS (Decision Support System for Chemical/Petrochemical Manufacturing Processes) is an initiative of the European Community (EC) Intelligent Manufacturing Systems consortium in collaboration with Japan and Korea [7]. The aim of the CHEM-DSS project is to develop and implement an advanced Decision Support System (DSS) for process monitoring, data and event analysis, and operation support in industrial processes, mainly in refining, chemical and petrochemical processes. The research instead focused on analyzing the properties of the individual techniques of the system such as FDI, planning, artificial intelligence, signal

processing, and scheduling. So, twenty three software toolboxes were developed during the project.

MAGIC (Multi-Agent-Based Diagnostic Data Acquisition and Management in Complex Systems) [8] is developed by a joint venture of several European universities and companies. The MAGIC system consists of several model-based and cause-effect diagnostic agents and a process specification agent to specify the process to be monitored and diagnosed. Depending on the process specifications, the appropriate data and knowledge acquisition is performed by another agent. A diagnostic decision agent and a diagnostic support agent propose a final diagnostic decision, which is displayed with other information to an operator interface agent. The MAGIC system prototype is developed for the metal processing industry.

However, knowledge of control systems mentioned above is not available in structured formats. For this reason, the new generation of decision support systems needs to tap into knowledge that is very broad combining learning, structured representations of domain knowledge such as ontologies and reasoning tools. It is in this context that joins our reflection.

3. INDUCTIVE LOGIC PROGRAMMING

Inductive Logic Programming (ILP) [9] was born at the intersection of Concept Learning and Logic Programming. It is a supervised machine learning technique that uses logic programming as a representation for examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system derives a hypothesized logic program which entails all the positive examples.

Though the use of background knowledge has been widely recognized as one of the strongest points of ILP when compared to other forms of Concept Learning, the background knowledge in ILP is often not organized around a well-formed conceptual model. This practice seems to ignore the growing demand for an ontological foundation of knowledge in intelligent systems. Rather, it highlights some difficulties in accommodating ontologies in ILP. Indeed the underlying Knowledge Representation (KR) frameworks (DLs (Description Logics) and HCL (Horn Clausal Logic) respectively) are deeply different in several respects but can be combined according to some limited forms of hybridization.

3.1. Induction in ILP

Induction in ILP generalizes valid hypotheses from individual instances/observations in presence of background knowledge. In Concept Learning, generalization is traditionally viewed as a search through a partially ordered space of inductive hypotheses [10]. According to this vision, an inductive hypothesis is a clausal theory and the induction of a single clause requires (1) structuring, (2) searching and (3) bounding the space of clauses [11].

3.2. Ontologies and relational learning in ILP

An ontology formally represents knowledge as a set of concepts of a specific domain and the relationships between these concepts. The widely accepted definition of ontology is “a formal, explicit specification of a shared conceptualization” [12].

Hybrid KR systems combining DLs and (fragments of) HCL have very recently attracted some attention in the ILP community. Three ILP frameworks have been proposed which adopt a hybrid

DL-HCL representation for both hypotheses and background knowledge: Carin-ALN, resorts to AL-log, and builds upon SHIQ+log.

3.2.1. Learning in Carin-ALN

The framework proposed in [13] focuses on discriminant induction and adopts the ILP setting of learning from interpretations. Hypotheses are represented as CARIN-ALN non-recursive rules with a Horn literal in the head that plays the role of target concept. The coverage relation of hypotheses against examples adapts the usual one in learning from interpretations to the case of hybrid CARIN-ALN BK. The generality relation between two hypotheses is defined as an extension of generalized subsumption. Procedures for testing both the coverage relation and the generality relation are based on the existential entailment algorithm of CARIN.

3.2.2. Learning in AL-log

In [11], hypotheses are represented as constrained Datalog clauses that are linked, connected (or range-restricted), and compliant with the bias of Object Identity (OI). Therefore the literal in the head of hypotheses represents a concept to be either discriminated from others (discriminant induction) or characterized (characteristic induction). The generality relation for one such hypothesis language is an adaptation of generalized subsumption, named B-subsumption, to the AL-log KR framework. It gives raise to a quasi-order and can be checked with a decidable procedure based on constrained SLD-resolution.

3.2.3. Learning in SHIQ+log

This ILP framework represents hypotheses as SHIQ+log rules restricted to positive Datalog [14] and organizes them according to a generality ordering inspired by generalized subsumption. The resulting hypothesis space can be searched by means of refinement operators either top-down or bottom-up. A decidable KR framework SHIQ+log is the most powerful among the ones currently available for the integration of DLs and HCLs.

4. ARCHITECTURE OF OUR SYSTEM

Our system is developed to identify the causes and provides operation suggestions when abnormal situations occur. Figure 2 shows the main architecture of our system. This architecture consists of two distinct parts, one used offline which includes a module for generating examples and a module for learning discriminative patterns, and the other is used on line which includes a module for semantic reasoning.

4.1. Steam boiler ontology construction

Several ontologies have been proposed in the field of process engineering for the design and operation of chemical and pharmaceutical processes. Among the earliest, multi-dimensional formalism was developed as a collection of interrelated ontologies for describing the plant structure, the materials involved, the behavior of the material based on physico-chemical transformations occurring in the plant, and concepts to specify typical tasks in plant operation. This collection of ontologies was proposed to support engineering activities across the life cycle of the plant.

Development cost of these knowledge-based systems is often high because knowledge bases are often constructed from scratch in a distributed and heterogeneous environment.

Relational databases are valuable sources for ontology learning. In this paper, we describe an approach for steam boiler ontology construction using heterogeneous databases. Our objective is to build an ontological resource, in a most automated way. The main data and information constituting our system come from disparate databases for equipment characteristics.

Methods and tools have been proposed to generate ontologies from such structured input. The mappings are the correspondences between each created ontology component (e.g., concept, property) and its original database schema concept (e.g., table, column).

The implementation of the proposed solution is realised using Protégé Plug-in DaTaMaster and it followed the steps below:

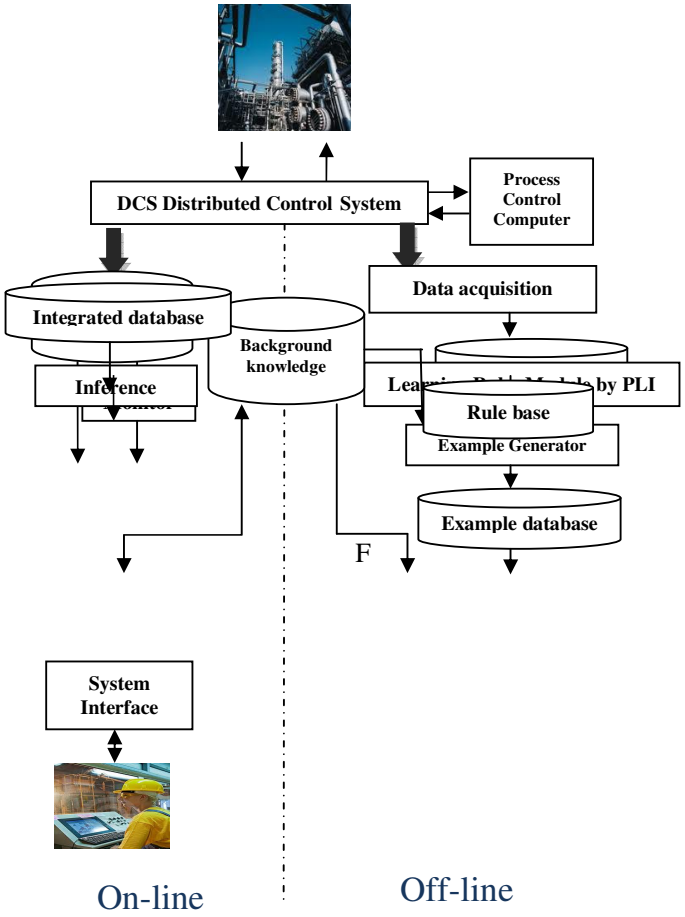


Figure 2. Architecture of the proposed system

- The choice of the connection driver types: Open Data Base Connectivity (ODBC) or Java Databases Connectivity (JDBC) and the data source.
- The selection of a given table activates the visualization of its content, then the user have the choice of importing the table or not.
- The chosen data base tables are activated and visualised, each table is transferred into one class or sub-class depending on the user's choice.

4.2. Distributed Control System (DCS)

A Honeywell Distributed Control System (DCS) operates in the Production Unit, which serves the data via the so-called Process History Database (PHD) module. This database contains the most important process variables and some technological variables calculated by the Advanced Process Control (APC) module of the DCS.

4.3. Data acquisition and Data Warehouse

Large and complex industrial processes such as chemical plants and petroleum refineries are typically equipped with distributed control systems (DCS) which allow users to vary the number of alarms for the purpose of better monitoring process-variables. As such industrial processes increase in size, the volume of alarm information being presented to the operator also increase.

They not only may be thousands of individual alarms, nuisance alarms could also distract the operator's attention from more important problems. Process DataWarehouse is a data analysis-decision support and information process unit, which operates separately from the databases of the DCS. It is an information environment in contrast to the data transfer-oriented environment, which contains trusted, processed and collected data for historic data analysis. The data collected into DW directly provide input for different data mining, statistical tools, like classification, clustering, association rules, etc.

The first phase in the knowledge discovery process is the generation of simulated data sets and data preprocessing.

4.3.1. Generation of simulated datasets

It is observed that the injection of disturbances will cause changes in plant measurements, while the measurements outputs of the model under normal operation conditions remain unchanged. So the differences between the distributed and normal measurement outputs are used to generate discrete alarm data.

4.3.2. Data preprocessing

Alarm databases in a gas plant consist of records of time-stamped event logs. An example of raw gas plant data showing different fields is presented in Figure 3. As a minimum requirement, each record contains fields that store information such as a unique identifier for each alarm tag, time/date, alarm priorities, alarm description, alarm settings and the possible states of an alarm tag which include activation [ALM], return [RTN] and acknowledge [ACK] event types. We assume that a possible alarm sequence could be {[ALM]→[RTN]}, or {[ALM]→[ACK]→[RTN]}butnot{[ALM]→[ACK]→[ALM]},or {[ALM] →[RTN]→[ACK]}.

During the data preprocessing phase it is extremely important to carefully investigate and prepare alarm data since real plant data tends to be inconsistent, with errors, missing values, outliners and duplicate values.

EVENT JOURNAL - PROCESS ALARMS

Time Stamp / Unit ID	Point Name / Point Description	Tripped Value	Alarm Type / Alarm Priority	Status / Event	Source / Sequence of Events Info.
10/08/2012 21:39:07 48	PM413 M/A POMPE GRAIS		OFFNORM HIGH	ACK 31	P2-APP56 D4801 3
10/08/2012 21:39:06 48	YHS455 LOCAL / DISTANCE		OFFNORM LOW	RTN 21	P2-APP56 DISTANCE DGPL03
10/08/2012 21:39:01 48	PM413 M/A POMPE GRAIS		OFFNORM HIGH	ALM 21	P2-APP56 ARRET D4801
10/08/2012 21:38:33 48	CM413 M/A COMPRES C413		OFFNORM HIGH	RTN 21	P2-APP56 MARCHE DGPL03
10/08/2012 21:38:01 48	PM413 M/A POMPE GRAIS		OFFNORM HIGH	RTN 21	P2-APP56 MARCHE D4801
10/08/2012 21:20:22 50	PI455 STOCKAGE TK-411	20.000	PVLO LOW	RTN 21	P2-APP56 21.008 D5001
10/08/2012 21:02:09 10	FI5513 DISTRIB VAP DEM 6-S-3303		BADPV HIGH	RTN 21	P2-APP56

Figure 3. An example of raw gas plant data showing different fields

4.4. Generator of examples

The simulation process associates to each breakdown situation a set of sequences of observations, building up a learning database. We recall that we are interested in cases of plant malfunctions. We want to model these unusual situations. Thus, the set of positive examples is made up of symbolic descriptions of the state of the plant that leads to a malfunction. We use the ontology as hypothesis language, and instances of this ontology as example language.

The generation process is an ontology population task. It lies in the acquisition from simulator of new extensional knowledge, i.e. ontology's instances. Each example of the base of the sequences alarms follows the following syntax: Example (I, C, O)

The parameter I is the identifier of example. We number the examples by integers from 1 to 200. The parameter C is a list of class that the example belongs. As for the last parameter, the parameter O corresponds to the object describing the sequence. In our simplified example, the background knowledge is made up of a symbolic description of the working environment that is the components of the plant (Cf. § 4.5).

For instance a positive example is defined, for instance, in the following way:

Example (1, default (V4-1, TRC 317), sequence ([increase (V4-1, pressure)], [closed (V4-1)], [increase (V4-1, flow)], [after (V4-1, P1)]))

4.5. Diagnosis rules learning by PLI

The ILP unit allows generating general diagnosis rules that guide the operator in an efficient way to handle future malfunctions. The main objective of our work is the automatic learning of diagnosis rules. The examples used for this purpose are sequences of events obtained by simulation, each one of them being labeled according to the malfunction having caused it. The learning technique is based on the inductive logic programming.

We consider the problem of learning rules from ontologies and relational data. We assume that the predicate in the rule head represents a concept to be characterized (characteristic induction). The data are represented as a SHIQ+log knowledge base B where the intensional part K (i.e., the

TBox T plus the set R of rules) plays the role of background knowledge and the extensional part (i.e., the ABox A plus the set F of facts) contributes to the definition of observations. Therefore ontologies may appear as input to the learning problem of interest. The observations are represented as a finite set of logical facts E. E could generally discomposed into the positive examples E^+ and the negative ones E^-.

The background knowledge is supposed to be insufficient to explain the positive observations and the logical translation of this fact is: $B \not\models E^+$ but there is no contradiction with the negative knowledge: $BUE^- \not\vdash \perp$.So an ILP machinery with input E and B, will output a program H such that $B \cup H \models E$. So H constitutes a kind of explanation of our observations E^+. The language L of hypotheses must allow for the generation of SHIQ+log rules. More precisely, we consider defined clauses of the form: $p(\boldsymbol{X}) \leftarrow r_1(\boldsymbol{Y}_1), \ldots, r_m(\boldsymbol{Y}_m), s_1(\boldsymbol{Z}_1), \ldots, s_k(\boldsymbol{Z}_k)$. Where $m \geq 0$, $k \geq 0$, each $p(X)$, $r_j(Y_j)$, $s_l(Z_l)$ is an atom, and the literal $p(X)$ in the head represents the target concept.

Figure 4 reports the main procedure of an algorithm analogously to FOIL [15] for learning onto-relational rules. The outer loop learns new rules one at a time, removing the positive examples covered by the latest rule before attempting to learn the next rule. The inner loop searches a second hypothesis space, consisting of conjunctions of literals, to find a conjunction that will form the body of the new rule.

```
Hset := Ø
Pos:= E+
while Pos ≠ Ø do
   h := {p(X ) ←};
   Negh:= E-
        while Negh ≠ Ø do
        add a new literal L to specialize h
     end while
Hset := Hset ∪ {h};
Posh:= {e ∈ Pos | B ∪ h⊨e};
Pos:= Pos\Posh
endwhile
return Hset
```

Figure 4. General algorithm for learning onto-relational rules

We want to show, on a small example, the way ILP may be used to induce a model of plant malfunctions. We want to model these unusual situations. Suppose we have a SHIQ+log KB consisting of the following intensional knowledge K:

Valve (C) $\sqsubseteq$ Component (C)
Pump (C) $\sqsubseteq$ Component (C)
Valves (v1)
……..
Valve (v2)
Pump (P1)
Parameter (pressure)

We aim at inducing a definition for the predicate default/1 modeling the default of one component of the plant. In the entire system, the arity of this predicate may be greater than one depending on the number of components involved in the malfunction. If we run our system on the complete set of positive examples that describe the problem, the system induces the following definition for

the predicate default/1: **Default (A) ← increase (A, pressure), after (A,P1), closed(A)**, which means that if the pressure increase in A and A is located after the pump P1 and A is closed then the component A causes a malfunction of the plant.

4.6. Monitor

The monitoring module is used to supervise production process. It monitors data streams obtained from the control system, e.g. temperature, pressure, and flow. If a situation is judged to be abnormal by the module, the data are automatically transferred to the inference machine to solve the problem. At the same time the data are stored in the integrated database.

4.7. Inference machine

It controls and executes problem solving. Semantics search is based on the application of reasoning techniques to perform logical deductions. This one results from raisoners able to use the semantics of OWL ontology. Regarding the deficiency of this language, our ontology is extended by diagnosis rules to increase its expressiveness. These rules represent the deductive part of a domain. They permit the propagation of relations and deduce new facts starting from the existing ones.

A hybrid reasoning mechanism developed in our previous works [16], is applied in this system of supervision. It ensures the reasoning and searches the rules efficiently. We suggest a combination method of reasoning to improve the search of results. The method is based on the principle of hybrid combination of two reasoners in which each treats a distinct party of knowledge base: a logical description raisoner for the structural part (OWL-DL) and a rule engine for the deductive part (RULES).

4.8. Explanation instrument

Tracks the route of reasoning and explains the results of the reasoning of system to the user upon request.

4.9. Integrated database

Stores facts, intermediate results of reasoning processes, real-time data, and historical data from the distributed computer system installed in the plant.

5. CONCLUSION

Complex processes involve many process variables, and operators faced with the tasks of monitoring, control, and diagnosis of these processes. They often find it difficult to effectively monitor the process data, analyze current states, detect and diagnose process anomalies, or take appropriate actions to control the processes.

To assist plant operators, decision support systems that incorporate artificial intelligence (AI) and non-AI technologies have been adopted for the tasks of monitoring, control, and diagnosis. In this paper, a real-time system is proposed for monitoring and diagnosing of chemical processes. The representation of knowledge base, inference machine and the relations among them are considered in this paper according to the characteristics of chemical processes.

This system helps the operators a lot to eliminate potential and disasters faults. The system also decreases the loss brought by unstable process situations and the loss if the time used for eliminating faults is too long. When new fault occurs, the stored data helps the domain expert to analyze the reason of the fault, and give earlier prediction of the trend. Our system design approach can be exploited to develop and rapidly prototype real time distributed multi-agent systems.

ACKNOWLEDGEMENTS

The authors are very grateful toward all the personnel of the natural gas liquefaction complex of Skikda for providing all information related to the steam boiler equipments enabling the present study to be achieved.

REFERENCES

[1] H.L Chiang, L.E. Russell and D.R. Braatz, (2001) "Fault Detection and Diagnosis in Industrial Systems". Springer, London, Great Britain.
[2] Technical operations manual of the industrial steam boiler ABB ALSTOM (2000).
[3] D. R. Tucakovic, V. D. Stevanovic and T. Zivanovic, (2007) "Thermal hydraulic analysis of a steam boiler with rifled evaporating tubes." Applied Thermal Engineering, 27, 509–519.
[4] S. Yoon, F.J. MacGregor, (2004) "Principle-component analysis of multiscale data for process monitoring and fault diagnosis." AIChE Journal 50 (11), 2891–2903
[5] A.E. Garcia, M.P. Frank, (1996) "On the relationship between observer and parameter identification based approaches to fault detection." In: Proceedings of the 13th IFAC World Congress, vol. N. Piscataway., New Jersey, pp. 25–29.
[6] X. Luo, C. Zhang and R.N. Jennings, (2002) "A hybrid model for sharing information between fuzzy, uncertain and default reasoning models in multi-agent systems". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (4), 401–450.
[7] S. Cauvin, (2004) "CHEM-DSS : Advanced decision support system for chemical/petrochemical industry", Fifteenth International Workshop on Principles of Diagnosis (DX'04), AAAI, Carcassonne, France
[8] B. Kőppen-Seliger, T. Marcu, M. Capobianco, S. Gentil, M. Albert, and S. Latzel, (2003) "MAGIC: An integrated approach for diagnostic data management and operator support," in Proceedings of the 5th IFAC Symposium Fault Detection, Supervision and Safety of Technical Processes - SAFEPROCESS05, Washington D.C.
[9] S.Muggleton, (1999) "Inductive Logic Programming: Issues, Results and the Challenge of Learning Language in Logic", Artificial Intelligence, 114(1-2), pp.283--296.
[10] T. Mitchell, (1982) "Generalization as search. Artificial Intelligence" 18, 203–226.
[11] F. Lisi, (2008) "Building Rules on Top of Ontologies for the Semantic Web with Inductive Logic Programming.", Theory and Practice of Logic Programming 8(03), 271–300.
[12] A. Borgida, (1996) "On the relative expressiveness of description logics and predicate logics.", Artificial Intelligence 82(1–2), 353–367.
[13] C. Rouveirol, V. Ventos, (2000) "Towards Learning in CARIN-ALN." In: Cussens, J., Frisch, A.M. (eds.) ILP 2000. LNCS (LNAI), vol. 1866, pp. 191–208. Springer, Heidelberg.
[14] F. Lisi, F. Esposito,(2007) "Building Rules on top of Ontologies? Inductive Logic Programming can help!" SWAP 2007
[15] J.R. Quinlan, (1990) "Learning logical definitions from relations." Machine Learning, 5:239–266.
[16] S. Bouarroudj, Z. Boufaida, (2010) " A multi-reasoner system for semantic search of annotated images", EGC-M 2010, Algiers , Algéria, pp. 118-129.

Authors

Samiya BOUARROUDJ is a doctoral student of Computer Science at University of Constantine 2, Algeria. Her current research activities are conducted at the LIRE Laboratory. Here research interests are knowledge representation and reasoning, semantic web technologies and advanced monitoring systems of the processes.

Zizette BOUFAIDA is a Professor of Computer Science at University of Constantine 2, Algeria and the co-head of the SI&BC research group at the LIRE laboratory. Her research interests include knowledge representation and reasoning; formal knowledge representation for Semantic Web, ontology development,...

INTENTIONAL BLANK

UNDER WATER NOISE REDUCTION USING WAVELET AND SAVITZKY-GOLAY

Selva Balan[1],Arti Khaparde[2],Vanita Tank[2],Tejashri Rade[2]and Kirti Takalkar[2]

[1]Central Water and Power Research Station,
Khadakwasla, Pune, Maharashtra, India
[2]Department of E &TC, Maharashtra Institute of Technology, Pune,
Maharashtra, India

ABSTRACT

A precise, linear indication of the depth of water in a specific part of water body is what always required. Presently there are a wide variety of ways to produce a signal that tracks the depth of water.The Ultrasonic signal is most commonly used for the depth estimation. This signal is affected by various underwater noises which results in inaccurate depth estimation. The objective of this paper is to provide noise reduction methods for underwater acoustic signal.In present work, the signal processing is done on the data collected using TC2122 dual frequency transducer along with the Navisound 415 echo sounder. There are two signal processing techniques which are used: The first method is denoising algorithm based on Stationary wavelet transform (SWT)and second method is Savitzky-Golay filter. The results are evaluated based on the criteria of peak signal to noise ratio and 3D Surfer plots of the dam reservoir whose depth estimation has to be done.

KEYWORDS

Savitzky-Golay filter, Wavelet transform, PSNR

1. INTRODUCTION

Dams in any country are a part of several multi-purpose projects to serve a variety of needs. Basically, dams are built to harness the river water so that it can be utilized according to the needs. A multipurpose project is often launched for storing water for irrigation purposes, providing drinking water and generating hydro-electricity. The water stored by the dams can also be used to prevent floods and facilitate forestation in the catchments areas of the reservoirs. Depth measurement of water bodies has developed remarkably in the last few decades with the adaptation of new techniques and technologies. Several methods have been studied and introduced by researchers for assessing the depth of the reservoirs using Ultrasonic signal.

Ultrasound wave is basically cyclic sound pressure whose frequency ranges from 15 kHz to 200 kHz [1]. The depth measurement is quite sensitive to variations of the sound velocity profile. The sound velocity profile is affected by factors such as, variation of one degree Celsius in temperature, salinity which is a measure of the quantity of dissolved salts and other minerals in water and the total amount of dissolved solids in water. The pressure also has a significant impact on the sound velocity variation and has a major influence on the sound velocity in deep water [2].

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 243–250, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4421

When an ultrasonic wave is transmitted through water, it is expected to reach the bottom and then reflect back, but instead of this, it gets contaminated with noise and gets reflected back by the obstacles such as stones, waste thrown in water or the creatures living under the water. This gives a false bottom anticipation which doesn't provide the accurate results.

This paper deals with depth analysis of the water reservoir using reflected ultrasound waves.The bed of the dam is mostly even due to deposition of silt by the water current. When the reflected signal of sensor is plotted in a 3D image, it is shown as a bed containing sharp peaks which is not expected at the bottom. These sharp peaks could be the reflections from the suspended obstacles which come in the path of the transmitted ultrasonic signal.

The data was collected using sensor Reson's TC2122 dual frequency survey echo sounder transducer which works on two resonant frequencies 33kHz and 200kHz and Reson's Navisound 415 hydrographic single beam echo sounder.General assumption is that the noise present is white Gaussian noise but the underwater noise does not full fill the classical white noise assumption [3] and hence Non-white noise is assumed. To reduce noise from the given data and to estimate approximate depth, two techniques are applied- Denoising based on Stationary Wavelet Transform and Savitzky-Golay filter.

This paper is organized as follows:-Section 2 deals with wavelet transforms. Savitzky-Golay filter is explained in section 3, Section 4 & Section 5 deal with results & conclusion respectively.

2. WAVELET TRANSFORM

Wavelet transforms have become one of the most important and powerful tool for signal denoising [4]. Discrete Stationary Wavelet Transform is undecimated versions of discrete wavelet transform which is used for signal denoising [5] and pattern recognition. The main idea is to average several detailed coefficients which are obtained by decomposition of the input signal [6].

There are a number of wavelets that can be used for noise removal: Harr, Daubechies, Symlet, Coiflet, Biorthogonal, Reverse Biorthogonal, Meyer [7][8] to name few. In order to use the wavelet transform effectivelythe details of the particular application should be taken into account and the appropriate wavelet should be chosen. They are chosen based on their shape and their ability to analyze signal in particular application [9]. The performance of wavelet based denoising depends on wavelet decomposition structure.

For selecting particular type of wavelet, performance comparison of some known wavelet familieswas done and their effect on the given signalwas observed.In present case, as explained earlier smoothness of the surface is the basic criteria for depth estimation, so accordingly one wavelet from each wavelet family was selected. These are shown in Table 1.

Table 1.Wavelet selected from respective wavelet family.

Wavelet Family	**Selected wavelet**
Harr	harr
Daubechies	db8
Symlet	sym5
Coiflet	coif5
Meyer	dmey
Biorthogonal	bior2.2
Reverse biorthogonal	rbior2.2

The detailed and approximation coefficients are obtained using signal decomposition. Further decomposition of approximation coefficients up to specified level is done.The maximum decomposition level depends on number of data points contained in a data set. Present depth analysis 5 decomposition levels were found to be appropriate.

Thresholding of data in wavelet domain is done to smooth out or to remove some of the coefficients of wavelet transform of measured sub-signal introduced due to noise or obstacles in water bodies [10]. Two commonly used types of thresholding are hard and soft thresholding. In hard thresholding [10][11] if any coefficient (x) less than threshold value(t) then it is set to zero otherwise it remains unchanged.

$$hard(x) = \begin{cases} x; & |x| > t \\ 0; & otherwise \end{cases} \quad (1)$$

Soft thresholding [10][11] is similar to hard thresholding with a little difference i.e. no coefficient remains unchanged instead it is shrunken by threshold value(t). The present analysis is done using soft thresholding technique.

$$soft(x) = \begin{cases} x - t; & |x| > t \\ 0; & otherwise \end{cases} \quad (2)$$

3. SAVITZKY-GOLAY FILTER

The Savitzky-Golay filter is a particular type of low-pass filter. It is well-adapted for data smoothing. It is also referred to as least-squares or Polynomial Smoothing filter [12]. Rather than having their properties defined in the Fourier domain, and then translated to the time domain, Savitzky-Golay filters derive directly from a particular formulation of the data smoothing problem [7] in the time domain. These filters are of type-I FIR low pass filters with nominal pass band gain of unity [14].Savitzky and Golay proposed the method of data smoothing based on local least-squares polynomial approximation [14]. Polynomial smoothing is the process which replaces the noisy samples by the values that lie on the smooth polynomial curves drawn between the noisy samples. For every polynomial order, the coefficients must be determined optimally such that the corresponding polynomial curve best fits the given data [12]. Instead of applying averaging filter it is better to perform least squares fit of a small set of consecutive data points to a polynomial. So Least-squares fit technique is used to choose the polynomial coefficients such that they give minimum mean square error [12][15]. The output smoothed value is taken at the center of the window to replace the original data.

In Savitzky-Golay filter, the odd-indexed coefficients of the impulse response design polynomial are all zero. The nominal normalized cutoff (3 dB down) frequency depends on both the implicit polynomial order and the length of the impulse response. The impulse response of filter is symmetric,so the frequency response is purely real. These filters have very flat frequency response in their pass bands and fair attenuation characteristics in their stop band regions [14].

Following are the constraints on polynomial fitting [14]

- The number of data points must be strictly greater than the number of undetermined coefficients to achieve smoothing by the Savitzky-Golay process.
- If the order of the polynomial is too large, the solution will be of no value.

Generalize algorithm is as follows:

Consider frame size or filter length N is odd, $N = 2M + 1$ and $N \geq d + 1$ where d is order of polynomial.

If x is noisy signal with noisy samples x_n, $n = 0,1,\ldots\ldots,L\text{-}1$ and it is supposed to be replaced by its smoothed output version y which contains y_n, $n = 0,1,\ldots\ldots,L\text{-}1$ then input vector has $n = L$ input points and $x = [x_0, x_1 \ldots x_{L-1}]^T$ is replaced by N dimensional one, having M points on each side of x.

$$x = [\, x_{-M}, \ldots x_{-1}, x_0, x_1, \ldots, x_M]^T \quad (3)$$

There are 3 cases, for calculating the output result. These cases are explained in [16]. Smoothed output y is calculated as

$$y = Bx \quad (4)$$

The Savitzky-Golay filter coefficients $b_0, b_1, \ldots\ldots$ are the elements of matrix B.

$$B = [b_{-M}, \ldots . b_{-1}, b_0, b_1, \ldots . b_M] \quad (5)$$

$$B = GS^T \quad (6)$$

Where $G = S(S^TS)^{-1}$, $S = [s_0, s_1, \ldots . , s_d]$, $s_0(m) = 1, s_1(m) = m, s_2(m) = m^2, \ldots , s_d(m) = m^d$ where $-d \leq m \leq d$ [12][13].

4. RESULTS AND DISCUSSION

The peak signal to noise ratio represents the measure of peak error [9]. It is given as,

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE}\right) dB \quad (7)$$

Where

$$MSE = \frac{1}{m} \sum_{i=1}^{m} [I(i) - O(i)]^2 \quad (8)$$

MSE is Mean Square Error
I = original value
O = output value
R = maximum input value

Generally PSNR should be greater than 30dB in order to reduce noise effectively.

For comparing results of Savitzky-Golay filter, another parameter used is Time Constraints which is time required for execution of program.

Table 2. Values of PSNR for different types of wavelets.

FileNo.	Harr	Bior	Rbior	Sym	Coif	Dmey	Daub
1	41.93	41.79	42.04	41.22	40.66	40.48	40.46
2	51.74	50.51	49.41	49.25	48.68	48.34	48.26
3	43.79	43.77	43.23	43.02	42.73	42.59	42.78
4	43.98	43.59	42.84	43.02	42.96	42.86	42.59
5	40.07	39.24	39.17	39.89	39.03	38.59	39.76
6	37.07	37.34	37.07	36.87	36.84	36.77	36.90
7	40.00	40.37	40.39	39.77	39.80	39.64	39.91

The results presented in Table 2 show PSNR values for different wavelets. It can be seen that Harr wavelet is giving better result than other wavelets in this case.

Table 3. Values of PSNR by varying order and with fixed frame size for Savitzky-Golay filter.

File	Or1_31	Or2_31	Or3_31	Or4_31
File1	40.6843	42.5433	42.6120	**44.0280**
File2	46.7581	47.8004	47.7967	**49.4808**
File3	42.0438	43.0886	43.1009	**44.4367**
File4	42.0823	43.8442	43.8360	**44.7824**
File5	39.0119	39.7840	39.7839	**40.8566**
File6	36.4540	36.7830	36.8013	**37.2930**
File7	39.2207	40.6914	40.7429	**41.1404**
Avg. Time(sec)	2.59	2.41	2.62	**2.41981**

From Table 3 it can be seen that as the order of polynomial increases, PSNR value also increases. So PSNR is directly proportional to order of polynomial for Savitzky-Golay filter. Computational complexity is less for higher order. (Processer used-Intel core i5)

Table 4. Values of PSNR by varying frame size and with fixed order for Savitzky-Golay filter

File	Or4_31	Or4_33	Or4_41	Or4_49
File1	**44.0280**	43.8143	43.0032	42.6637
File2	**49.4808**	49.1640	48.2272	47.8468
File3	**44.4367**	44.0604	43.4561	43.1529
File4	**44.7824**	44.6722	44.1377	43.7461
File5	**40.8566**	40.6501	40.0252	39.9840
File6	**37.2930**	37.0813	37.0102	36.8819
File7	**41.1404**	41.1594	40.9853	40.5424
Avg.Time (sec)	**2.41**	2.55	2.53	2.56

From Table 5 it can be seen that as the lesser the frame size, more is the PSNR. So PSNR is inversely proportional to frame size for Savitzky-Golay filter. Computational complexity is less for smaller frame size.(Processer used-Intel core i5)

3D plots of signal are obtained using software surfer11 which are shown below:

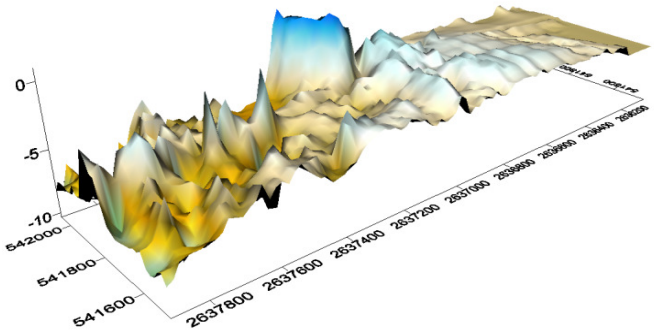


Figure 1.Original signal

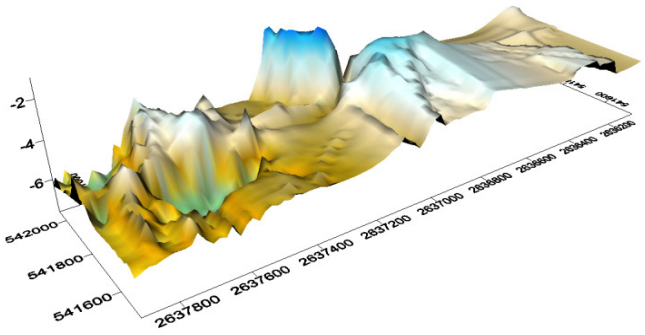


Figure 2. Signal processed using Harr wavelet

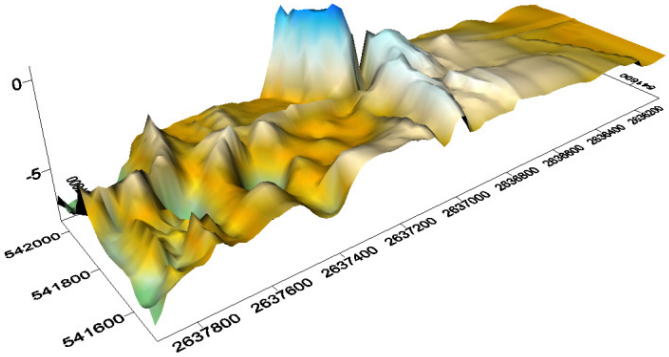


Figure 3.Signal processed using Savitzky-Golay Filter

5. CONCLUSION

Henceanalysis of Ultrasonic signal is done using two techniques:Harr wavelet Transform and Savitzky-Golay filter.It is found that out of all wavelet transforms, harrwavelet is most suitable for noise reduction in ultrasonic signal because PSNR value is high among all the wavelets used. The present paper gives a good comparison of wavelet based denoising of acoustic data.

In Savitzky-Golay Filter, analysis is done for different orders of polynomial and frame sizes which show that with higher order of polynomial and lesser frame size PSNR is high.

The results from surfer plots show that the harr wavelet with decomposition level up to 5 and Savitzky-Golay filter with order 4 and frame size 31 can be effectively used for smoothing the data obtained which can lead to estimation of depth with minimum error using empirical formula designed for a particular application.

ACKNOWLEDGEMENT

We are thankful to Central Water and Power Research Station for providing the data of reservoir in India.

REFERENCES

[1] Sabuj Das Gupta, Islam Md. Shahinur, Akond Anisul Haque, Amin Ruhul, Sudip Majumder,(October 2012)"Design and Implementation of Water Depth Measurement and Object Detection Model Using Ultrasonic Signal System",International Journal of Engineering Research and Development, Volume 4, Issue 3, pp.62-69.

[2] International hydrographic Bureau, (2005)"Manual on hydrography", M-13, pp.126.

[3] Arnaud Jarrot, Cornel Ioana, Andr´e Quinquis, (2005)"Denoising Underwater Signals Propagating Through Multi–path Channels", Oceans - Europe (Volume:1) pp.501-506.

[4] SJS Tsai, (2002)"Chapter 4 Wavelet Transform and Denoising".

[5] Chu-Kueitu, Yan-Yao Jang, (2004)"Development of Noise Reduction Algorithm for Underwater Signals", Underwater Technology, International Symposium on, pp.175-179.

[6] V. Matz and J. Kerka, "DIGITAL SIGNAL PROCESSING OF ULTRASONIC SIGNALS", pp.3.

[7] wavelets.pybytes.com by Filip Wasilewski.

[8] M. Kreidl, P. Houfek, (2002)"Reducing Ultrasounic Signal Noise by Algorithms based on Wavelet Thresholding", Acts Polytechnica Vol. 42, pp.60-65.

[9] S.Kumari, R.Vijay, (January 2012)"Effect of Symlet Filter Order on Denoising of Still Images", Advanced Computing :An International Journal(ACIJ).Vol.3.No.1, pp.137-143.

[10] K.Mathan Raj, S.Sakthivel Murugan, V. Natarajan, S.Radha, (2011)"Denoising Algorithm using Wavelet for Underwater Signal Affected by Wind Driven Ambient Noise", Recent Trends in Information Technology (ICRTIT), pp.943-946.

[11] Bernhard Wieland, (October 2009) "Speech Signal Noise Reduction with Wavelets", pp.55-56.

[12] Sophocles J. Orfanidis, (2010)"Introduction To Signal Processing", Pearson Education, Inc., pp.427-451.

[13] William H. Press, Brian P. Flannery, Saul A. Teukolsky, William T. Vetterling, (1988-1992)"Numerical Recipes in C:The Art of Scientific Computing", Cambridge University Press, pp.650-651.

[14] Ronald W. Schafer, (July 2011)"What is a Savitzky-Golay filter?", IEEE SIGNAL PROCESSING MAGAZINE, pp.111-115.

[15] Savitzky A., and Golay, M.J.E. (1964)"Analytical Chemistry", Volume 36, pp.1627-1639.

[16] Md. Abdul Awal, Sheikh Shanawaz Mostafa and Mohiuddin Ahmad, (2011)"Performance Analysis of Savitzky-GolaySmoothing Filter Using ECG Signal", IJCIT, VOLUME 01 ISSUE 02, pp.24-29.

Authors

Mr. Selva Balan is currently working as a Chief Research Officer at Central Water and Research Power Station (CWPRS), Khadakwasla, Pune, India. At CWPRS he is associated with various national important projects under hydrology projects, development projects related to Ports, Coastal development, River Engineering, Dam Instrumentation, Remote sensing, Real time systems and many other projects. He is Faculty at National Water Academy.Published more than 50 conference, 20 journal papers. He is an Expert Member nominated under WORLD BANK projects in hydrology group.He is an Editor for International Standards on Hydrometry, and at BIS. His research interests include ANN, Fuzzy logic, Microcontrollers, Image processing, Wavelets etc

Dr. Arti Khaparde is currently working as Professor in the Department of Electronics and Telecommunication with Maharashtra Institute of Technology, Pune. She has vast experience of 14 years in teaching and research and almost 20 Publications on her name in reputed International Conferences and Journals. She received her Ph.D. in Electronics and Communication Engineering, from JNTU, Hyderabad in 2009. Her research interests include Digital Signal processing, noise removal techniques, Digital Image and video processing, Pattern recognition, Programming of microprocessors and microcontrollers.

Ms. Vanita Tank received her B.E degree from Barktullah University and M.Tech from Maulana Azad National Institute of Technology, Bhopal. Currently she is working as Assistant Professor in the Department of Electronics and Telecommunication with Maharashtra Institute of Technology, Pune and pursuing her Ph.D in audio and speech signal processing from University of Pune. She has 7 years of experience in teaching. Her current interest includes applications of signal processing and signal coding.

Ms. Tejashri Rade is a final year student of B.E Electronics & Telecommunication in Maharashtra Institute of Technology, Pune. Presently she is working under Mr. Selva Balan and Ms. Vanita Raj Tank for her final year project on Noise reduction from Ultra sound waves.

Ms. Kirti Takalkar is a final year student of B.E Electronics & Telecommunication in Maharashtra Institute of Technology, Pune. Presently she is working under Mr. Selva Balan and Ms. Vanita Raj Tank for her final year project on Noise reduction from Ultra sound waves.

EFFECT OF MOBILITY ON PERFORMANCE OF MACAW MECHANISM OF IEEE 802.11 ADHOC NETWORKS

Ghadeer Hassan Mustafa[1], Mohamed Essam Khedr[2] and Ramy Eltarras[3]

[1,2]Electronics and Communication Engineering Department
Arab Academy for Science and Technology Alexandria, Egypt
ghadeerhassan2009@gmail.com
khedr@aast.edu
[3]Computer Science Department
Arab Academy for Science and Technology Alexandria, Egypt
ramy@aast.edu

ABSTRACT

Packet Collisions in wireless communication are one of the most causes of network performance degradation. MACAW mechanism is a popular and widely used mechanism to reduce packet collisions that arise due to the hidden terminal problem in mobile ad-hoc networks. However; Mobility is considered a vital issue in mobile ad-hoc networks that has not been thoroughly investigated. A real time simulation is carried out to measure the Throughput and the Packet Delivery Ratio of an ad-hoc network under Shadowing propagation model using MACAW mechanism with different mobility scenarios.

KEYWORDS

Mobile Ad-hoc network, MACAW mechansim, IEEE 802.11b, Shadowing

1. INTRODUCTION

Wireless Communication has impacted our daily lives and has become significant for our modern existence. Drastic changes have taken place since the early development of radio telephony to current smart devices such as tablets, iPhones that support communication with higher data rate, coping with the increasing demand for mobility and flexibility in nowadays lifestyle. In addition to wireless multimedia that is becoming increasingly popular as they offer users the accessibility to information and multimedia services anytime. As wireless communication continues to evolve, the future holds many more possibilities in this field.

IEEE 802.11 is considered to be an important standard for Wireless Local Area Networks (WLANs) which is adopted by many manufacturers of WLAN products. IEEE 802.11 is concerned with the physical layer and the MAC layer. The physical layer is in charge of transmitting the raw data over Radio Frequency (RF) .Whereas, the MAC layer coordinates access

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 251–259, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4422

to a shared radio channel and guarantees the privacy of the transmitted data as well as the reliability of the data services. In addition to management protocols that ensure authentication and data delivery. In contrast to other LAN standards, wireless standards address unique issues such as node mobility, limited Bandwidth availability, error prone broadcast channel, power management, link reliability management and hidden and exposed terminal problems, none of which was a concern for other standards in IEEE 802.

The IEEE 802.11 has attracted many researchers to analytically model its access mechanism due to its wide popularity. Network designers were able to ascertain proper values for different parameters that lead to the best performance; also they were able to decide efficiently on the network design taking into consideration the required performance and the expected traffic. Finally, researchers have proven the efficiency of new mechanisms that were proposed to enhance the 802.11 performance. [7]

One of the major challenges that faced researchers was the impact of mobility on wireless ad-hoc networks. Mobile ad-hoc networks lack fixed topology due to node mobility, interference, path loss, shadowing. As a result of this challenging environment, rapid and random changes of the network topology at unpredictable times take place.

The main contribution of this paper is to provide a real time simulation of a randomly distributed mobile ad-hoc networks under Shadowing propagation model in order to compute the Throughput and the Packet Delivery Ratio with different mobility scenarios and different network loads.

This paper is organized as follows. First, Section II gives a review study on related work. Section III describes the system implemented. Section IV demonstrates the system design with simulations and finally conclude the paper in Section V.

2. RELATED WORK

In [7], Mohand and Louiza introduced an extension of the Bianchi's Markov chain model with Packet Fragmentation Mechanism (PFM) and the Packet Error Rate (PER). A mathematical model was presented to compute the overall throughput and the mean response time of the IEEE 802.11b and the results were validated by simulation. The authors focused on how to improve the performance achieved in IEEE 802.11b Distributed Coordination Function (DCF) network by using the packet fragmentation mechanism under the assumption of non- ideal channel and infinite load conditions. The introduced analytical results have been carried out for different Bit Error Rate (BER) values, packet lengths, and network sizes and data rates.

In [6],Guowang, Ye and Ananthram presented a multistage channel aware aloha scheme for the contention period that allows users with relatively better channel states to have a higher probability of contention success while assuring fairness among users. The authors were able to analytically prove that the proposed scheme resolves network contention and attains throughput close to that of centralized schedulers. The scheme is also robust to any uncertainty in channel estimation. Simulation results were carried out to demonstrate that the proposed scheme significantly improves network performance in comparison to other existing scheme such as 802.11 RTS/CTS scheme.

In [3],Abderrahim and Abderrezak introduced a scheme called Relative Fairness and Optimized Throughput (RFOT) that ensures fairness and allows each node to adapt its transmission rate and contention window size to channel quality. The propsed scheme was validated using analytical model based on a 3-dimension Markov chain. The authors were able to prove that their scheme achieved fairness without comprising the throughput even under high load network.

Through the reviews presented previously, it is obvious that many researchers were interested to analyse and enhance the performance of IEEE 802.11.

3. SYSTEM DESCRIPTION

The performance of wireless network is dependent on medium access control (MAC) protocol used. Carrier sense multiple access (CSMA) is widely used due to its scalability and simplicity. CSMA is prone to hidden terminal problem especially in ad-hoc networks where there is a direct communication between nodes. Hidden nodes result in collision of packets which leads to degradation in network performance. As a solution to such problem, MACAW mechanism was introduced. [8]

In MACAW mechanism; when a node attempts to transmit data, it first sends a short control packet called Request To Send (RTS) to the receiver. The RTS packet includes the source, destination and the duration of the whole data transmission. This packet is received by all users that lie within the transmitter's range. Every node upon receiving the RTS sets it's virtual carrier sense indicator called Network Allocation Vector (NAV). The NAV has a time value that represents the duration up to which the wireless medium is expected to be busy because of an ongoing transmission. Therefore; every packet has the duration information for the remainder of the message. Every node overhearing a packet continuously updates its own NAV. After waiting for Short Inter-frame Spacing (SIFS), the receiver answers with a response control packet called Clear To Send (CTS).[8]

All users receiving either RTS or CTS automatically set their NAV for the given duration and then use it together with the Physical carrier sense to sense the medium. Once the RTS packet has been sent and CTS packet has been received successfully, all nodes within receiving distance from the sender and from the receiver are informed that the medium is reserved for one sender exclusively.[3] After waiting for SIFS, the sender starts data packet transmission. Upon receiving the data the receiver waits an additional SIFS and then sends and an ACK. After this four handshake process the transmission is over and the NAV in each node is marked as free and the process can repeat again.

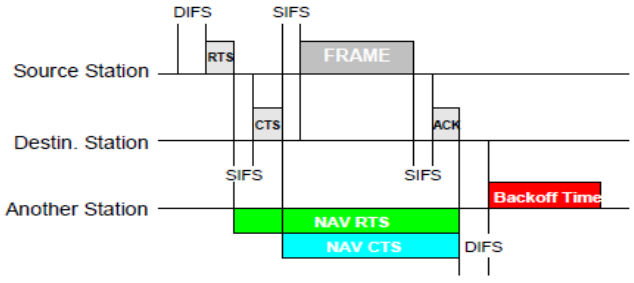


Fig.1. MACAW mechanism

Though; MACAW mechanism ensured packet delivery through the addition of the RTS/CTS packets, mobility of users was still a major issue that questioned such efficiency. Therefore, Mobility management becomes an interesting area of research as the mobile user should benefit from smooth handoff and low packet drops. It is considered a key participant in efficiently delivering data to mobile users. The main use of mobility management is to track, store and update the location information of the mobile user that move from one place to another within a certain coverage area of single wireless mobile ad-hoc network or within multiple wireless mobile ad-hoc networks in order to allocate the right amount of resources.

Fading is a vital part of a wireless communication design. Fading can be defined as a rapid fluctuation of the received signal strength over short time intervals. As the signal propagates through the wireless medium, it is faced with several obstructions, for example buildings, walls and other objects. These so called physical obstructions make the transmitted signal encounters signal attenuation. Therefore, Shadow fading can be defined as the variation in the received signal power due to these obstructions. Shadow fading is a fading that occurs on a large scale. In general, the variation in the received power due to shadow fading follows a Gaussian distribution.

$$\textbf{From Equation} P_{r=}\emptyset_r a = \left(\frac{P_t}{4\pi d^2}\right) a$$

WhereP_t is the radiated power from an isotropic transmitter, ais the effective area of the receiving antenna, P_r is the captured power and finally d is the distance where the receiver is located. Even though d is the same in the previous equation, the received power still varies since some locations face greater shadow fading than do others. In order to resolve this issue, the transmitted power should be increased to compensate for the shadow fading effect.

In our wireless system design, we were able to simulate the MACAW mechanism through designing a network with randomly distributed mobile users under Shadowing propagation model. Each user has a 40 m range in a 100 m square area as shown in figure

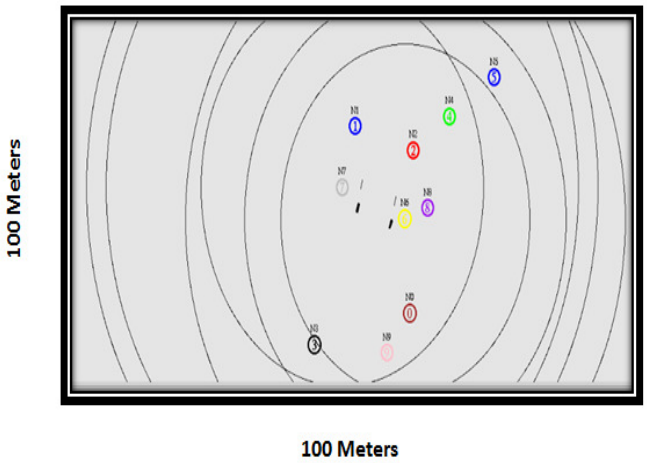


Fig.2. Randomly Distributed Mobile Users

4. SIMULATION RESULTS

In the network research field, deploying a complete test beds that are composed of multiple networked computers, routers and data links to endorse and substantiate a certain network protocol or a specific network algorithm are extremely expensive. Therefore, the use of network simulators in such cases saves a lot of money and time in achieving this task. In this section, we demonstrate the performance of MACAW mechanism using Network Simulator Tool (NS2).[4]

NS2 is considered one of the most popular open source network simulator tool used among researchers.[9] It is an object oriented, discrete event driven network simulator that was developed at UC Berkeley as a part of Virtual Internet Test beds (VINT) project. Ns2 uses two languages C++ and OTCL (TCL script language with Object-oriented extension developed at MIT). [4] Therefore, the presence of these two languages has proven to be more effective. A simplified user's overview of Ns2 is shown below.

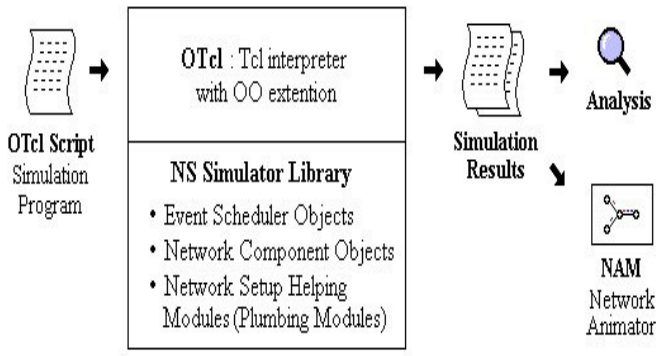


Fig.3. A Simplified User's Overview of Ns2

Simulation results are stored as trace files in Ns2 and a visual overview of the network is displayed using the Network Animator (NAM). A flow chart that highlights the flow of events for a Tcl script file runs in Ns2.

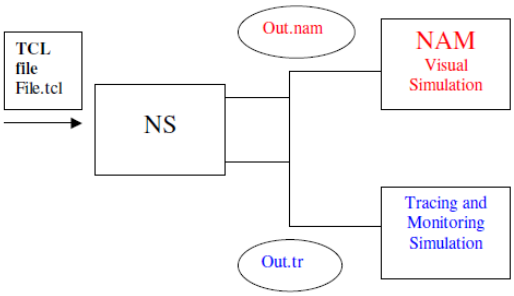


Fig.4. Flow of events for a Tcl file Run in Ns2

The performance of MACAW mechanism is determined by certain parameters such as Throughput and Packet Delivery Ratio (PDR).

Throughput is the measure of the actual data sent per unit time across a network in the real world.

$$Throughput = \left(\frac{Total\ number\ of\ Bytes\ received}{Total\ Time\ duration\ of\ each\ packet}\right)$$

PDR is the ratio of the data packets successfully delivered to the destination to the packets generated by the sender. [2]

$$PDR = \left(\frac{Received\ Packets}{Sent\ Packets}\right) * 100\%$$

Consider the network topology given in Fig.2.In each simulation trial, mobile users are randomly distributed under Shadowing propagation model in a 100m square area. Each user has a 40m transmission range and sender-destination pair is selected randomly among users for data transmission. The different parameters that were used to set up our network design are illustrated by both Table.I. and Table.II.

TABLE.I Simulation Setup Parameters

Parameters	Values
Area	100x100m^2
Transmission Range	40m
Number of randomly distributed users	5,10,15,20 users
Physical/MAC layer	IEEE 802.11b at 1Mbps
Mobility Model	Random way point with pause time
Pause Time	2.0s
Maximum Mobility Speed	10m/s
Packet size	512 Bytes
Traffic Type	TCP(Transmission Control Protocol)
Simulation Time	200s
Number of Simulation Runs	30

Fig.5. investigates the influence of network load on Throughput; we run simulation with 5, 10, 15, and 20 randomly distributed mobile users. For each case, we run 30 trials for 200s; each packet has a size of 512 bytes. We observe that network designed with 5 users had the least throughput of 0.3608Mb/s. Upon increasing the number of users to 10, a significant increase in the throughput of 0.4149 Mb/s was noticed. A further increase in the network load where the numbers of users were increased to 15 users led to a slight decrease in throughput value that reached 0.4148Mb/s. Finally, the maximum throughput of 0.4194 Mb/s was achieved at a heavier network load of 20 Users.

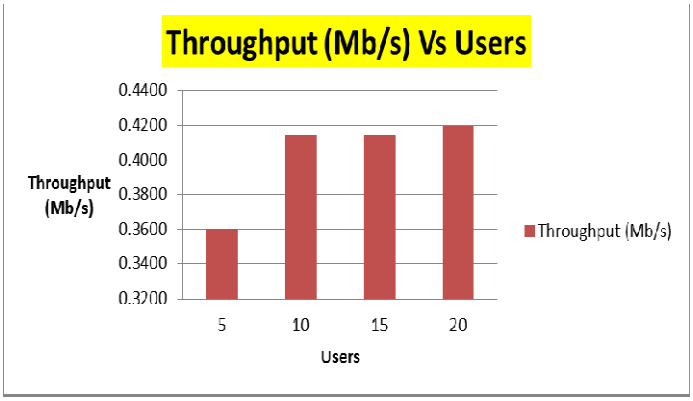


Fig.5.Average Throughput versus Randomly Distributed Mobile Users

Fig.6. shows the impact of network load on the Packet Delivery Ratio. In order to verify this impact, we run simulation with 5, 10, 15, and 20 randomly distributed mobile users. For each case, 30 trials were carried out for 200s, we see that, network designed with the least number of users had the least PDR of 91.9238%. As we increase the number of users to 10 users, a significant rise in the PDR value that reached 93.0086 %. A noticeable drop in the PDR value of 92.7957% took place when the number of users was increased to 15 users and this reflects the drop in throughput that took place in Fig.5. Finally, the maximum PDR was reached at 93.0298% when the maximum number of users was used.

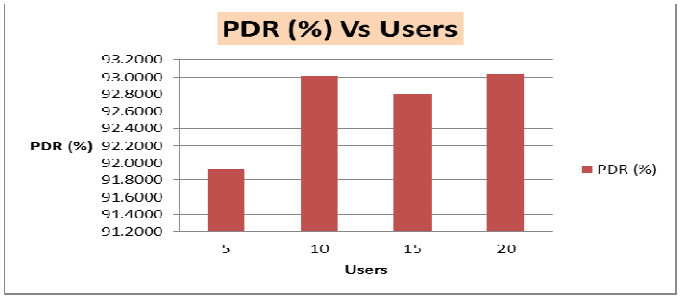


Fig.6.Average PDR versus Randomly Distributed Mobile Users

Fig.7. shows how mobility of nodes has a major impact on the overall network throughput. We run simulation with maximum speed of 5, 10, 15, and 20m/s. For each case, we compute the average speed of nodes and run our simulation for 30 trials, each with duration time of 200s. Fig.7. shows that at average speed of 2.5m/s, a maximum throughput of 0.4257Mb/s was achieved. A further increase in the average speed of nodes led to a drop in the overall network throughput that has reached a value of 0.4081Mb/s at 5.0m/s. A slight increase in the overall network throughput value of 0.4097Mb/s was noticed at average speed of 7.5m/s. Finally, the minimum throughput of 0.4007Mb/s took place at average speed of 10m/s.

TABLE.II Simulation Setup Parameters

Parameters	Values
Area	100x100 m^2
Transmission Range	40m
Number of randomly distributed users	10 users
Physical/MAC layer	IEEE 802.11b at 1Mbps
Mobility Model	Random way point with pause time
Pause Time	2.0s
Maximum Mobility Speed	5, 10, 15, 20m/s
Packet size	512 Bytes
Traffic Type	TCP(Transmission Control Protocol)
Simulation Time	200s
Number of Simulation Runs	30

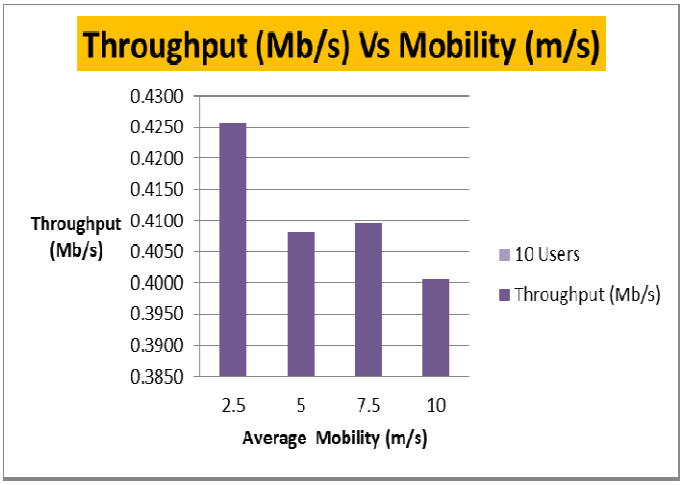


Fig.7. Average Throughput Versus Average Mobility

Fog.8.shows the effect of mobility of nodes on the PDR.We run simulation with maximum speed of 5, 10, 15, and 20m/s. For each case, we compute the average speed of nodes and run our simulation for 30 trials, each with duration time of 200s shows that at average speed of 2.5m/s, the PDR reached its maximum value of 94.0059%. A noticeable drop in the PDR of 92.7323% took place at average speed of 5m/s. A slight increase in the PDR of 92.8258% at average speed 7.5m/s. The least PDR took place at average speed of 10m/s.

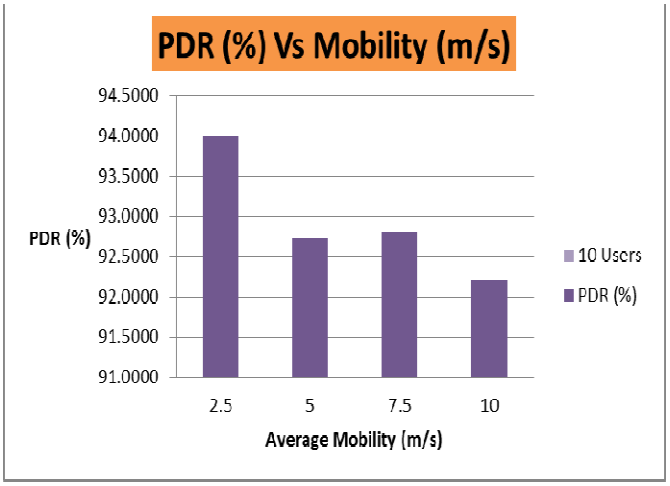


Fig.8. Average PDR versus Average Mobility

5. CONCLUSION

The degree of variability of the channel state of the wireless networks is quite high in comparison to wired networks. This variability is a result of several factors, one of which is the mobility of nodes. In this paper, a real time simulation was carried out with different mobility scenarios and different network loads to investigate the effect of such factors on the overall network performance. Throughput and Packet Delivery Ratio were computed to validate our simulation.

REFERENCES

[1] I.F. Akyildiz,J.S.M. Ho, and W. Wang"Mobility Management in Next-Generation Wireless Systems,"PROCEEDINGS OF THE IEEE, vol. 87, August 1999.

[2] N.P. Makwana, S.K. Vithalani, and J.D. Dhanesha, "Intrusion Detection-Watchdog: For Secure AODV Routing Protocol in VANET," International Journal of Engineering Trends and Technology, vol.4, May 2013.

[3] A. Benslimane, A. Rachedi, "Rate Adaptation scheme For IEEE 802.11-based MANETs," Journal of Network and Computer Applications, vol. 39, March 2014

[4] S.Siraj,A.K. Gupta, and R. Badgujar "Network Simulation Tools Survey,"International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, June 2014.

[5] G. Anastasi, M. Conti, E. Gregori, "IEEE 802.11 AdHoc Networks, Protocols, Performance and Open Issues,".

[6] G. Miao, Y.G. Li, and A. Swami, " Channel-Aware Distributed Medium Access Control ," IEEE/ACM TRANSACTIONS ON NETWORKING,vol.20, August 2012.

[7] M. Yazid, L.B. Medjkoune, D. Aı¨ssani, and L.Z. Khodja, "Analytical analysis of applying packet fragmentation mechanism on IEEE 802.11b DCF network in non ideal channel with infinite load conditions," Springer, October 2013

[8] P. Srivastava, and D. Singh, "A Survey on Modified RTS/CTS Mechanism,"International Journal of Computer Networks and Wireless Communications, vol. 3, February 2013.

[9] http://www.isi.edu/nsnam

[10] N. Sah, N.R. Prakash, and D. Bagai, "Analysis of Bandwidth Utilization for Wireless Mesh Networks,"vol.4, October 2013.

INTENTIONAL BLANK

TOWARDS A SOLUTION FOR INTEROPERABILITY OF SMARTHOMES DEVICES

Héldon José O. Albuquerque[1] and Gibeon S. Aquino Junior[1]

[1]Department of Informatics and Applied Mathematics-DIMAP Federal University of Rio Grande do Norte, Natal, Brazil

ABSTRACT

During the recent years, we can observe how mobile devices entered in the lives of people, becoming their main personal assistants and helping in various daily tasks. However not just mobile devices have evolved. And others electronic devices we use every day also experienced changes that have become smarter as the example of home devices. All these devices interconnected in the same environment or the same network, making use of services and exchange information with other devices, characterize a smart environment. smarthomes are special class of such environments and, increasingly, has become a scene with a variety of heterogeneous devices. However, due to the rapid progress of technology and the rise of a large number of heterogeneous devices, a variety of independent communication protocols were created, establishing a complex scenario to ensure interoperability between them. In this context, the main objective of this paper is to analyze the state of the art with an emphasis on the interoperability of current Mobile Devices with other digital devices present in a smarthome. Thus, it will be defined which protocols are currently most commonly used for this purpose, what are the current ongoing projects, what are the limitations of the solutions found in the research and, finally, it will be proposed an alternative solution for interoperability between the devices at a smarthome..

KEYWORDS

smarthome, systematic review, proxy, connectivity, communication protocol, interoperability, mobile devices.

1. INTRODUCTION

In 2005, [15] assumed that mobile computing would, in a not so distant future, replace computing based on "desktop computers", which were, on the other hand, already an improvement compared to the "mainframe computers". This supposition, according to [11, 13, 18], actually became true nowadays. Furthermore, they indicate that ubiquitous computing is the natural successor of mobile computing and, therefore, a reality.

Mobile devices began to emerge in the 80s [3], however, it was only in early 2000 that the first devices with a considerable processing capacity appeared [7]. Among these were smart-phones and tablets. These devices gained more notability than personal computers because of the mobility that they allow and provide.

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP- 2014
pp. 261–272, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4423

Over the past few years, mobile devices have become essential items in daily life. They became everybody's main personal assistants, used as a tool to aid in various tasks of day to day life.

However, mobile devices are not the only ones that have progressed. Kaed et al [11], points out that other devices and electronics that were not mobile or portable underwent changes that made them more "smart". An example of such are appliances like televisions that display internet content, which have the ability to connect with other devices to allow audio and video transmissions. There are refrigerators that make shopping lists, air conditioners and lights that are controlled over the Internet, among other examples of smart devices that function without mobility. All of these devices interconnected in the same environment, or on the same network, and inter-acting with other devices features an intelligent environment [5].

In September 2003, "Housing Learning & Improvement Network" created a project called DTI Smart Homes Project, offered by INTERLEK. This work is the first definition of an intelligent environment. In particular, there is a smarthome, also known as smart house, automated home or home automation [20].

A smarthome is an environment with a diversity of devices that use heterogeneous communication protocols, and in which the interoperability between them is a critical part to their contribution, but also one of the main problems found [1]-[5]-[18]. Most devices are independent and offer specific protocols so that they can be discovered in a home network.

In this context, the main objective of this paper is to analyze the state of the art with an emphasis on the interoperability of current Mobile Devices with other digital devices present in a smarthome. Thus, it will be defined which protocols are currently most commonly used for this purpose, what are the current ongoing projects, what are the limitations of the solutions found in the research and, finally, it will be proposed an alternative solution for interoperability between the devices at a smarthome.

The remainder of this paper is organized as follows: Section 2 defines the systematic review to find out the state of the art protocols involving this context smarthome. Section 3 defines the work related to the objective of this research. Section 4 highlights potential problems with the main conclusions. In Section 5, a generic suggestion will be done to effectively solve the problem that was addressed. Finally, in Section 6, the final considerations resulting from this research will be presented.

2. SYSTEMATIC REVIEW

To perform the investigation proposed in this work it was used as methodology, a systematic review, in order to ensure a theoretical and practical background in the state of art of the proposed theme. Second [14], "a systematic review is a form to identify, evaluate and interpret all the available and relevant research to a specific research question, or topic area, or interest phenomenon".

The main goal of a systematic review is to integrate the information of a number of studies performed separately about a specific topic or issue, with the purpose of revelling conflicting and/ or coincident results, as well as to identify themes that need evidences and help in the orientation to investigations for future work [10].

2.1. Systematic Review Protocol

The Systematic Review Protocol defines the rules and steps to be adopted to the performance of the research in the state of the art. In the following subsections the principal steps that constitute the protocol of elaboration of this systematic review are briefly described.

2.1.1. Objectives of the study and Research questions

Based on the proposed theme, the objective for this systematic review is **to perform an analysis of the state of the art of interoperability of the current Mobile Devices with the digital devices of a smarthome**. In particular, it is intended to investigate the connectivity protocols and technologies and the developed and proposed solutions in this environment.

Based on the established objective and following the recommendations approached in [14], it was prepared the following research questions:

- **RQ1** - What are the protocols and technologies that permit the connectivity among the current Mobile Devices and the devices of a smarthome?

2.1.2. Search and selection of the articles

Approaching the keywords contained in the research questions, contextualized to the ambience of mobile computing, Ubiquitous and Pervasive, it was defined the following string of search.

- (((Ubiquitous OR Pervasive) AND Computing) OR Internet of Things) AND ((Mobile Device) AND (Interoperability OR Connectivity OR Communication OR integration) AND (smarthome))

Articles that are available only on a paid form, along with articles that have more than ten years from the date of publication, were discarded. At last, the search vehicles used for this research were:

- ACM Digital Library;
- IEEEXplore Digital Library

2.1.3. Search and selection of the articles

According to [14], the form to evaluate and extract data from the research is defined in the Execution phase. In this one, the tests of execution are performed, that is, some of the publications based on the selection procedure of the review protocol are chosen. During this step, the primary studies are identified, selected and evaluated regarding the selection procedure.

Table I exposes the quantity of references retrieved according to the search vehicles, the quantity of articles selected for the reading and the quantity of articles accepted regarding the selection procedure. To avoid an exhaustive analysis e based on the fact that the results are ordered by relevance, the maximum of the first 50 works returned in the search in the vehicles were selected.

Table 1. Execution Table

Vehicle	Returnees	Selected	Accepted
ACM	46	46	9
IEEE	97	50	14

2.1.4. Information Extraction of the Systematic Review.

The main goal of the protocols of interconnection in a smarthome is to provide flexibility in information sharing, in a transparent manner among the devices, being mobile or electronic, and especially for the final user [17].

The connectivity protocols facilitate the interaction among devices, electronics and computers, dismissing the problem with the configuration of the entrance of a new component in the ambience and, consequently, freeing the user to do it. Thus, a user without technical knowledge will be able to monitor and control equipment's such as fridges, air conditioner, stove, lights, windows, doors, TVs, ventilators, printers and other computing devices, simply from another device [4, 11, 13, 7].

Due to the accelerated advancement of technologies and the emergence of a great quantity of heterogeneous devices in the same smarthome, a variety of protocols of independent communication were created [1, 11, 23], establishing a complex scenario of ensuring the interoperability among them.

This section presents the result of the research about the state of the art of the protocols of communication and connectivity in a smarthome, answering the research question **RQ1**, defined in section 2.

Table 1 presents the protocols used in the 23 articles analysed. It is noteworthy that, some of these works just mentioned protocols capable to manager devices present in a smarthome, but didn't use them in their researches. For our analyses, we consider only the protocols directly used in the research works.

Table 2. Connectivity Protocols

Protocols	References
UPnP	[1]-[5]-[6]-[8]-[9]-[11]-[12]-[13]-[16] [18]-[21]-[23]-[24]
DLNA	[2]-[5]-[8]-[17]-[19]-[21]
ZigBee	[13]-[17]-[18]-[23]-[25]
DPWS	[11]-[12]
JINI	[1]
Bluetooth SDP	[9]

The Universal Plug and Play (UPnP) was initialized by Microsoft in the year 1999 as a model of network architecture that provides connectivity ad-hoc among devices in a distributed manner, transparent, independent of driver or platform and with no need of any kind of configuration. One of the main characteristics of the UPnP, is the facility to provide services in the network, owning a protocol of device discovery and, consequently, of services. A main limitation of the

UPnP is that its structure is limited to an only network LAN. Companies as Nokia, Intel and HP, collaborated with this protocol, emerging from this partnership the UPnP Forum[1].

The Digital Living Network Alliance[2] (DLNA) is a standard established in 2003 by SONY and adopted by many companies in the industry. It is a connectivity protocol which has as objective to promote the interoperability among equipment's and electronic devices, mobile devices and personal computers. This builds upon the UPnP protocol, leveraging the power of its discovery, entreated, DLNA define particular characteristics for the exclusive use of its protocol.

The ZigBee, is a wireless network protocol developed by Alliance ZigBe[3] with techniques of small energy consumption wireless communication and short distance frequency. The main characteristics of ZigBee are the low transmission of data and the low cost of the devices. One of the projects developed by Alliance ZigBee is the ZigBee Home Automation, which offers an interoperability standard for domestic and digital equipment's, and devices with platforms of development, allowing, this way, the integration with devices in other ZigBee networks.

The Device Profile for Web Service (DPWS) emerged in 2004 and was acknowledged and standardized in 2008. Maintained by OASIS[4], the DPWS is a stack of protocols that define a minimum set of functionalities so that resources limited devices can adopt Web Service (WS) standard. Its main characteristics are the event notification, message exchange, discovery service and description of services, developing mechanisms of high level communication for interoperability among devices present in a smarthome.

The JINI[5] protocol was created in 1998 and maintained by Sun Microsystem. As well as the protocol UPnP, DPWS and DLNA, the JINI is a protocol that enables devices to connect to each other and share resources. The devices involved can be, from personal computers to mobile and electronic devices. However, it is limited to the JAVA platform, can be used in any mobile device that uses this platform.

At last, the protocol Bluetooth Service Discovery Protocol (BSDP) is used to allow that, devices, through the Bluetooth technology, can discover which services other devices bear or provide. The BSDP has as basis the protocol Service discovery Protocol (SDP), which is the same used by UPnP and DLNA to discover devices in a network. The BSDP is maintained by the organization Bluetooth SIG[6].

Some of the analysed articles use a combination of proto-cols in their answers in order to create a middleware among them, as the example of [8]. In this work, it is proposed a module for transparent replication of a DIGITAL TV in networks based on UPnP and DLNA protocols for a smarthome. Devices that endure only the UPnP protocol, can receive information from a DIGITAL TV that only uses the DLNA protocol and vice versa.

As can be seen in the graphic of Figure 1 of the found protocols, the UPnP stands out in relation to the others, in other words, 46% of the articles prefer to use the UPnP protocol to perform the communication among the Mobiles Devices with the devices in a smarthome. Coming in second the DLNA protocol with 26%, and in third, the protocol ZigBee with 18%.

---

[1] http://www.upnp.org
[2] http://www.dlna.org
[3] http://www.zigbee.org
[4] www.oasis-open.org
[5] http://www.jini.org
[6] https://www.bluetooth.org

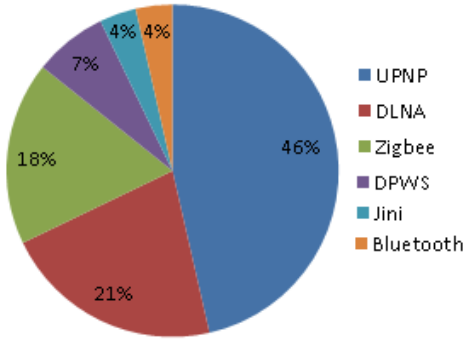


Figure 1. Connectivity Protocols.

The UPnP protocol stands out in the selected works for being an open standard and mainly for being based on the IP protocol, becoming a compatible protocol with any equipment turned to the network connection [23]. It has a light architecture that allows its use either for high technology devices, such as the micro-controllers, being able to be used to perform simple operation of turn on-turn off (Light- UPnP), or perform procedure to make specific decisions (Presence- UPnP) [16]. Besides, it is a protocol that operates using the peer-to-peer technology, not requiring any equipment or technology present among the devices participants to integrate them. At last, one the main reason for its use is that it is protocol based on the Plug and Play model, facilitating the installation and configuration of devices [6].

Examples of other connectivity protocols used in a smarthome, X10 [13], HAVI [22], IGRS [11], Salutation [2], GENA [16], Z-Ware [13], Ninja [1], were also mentioned in the articles, although they were not used with main focus.

3. RELATED WORKS

The variety of functionalities desired in a smarthome is increasing the complexity of the systems involved in this area. This complexity has motivated research groups to discover solutions in this context. One of the well mentioned and discussed problem is the one in intercommunication among devices [24]. With the objective of answering the Research Question **RQ2** proposed in section 2, the research projects reported in the articles were evaluated.

In [18], it is developed a bridging protocol, for communication among the protocols UPnP and ZigBee. To the development of this bridge, it was also used the COAP protocol (Constrained Application Protocol) which is used in electronic devices. This solution will provide flexibility in communication among devices with the UPnP and ZigBee protocol, making their particularities transparent among them.

[11] has developed the middleware INSIGHT, with capacity to provide the interoperability and services management in a smarthome over the UPnP and DPWS protocols.

In [25], it is presented the InfoPods system. An open architecture composed by a controller based on the ZigBee protocol, allowing that any device which uses this protocol can control the electronic devices belonging to a smarthome.

Besides these ones, other projects were also identified in a direct manner, as the articles were read and mentioned research works that were not present in the articles resulting from the execution of the search String. Among these projects, we can detach the followings:

The Service Bus Device (SBD) is a middleware which has as function to integrate heterogeneous devices in a Ubiquitous environment. This project was developed in a portable platform with the purpose of being executed in any kind of platform, being mobile or electronic. It has as the main objective to create a communication bus so that the equipment's with specific technologies, like the RFID , Bluetooth, Wi-fi, can offer services in the network. This one uses the DPWS protocol to perform the interoperability among the devices present in the ambience.

The Accessing We-based Applications on Consumer De-vices project (Web4CE) defines a network architecture that allows the devices present in a smarthome to directly access the web, without depending on other technology for that, being possible to access its individual configurations. This project uses the DLNA protocol for the interoperability of its network with other devices.

The MavHome is a project of a smarthome, developed by the Texas University. Its objective in to make the ambiences adaptive, that notices the state of the house with the help of sensors. In it, intelligent devices control the ambience ensuring the user's comfort by the recognition of activity, using video cameras and sensors in general. These devices process data to acknowledge what the users are doing, therefore, can take some initiatives to foresee what the user would do then, helping him somehow and minimizing the effort of the user.

Other proposes, like [12, 17, 18, 21], have as objective to allow that distinct protocols can interconnect to each other in a smarthome, avoiding the particularities of each protocol. The main motivation for the majority of the accepted works in this research is to develop middleware platforms of integration, with the purpose of interconnecting all kinds of devices with different communication technologies in the same ambience.

Although the solution with the use of middleware's being accepted, works like [16, 19, 25] defends a solution with the use of an only connectivity protocol in a smarthome. This approach would decrease the complexity offered to the integration of different kinds of protocols and would be certainly more performativity and economic.

The fact is that in the current estate of technologies development, an integrative solution, based on middleware's, is the faster and flexible answer to ensure the interoperability among the devices. Although, in a long-term vision, the most appropriate way would be that only one standard predominate. This standard would be evolved enough to incorporate the characteristics and advantages of the others, besides being open and maintained by a large number of representatives of industry and academy. Probably, the UPnP is the protocol nearest this reality, but still needs to evolve a lot from the technical point of view to be considered substitute of the others.

4. LIMITATIONS FOUNDING SOLUTIONS

An intelligent environment contains many interconnected heterogeneous devices. They use services or exchange in-formation in a transparent and dynamic way [5]. According to [12] and [17], the ability to be automatically recognized in an environment is fundamental to the achievement of an intelligent environment and is also the most fascinating role of the protocols identified in this paper. Moreover, the diversity of these protocols and technologies that are available nowadays makes the seamless interoperability between devices inhibitive and, thereafter, prevents the full implementation of smart home environments.

As an overview of the architectural behaviour of the integrated protocols in a smarthome, we can identify two common technical features between the protocols.

First, these protocols use the communication protocol based on TCP/IP, or, in other words, any device that supports a protocol for a smarthome can be integrated using any means of communication, whether it is Ethernet, Wi-Fi, etc. However, it is necessary that the machine is able to acquire an IP address.

Second, the components of a smarthome [11] can be classified in two different ways: controllable devices – those who provide the services that the devices offer; and control point – those responsible for using the services provided by the devices.

Also according to [11], other components may arise de-pending on the protocol, but in general, these protocols are based on the client-server concept, where the client is known as Control Point and the server is known as the Device.

Although these protocols have similar high level goals, they are composed of very different architectures. Each machine using the discovery of service will use only one of these protocols; for example, the UPnP control point will only find in a network the UPnP devices, and so it happens for the other protocols. This means that clients and services using different technologies will not be able to cooperate. Since it is likely that several protocols will be widely used in the environment, there is a need for a structure of interoperability that allows clients and services written using different protocols to cooperate.

Until now, it was realized that there are two possibilities to integrate a new device with a protocol not yet supported by the environment, as shown in Figure 2.

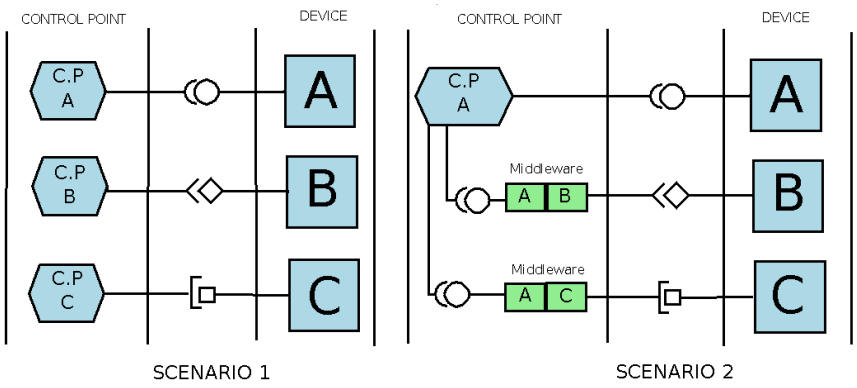


Figure 2. Scenarios insertion of a new protocol

It is observed in Scenario 1 that, in the integration of multiple devices with different protocols in a smarthome, it is necessary to add control points to interact with the services provided by the device. For each device with protocols A, B and C, control points C.P-A, C.P-B and C.P-C must be added to the environment. The main problems related to this scenario is the amount of control points managed by the user, creating a non-scalable solution, since the main goal of a smarthome is to reduce the user's effort in controlling the devices in the environment; and the non-interaction between devices, since each device can only support a single protocol, and these protocols do not recognize each other in the same environment.

In Scenario 2, on the other hand, it is observed that the middleware's are solutions widely used for interoperability between different communication protocols in smarthomes. More concretely, the problem occurs when we add devices with connectivity protocol B and C in an environment that supports only protocol A. middleware's must be developed and integrated directly into the

control point A to interconnect the new protocols on the environment. In this case, we observed as main advantage the minimization of the overhead produced by the conversion of packages between the protocols, since there is more than one layer of middleware on a single control point. In contrast, the existing solutions tend to be intrusive in control points. In other words, to integrate any middleware layer in control points, which would be the most ideal solution for the reuse of existing control points, the direct access to their source code should be allowed. However, many existing solutions do not provide this access to the code.

However, [16] emphasizes that the diversity of middlewares present in a smart environment will provide, at some point, incompatibilities when they interconnect with each other. Moreover, such solutions produce a gradual increase in the complexity of the environment and these solutions found so far are directly linked to a change in the context of the control point.

5. PROPOSAL FOR A GENERIC SOLUTION

In this section, a proposal of interoperability solution between the protocols of a smarthome will be presented. The proposal is based on the introduction of a proxy service that will allow the interoperability of different protocols without changing any of the existing components in the environment. As can be seen in Figure 3, we will add a physical device to the communication infrastructure, which will make the translation of communication packages between protocols possible.

Thus, the control point A (PC-A) will be able to communicate normally with devices with protocol A and also with the other protocols in the environment through the proxy, without the need of any programmatic modification in its structure.

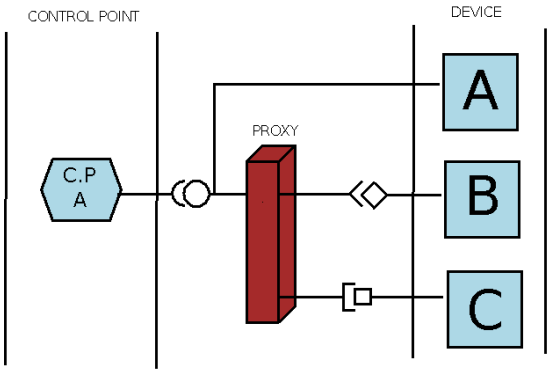


Figure 3. Proposal for the interoperability between the protocols

In this article, we will focus on the interoperability between UPnP and DLNA protocols because they are the most important protocols in the current scenario. Despite the high-level similarities between these protocols, since they both use consolidated and standardized internet protocols such as TCP/IP, HTTP, SOAP and others as their basis, the exchange of information of services provided by devices and the interoperability between them is not simple. By using the same language for data exchange of services provided by devices, each protocol uses an independent XML scheme, in other words, the XML scheme used by each protocol defines different validation rules, this being the main reason for the inability of these protocols to interact with each other.

As a way to assess the applicability of the proposed solution, a case study was implemented considering a scenario where already developed UPnP Control Points interact with DLNA devices existing in the environment.

This proposal essentially consists of "virtual services" that will be automatically instantiated by the proxy in response to the discovery of DLNA devices in the network. To achieve this goal, we need to "trick" the UPnP Control Point, making it believe that it is interacting with the services provided by an UPnP device, but which are in fact services provided by DLNA devices.

In the scenario above, the proxy creates virtual UPnP devices for each DLNA device found in the environment, recording the services provided by each device as illustrated in Figure 4. For each instance of service supported, virtual de-vices will be responsible to make a bridge between protocols, translating the XML file format of the virtual UPnP device's services to the DLNA and performing the interoperability between the protocols. When a DLNA device is removed from the network, the proxy will automatically destroy the corresponding virtual services.

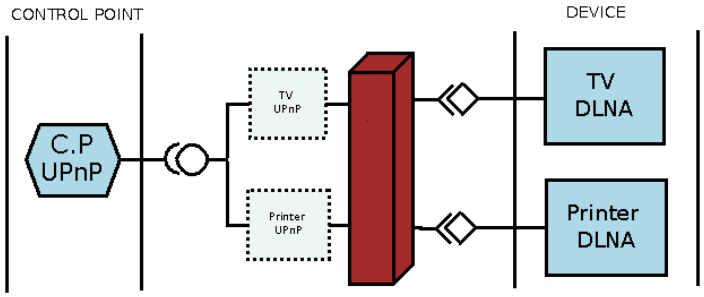


Figure 4. Proxy for interoperability between UPnP and DLNA protocol

To achieve this goal, the proxy architecture is divided into three basic components:

- **Discovery Module:** Provides basic functions of a DLNA Control Point, like the discovery of the services provided by devices in the environment and detection of change in the values of the state variables of the service provided. The first function is needed to transmit the commands to a DLNA service through the virtual UPnP device, while the second is responsible for updating the values of the variables of the virtual device;
- **Analysis Module:** Responsible for managing the other modules of the system, managing the services provided by devices, requesting the database of the UPnP devices the most adequate device to perform the translation between protocols, managing the state variables of the services provided by the Discovery Module and by the virtual UPnP devices;
- **Virtualization Module:** Component used to implement the virtual UPnP devices. This module is also responsible for ensuring that the same device is not duplicated when a state variable of a service is modified.

For each new device in the environment that is not sup-ported by proxy, a modest amount of code must be written to adapt the proxy for its recognition so UPnP/DLNA and its Control Points do not need any modification. One of the main goals of this "proxy" is the reduction of the effort in the use of already developed applications.

6. CONCLUSIONS

We present a proposal for the interoperability between the protocols of a smarthome and demonstrate the viability of the proposal by applying it to a context involving DLNA and UPnP protocols. Our proposal will introduce virtual services that allow DLNA customers to interact with UPnP Control Points already developed, aiming at the reuse of existing solutions in the market and, consequently, at a reduction of efforts to create or adapt existing middleware solutions. The efforts of this solution are important because, in the short term, with the lack of both academic and industry initiatives for the development of a standard that involves the field of smart environments, a number of technologies and protocols will compete for space, thus overloading an increasing number of independent protocols in an environment, making them increasingly heterogeneous.

REFERENCES

[1] Sameh, A.; El-Kharboutly, R., "Modeling Jini-UPnP Bridge using Rapide ADL,"Pervasive Services, 2004. ICPS 2004. IEEE/ACS International Conference on, vol., no., pp.237,237, 19-23 July 2004.

[2] S.V. Anand. A dlna framework for next gen mobile terminals connecting ims networks for human-centered digital home environment. In IP Multimedia Subsystem Architecture and Applications, 2007 International Conference on, pages 1–5, 2007.

[3] Daniel Barbar´ a. Mobile computing and databases-a survey. IEEE Trans. on Knowl. and Data Eng., 11(1):108– 117, January 2009.

[4] C. Beckel, H. Serfas, E. Zeeb, G. Moritz, F. Gola-towski, and D. Timmermann. Requirements for smart home applications and realization with ws4d-pipesbox. In Emerging Technologies Factory Automation (ETFA), 2011 IEEE 16th Conference on, pages 1–8, 2011.

[5] Wally Chen, Sy-Yen Kuo, and Han-Chieh Chao. Service integration with upnp agent for an ubiquitous home environment. Information Systems Frontiers, 11(5):483– 490, November 2009.

[6] R. Chowdhury, A. Arjona, J. Lindqvist, and A. Yla-Jaaski. Interconnecting multiple home networks services. In Telecommunications, 2008. ICT 2008. International Conference on, pages 1–7, 2008.

[7] M.N. Cortimiglia, A. Ghezzi, and F. Renga. Mobile applications and their delivery platforms. IT Professional, 13(5):51–56, 2011.

[8] Giliard Brito de Freitas and Cesar Augusto Camillo Teixeira. Ubiquitous services in home networks offered through digital tv. In Proceedings of the 2009 ACM symposium on Applied Computing, SAC '09, pages 1834– 1838, New York, NY, USA, 2009. ACM.

[9] A. Delphinanto, A. M J Koonen, M. E. Peeters, and F. T H Den Hartog. Proxying upnp service discovery and access to a non-ip bluetooth network on a mobile phone. In Communications and Vehicular Technology in the Benelux, 2007 14th IEEE Symposium on, pages 1–5, 2007.

[10] O. Dieste, M. Lopez, and F. Ramos. Formalizing a sys-tematic review updating process. In Software Engineering Research, Management and Applications, 2008. SERA '08. Sixth International Conference on, pages 143–150, 2008.

[11] Charbel El Kaed, Antonin Chazalet, Lo¨ ıc Petit, Yves Denneulin, Maxime Louvel, and François Gaël Ottogalli. Insight: interoperability and service management for the digital home. In Proceedings of the Middleware 2011 Industry Track Workshop, Middleware '11, pages 3:1– 3:6, New York, NY, USA, 2011. ACM.

[12] Charbel El Kaed, Yves Denneulin, and Francois-Ga¨ el Ottogalli. Dynamic service adaptation for plug and play device interoperability. In Proceedings of the 7th International Conference on Network and Services Management, CNSM '11, pages 46–55, Laxenburg, Austria, Austria, 2011. International Federation for Information Processing.

[13] Ji Eun Kim, G. Boulos, J. Yackovich, T. Barth, C. Beckel, and D. Mosse. Seamless integration of heterogeneous devices and access control in smart homes. pages 206– 213, 2012.

[14] B. Kitchenham, R. Pretorius, D. Budgen, Pearl B. O., M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering - a tertiary study. Inf. Softw. Technol., 52(8):792–805, August 2010.

[15] J. Krikke. T-engine: Japan's ubiquitous computing architecture is ready for prime time. Pervasive Computing, IEEE, 4(2):4–9, 2005.

[16] B. Kumar and M. Rahman. Mobility support for Universal Plug and Play (UPnP) devices using session initi-ation protocol (sip). In Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE, volume 2, pages 788–792, 2006.

[17] Chin-Feng Lai and Yueh-Min Huang. Context-aware multimedia streaming service for smarthome. In Proceedings of the International Conference on Mobile Technology, Applications, and Systems, Mobility '08, pages 106:1–106:5, New York, NY, USA, 2008.

[18] Jin Mitsugi, Shigeru Yonemura, Hisakazu Hada, and Tat-suya Inaba. Bridging upnp and zigbee with coap: protocol and its performance evaluation. In Proceedings of the workshop on Internet of Things and Service Platforms, IoTSP '11, pages 1:1–1:8, New York, NY, USA, 2011. ACM.

[19] K.J. Patel, S.V. Anand, and S.P. Sumant Kumart. A robust qos framework on android for effective media delivery to dlna enabled home gateway in smart home environment. In Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference on, pages 217–222, 2010.

[20] V. Ricquebourg, D. Menga, D. Durand, B. Marhic, L. Delahoche, and C. Loge. The smart home concept : our immediate future. In E-Learning in Industrial Electronics, 2006 1ST IEEE International Conference on, pages 23–28, 2006.

[21] C. Rus, K. Kontola, I.D.D. Curcio, and I. Defee. Mobile tv content to home wlan. Consumer Electronics, IEEE Transactions on, 54(3):1038–1041, 2008.

[22] A. Uribarren, J. Parra, R. Iglesias, J.P. Uribe, and D. Lopez-de Ipina. A middleware platform for application configuration, adaptation and interoperability. In Self-Adaptive and Self-Organizing Systems Workshops, 2008. SASOW 2008. Second IEEE International Confer-ence on, pages 162–167, 2008.

[23] K.I.-K. Wang, W.H. Abdulla, Z. Salcic, N. DeSouza, and S. Ramkumar. Multiagent control system with mobile ubiquitous platform for ambient intelligence. In Intelligent Environments, 2008 IET 4th International Confer-ence on, pages 1–7, 2008.

[24] E.U. Warriach, E. Kaldeli, J. Bresser, A. Lazovik, and M. Aiello. Heterogeneous device discovery framework for the smart homes. In GCC Conference and Exhibition (GCC), 2011 IEEE, pages 637–640, 2011.

[25] I.A. Zualkernan, A. R. Al-Ali, M.A. Jabbar, I. Zabalawi, and A. Wasfy. Infopods: Zigbee-based remote information monitoring devices for smart-homes. Consumer Electronics, IEEE Transactions on, 55(3):1221–1226, 2009.

AUTHOR INDEX