SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

ISSN: 1870-4069

RCS

Research in Computing Science

Vol. 63

Advances in Computing Science

E. A. Santos Camacho,
J. C. Chimal Eguía,
L. Cabrera Rivera,
E. Castillo Montiel, (Eds.)

Advances in Computing Science

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 63,** Noviembre, 2013. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. : 04-2005-121511550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 63**, November, 2013. Printing 500. The authors are responsible for the contents of their articles.

Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Volume 63
Volumen 63

Advances in Computing Science

Editores del volumen:
Volume Editors

Evelia Araceli Santos Camacho,
Juan Carlos Chimal Eguía,
Luis Cabrera Rivera,
Erandi Castillo Montiel, (Eds.)

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA
Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2013

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX and Periodica / Indexada en LATINDEX y Periódica

Printing: 500 / Tiraje: 500

Printed in Mexico / Impreso en México

Preface

The purpose of this volume is to present the most recent advance in selected areas of Computer Science, The works included in this volume were carefully selected by the editors on the basis of the blind reviewing process and the main criteria for selection were originality and technical quality. This issue of Research in Computing Science will be useful for researches and students working in the different areas of Computer Science, as well as, for all readers interested in enrich the knowledge of this document.

In total, we received 55 papers that were submitted for evaluation; each submitted paper was reviewed by 2 independent members of the editorial board or additional reviewers and the acceptance rate is 60%. This volume contains the revised version of 17 accepted papers, divided into two sections corresponding to the areas of: Artificial Intelligence, Communications and Computer Networks. The others papers will be published in the volume number 67 of the journal Research in Computing Science.

We would like express our gratitude to all people who help to elaborate this volume. First to the authors of the papers for the technical excellence of their works, that guarantees the quality of this publication. We also want to thanks the members of the editorial board for their hard work in evaluating and selecting the best's papers out of many submissions that we received. We express sincerely our gratitude to the Sociedad Mexicana de Inteligencia Artificial (SMIA) for their collaboration in the elaboration of this publication. Also we want to give special recognition to the Centro de Investigación en Computación of the Instituto Politécnico Nacional (CIC-IPN) for the facilities given to achieve the success in the publication of this volume. The submission, reviewing, and selection process was supported for free by the EasyChair system, www.EasyChair.org. Also we want to give special recognition to ORACLE, for its participation as a sponsor.

November 2013

Evelia Araceli Santos Camacho
Juan Carlos Chimal Eguía
Luis Cabrera Rivera
Erandi Castillo Montiel

Table of Contents

Índice

Artificial Intelligence

Communications & Computer Networks

Artificial Intelligence

Benchmark between Different Feature Extraction Methods Applied to Face Recognition

Giovanni Mejía, Mauricio Martinez, Roberto A. Vázquez

Universidad La Salle, Benjamin Franklin 47, Mexico City, Mexico.
rgmb@lasallistas.org.mx, mauricio.martinez@lasallistas.org.mx, ravem@lasallistas.org.mx

Abstract. Face identification systems rely on feature extraction from still images. There exist many techniques for this purpose, each one having its own inherent limitations and impact on the system's performance in terms of recognition rate. In this paper, we describe a facial recognition system that can be used with different signal preprocessors for feature extraction, and compare different well known approaches, including the use of neural networks. We also introduce the use of two -dimensional wavelets as feature extractors. Our results show a competitive performance. We conclude that wavelet based signal preprocessors constitute an interesting alternative for facial recognition, when combined with neural network classification.

Keywords. Face recognition, wavelet transform, DFT, DCT, PCA, DWT, neural networks, image classification, computer vision.

1 Introduction

The field of image processing and computer vision is an expanding area, which presents many unsolved problems. One of the relevant problems in engineering is adapting computer vision to specific systems that imply first of all to identify a problem that can be solved with automated artificial vision. It is first necessary to comprehend the fundamental capabilities and limitations of the image processing and computer vision systems, as well as the knowledge of the fundamental image processing techniques and algorithms.

Human beings have always developed different systems to confirm the identity of people with techniques like personal identification numbers, signatures, etc. But the applications of biometric techniques such as fingerprint and retina analysis were implemented until the 1980s [1]. In this work, we use the preprocessing of images applied to face recognition; nonetheless, the techniques exposed can be applied to the solution of similar problems like object recognition, crop recognition and other areas related to digital images.

A general statement to describe the problem we are studying could be formulated in an axiomatic manner: “Given a set of images, identify or verify a person on the

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 3–12
Paper Recived 16-08-2013 and Acepted 21-10-2013

scene using a given database of faces". We pretend to find the performance of each wavelet family. To achieve this, we have proposed the following particular objectives.

1. Evaluate the performance of the wavelet transform as a feature extraction technique with different classification systems of face images.
2. Explore the representation space of face images by the application of different wavelet families.
3. Develop distance classifiers (Euclidian, Mahalanobis and Manhattan) and neural networks to perform face recognition.
4. Use the previously proposed classification techniques with preprocessed images by wavelet families, a standard database will be used (ORL database) and their performance will be evaluated.
5. Different strategies of feature extraction will be compared like DFT (Discrete Fourier Transform), PCA (Principal Component Analysis), DCT (Discrete Cosine Transform), with the selected wavelet families.

In table 1 we present some investigations with relevant results on the face recognition area.

Table 1. Comparison of remarkable results on the state of the art.

Method	*Recognition Percentage*	*Characteristic*	*Ref*
PCA	85%	High Computational Cost	[2]
Kernel PCA	84%	High Computational Cost	[2]
Kernel PCA Gauss	85%	High Computational Cost	[2]
DCT Geometric Regions	97%	Hybrid Process	[3]
Wavelet + PCA	92%	Hybrid Process	[4]
Wavelet Sym + PCA	90%	Hybrid Process	[5]
Wavelet + DCT + PCA	95%	Hybrid Process	[6]

2 Feature extraction techniques

In this section, we describe the techniques used to develop this work. If the input data to an algorithm cannot be directly processed, we have to transform it to a reduced set of characteristics. This process is known as feature extraction. If the characteristics are chosen on a careful way, we can expect those to have relevant information from the input data, which allows to perform the desired actions using this reduced representation. The specific techniques used for this investigation are as follow.

Because we will be using static images we require a version of the Fourier transform that can handle this, that is why we used the two-dimensional ***Discrete Fourier Transform***, this is possible because the pixels have indexed bidimensional coordinates. Fourier Features are used as a representation instead of using the complete image. This is very useful to process a smaller quantity of characteristics [7].

The ***Principal Component Analysis*** is a feature extraction technique used to perform pattern recognition in a given data set, and to express that data set in a way that the similarities and differences are highlighted. It is a very useful tool to transform data and make it easier to be separated on different classes, which is equivalent to transform its representation space. This technique also allows a high data compression rate [8].

The ***Discrete Cosine Transform*** is a linear transformation that can express a finite sequence of speech data as a sum of cosine functions spanning different frequencies. After the original signal has been transformed, its DCT coefficients reflect the importance of the frequencies present on it, the first coefficient refers to the lower frequency on the signal and it usually contains most of the relevant information of the signal, and the last coefficient refers to the higher frequencies of the signal. For this work we used the two-dimensional DCT, which is possible because every pixel has bidimensional indexed coordinates [7].

As with the DFT, we have the ***Discrete Wavelet Transform***, which can be calculated at some point on the time – scale space, there is no need to know the values of the function for the axis at all times, since all we need is the function values on the point in which the value is different from zero. Consequently the evaluation of the wavelet transform can be done almost in real time. The wavelet families used on this research are Daubechies, Symlets, Coiflets, Biorthogonoal and rBiorthogonal.

The discrete values of the scalar parameter a and traslation parameter b are taken in a different manner. a will be taken as 2^{-s} and b as $k2^{-s}$, where $k, s \epsilon \mathbb{Z}$ with the values of a and b, the integral of the continuous wavelet transform is transformed:

$$W_{\psi}f(k2^{-s}, 2^{-s}) = 2^{s/2} \int_{-\infty}^{\infty} f(t)(2^{s}t - k)dt \qquad (1)$$

Now the function $f(t)$ must be discrete. By simplicity we assume the sampling rate is 1. In that case the expression can be written on this way:

$$W_{\psi}f(k2^{-s}, 2^{-s}) \approx 2^{s/2} \sum_{n} f(n)\psi(2^{s}n - k) \qquad (2)$$

To calculate the wavelet Transform of a function at some point on the time – scale space, it is not necessary to know the values of the function for the complete time axis. All we need to know is the function on that values of time in which the wavelet is different from zero. Consequently the evaluation of the wavelet Transform can be done almost in real time. [5] Each wavelet Transform depends of its wavelet or mother function represented by “ψ" [9].

3 Classifiers

As it has been mentioned, pattern recognition is a discipline with the objective of classifying objects in a finite number of classes or groups. Such objects can be audio files, pictures or any type of data that can or needs to be classified. The purpose of the classifiers in this work is to separate on different classes (also known as clusters) the groups of images of the desired database. We will use three distance classifiers which are Euclidian, Mahalanobis and Manhattan as well as a Backpropagation Neural Network.

Euclidian Distance refers to the ordinary distance between two points generated when they are joint by a straight line, it is based on the Pythagoras' theorem, and when it is used as a distance it becomes a metric [10].

Mahalanobis Distance measures the distance of two groups of objects based on its means. Groups must have the same number of characteristic features but no necessarily the same quantity of elements [11].

Manhattan Distance is a metric in which the distance between two points equals the sum of the differences of its coordinates. It is also known as distance 1, norm 1 or Manhattan distance. The distance between two vectors in a vector space is the sum of the distances of the projections of line segments between its coordinate axes. [11]

The ***Artificial Neural Networks*** are an information processing paradigm inspired on the human brain. The key element on this is its structure, the Artificial Neural Networks are composed of a number of processing elements or neurons which work together to resolve a specific problem. The ARN used on this work are based on the McCulloch & Pitts model. Inside of a neural network there exist a great number of connections between the neurons, because each neuron has synaptic weighs which simulate the neural interconnections of the brain [12].

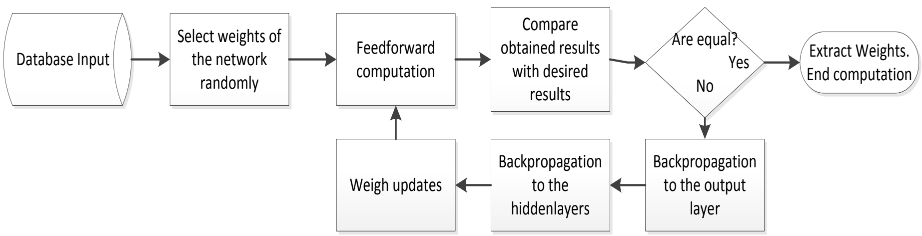


Fig. 1. Artificial Neural Network Algorithm

4 Proposed methodology

On this section we explain which tools were used to the development of this work, a general methodology is presented for wavelet feature extraction process and neural network classifier having created a benchmark based on classical feature extraction techniques (PCA, DCT and DWT). The methodology used in our research is summarized as follows:

1. Select database.
2. Apply different wavelet transforms to the database.
3. Develop and train recognition systems using a portion of the preprocessed database.

4. Test the systems with the remaining portion of the database.
5. Compare the results with the classical transforms (PCA, DCT and DWT).
6. Demonstrate based on results which wavelet transform presents the best performance.

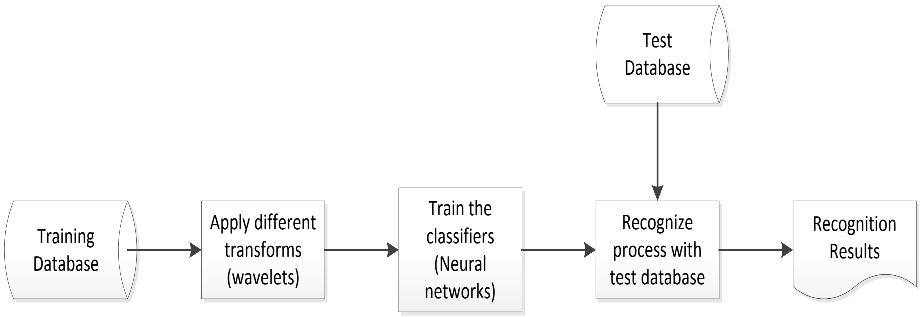


Fig. 2. General process for recognition systems.

This project is limited to implement offline face recognition on static images, with a standard database. All other implementations as a real system will be considered as future work. It should be noted that the intention of this research is to suggest a benchmark for the usage of different feature extraction techniques on images, and to test their performance with different classification systems. The main techniques used to establish the results of this work are described in the next two algorithms.

Wavelet algorithm

```
1. Obtain database specifications.
2. Obtain the list of images to transform.
3. Create the loading architecture of the images. In this
case we will use an array of three dimensions (image,
row, column).
4. Select wavelet and iteration depth.
5. Obtain the discrete wavelet transform of each image
and place it on the same architecture of three dimen-
sions.
6. If iterating, do it with the approximation coeffi-
cients.
7. Define which images will be used for training and
which will be used for testing.
8. Transform the images to row vectors.
9. Create two matrixes, the first one will contain the
vectors of the training images and the second will con-
tain the vectors of the test images.
```

Neural Network

```
1. Transform the input data on a matrix where rows are
the images and the columns are the features of the imag-
es.
2. Normalize the features between zero and one
3. Select training set.
4. Initialize weights and biases.
5. Execute in parallel the training procedure with a min-
imum of four processes with the same data but different
learning rate with the training set.
6. Iterate until the output is equal to the desired out-
put, while iterating, transfer the weights with best re-
sults to the other neural networks.
7. Insert test group to the Neural Network and classify.
Compare obtained results with desired.
```

5 Experimentation and results

This section has the objective to perform a series of experiments with the purpose of generating a benchmark between the classical feature extraction techniques (PCA, DCT, DFT) and wavelet transform families, all this is evaluated with distance classifiers and neural networks. First of all, all feature techniques were carried on to determine the quantity of features required for each technique according to a representation percentage or an iteration level. Then, using the necessary quantity of features, all distance classifiers were tested and we obtained a matrix of performance for each feature extraction process and classifier tested, based on this, the wavelets with the best results were picked and processed with a neural network classifier, with the obtained results, a specific wavelet is purposed to solve the face recognition problem with a specific database.

Our experiments have been carried out, using the ORL database. It has 400 images divided on 40 subjects with ten images each. Some images have been shot with different kinds of illumination, facial expressions, and face details like with or without sunglasses. All the images were shot with a black homogeneous background with the subjects on frontal position with soft lateral variations.

The experiments were carried using five images to train the classifiers and the five remaining to test the system. For each series, the training images were chosen randomly, and the remaining five will integrate the test group, the results reported are an average of 30 repetitions for each experiment.

First of all, the feature extraction processes DCT, DFT and PCA were used. The results showed that the quantity of characteristics to represent an image can vary from the 0.25% of the original to 14.91% depending on the technique and the percentage of representation needed.

Table 2. Quantity of features required to represent images according to the representation percentage of the image

	Quantity of features required to represent images					
	Quantity of features			*Percentage of features*		
%	DCT	DFT	PCA	DCT	DFT	PCA
75	55	55	26	0.53%	0.53%	0.25%
80	64	64	36	0.62%	0.62%	0.35%
85	74	73	52	0.72%	0.71%	0.50%
91	104	112	82	1.01%	1.09%	0.80%
92	122	141	89	1.18%	1.37%	0.86%
93	142	169	97	1.38%	1.64%	0.94%
94	168	214	106	1.63%	2.08%	1.03%
95	202	265	116	1.96%	2.57%	1.13%
96	259	365	127	2.51%	3.54%	1.23%
97	353	533	139	3.43%	5.17%	1.35%
98	535	819	155	5.19%	7.95%	1.50%
99	1035	1536	173	10.04%	14.91%	1.68%

Feature extraction with the different wavelet families can be done with 0.04% of features from the original image to 34.70%. The results with the test sets of images show that PCA method offers best percentage of face recognition (Table 3), the PCA method requires less quantity of features to train each classifier; DFT and DCT methods have high error percentages. To illustrate this, we can use the Manhattan classifier with data preprocessed by the PCA technique. With this, we were able to recognize 89% of the subjects from the database, from a representation level of 80% of the image. The disadvantage of this method is that it presents a computational complexity $O(n^2)$. For this reason, it is not functional for huge databases.

We carried out the wavelet experiments by using the DWT [9] for different wavelet families. From these experiments, we see that the methodology achieves great data compression and have less error percentage on the training step. We have established as benchmark PCA technique which has given a 6.5% of error in classification using only 173 features per image. The results according this benchmark go between a 56.5% of recognition error and 10% depending on the selected classifier and wavelet family.

We have selected the best results for each wavelet family based in the previous tests and using the classifier system ANN. In Table 4 we show the number of required features needed to represent each image and to achieve the smaller recognition errors presented in Table 5.

Table 3. Smaller error percentage obtained for each feature extraction technique using DWT with different wavelet families.

	DCT	*DFT*	*PCA*	*db 2,4*	*sym 2,4*	*coif 1,4*	*bior 2.2,4*	*Rbio 3.1,4*
Euclidian	47%	37.5%	9%	10%	10.5%	11%	10.5%	10%
Mahalanobis	78%	68%	12.5%	15%	24%	25%	15.5%	15%
Manhattan	47%	37%	6.5%	9.5%	12%	12%	10.5%	10%

Table 4. Number of features required by wavelet family to successfully represent an image.

Iter	*db2,4*	*sym2,4*	*coif1,4*	*bior2.2,4*	*rbio3.1,4*
4	0.699%	0.699%	1.068%	1.068%	0.699%

Table 5. Smaller recognition errors per wavelet family using DWT

Iter	*db2,4*	*sym2,4*	*coif1,4*	*bior2.2,4*	*rbio3.1,4*
4	15.500%	15.500%	7.000%	12.000%	13.000%

Based on these results, we can affirm that the wavelet who has the best performance to carry out face recognition is the Coiflets1on iteration level 4, this configuration only reported 7% of classification errors on the test set.

6 Conclusion

It can be concluded that based on the recognition results that images processed with PCA offer better results. Still, this procedure has a high computational cost because it requires calculating the covariance matrix of the features each time a subject is added to the database, which possesses a quadratic complexity. This cannot be applied to real applications because the quantity of features shall significantly increase, and the processing time of that matrix would be too high for real time applications with current technology.

The performance of wavelets as feature extraction technique was completely satisfactory, because superior percentages to 90% in recognition on the test sets were reached. This was achieved using a reduced quantity of features, in relation with the original image; only 1.1% of features were used to obtain excellent results on face recognition. The classifiers used in this work have been the fundamental tool to establish the comparatives between the different feature extraction methods. Its development and implementation have impacted directly on the results because on the development of this work, we have proposed a training method for neural networks which maximizes the usage of the processor, and considerably reduces the training time.

We can conclude that the neural networks method combined with wavelet transform are applicable to the implementation on real applications because, if we increment the database, we only have to train the neural network again. Nonetheless, to increment the number of subjects on PCA we have to recalculate the covariance matrix which might become incalculable. The purposed methodology shows better performance than the one presented on the state of the art, because it does not combine feature extraction techniques like wavelets + PCA, DCT + PCA to mention some. It only uses one technique of feature extraction at the time and with that the percentages of error obtained are considerably low. Based on the results, it has been established that the wavelet with best face recognition results using traditional classification systems and with neural networks is the Coiflets 1, followed by the Biorthogonal 2.2, in third place we can find rBiorthogonal 3.1 and in fourth place are Daubechies 2 and Symlets 2. For all this wavelet Families the iteration level used was 4 which can achieve a reduction of features to process inferior to 2%.

Acknowledgements

Giovanni Mejía wants to thank to CONACYT for providing the financial resources as a scholarship to pursue my Master of Science degree. Also I would like to thank to my advisor and co-advisor for their commitment in the development of this research.

The authors would like to thank to Universidad La Salle for providing the support under the projects I-064/12 and I-061/12.

References

1. Wayman J. A., Maltoni D., Maio D.: Biometric Systems. Springer, UK (2005).
2. Hala M. E.: Feature Extraction using PCA and Kernel -PCA for Face Recognition, The 8th International Conference on INFOrmatics and Systems (2012).
3. Omaia D., Poel J., Batista L.: 2D DCT Distance Based Face Recognition Using a Reduced Number of Coefficients, XXII Brazilian Symposium on Computer Graphics and Image Processing. (2009).
4. Xian L., Sheng Y., Qi W., Ming L.: Face Recognition Based on Wavelet Transform and PCA, Pacific-Asia Conference on Knowledge Engineering and Software Engineering. (2009).
5. Puyati W., Walairacht S. Walairacht A.: PCA in Wavelet Domain for Face Recognition, ICACT (2006).
6. Abouzar P., Yousefi S., Setarehdan S.: Hybrid WT Based DCT Based Face Recognition, IEEE International Conference on Signal Processing and Communications (2007).
7. Theodoridis S., Koutroumbas K.: Pattern Recognition. Elsevier, USA (2003).
8. Abdi H., Williams L. J.: Principal Component Analysis, USA (2010).
9. Mallat S.: A wavelet tour of signal processing: The sparse way, Academic Press, USA (2008).
10. Saad Y.: Iterative methods for sparse linear systems, Society for Industrial and Applied Mathematics; USA (2003).
11. Webb A. R.: Statistical Pattern Recognition, QinetiQ Ltd, UK (2002).
12. Rojas R.: Neural Networks, Springer, Berlin (1996).

Imperative Programming for Applications with built-in Intelligence

José Oscar Olmedo-Aguirre

Electrical Engineering, Cinvestav-IPN, Av. IPN 2508, Mexico City, Mexico 07360,
oolmedo@cinvestav.mx

Abstract. The convergence of diverse technologies has lead to increasingly complex applications that demand more built-in intelligence. Efficient programming languages like C lack of the right level of abstraction required for this purpose. The integration of the very high-level features of symbolic languages like Prolog combined with the efficiency of imperative languages like C would be welcomed with no doubt. DLProlog (Dynamic Logic Prolog) is an experimental language that integrates in a uniform programming model the elements required for the development of applications with built-in intelligence by coalescing the functional, logic and imperative programming paradigms. The key feature of DL-Prolog consists in the introduction of dynamic logic modalities into the head of the clauses. The modality encloses an efficient imperative program that uses high-level symbolic instructions like assignments based on E-unification. In this paper, the imperative fragment of DLProlog is formally described in the structured operational semantics style. A working research prototype has been built upon this design and it is available on request from the author.
Keywords. Logic Programming, Dynamic Logic, Knowledge Representation and Reasoning.

1 Introduction

The modern convergence of diverse technologies has lead to increasingly complex applications. This trend calls for an analogous convergence in programming concepts and models. Unified programming research attempts to identify the minimal set of the simplest and most fundamental programming concepts that coherently can be applied in as many as diverse areas of application and still can lead to reasonable efficient implementations. Among the most successful efforts in conciliating such a diversity of concepts, the functional logic programming paradigm is capable of embedding the notion of state that characterizes the imperative programming, along with constraint satisfaction, concurrent processing and object orientation, among others. Nonetheless, there is neither symmetry nor balance in their integration in the sense that a Java programmer, for example, needs to learn a number of sophisticated concepts that are uncommon in her experience to start programming even relatively simple programs.

In this respect, DLProlog [8], an extension to pure Prolog with dynamic logic modalities, has been designed by the author to provide a more balanced approach

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 13–23
Paper Recived 31-08-2013 and Acepted 21-10-2013

in the sense that the same Java programmer has not to learn too many concepts if she does not need to. Instead of placing the imperative programming at the top of the functional-logic programming paradigm as in Curry [4], the unified approach of DLProlog stands more balanced upon the imperative, functional and logic programming paradigms. Unfortunately, due to lack of space, the previous claim will only be justified by discussing some program examples that will also help to show the programming style of the imperative fragment of DLProlog.

This article is organized as follows. In section 2, a succinct review of the related work is presented. In section 3, a brief account of the programming style of the imperative fragment of DLProlog is shown. In section 4, a formal description in the structured operational semantics is given for the imperative fragment. Finally in section 5, some further research directions and concluding remarks are given.

2 Related Work

Classical papers on the integration of functional, equational and logic programming were collected in [1]. Among them, Kahn's Uniform language is the closest to ours in its use of unification though his proposal was developed for the functional language LISP, whereas DLProlog applies unification for its imperative fragment [2]. Uniform is an AI language that uses augmented unification to solve equational problems that cannot be solved by syntactic unification. Probably the most serious difficulty faced in the design of Uniform was the computational complexity of finding a solution in the presence of equality theories that lead to large chains of equalities. In fact, solving equational problems in presence of equality theories is known to be a NP-complete problem [3]. In DLProlog the combinatorial production of equalities has been drastically reduced by (i) incorporating a greater degree of control into the language by means of functions instead of equations, and (ii) avoiding comparing two functional terms, so any equation can relate at least one constructor-rooted term. Thus DLProlog unification prevents applying higher-order unification which is known to be a NP-complete problem [3].

As discussed by Kahn, Uniform lacked of more theoretical foundations in equational unification to deal with the combinatorial search involved in the generation of solutions. Taking advantage of the reasonably efficient implementations of functional languages, the integration efforts were later oriented to approach relations as Boolean functions. Curry [4] is a functional logic programming language based on the evaluation by need of expressions (a mechanism called *narrowing* used in lazy functional programming) along with the possible instantiation of free variables occurring in the expressions (a mechanism called *residuation* used in constraint-logic programming). Nonetheless, imperative programming is not directly available but through a monadic IO system that enable sequence of input-output actions. Despite its growing success and acceptance, the use of evaluation by need makes difficult to reason about program behavior and memory requirements. This means that a programmer needs to understand the not so intuitive IO monadic system in order to write even the simplest imperative programs.

Concurrent programming, particularly the so-called algebraic theory of processes has provided an appealing basis for its integration with the functional, imperative and object-oriented paradigms. The language PICT based on the π-calculus is an example of this approach [5]. However, besides the formidable task of implementing enough programming concepts as networks of interacting processes, the several layers that will emerge from such design makes no clear how efficient and of practical use for the most common applications would be this integration. Nonetheless, the concurrent programming approach has also leaded to one of the more successful pragmatic language designs as in Oz [6]. Oz is based on residuation, a simple yet powerful mechanism to solve concurrently constraints by processes that suspend their execution whenever a variable is undefined in a predicate and resume their execution soon after the variable becomes defined.

3 Equational Reasoning within Imperative Programs

The importance of unification and equational unification (E-unification) is due to the fact that is widely used in automated theorem proving and related fields like logic programming. The term unification generally stands for syntactic equality on terms, whereas semantic E-unification generally stands for syntactic equality modulo an equational theory. Augmented unification is a restricted version of E-unification to be used as a model of execution [2]. Hereinafter, DLProlog variable names are written starting with the last alphabet letters u, v, w, x, y, z and lists of terms are written in the Prolog style. For a brief account of unification and E-unification, consider the following DLProlog program that involves lists of terms as in Prolog. The program simply checks that the well-known property of lists $rev(app(xs, ys)) = app(rev(ys), rev(xs))$ holds for the lists $[a, b, c, d]$ and $[e, f, g, h]$.

$$
\begin{array}{l}
[\texttt{var}\, ys, zs :\ xs = [a, b, c, d]; ys := [e, f, g, h]; \\
\quad zs := rev(app(xs, ys)); us = zs; writeln(revapp : us); \\
\quad zs := app(rev(ys), rev(xs)); vs = zs; writeln(apprev : vs) \\
](us = vs).
\end{array}
$$

This DLProlog program consists of two parts: (i) the modal connective that encloses in brackets the *action* (i.e. a well-formed imperative program fragment): $\texttt{var}\, ys, zs : xs = [a, b, c, d], \cdots$, and (ii) the postcondition $us = vs$ that states the properties of the action that hold after its execution upon the values bound to the variables us and vs. The property is valid in a theory of lists with the usual operations $rev(xs)$ that reverses the order of the elements of list xs and $app(xs, ys)$ that appends the elements of list ys at the end of list xs. In the action, the logical variable xs is defined with list $[a, b, c, d]$, whereas the imperative variable ys is defined with list $[e, f, g, h]$. Logical variables are pure Prolog variables that can be defined by equality at most once, whereas imperative variables are DLProlog variables that can be arbitrarily redefined as many times as needed by means of the destructive assignment of imperative programming. Furthermore, logical variables can be introduced with no declaration as in Prolog, whereas imperative variables can be only introduced by declaration with the `var` binder. In any case, the occurrence of either a logical or imperative variable in an expression denotes the value bound to it if any.

In DLProlog, imperative instructions are sequentially executed as usual from left to right of the semicolon connective. In the action, the assignment $zs := rev(app(xs, ys))$ defines the local variable zs with a list that results from reversing the concatenation of lists xs and ys. This list is used to define the logical variable us that is next written in the terminal output preceded by label *revapp*. The assignment $zs := app(rev(ys), rev(xs))$ redefines the local variable zs with the concatenation of the reversed lists of ys and xs. This list is used to define the logical variable vs that is next written preceded by label *apprev*. Beyond this point in the text, the scope of the binder `var` reaches its end and the imperative variables ys and zs cease their existence. Nonetheless, the values bound to the logical variables us and vs remain unaltered at the action postcondition $us = vs$. The partial correctness of the action ensures that the postcondition is always satisfied upon its termination whenever the input variables satisfy a suitable precondition. Hence there is no need to test the postcondition whenever the partial correctness of the action can be formally proved.

3.1 The Reverse Program

The DLProlog program for the reverse function $rev(xs) = zs$ is presented next in a stylized format that helps to show its syntactic structure:

$$\left[\begin{array}{l} \texttt{var}\, ys_1, ys_2\colon \\ \quad (ys_1, ys_2) := ([\,], xs); \\ \quad \texttt{while}\, ys_2 \neq [\,]\ \texttt{do} \\ \qquad \texttt{var}\, y\colon (y, ys_1, ys_2) := (\texttt{hd}(ys_2), [y \mid ys_1], \texttt{tl}(ys_2)) \\ \quad \texttt{od}; \\ \quad zs := ys_1 \end{array}\right] rev(xs) = zs \Leftarrow list(xs)$$

where `hd` and `tl` are predefined functions that obtain the first and the rest of a list. The above program is in fact the partial correctness property of the program that computes the reverse function. Like Prolog, the reverse program is a Horn clause consisting of a single conclusion predicate called the *head* of the clause and an antecedent formed by a conjunction of predicates called the *body* of the clause. Unlike Prolog, the head of the clause is a dynamic logic assertion consisting of the reverse action enclosed in brackets followed by the postcondition $rev(xs) = zs$. A DLProlog action fails either if it does not terminate, or if upon its completion, it cannot satisfy its postcondition. As illustrated by the previous example, a DLProlog clause has the following structure that states the partial correctness relation of the action A:

$$[A]R \Leftarrow P$$

where P and R are the pre- and post-conditions of A. All the logical variables occurring in P are considered *input variables* in A, whereas the variables occuring only in R (and not in P) are considered *output variables*. Following the reverse program example, the only input variable is xs, whereas the only output variable is zs. In DLProlog, all input variables are treated as constants in action A, whereas the output variables can be defined at most once there. The input and output variables are visible to any instruction of A. DLProlog relies on the usual block programming construct to ensure both restricted lexical visibility for the instructions within the block and controlled allocation of local variables. The block ensures that new local variables become into existence when the execution enters into it and cease their existence when the execution leaves it, keeping them appart from all the input and output variables.

In the imperative programming model of DLProlog, imperative variables can be redefined, whereas all the logical variables cannot. However, both logical and imperative variables may freely occur in expressions, because the occurrence of a variable name denotes the value bound to it according to the current state. E-unification, assignment and resolution on a predicate goal may modify the state of the computation, whereas all the other programming constructs either evaluates expressions or alters the course of the action. In the block structure of DLProlog, all computations performed by the modality $[A]$ can only modify either the imperative variables or the output logical variables. Because output variables can be defined at most once, they are only used to assign the final values computed in A. Thus, the behavior of the imperative program fragment becomes encapsulated within the modal connective, being only visible its effects through the modification of the output variables occurring in R.

4 DLProlog Formal Description

Because most of the modern imperative languages are type-checked, DLProlog is formalized as a many-sorted first-order logic language to ensure a type checking discipline. Given a (Σ, Ξ)-structure, $\Sigma = \bigcup_\alpha \Sigma_\alpha$ is a family of *constructor* (constant) names and $\Xi = \bigcup_\beta \Xi_\beta$ is a family of *variable* names, each partitioned by the basic types (sorts) `bool` and `nat`, among possibly others. In the following presentation, the name `var` can be replaced by any of the basic types. The set $T(\Sigma, \Xi)$ of *terms with variables* is the minimal set of phrases that is closed under composition of a constructor with a (possibly empty) sequence of terms. The set $T(\Sigma) = T(\Sigma, \emptyset)$ of *ground terms* consists of terms with no variables. *Type inference* is embedded in the grammar rules of the language to ensure that clauses and programs are well-formed. In particular, the grammar rule shown below describes the syntactic structure of a well-formed term T_β of type β:

Type inference $T_\beta ::= x_\beta \mid c_\beta \mid c_{\beta_1 \cdots \beta_n \to \beta}(T_{\beta_1}, \ldots, T_{\beta_n}) \mid f_{\beta_1 \cdots \beta_n \to \beta}(T_{\beta_1}, \ldots, T_{\beta_n})$

In that follows, all type annotations are ommited for the sake of brevity. The set $P(\Sigma, \Xi)$ of *atomic predicates with variables* is the minimal set closed under composition of predicate symbols with (possibly empty) sequences of terms. The set $P(\Sigma) = P(\Sigma, \emptyset)$ of *ground atoms* consists of all atoms with no variables. A *literal* is an atom or a negated atom. A *clause* is a disjunction of literals. The set of clauses with variables is denoted $C(\Sigma, \Xi)$. Clauses are usually written in implication form $P \Leftarrow Q$, where P, called the *consequent,* is a disjunction of atoms and Q, called the *antecedent*, is a conjunction of atoms. A *unit clause* contains only one literal. A *positive clause* contains no negated atoms, whereas a *negative clause* contains no positive atoms. A *Horn clause* contains at most a positive atom. A *goal* consists only of negative atoms that can be represented by an implication with false as consequent. The logical variables occurring in a clause are universally quantified. The *set of goals with variables* is denoted $G(\Sigma, \Xi)$. In that follows, in any context where the term list $T_1, \ldots, T_n$ may occur, the restriction $n \geq 0$ is always assumed for n.

Terms	$T ::= x \mid c \mid c(T_1, \ldots, T_n) \mid f(T_1, \ldots, T_n) \mid (T_1, \ldots, T_n)$
Lists	$L ::= [\,] \mid [\, T_1, T_2, \ldots, T_n \mid L \,]$
Predicates	$P ::= \texttt{false} \mid \texttt{true} \mid T_1 = T_2 \mid p(T_1, \ldots, T_n)$
Goals	$G ::= P \mid G_1 \wedge G_2$
Clauses	$B ::= P \mid P \Leftarrow G \mid \forall x.B$
Actions	$A ::= \texttt{skip} \mid \texttt{fail} \mid T_1 := T_2 \mid G\texttt{?} \mid$
	$A\texttt{*} \mid A_1\texttt{;}A_2 \mid A_1 \cup A_2 \mid \texttt{var}\ x_1, \ldots, x_n :\ A \mid (A)$
Modal actions	$M ::= P \mid\ [A]\ M \mid \langle A \rangle\ M$
Modal clauses	$F ::= M \mid G \Rightarrow M \mid \forall x.F$

The set $A(\Sigma, \Xi)$ *of actions with variables* is the minimal set of actions that is closed under composition according to the following action connectives. A basic action is either the null action (`skip`), the failure (`fail`), the assignment of terms to variables by using the binary operator (:=), or the postfix unary operator for condition testing (`?`). The action connectives, in descending order of precedence, are the postfix unary operator for iteration (`*`), the infix binary operator for sequential composition (`;`), and the infix binary operator for non-deterministic choice ($\cup$). The declaration of the imperative variables $x_1, \ldots, x_n$ in A by the binder `var` introduces variables in the current block with a scope that extends as far as possible to the right. The precedence of the action connectives and the extension of the scope can be altered by grouping (A). The mixfix modal connectives of modal necessity ([ ]) and possibility ($\langle\,\rangle$) compose actions along with their postconditions. The following equalities between actions provide an interpretation of the usual imperative programming constructs:

$$\texttt{skip} \equiv \texttt{true?}\ ,\ \ \texttt{fail} \equiv \texttt{false?}$$
$$\texttt{if}\ F\ \texttt{then}\ A\ \texttt{fi} \equiv F\texttt{?;}A$$
$$\texttt{if}\ F\ \texttt{then}\ A_1\ \texttt{else}\ A_0\ \texttt{fi} \equiv (F\texttt{?;}A_1) \cup (\neg F\texttt{?;}A_0)$$
$$\texttt{while}\ F\ \texttt{do}\ A\ \texttt{od} \equiv (F\texttt{?;}A)\texttt{*;}\neg F\texttt{?}$$

The unification $T_1 = T_2$? between terms T_1 and T_2 is a test for equality and it can be written with no test operator `?` if followed by a sequential composition connective. The assignment notation $T_1 := T_2$ generalizes the single assigment

$x := T$ and the multiple assignment $x_1, \ldots, x_n := T_1, \ldots, T_n$ to arbitrary equally structured terms T_1 and T_2. Their precise meaning will be given later on by means of the SOS semantics of DLProlog.

In a signature (Σ, Ξ) with variables, a *substitution* is a partial function $\sigma : \Xi \to T(\Sigma, \Xi)$, where $\sigma(x) \neq x$ for any variable $x \in \Xi$. $\{\}$ denotes the empty substitution. A *ground substitution* is a substitution $\sigma : \Xi \to T(\Sigma)$ valued on ground terms. For any variable $x \in \Xi$ and any substitution σ, let $x\sigma = \sigma(x)$ if $x \in \mathrm{dom}(\sigma)$ and $x\sigma = x$ otherwise. For any term $t \in T(\Sigma, \Xi)$, let $t\sigma$ be the term obtained by substituting any variable x appearing in T by $x\sigma$:

$$\begin{aligned} x\{\} &= x \\ x\,\sigma &= \begin{cases} x & \text{if } x \notin \mathrm{dom}(\sigma) \\ \sigma(x) & \text{if } x \in \mathrm{dom}(\sigma) \end{cases} \\ c\,\sigma &= c \\ c(T_1, \ldots, T_n)\,\sigma &= c(T_1\,\sigma, \ldots, T_n\,\sigma) \\ [A]\,p\,\{\} &= [A]\,p \\ [A]\,p\,\sigma &= [\sigma_{:=}\,; A]\,p \end{aligned}$$

where notation $[\sigma_{:=}]$ stands for the multiple assignment $x_1, \ldots, x_n := T_1, \ldots, T_n$ obtained from the substitution $\sigma = \{x_1 \mapsto T_1, \ldots, x_n \mapsto T_n\}$. Thus the substitution for a modal action A is defined as the initial value that the variables take before the action starts its execution. The *composition* of two substitutions $\sigma_0, \sigma_1 \in \Xi \to T(\Sigma, \Xi)$, written $\sigma_0 \cdot \sigma_1$, is defined as

$$\sigma_0 \cdot \sigma_1 : x \mapsto \begin{cases} (x\sigma_0)\sigma_1 & \text{if } x\sigma_1 \notin \mathrm{dom}(\sigma_1) \\ x\sigma_1 & \text{if } x \in \mathrm{dom}(\sigma_1) - \mathrm{dom}(\sigma_0) \\ failure & \text{otherwise} \end{cases}$$

Besides the natural extension to terms $T(\Sigma, \Xi) \to T(\Sigma, \Xi)$, substitutions are also extended to predicates, goals, and both backward and forward rules.

The unification of two terms either finds a unifier, i.e. a substitution that solves the equational problem or terminates in failure. Unification is a step by step transformation process between sets of equations until no further transformations can be applied. The unification procedure is defined through the following rules:

$$\begin{aligned} (\{t = t\} \cup P, S) &\to (P, S) \\ (\{f(t_1, \ldots, t_n) = f(s_1, \ldots, s_n)\} \cup P, S) &\to (\{t_1 = s_1, \ldots, t_n = s_n\} \cup P, S) \\ (\{f(t_1, \ldots, t_n) = g(s_1, \ldots, s_m)\} \cup P, S) &\to (\{\}, \{\bot\}) \\ (\{x = t\} \cup P, S) &\to (\{\}, \{\bot\}) \;\text{ if }\; x \in fv(t) \text{ and } x \neq t \\ (\{x = t\} \cup P, S) &\to (P\,\{x \mapsto t\}\,, S\,\{x \mapsto t\} \cup \{x = t\}) \\ &\qquad \text{if }\; t \notin \Xi, x \notin fv(t) \end{aligned}$$

By applying the above unification rules to the equational problem $t = s$, the pair $(t = s, \{\})$ is transformed into the pair $(\{\}, S)$, meaning that the set S of equations is in solved form, being S the solution to the problem; otherwise, the pair $(t = s, \{\})$ is transformed into the pair $(\{\}, \{\})$, meaning that the equational problem has no solution.

The semantic description of DLProlog is established according to the structural operational semantics [7] by defining program descriptions and transitions between program descriptions. Assuming the termination of the program, a *program description*, or simply a *description*, can be either instantaneous or final. An *instantaneous description* is a pair $([A]M, \sigma)$ relating a program modality $[A]M$ and a program state σ, meaning that the program A starts its execution in the state σ, reaching a state that is assumed to satisfy M whenever A terminates. Note that M may be either another program modality or a predicate. For the former case, a new instantaneous description can be established, whereas for the latter case, a final description is reached. A *final description* corresponds to the state reached when the program either terminates or fails. The final description of a program that terminates successfully is a state $\sigma = \{x_1 \mapsto t_1, \ldots, x_n \mapsto t_n\}$, containing the bindings of the variables x_i with their final values t_i for $i = 1, \ldots, n$. Instead, the final description of a program that terminates in failure is represented by $\perp$. A *transition* is a relation $\triangleright$ over pairs of program configurations. An *execution* of a program corresponds to the reflexive and transitive closure of the transition relation. The execution of a terminating program is a finite sequence of instantaneous de-scriptions ended by a final description. For the successful program execution, we have, for some $n \geq 0$:

$$(M_0, \sigma_0), (M_1, \sigma_1), \ldots, (M_i, \sigma_i), (M_{i+1}, \sigma_{i+1}), \ldots, (M_n, \sigma_n), \sigma_n,$$

the program starts with the initial modality $M_0 = [A]P$ along with the variables initialized according to σ_0 and the program terminates reaching the state σ_n, satisfying the partial correctness property. Thus, $[A]P\sigma_0$ implies $P\sigma_n$. This property follows by induction on the length n of the sequence of instantaneous descriptions, where $M_i\sigma_i$ implies $M_{i+1}\sigma_{i+1}$ for all $i = 1, \ldots, n$. For the unsuccessful program execution:

$$(M_0, \sigma_0), (M_1, \sigma_1), \ldots, (M_i, \sigma_i), (M_{i+1}, \sigma_{i+1}), \ldots, (M_n, \sigma_n), \perp$$

for some $n \geq 0$. The program starts like before, though it now terminates in failure, meaning that nothing can be asserted about the final program state.

The following inference rules define the SOS semantics of the imperative fragment of DL Prolog:

$$\frac{(\{f(t_1, \ldots, t_n)\sigma = f(s_1, \ldots, s_n)\sigma\}, \{\}) \to (\{\}, \{x_1 = r_1, \ldots, x_m = r_m\})}{([f(t_1, \ldots, t_n) = f(s_1, \ldots, s_n)]\, M, \sigma) \triangleright (M, \sigma\,\{x_1 \mapsto r_1, \ldots, x_m \mapsto r_m\})} \tag{1}$$

$$([f(t_1, \ldots, t_n) = g(s_1, \ldots, s_m)]\, M, \sigma) \triangleright \perp \tag{2}$$

$$\frac{(\{s = t\sigma\}, \{\}) \to (\{\}, \{x_1 = r_1, \ldots, x_n = r_n\})}{([f(s_1, \ldots, s_m) = t]\, M, \sigma) \triangleright (M, \sigma\,\{x_1 \mapsto r_1, \ldots, x_n \mapsto r_n\})} \tag{3}$$

$$([t := s]\, M, \sigma_1\sigma_2) \triangleright ([t = s\sigma_1\sigma_2]\, M, \sigma_1) \tag{4}$$

$$([t := s]\, M, \sigma) \triangleright \perp \tag{5}$$

$$([\top?]\, M, \sigma) \rhd (M, \sigma) \tag{6}$$

$$([\bot?]\, M, \sigma) \rhd \bot \tag{7}$$

$$([A_1;A_2]\, M, \sigma) \rhd ([A_1]\, [A_2]\, M, \sigma) \tag{8}$$

$$([A_1 \cup A_2\,]\, M, \sigma) \rhd ([A_1]\, M, \sigma) \tag{9}$$

$$([A_1 \cup A_2\,]\, M, \sigma) \rhd ([A_2]\, M, \sigma) \tag{10}$$

$$([A{*}]\ M, \sigma) \rhd ([\top?\ \cup A; A{*}]\, M, \sigma) \tag{11}$$

$$([\,]\, P, \sigma) \rhd \sigma \tag{12}$$

Rules from (1) to (3) describe unification and E-unification. In (1), the unification of two constructor-based terms with identical constructor is converted into the equational problem $f(t_1, \ldots, t_n)\sigma = f(s_1, \ldots, s_n)\sigma$ that is the instance of the equation under σ. Whenever the unification problem has a solution, the solution is composed with the current state. In (2), the equation $f(t_1, \ldots, t_n) = g(s_1, \ldots, s_n)$, has no solution if either f and g are different constructors or there are not defined functions for at least one of them. In (3), the equation $t = g(s_1, \ldots, s_n)$ can be solved if there is a functional definition of g and both the application $g(s_1\sigma, \ldots, s_n\sigma)$ and the term $t\sigma$ are equal to some ground term s. Rules (4) and (5) describe the assignment instruction. In (4), if there is a function defined for f and if $s\sigma_1\sigma_2$ is a ground term $\sigma_1 = \{x \mapsto r | x \notin fv(t)\}$ and $\sigma_2 = \{x \mapsto r | x \in fv(t)\}$, the assignment instruction succeeds if the instance of the term at the right-hand side under $\sigma_1\sigma_2$ is a ground term and all the variables occurring in the term at the left-hand side have been unbound from their previous values if any. In (5), the assignment fails either if $s\sigma$ is neither a ground term nor $s\sigma$ unifies with t. In (6), the action `skip` corresponds to the guard `true`?, always terminates successfully, whereas in (7) the action `fail` corresponds to the guard `false`? does not terminate and if it does, it reaches an state where no condition holds. In (8), the sequential execution of $A_1; A_2$ corresponds to the execution of A_1 followed after its completion by the execution of A_2; otherwise $A_1; A_2$ fails if either A_1 or A_2 fails. In (9) and (10), the non-deterministic execution $A_1 \cup A_2$ of A_1 and A_2 corresponds to the selection of the unfailing execution of any of them. In the case that both succeeded, one of them is arbitrarily chosen, whereas if both fail, the instruction fails too. In (11), the repetitive execution of A is obtained by non-deterministically choosing between `skip` and unwinding the iteration by means of the sequence $A; A{*}$. Finally, in (12) the program modality terminates when no instructions remain to execute. In this case, if the program executed is partially correct w.r.t. postcondition P, then $P\sigma$ holds.

5 Conclusions and Further Work

In this paper we have described some of the programming features of DLProlog, a dynamic logic modal extension to the pure Prolog programming language and model, in order to preserve its declarative nature without compromising its expressiveness. An experimental compiler and interpreter has been already written in LPA Prolog [9] to provide some evidences about the improvements on the readability and efficiency of DLProlog over both standard and pure Prolog. There is a number of appealing research directions from the proposed DL modal extension:

1. Extending the unification procedure to other constraint satisfaction methods including other domains like finite domain constraints seems to be possible. However, for such an extension, don't known non-determinism must be included in the semantics of actions.
2. Extending the imperative language and programming model with mechanisms for encapsulation and inheritance to introduce the fundamental notions of objects and classes.
3. Using popular imperative languages like Java or C# instead of the algorithmic language used here that is reminiscent of the Algol68 programming language.
4. Designing and implementing an extension of the standard WAM to include low level instructions like destructive assignment and conditional jumping. A substantial contribution may arise from this research direction.

Due to its inherited symbolic processing capabilities, DLProlog is a very high-level programming language thay brings the best of the successful imperative and functional-logic paradigms for programmers that need to develop modern applications with built-in intelligence.

6 References

1. DeGroot, D., Lindstrom, G. (eds.): Logic Programming: Functions, Relations and Equations. Prentice-Hall, 1986.
2. Kahn, K.M.: Uniform : A Language Base upon Unification which unifies much of Lisp, Prolog and Act 1. In [1], 411-440.
3. Baader, F., Nipkow. T.: Term Rewriting and All That. Cambridge, 1998.
4. Antoy, S., Hanus, M.: Functional Logic Programming. Communications of the ACM, 53(4):74-85, 2010.
5. Sewell, P., Wojciechowski, P.T., Unyapoth, A.: Nomadic PICT: Programming Languages, Communication Infrastructure Overlays, and Semantics of Mobile Computation. ACM Transactions of Computer Languages, 32(4):12:1-12:63, 2010.
6. Van Roy, P.: Multiparadigm programming in Mozart/Oz. Second International Conference MOZ 2004, 2005.
7. Reynolds, J.C.: Theories of Programming Languages. Cambridge, 1998.
8. Olmedo-Aguirre, J.O., Morales-Luna, G.: A Dynamic-Logic-based Modal Prolog. In proceedings MICAI 2012, CPS IEEE Computer Society. pp. 3-9. ISBN: 978-0-7695-4904-0.
9. LPA Prolog, Logic Programming Associates, url: http://www.lpa.co.uk/

Automatic Detection, Classification and Counting of Leukocytes in Smears of Peripheral Blood

J. Martínez-Castro [a], S. Reyes-Cadena [b], E. Felipe-Riverón [a]
[a] Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico
[b] Instituto Nacional de Rehabilitación, Mexico City, Mexico
macj@cic.ipn.mx, susana.reyes.cadena@gmail.com, edgardo@cic.ipn.mx

Abstract. Using the *k*-NN classifier in combination with the first Minkowski metric, in addition to techniques of digital image processing, we developed a computational system platform-independent, which is able to identify, to classify and to count five normal types of leukocytes: neutrophils, eosinophils, basophils, monocytes and lymphocytes. It is important to emphasize that this work does not attempt to differentiate between smears of leukocytes coming from healthy and sick people; this is because most diseases produce a change in the differential count of leukocytes rather than in theirs forms. In the other side, the system could be used in emerging areas such as the topographic hematology and the chronobiology.

Keywords. Automatic classifier of leukocytes, k-NN, minkowski metric, pattern recognition, digital image processing.

1 Introduction

Hematology is the branch of medical science responsible for the study of the elements integrating the blood as well as their precursors, and the structural and biochemical disorders of these elements that can lead to a disease. The hematologic diseases affect mainly the production of blood and its components, such as the red blood cells (erythrocytes), white blood cells (leukocytes), hemoglobin, plasma, platelets and the mechanism of blood clotting.

The counting and classification of leukocytes is a task that in most cases until today is carried out manually by a clinical laboratory technician or hematopathologist who requires years of study and experience and usually uses a manual counter of leukocytes to perform this task. The repetitive procedure proved to be very tedious and in the long run can cause disruptions in the visual health of the people responsible for carrying out it. Additionally the reliability of the results depends largely on the ability of the person that performs the study.

Fortunately, the science of electronics and computing has moved closer to other branches of knowledge, as it is the case of the hematology, in order to support and provide solutions to the problems they face. These solutions are built, in its majority, based on the introduction of mathematical methods, algorithms and software tools.

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 23–33
Paper Recived 01-09-2013 and Acepted 21-10-2013

There are instruments such as flow cytometer and automatic counting machines that examine white blood cells quantitatively but not qualitatively. Computer-aided automatic analysis not only saves manpower and time, but also reduces the effects of human errors. Up to now, automatic cell classification systems cannot meet the complexity of real clinical demands.

Segmentation subdivides an image into its constituent regions or objects. The level to which the subdivision is carried depends on the problem being solved. Segmentation should stop when the objects of interest in an application have been isolated [1]. In order to validate segmentation output it is necessary to know the ground truth that is the true features of each object of interest. However, the ground truth in medical imaging is an elusive concept because there is not a clear and absolute way to define it. The only way and often the best option is to have a human expert observer (or more than one) to define the ground truth by generating manual outlines [2]. The final objective of blood cell segmentation is to extract each constituent "object" and segment every cell into its morphological components such as nucleus, cytoplasm, holes and other organelles [3].

The paper is organized as follows. Section 2 briefly introduces some general characteristics of leukocytes and their different types. Section 3 includes an analysis of the antecedents and the state of the art in segmenting the leukocytes and classifies them in their five different types. In Section 4 is described the methodologies carried out in the segmentation of each type of leukocyte. Experimental results are included in Section 5 with some discussion about them and finally the conclusions of the paper are given in Section 6.

2 Background

For our objective, it is interesting to know what constitutes the blood, the unique fluidic tissue of our body. For this, the Hematology is supported by the process of hematopoiesis, that is, the process of production, maturation and growth of the cell lines present in the blood and tissues, in order to be aware of the morphologies and components of the blood cells during the process of maturation and growth. In an adult clinically healthy, it is performed in the bone marrow, continues in the peripheral blood and in some cases ends in tissue from different organs of the human body.

The hematopoiesis presents three major cell lines: The erythrocyte cell line, the leukocyte cell line and finally the megakaryocytic cell line producing platelets. Below we will shortly describe the erythrocyte and leukocyte lines.

Erythrocytes are anucleated cells that belong to the series of red blood cells. Its role is to carry out the exchange of gases in the body, i.e. carry oxygen to the body and discard carbon dioxide. The erythrocytes are characterized by a natural pigmentation in reddish tone related to the amount of iron present in the cells.

Leukocytes are cells that represent the main mechanism of the body's defense against infectious processes and inflammatory and allergic reactions. The leukocytes, better known as white blood cells, lack of pigment.

In order to differentiate white blood cells in the microscope, it is needed to process the samples and apply a specific stain. The tincture most frequently employed for dying leukocytes is called Wright and Giemsa. In contrast to the erythrocytes, leukocytes contain abundant cytoplasmic granulations among other characteristics that let us clearly distinguish them from each other. Different types of white blood cells vary in color, granularity and texture, especially in the cytoplasm. However, in many cases leukocytes must be extracted using only the nuclei since the nucleus is a relatively stable structure. In blood smears, the number of red cells is many more than white blood cells. Platelets are smaller particles not clinically so important.

The main goal of this paper is related to those leukocytes present in the peripheral blood; they constitute an important component of the overall immunity mean that defends the body against infectious diseases and foreign substances. White blood cells are classified into five main groups: phagocytic polymorph nuclear leukocytes (neutrophils), eosinophils, basophils mononuclear phagocytes (monocytes) and lymphocytes [4]. Neutrophils, eosinophils and basophils are known as granulocytes, because they present a granulated nucleus, whereas monocytes and lymphocytes present a well-defined one-piece nucleus. On the other hand, neutrophils and eosinophils have a greater size than basophils, and have a clear differentiation in the amount of cytoplasm with respect to the monocytes and lymphocytes.

The objective of this work is to describe a computational system, platform independent that enables the identification, counting and classification of the five normal forms of leukocytes present in the peripheral blood. Simultaneously, it is intended that the computational system manages in a secure and efficient way the data of users and patients introduced to the system.

To achieve our goal, original microscopic images are previously preprocessed in order to enhance their quality. Images are initially normalized in size. Then, the processes for the reduction of the additive noise and the contrast improvement are carried out. With this, we can get images of higher quality ready to be analyzed. This work uses color images in the frequently used RGB (Red-Green-Blue) color space.

3 Antecedents and state of the art

In recent times there have been efforts to develop computer systems that perform the leukocitary differential count automatically, looking for shorten the waiting times and help to give a more accurate diagnosis as soon as possible and independent of the human factor. Hitherto, automatic cell classification systems cannot meet yet the complexity of real clinical demands. Nevertheless, yet 21% of all processed blood samples still require microscopic review by experts [5]. In 2002 Liao and Deng [6], were the first to introduce a shape analysis for white blood cell segmentation. Through basic segmentation using simple thresholding the borders of white blood cells are identified using a shape analysis step. However, despite the simplicity and effectiveness, this method only applies to circular-shaped white blood cells such as lymphocytes. Paper [7], an approach that actively selects efficient samples by simulating visual attention, uses a two-stage method (bottom-up and top-down) via learning by on-line sampling to automatically segment complex leukocyte images. Adollah et al., in [8], present a review of segmentation methods that have found application in classification in biomedical-image processing especially in blood cell image processing up to 2008. More recently Ghosh et al. [9] developed an automated leukocyte recognition approach using fuzzy divergence and modified thresholding techniques. They concluded that Cauchy functions provide better segmentation, particularly for leukocyte recognition, in comparison with Gamma function-based divergence. Morphological operators have been developed to detect and classify other objects, i.e. malaria parasites, in infected blood cell images [10]. Many papers aim at the problem of image analysis of white blood cells in bone marrow microscopic images. One of them [11] proposed the use of multispectral imaging techniques with spectral calibration method to acquire device independent images. For image segmentation they applied Support Vector Machines directly to the spectrum of each pixel, and sequential minimal optimization algorithm for feature selection to reduce the time of training SVM classifier. The use of a differential evolution algorithm has been used also to segment the leukocytes from the images of mice skin tissue images stained with hematoxylin and eosin [12]. In [13] image processing algorithms are proposed to recognize five types of white blood cells in peripheral blood automatically. First, a method based on Gram-Schmidt orthogonalization is proposed along with a snake algorithm to segment nucleus and cytoplasm of the cells. Then, a variety of features are extracted from the segmented regions. Next, most discriminative features are selected using a Sequential Forward Selection algorithm and performances of two classifiers, Artificial Neural Network and Support Vector Machine, are compared.

4 Methodology

The main objective of the system developed is to identify, count and classify the type of leukocytes present in a digital image of a peripheral blood sample obtained through a microscope (smear). A typical image of it is shown in Fig. 1.

In general, papers dealing with automatic means to isolate leukocytes from erythrocytes and other organelles in peripheral blood are based on the color and size of these objects. The clear difference in color of leukocytes is achieved by using a particular tincture to produce the smear. The task of segmenting them is carried out after the contrast enhancement of the adequately prepared smear and the prior transformation to a gray-level image of the commonly used color RGB image. Now the nuclei and its corresponding membrane are segmented. The objects of interest will be sets with nucleus and cytoplasm.

4.1 Nuclei segmentation

The first step was to separate the white blood cells from the red blood cells. To segment the nuclei of the leukocytes we used the green channel (G), since in this channel the image presents more details and less additive noise. In this image the nuclei of interest presented a darker grey level than the rest of the objects.

Subsequently, we threshold the image obtained from the previous process by means of an adaptive thresholding of the histogram of the image. The adaptive thresholding [13] is a method that enables us to select the optimal threshold based on the probability distribution of gray levels represented in the histogram of the image.

The result of the adaptive thresholding is shown in Fig. 2. The image is negated for smoothing the noise and filtering the artifacts.

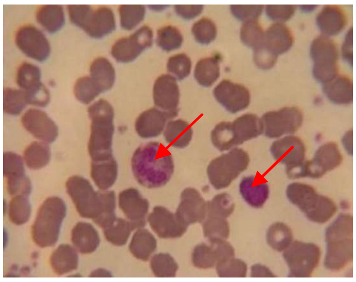


Fig. 1. Smear image containing two leukocytes (indicated by the red arrows).

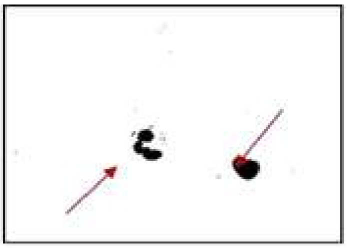
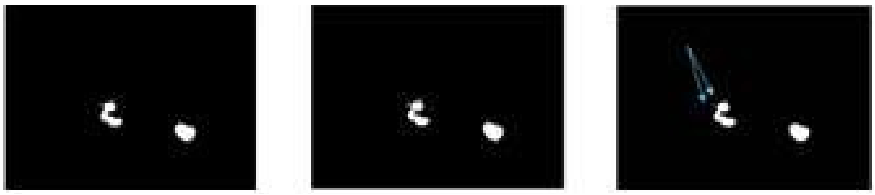


Fig. 2. Objects obtained through the adaptive thresholding of the green plane of the image (presence of nuclei and some artifacts).

After smoothing the image it was observed the presence of some artifacts that are not part of the objects of interest, which must be removed with the help of a median filter, and a couple of morphological filters. With this, we obtained an image with the objects of interest more defined (nuclei), Fig. 3. The morphological filters used were the closing (Fig. 3b) and the opening (Fig. 3c), which helped us to strengthen the weak links between objects, increase the definition of forms and remove the small unwanted dark objects; also the contours of the objects were finely smoothed.



(a) Median filter (b) Closing (c) Opening

Fig. 3. Result of noise filtering. At this step we have some artifacts yet (see the arrows).

As it can be noted in Fig. 3c, filters applied so far did not completely eliminate the unwanted artifacts in the image. For this reason, we calculated the area of objects counting the pixels of each object. Those objects having an area smaller than the average area of the nuclei were removed, obtaining the clean nuclei shown in Fig. 4.

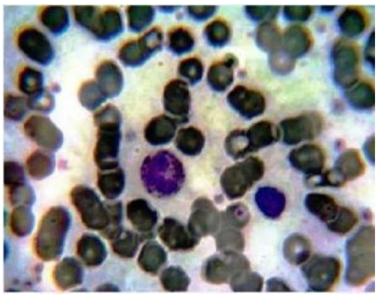
Fig. 4. Clean cells nuclei.

4.2 Leukocytes segmentation

After segmenting the nuclei to isolate them from the erythrocytes, we proceed to separate them from the nuclear membrane. To do this, we equalize the histogram of the original image (Fig. 1). The result of this operation is shown in Fig. 5. Note the difference in color of leucocytes respect to the erythrocytes color and background.

Next, the binary image is obtained by using adaptive thresholding over the green plane. In this way we could get the segmented cells. The result is shown in Fig. 6a. Similarly the median filter, and closing and opening operators are applied in order to obtain an image with the least possible noise and the cell membrane well defined (Fig. 6b). Finally, we got the image without noise and artifacts as shown in Fig. 7.

At this stage of the process, we already have the binary images of the nuclei (Fig. 4) and of the segmented cells (Fig. 9). However, it is still necessary to obtain the objects of interest in its original color in order to extract from them the characteristics that will help us to classify them. To get these objects in their original color we apply the logical AND between the original image (Fig. 1) and the segmented leukocytes without noise (Fig. 7). The fully segmented objects are shown in the Fig. 8 (left) and separately in Fig. 8 (right).



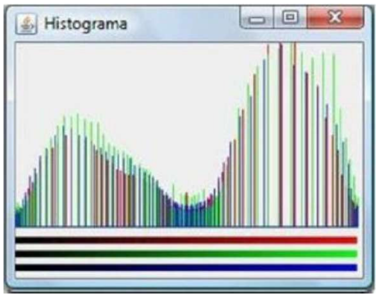


Fig. 5. Original image equalized (left) and histograms (right) after the equalization.

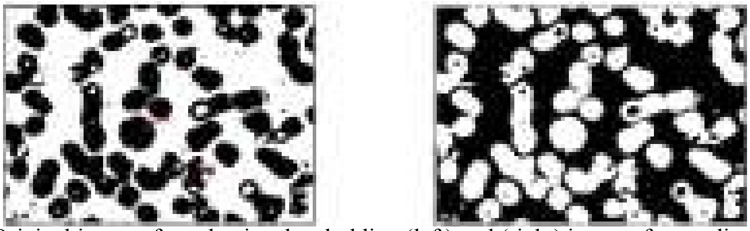


Fig. 6. Original image after adaptive thresholding (left) and (right) image after median filtering, closing and opening operations.

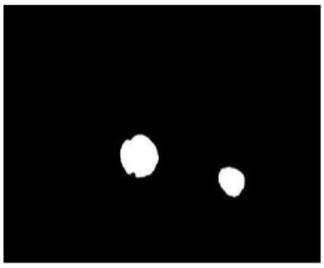


Fig. 7. Clean image of the leukocytes.

Once the cells are fully segmented, the classifier *k*-NN was applied using the first Minkowski[1] metric to perform the count of the five normal form objects of study.

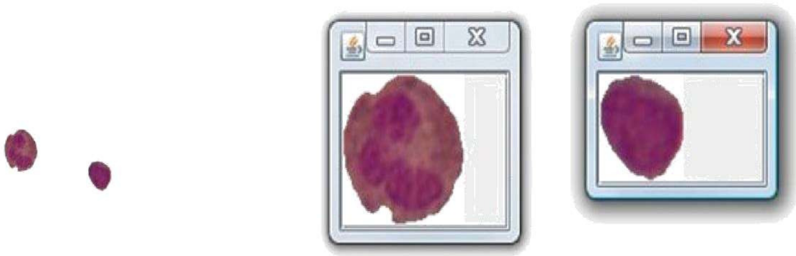


Fig. 8. (Left) leukocytes segmented with its original colors, and (right) leukocytes shown in separated images.

5 Results and discussion

To test the effectiveness of the proposed method, we compared the results obtained by our system against those obtained by a hematologist. The test consisted in classifying and counting the leukocytes present in a sample of 45 smears obtained from patients

---

[1] The first Minkowski metric between two points (also known as Manhattan) is defined as $L_1 = \sum_{i=1}^{d} |x_i - y_i|$.

of the National Institute of Rehabilitation, in Mexico City. The sample contained 21 neutrophils, 9 eosinophils, 2 basophils, 7 monocytes and 21 Lymphocytes previously classified by the hematologist. Table 1 shows the number of leukocytes of each type present in 45 smears previously classified by a specialist vs. the number of object detected and classified by our application.

During the recognition process many data were obtained related to the five types of leukocytes detected with the application. The most definitories were the Mean Value and the Standard Deviation, which are shown in Table 2. The *k*-NN classifier in combination with the first Minkowski metric was used for classifing the type of leukocytes amongst the five possible types. The features used to classify the objects were the nucleus area, the total object area and the compacity [1]. The efficiency of the application was calculated based on the number of detection included in Table 1. Table 3 shows the values achieved with the common equations related to the confusion matrix. Figure 11 shows the ROC (Receiver Operating Characteristic) curve plotted from the values of Table 3.

Table 1. Comparison of detected leukocytes versus real data classified by a specialist.

Leukocyte type	Real data	Detected
Neutrophils	21	23
Eosinophils	9	7
Basophils	2	3
Monocytes	7	5
Lymphocytes	21	22
Totals	**60**	**60**

Table 2. Values that discriminate the five types of leukocytes.

	Nucleus area	Total area	AN/AC Rate	Compacity
Lynphocytes				
Medium Value	15.38	21.66	2.88	0.79
St. Deviation	2.61	3.17	1.12	0.02
Neutrophils				
Medium Value	15.82	37.89	0.74	0.90
St. Deviation	2.58	5.76	0.22	0.10
Basophils				
Medium Value	27.39	31.69	6.83	0.80
St. Deviation	6.00	7.83	1.71	0.03
Monocytes				
Medium Value	17.19	27.77	2.17	0.87
St. Deviation	4.04	4.23	1.08	0.11
Eosinophils				
Medium Value	29.37	41.6	3.17	0.82
St. Deviation	3.55	3.89	1.73	0.02

Table 3. Values calculated from the results obtained and shown in Table 1.

	True positive rate	False positive rate
Monocytes	0.71	0
Eosinophils	0.78	0
Basophils	1.00	0.02
Lymphocytes	1.00	0.03
Neutrophils	1.00	0.05

The point (0, 1) represents the perfect classification. That means 100% sensitivity (no false negatives) and 100% specificity (no false positives). Note the diagonal that divides the ROC space; this diagonal represents the random classification. Points above the diagonal represent good classification and points below represent poor classification. From our test we obtained results that could be considered very near the perfect classification.

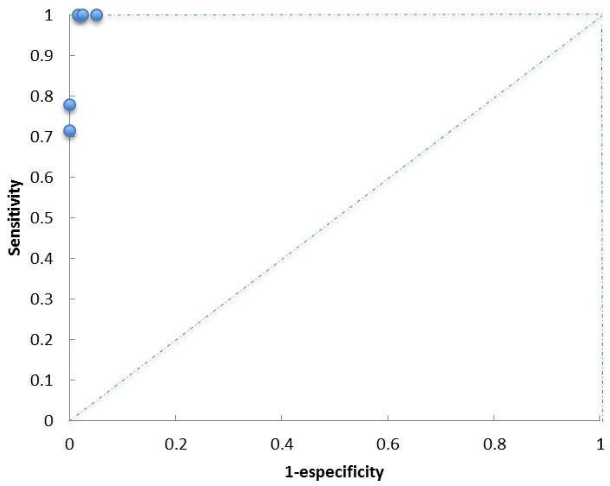


Fig. 11. ROC curve plotted from the results in Table 4.

6 Conclusions

In this work we presented a methodology to detect, classify and count five normal forms of leukocytes, namely, neutrophils, eosinophils, basophils, monocytes and lymphocytes. The effectiveness of our method was tested with a blind experiment. Using the *k*-NN classifier in combination with the first Minkowski metric, in addition to techniques of digital image processing, a computational system platform-independent was able to identify, to classify and to count the five normal types of leukocytes. The paper does not attempt to differentiate between smears of leukocytes coming from healthy and sick people; but only to detect their presence in peripheral blood independently of their number. The system could be used in emerging areas such as the topographic hematology and the chronobiology.

Acknowledgements

Jesús Martínez-Castro and Edgardo Felipe-Riverón thank to the Centro de Investigación en Computación of the Instituto Politécnico Nacional, COFAA-IPN, and SNI-CONACYT programs, Mexico, for the support given to this work. Susana Reyes-Cadena thanks to the National Institute of Rehabilitation, Mexico City, Mexico, for the support given to this work.

References

1. R. C. Gonzalez and R.E. Woods. Digital Image Processing. Pearson/Prentice Hall, (2008).
2. L. Costrarido. Medical Image Analysis Methods: Evaluation Strategies for Medical-image Analysis. Taylor & Francis, pp. 433-471, United States of America, (2005).
3. B. Ravi Kumar, D.K. Joseph, T.V. Sreenivas. Teager Energy Based Blood Cell Segmentation, Teager Energy Based Blood Cell Segmentation, Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India, (2002).
4. A. Bello González. Hematología Básica. Prado, (2001).
5. H. Ceelie, R.B. Dinkelaar, W. Gelder. Examination of peripheral blood films using automated microscopy: evaluation of Diffmaster Octavia and Cellavision DM 96. J. Clin. Pathol. 60, 72-79, (2006).
6. Q. M. Liao, Y.Y Deng. An accurate segmentation method for white blood cell images. In: Proceedings of the IEEE Intern. Symposium on Biomedical Imaging, pp. 245-248, (2002).
7. Chen Pan, Dong Sun Park, Sook Yoon and Yu Cheng Yang. Leukocyte image segmentation using simulated visual attention. Expert Systems with Applications 39, 7479-7494, (2012).
8. R. Adollah, M.Y. Mashor, N.F. Mohd Nasir, H. Rosline, H. Mahsin, H. Adilah. Blood Cell Image Segmentation: A Review, Biomed 2008, Proceeding 21, pp. 141-144, (2008).
9. M. Ghosh, D. Das, C. Chakraborty, A.K. Ray. Automated leukocyte recognition using fuzzy divergence. Micron 41, pp. 840-846, (2010).
10. C. Di Ruberto, A. Dempster, S. Khan, B. Jarra. Analysis of infected blood cell images using morphological operators, Image and Video Computing 20, pp. 133-146, (2002).
11. N. Guo, L. Zeng, Q. Wu. A method based on multispectral imaging technique for White Blood Cell segmentation, Computers in Biology and Medicine 37, pp. 70-76, (2006).
12. M. Saraswat, K.V. Arya, H. Sharma. Leukocyte segmentation in tissue images using differential evolution algorithm, Swarm and Evolutionary Computation, http://dx.doi.org/10.1016/j.swevo.2013.02.003, (2013).
13. S. H. Rezatofighi, H.Soltanian-Zadeh. Automatic recognition of five types of white blood cells in peripheral blood, Computerized Medical Imaging and Graphics 35, pp. 333-343, (2011).

Clustering Time Series by Wavelet and Evolutionary Computing

Antonio Ramírez, Ricardo Barrón, Giovanni Guzmán, Edgar. A. García

Artificial Intelligence Laboratory, Center for Computing Research
National Polytechnic Institute, Mexico
{tonotron, barron3121, giovanni.guzman, zavael}@gmail.com

Abstract. The classification of information can take various hues, depending on the approach to be performed [1], however to achieve this classification is one of the most important tasks of computational science and several of its branches [2]. In this paper a classification of time series by calculating numerical wavelet coefficients is presented, since it has been shown to be a very good tool for signal analysis [3]. A new approach in which tuning parameters of a classification function are optimized in a supervised way by an evolutionary algorithm proposed is also presented. With the use of different wavelets the best way to get descriptors for each data set was deducted, then sorted and grouped by the use of the evolutionary algorithm, combining these two tools achieved a better classification of the information presented in numeric strings, considered as time series.

Keywords. Wavelets, classification, time series, evolutionary computing, metaheuristics.

1 Introduction

Over the last five decades, preferences concerning with evolutionary algorithms have been growing. This framework, offers a wide set of techniques for solving the problem of searching optimal values by using computational models inspired by Nature, particularly evolutionary processes, the called Evolutionary Algorithms (EAs) which are population based optimization techniques and are designed for searching optimal values, in big data sets [5]. On the other hand, the signals analysis has been optimized thanks to employ of wavelets, which are considered one of most accurate ways to locate minutiae over signals [4]. Techniques, EA and wavelets are the most robust methods to acquire accurate information from a signal, which able us, to make a better interpretation of such data [16]. This paper applies the combination of EA and wavelets to generate a spectral signature descriptor based on the wavelet calculation [2], taking into account the time series of productivity of some oil wells. The information of such wells, which is grouped by a bio-inspired PSO algorithm [3], is the basis for characterization and classification.

Paper Recived 13-09-2013 and Acepted 21-10-2013

2 Theoretical backgrounds

This section provides some preliminaries about the state of the art of the tools used in this work. The different methods or algorithms for clustering, and their relationship with the time series, are also treated. It will be established some clustering algorithms from single to widely used, even the most modern and sophisticated based on evolutionary algorithms. Finally, we present a brief introduction to wavelet theory, which combined with meta-heuristics, serve to optimize the time carried out by the clustering algorithm. Also we conducted a comparison against the brute force algorithm proposed by [17], for classification of big data sets, in which good results were obtained.

2.1 Clustering

One of the best ways to handle large concentrations of data is clustering. A cluster is a collection of elements that share common characteristics, which can be separated and also a part of a group. The study area now known as clustering, is a method of unsupervised learning, and such clustering without a previous pattern separation or pre-classification. One of the first clustering algorithms is the classic K-means [14], which is quick and easy to be implemented. In contrast, such algorithm is also very sensitive to the initialization of the centroids and easily falls into the so-called "local optimum". A less sensitive algorithm initialization was proposed by Zhang, B. [15] called K-harmonic means (KHM), on a technical report for HP Labs, and this algorithm employs harmonic means for calculating centroids and is in some cases more efficient. For its part, the classical Pattern Recognition algorithm, also been successfully applied for the definition of various strategies for regulating the exploration of large search spaces.

Aforementioned algorithms are basically processes that partitions a population into k n-dimensional sets, based on calculations of average distance between each of the elements and the number of partitions obtained according to its variance. The difference between them is that one is based on the calculation and their means distances (KM) and the other based in their harmonic means (KHM), in order to determine their membership in a given cluster. Clustering is considered as an unsupervised learning method, and it is commonly used for the analysis of statistical data in many disciplines, including data mining, pattern recognition, image analysis, machine learning and bioinformatics to mention only some [3 7 8].

2.2 Particle swarm optimization

Particle Swarm Optimization (PSO) is a meta-heuristic based on populations of individuals as those found in nature, as swarms of ants, banks of fishes, flocks of birds, etc. The algorithm developed in 1995 by Kennedy and Eberhart, is based mainly on psychosocial approach known as "social metaphor" [16], PSO algorithm can be summarized as: "every individual can change their opinion and participate or not in a group, with based on three factors, environmental knowledge, experience, and knowledge of experiences of individuals in his neighborhood. This individual then adapts his schemes, beliefs and opinion in accordance with individuals who have better experiences in their environment and based on interaction rules laid down"[13].

The heuristic techniques for exploring large search spaces, used by these two disciplines have been used successfully to define data clustering algorithms, allowing troubleshooting data clustering and classification of large volumes of data, or in a dynamic context [8].

Using heuristic algorithms for defining cluster has some important advantages over the use of traditional construction algorithms [7]. The use of heuristics requires a programmer to stake a grouping or classification problem as one of optimization (either looking for the optimal position of the centroid of each group, or optimizing the topological properties of the final clustering) which in turn permits to define clustering algorithms that provide a certain probability of quality response as data volume expands.

2.3 Wavelets

The wavelets are short wave functions with compact support that can represent both a signal or time series in either the time or the frequency domain [5]. Unlike the Fourier transform, wavelet transform consists of a set of basic functions, often called mother functions, which allows a local description of the behavior in frequency (spectrum of the signal). Moreover, wavelets can capture transient data and not only the average frequency behavior, as occurs in the Fourier transform.

The wavelet transform of a function $\varphi(t)$, corresponds to the decomposition of that function, in a group of daughters functions with the form $\psi_{(m,n)}(t)$, called wavelets. As it occurs in the continuous case, there is also the wavelet transform for the discrete case that can be defined as.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t-b}{a}\right) \tag{1}$$

where b ∈ R is a translation parameter and ∈ R, (a > 0), is regarded as a scale factor. In the development of this work, tests were conducted with the families of Haar wavelet [6] and Daubechies [7], using the criterion of energy minimum variation. This criterion refers that the coefficients resulting from the scaling factor of any signal must preserve energy from the signal.

2.4 Time series properties

For concreteness, we begin with a definition of our data-type interest of a time series. Definition: A time series $T = t_1, \dots, t_j$ is an ordered set or not, of j real-valued elements. Assume a time series $\vec{X}\ (\vec{X} \in \mathbb{R}^n)$ is located in the J scale, this series can be decompose into a j specific scale $(j \in [0, 1, \dots, J-1])$, then the coefficients $H_j(\vec{X})$ correspond to the scale j and can be represented by a series $\{\vec{A_j}, \vec{D_j}, \dots, \vec{D_{J-1}}\}$; where the $\overrightarrow{A_J}$ must be considered an approximation coefficients which are the projection of $\vec{X}$ in V_J and the $\overrightarrow{D_J}, \dots, \overrightarrow{D_{j-1}}$ are the wavelet coefficients in $W_J, \dots, W_{j-1}$ representing the detail information of $\vec{X}$, where W_{j-1} is considered a sub-space or complement of $V_j = V_{j-1} \oplus W_{j-1}$, since by defining the $\varphi_{j,k}$ and $\psi_{j,k}$ are the orthogonal basis of W_j; $\{\varphi_{j,k}, \psi_{j,k}; j \in \mathbb{Z}, k \in \mathbb{Z}\}$ [6 17].

2.5 Spectrar signature

Each element over earth's surface such as forests, crops, rivers, lakes, buildings, plains, human default erosions, etc., have the ability to transform differently the electromagnetic radiation receiving of the sun. Each type of these objects, have a level of specific response and this is measured as spectral signature.

The amount of energy variation that one object can reflect (reflectance), is a function of the wavelength that each emits body and is called spectral signature or spectral firm [11]. The spectral signature is then a quantitative measure of the spectral properties of an object in one or more spectral bands. Also known as spectral behavior. Thereby, any object can vary in quality and / or quantity of their spectral signature, depending on weather conditions, seasons of the year and essentially lighting conditions. Thereby, any object can vary in quality and / or quantity of their spectral signature, depending on weather conditions, seasons of the year and essentially lighting conditions. So it is possible to identify in a image for instance, the organic and physical nature of a particular object. In the analysis of data sets may also obtain a spectral response of the objects that represent such information. It means in a data set, we could identify small variations and to interpret info about them.

3 Proposed algorithm

For spectral characterization, which we call 'signature', calculate the wavelet transform of the signal to identify active frequencies and then choose a representative part of the signal, based on an energy criterion or threshold limit. Taking into account the information provided for the analysis and implementation of present work, we considered a data matrix in which first column is the code name for each well and the rest of the information on each line corresponds to data supplied each of them. In order to apply the proposed wavelet family, it was necessary to standardize the data in a matrix in which the data should be calculated in 2n basis, always considering the importance of using as closely as possible the information provided. Following the next criteria made this:

(a) The data elements of each row, represent one record of monthly production reading date from October 1980 to April 2011 yielding 367 columns of data.
(b) With intention that the data table was uniform and his columns were also a multiple of 2n, this matrix was completed with a zero in the empty positions (before and after the column 368).

(c) When data (columns) nonzero did not exceed 256, that figure was adjusted to apply the wavelet calculation.

The treatment of the lack of information that occurred in nearly 50% of the data matrix, was performed with the algorithm that shows in Table 1.

The heuristic technique used during this first phase of experimentation is Particle Swarm Optimization (PSO) [10]. The choice was due to the proven ability of this technique to explore large search spaces using a relatively small number of particles.

Furthermore, this technique allows to combine various search strategies, both on global and local criteria, which, inevitably leads to the possibility of running the parallel computational model with the consequent increase in performance.

Table 1. Pseudo code of treament of lack of information.

```
Read line
  count Data;
    if Data > 255
      For each blank,
            delete blank
            if Data <= 256 break
          end for
    else
       while Data < 512 add '0'
    end if
End
```

When the volume of data is considerably large in size according to the type of information that is analyzed, it is advisable to group the data by calculating the standard deviation typical and obtain uniformity coefficients, which will give rise to patterns leaders or centroid by calculating the variance. Overall building a digital distribution commonly consists of three stages: 1) to determine the "classes" and their intervals, 2) classifying (or distribute) the data in the appropriate classes, and 3) count the number of cases of each class, and set the corresponding "class mark" or centroid.

The algorithm implemented in this work uses two cycles (external and internal) that optimize the number of clusters and clustering quality using a cluster validation index and overall variance to form the groups. The outer loop of the algorithm, which measures the quality of clustering and optimizes the number of groups, used a model of Particle Swarm Optimization. The inner loop uses a hybrid algorithm between microPSO, and local adjustment algorithm. This hybrid algorithm form groups passing to outer loop. The solution that delivers the inner loop is a grouping, which passes to outer loop and this last one, measures the group quality and this way optimizes the number of groups.

3.1 External cycle

The fitness function of PSO algorithm performed by external cycle use the index j, which measures the quality of clustering and thus optimize the number of groups. The index j is described as follows:

$$J(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^P \quad (2)$$

where K is the number of groups, and P is a real number greater than or equal to 1, that controls the contrast between different cluster configurations. In the above equation E_k y D_k are described as follows:

$$E_K = \sum_{k=1}^{K} \sum_{j=1}^{n} u_{Kj} \left\| x_j - z_k \right\| \quad (3)$$

$$D_K = max_{i,j=1,\dots k} \left\| z_i - z_j \right\| \quad (4)$$

Where n is the total number of points in the data set, z_k is the centroid of *k-th* group, x_j is the *j-th* item from *k* group. The *index I* is a three factors composition called $\frac{1}{K}$, $\frac{E_1}{E_k}$ y D_k. The first factor decreases linearly as the value of *k* increases. Thus, this factor is to reduce the value of the *index i* to increase the value of *k*. The second factor is the radius of E_1, which is a constant value. The factor E_k, decreases when *k* value increases. The *index I* value increases with E_k decreasing. This, indicates that more groups should be formed that should be compact in nature. The third factor (which measures the maximum separation between two groups) will increase with the value of *k*. So while the first factor is decrease the value of *k*, the second and third factors try to increase the value of *k* favoring compact well-separated groups.

3.2 Internal cycle

The hybrid algorithm microPSO and local adjustment, used the fitness function as global variance, which is defined as follows:

$$f(x) = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{n} D_E(o_i, C_j) \quad (5)$$

where *K* is the group number, *n* is the number of elements belonging to the group; D_E is the function of similarity / difference between patterns; O_i is the *i-th* pattern that belongs to the class *j* and C_j is the centroid of the class. A pseudo code of this proposal is shown in Table 2.

3.3 Comparing algorithms for times series classification

The brute force algorithm let us to appreciate that we simply take each possible sequence of numerical data and find the distance to the nearest (any other), is match and the subsequence that has the greatest such value is the discordant. This is achieved with nested cycles, where the external cycle considers each possible candidate subsequence, and the internal cycle in this case, is a linear scan to identify the candidate's nearest any-other match.

In the Table 3, we can see the pseudo code of the algorithm, by [17]. Such, requires just one parameter, the length of subsequences to consider. The algorithm is easy to implement and produces exact results. However, the algorithm that we propose has not the excessive spending flaw, for data mining as brute force algorithm which has O(n^2) time-complexity.

Table 2. Pseudo code clustering cycles (external and internal).

```
GenerateInitialPopulation(P1)
Evaluate(P1)
while not ExternalConvergence(P1) do
        foreach particle (p_i) in P1
                GeneratePopulation(P2)
                Evaluate(P2)
                while not InternalConvergence (P2) do
                        foreach particle (x_i) in P2
                                CalculateNewPosition(x_i)
                                LocalAdjustment(x_i)
                                Evaluate(x_i)
                                ReplaceGlobalBest(P2)
                        end foreach
                end while
                SelectInternalBestIndividuals(P2)
                CalculateNewPosition(p_i)
                Evaluate(p_i)
                ReplaceGlobalBest(P1)
        end foreach
end while
SelectBestSolution(P1)
```

Table 3. Pseudocode code Brute Force Discord algorithm.

```
Function [dist, loc ]= Brute_Force(T,n)
  best_dist = 0
  best_local = N
  for p = 1 to |T|-n+1
    nearest_neighbor_dist = infinity
    for q = 1 to |T | - n + 1
      if p–q ≥ n
        if Dist (t_p,…,t_{p+n-1}, t_q,…,t_{q+n-1}) < nearest_neighbor_dist
          nearest_neighbor_dist = Dist (t_p,…,t_{p+n-1}, t_q,…,t_{q+n-1})
        end if
      end if
    end for
    if nearest_neighbor_dist > best_dist
      best_dist = nearest_neighbor_dist
      best_local = p
    end if
  end for
Return[ best_dist, best_local ]
```

3.4 Configuration parameters for experimentation

In the development of a configuration, experiments used for all tests, both for internal and external cycle. In Tables 4 and 5 shows the configuration data of their parameters.

Table 4. Configuration parameters of the external-cycle algorithm.

[1] Parameter	[2] Value
[3] Swarm size	[4] 20
[5] Search space dimensions	[6] 1
[7] Inertia factor	[8] 0.9
[9] Cognitive factor weight	[10] 1.2
[11] Social factor weight	[12] 1.4
[13] Limits to search space	[14] [2 N] where N is total number of data
[15] Max velocity	[16] 1.2

Table 5. Configuration parameters of the inner loop hybrid algorithm.

[17] Parameter	[18] Value
[19] Swarm size	[20] 3
[21] Search space dimensions	[22] 2
[23] Inertia factor	[24] 0.9
[25] Cognitive factor weight	[26] 1.6
[27] Social factor weight	[28] 1.8
[29] Limits to search space	[30] [0.0 3428835.7]
[31] Max velocity	[32] 1.5

3.5 Classification of wells using historical data

According to the information that had at that moment, was determined use the periodic information of each record, which in turn represent the wells production, as time series that can be analyzed as with digital signals. It was considered for this purpose, using wave-signal analysis, best known as wavelets analysis. This information should be standardized as a matrix, where each row corresponds to the registration of a well and this in turn should be composed of a number of 2^n-based data.

Thus, a small record an array of data, representing a signal from a well. At this signal (numerical series), was applied to the wavelet calculation for obtain a tuple of representative features, to this tuple (two values) we call descriptor. These descriptors are compared to each other to find the class it belongs to each well. This is determined by a minimum proximity criterion (calculating the Euclidean distance) with a centroid. This centroid, is a representative element for each cluster, is also a tuple of two numerical data. At this stage of development, the centroids are defined as statics.

Once you have found the optimal clusters of wells according to their productivity and their proximity to class centroids is possible to encode these classes and placed it in any of them. The production time series from wells, gives the grouping with formed clumps by algorithm known as ("clustering").

Algorithm to classify a well given their time series:
1. Standardize the sample size
2. Calculating the descriptor or spectral signature using wavelet transform.
3. Calculating the distance of the spectral signature of the well with respect to the centroid.
4. Assign the well label the corresponding class (classification).
5. Perform static grouping of wells (clustering)

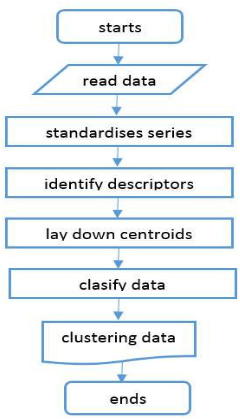


Fig. 1. Total classification algorithm flowchart.

4 Experimental results

To get the pooling of data from time series, were translated the original file format spreadsheet format to a flat file comma separated text (file.csv). Those data organized as a matrix then we applied the calculation algorithm based on wavelet spectral signature.

In the test phase we used two families of orthogonal wavelets: Haar and Daubechies, obtaining very similar results. We decided to use only the calculation of the spectral signature with the Haar, due this one was faster and offered better results.

With the spectral signature (time series analysis of the data) was obtained identification record for each of the wells then information was used as sample learning for clustering algorithm, generating different groups of data and different classes or groups.

Flat file was generated with each sediment records placed (Table 6). The final information is presented in four columns, the number of well, the coefficients calculated with the wavelet, which in this case are considered traits descriptors each well and the number of class to which they belong. These results may not be less than ideal, since the information available to the data file is very limited. In some cases it was necessary to assign a zero to the missing data.

Table 6. View the data analysis and classification.

POZO ID	DESC 1	DESC 2	CLASE
POZO 1	2261009.45965523	645128.454706114	CLASE 3
POZO 2	1114.4373721754	18696.6893687409	CLASE 6
POZO 3	0	17977.5910761071	CLASE 6
POZO 4	1263766.3033614	68687.3132890685	CLASE 5
POZO 5	1263766.3033614	164943.967887151	CLASE 5
POZO 6	0	0	CLASE 4
POZO 7	125.818652279295	1995.80321136723	CLASE 4
POZO 8	162.346705130885	2575.23126180834	CLASE 4
POZO 9	191603.720039493	42381.2007671788	CLASE 7
POZO 10	1205.24246354158	0	CLASE 6
POZO 11	109829.972788095	1006.87036987241	CLASE 6
POZO 12	1103145.18273855	43832.0783253027	CLASE 5

The experimental data was contained in a flat file, was treated as numerical series and subjects to mathematical calculus that result in a table of two columns in which are represented the descriptors of those series. These descriptors are evaluated to know to which class they belong. In Table 7, it shows a list of classes as was obtained from the PSO algorithm, within one of the end files. The final data within on figure 3, correspond to identification from each well, into each class.

The algorithm described above in Table 2 determined these seven types. Fixed data were used in each experiment shown in Table 8. The unique parameter that changed on the experiments was the fitness value.

When we compare the clustering by wavelet and evolutionary algorithm, with brute force algorithm, we can see a substantial difference in the time-convergence of clustering over similar data sets as shown in Figure 2.

Table 7. Classification of centroids obtained from PSO algorithm, a) best fitness values, b) coordinates of centroids.

1. 2214702.150332	**C1** (751072.158462, 196043.994000)
2. 2314530.151289	**C2** (3247609.92500, 423875.33500)
3. 2333896.527851	**C3** (2334484.40000, 330180.57900)
4. 2073020.127509	**C4** (31269.16174, 221666.88895)
5. 2080099.692564	**C5** (1258644.26400, 131729.75784)
6. 2333896.527851	**C6** (16855.87879, 22789.45322)
7. 2338401.671122	**C7** (314861.60150, 20148.04506)

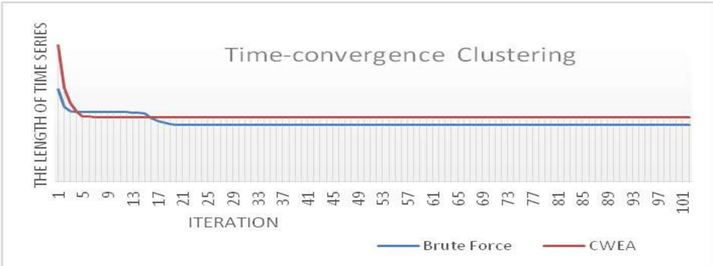


Fig. 2. Comparison of execution time between BF and CWEA algorithms.

Table 8. UGPSO Algorithm parameters.

1. Algorithm:	Ugpso
2. Adjust:	On
3. Groups:	7
4. Generations:	200
5. Popsize:	3
6. Inertia:	0.900000
7. Social:	1.800.000
8. Cognitive:	1.600.000
9. Vmax:	1.500.000
10. Range:	[0.000000, 3428835.700000]
11. PopKeep:	1
12. Fitness value	2073020.127509

5 Conclusions

In this paper, we presented a new approach of clustering, in which the tuning parameters a classification function is being optimized on supervised way by a combined heuristic technique. The proposed PSO-based technique being population-based random, then the search optimization technique does not require initialization of adjusting parameters.

Depending on some circumstances, big data sets would be considered like a temporary signal scannable as any other signal. In signal analysis area, there are several techniques to identify significant differences that allow us to isolate the unique traits it need to represent a pattern. When it is possible to identify patterns in the signals, they can be sorted and grouped. For this work we considered the information that was counted as time series data as represented.

The technique used in this work was able to separate into classes a list of series, based on calculations with wavelet and using the evolutionary algorithm to optimize a classification function, finding that with the proposed algorithm, it was finally possible to properly clustering the oil wells in seven categories, based on the analysis of time series representing historical record of production.

References

1. T. Babdagli, Development of mature oil fields a review, Journal of Petroleum Science and Engineering, Elsevier, (2007).
2. Qian Tao, Vai Mang, Wavelet Analysis and Applications, Applied and Numerical Harmonic Analysis, Springer Science + Business Media, Germany, (2007).
3. G. Luque, E. Alba, Parallel Genetic Algorithms, Theory and Real World Applications, Studies in Computational Intelligence, Springer-Verling, Spain, (2011)
4. L. Guan, Y. Du, L. Li, Wavelets in Petroleum Industry: Past, Present and Future, Society of Petroleum Engineers Inc. SPE 89952. (2004).
5. K. Premalatha. A New Approach for Data Clustering Based on PSO whit Local Search, CCSE, Computer and Information Science. Vol. l, No. 4, (2008)
6. Hui Zhan, et al, Unsupervised Feature Extraction for Time Series Clustering Usign Orthogonal Wavelet Transform, Informatics 30, (305-319), (2006).
7. Ten lectures on Wavelets. Daubechies I. Society for Industrial and Applied Mathematics (1992).
8. Michael W. Berry, Survery of Text Mining: Clustering, Classification and Retrieval, Springer, (2004).
9. Swagatam Das, Amit Konar, Meta-heuristic Clustering". Studies in Computational Intelligence 178. Springer, (2010).
10. Xin-SheYang, Nature-Inspired Metaheuristic Algorithms. Luniver Press, Second Edition, (2010).
11. Sean Luke, Essentials of Metaheuristics. Second Edition, Lulu Publishing, (2013).
12. Particle Swarm Optimization: Theory, Techniques and Applications (Engineering Tools, Techniques and Tables). Andrea E. Olsson (E). Nova Science Pub Inc, (2011).
13. Kennedy J., Russell C., Yuhui S., (Swarm Intelligence). The Morgan Kaufmann Series in Evolutionary Computation, USA, (2001).
14. MacQueen, J. B, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California, USA, (1967).
15. Zhang, B., Hsu, M., Dayal, U., K-harmonic means, a data clustering algorithm. Technical Report HPL-1999-124, Hewlett-Packard Laboratories, (1999).
16. Yuehui Chen, et al, A Local Linear Wavelet Neural Network, Proceedings of Congress on Intelligent Control and Automation, Hangzhou, P. R. China, (2004).
17. E. Keogh, J. Lin, HOT SAX: Finding the Most Unusual Time Series Subsequence: Algorithms and Applications. The Fifth IEEE International Conference on Data Mining, USA, (2005).

Evaluation and Limited Characterization of a Color Image Segmentation Method using Synthetic Images

Rodolfo Alvarado-Cervantes, Edgardo M. Felipe-Riveron and Rodolfo Alvarado-Martinez

Centro de Investigación en Computación, Instituto Politécnico Nacional, México

ateramex@gmail.com; edgardo@cic.ipn.mx; ralvarado@ateramex.com.mx

Abstract. We present an evaluation and a limited characterization of an own semi-automatic color image segmentation method using generated synthetic images with its associated ground truth. The synthetic images where designed to evaluate the efficiency of the resulting color information achieved from color segmentation algorithms. By the use of ROC curves and its analysis, we achieved the evaluation and some particular characteristics of our segmentation method, such as the level of sensibility related to the threshold selection and to the appropriate number of pixels to have by the color sample needed by the algorithm.

Keywords: Color segmentation evaluation; color segmentation characterization; synthetic color image generation; ROC curves evaluation.

1 Introduction

Image segmentation consists of partitioning an entire image into different regions, which are similar in some predefined manner. Segmentation is an important feature of human visual perception, which manifests itself spontaneously and naturally. All subsequent steps, such as objects recognition depend heavily on the quality of segmentation [1].

While the development of segmentation algorithms has attracted remarkable consideration, relatively fewer efforts have been spent on their evaluation [2], [3], [4] and [5]. Since none of the proposed automatic segmentation algorithms published are generally applicable to all types of images, and different algorithms are not equally suitable for particular applications, the performance evaluation of segmentation algorithms and its characterization are very important subjects in the study of segmentation [2].

For a long time the evaluation was limited to few real images acquired from particular application, that has as advantage that they are closer to reality although its intrinsic random nature makes them no suitable for analytical evaluation [2], [3], [6]. Many undetermined characteristics of those images prevent its use to compare different segmentation techniques because many phenomena are mixed and make it difficult to study its influence individually [2]. Another problem arrives from the lack

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 45–54
Paper Recived 14-09-2013 and Acepted 21-10-2013

of a Ground Truth (GT) that has to be obtained from 'experts' whose results always have intrinsic differences. This subjective and imprecise procedure is not appropriate for quantitative evaluation [3], [6].

So far segmentation evaluation methods can be divided in two groups: analytical and empirical. The analytical methods directly inspect and evaluate the segmentation algorithms themselves by analyzing their principles and properties. The empirical methods indirectly judge the efficiency of segmentation algorithms applying them to test images and measuring the quality of the results [2], [4], [5].

Several empirical methods have been proposed, the great majority can be classified in two types: goodness methods and discrepancy methods. In the first category some desirable properties of segmented images, often established according to human intuition, are measured by ‘goodness parameters’. The performances of segmentation algorithms are judged by the values of goodness measures. In the second group the GT that presents the ideal or expected segmentation result must be first found. The actual segmentation results obtained by applying a segmentation algorithm are compared with the reference to count their differences. The performances of segmentation algorithms under investigation are then estimated according to discrepancy measures [2], [3], [4] and [5].

Receiver operating characteristics (ROC) curves are useful for organizing classifiers and visualizing their performance. ROC curves are commonly used in medical decision making, and in recent years have been used increasingly in machine learning and data mining research [10].

In this work an evaluation and characterization of a semi-automatic color image segmentation method using synthetic images generated with its associated ground truth is presented. The synthetic images where designed to evaluate the efficiency of achieved color information from a given segmentation algorithms. The system was applied to our semiautomatic color segmentation method already presented in international conferences [1], [7]. By the use and analysis of ROC curves we obtained some proper characteristics of the segmentation method under study such as its stability related to the threshold selection and to the selection of an appropriate number of pixels required by the color samples. This system may be useful for assessing the quality of use of the color information inside the segmentation algorithms in general.

2 Previous works

The first comprehensive survey on evaluation methods of image segmentation is presented in [2]. It brings a coherent classification of existing methods at the time. An up to date of 5 years of progresses in the subject is presented in [4] after the first survey. Another actualization is presented 5 years later [5], embracing together the principal methods of segmentation evaluation available until 2007.

In [3] a comprehensive survey on unsupervised methods of segmentation evaluation is presented. The authors propose a hierarchy of published methods at that time by summarizing the main branches and locating the group of unsupervised methods on it. They mention its advantages such as not requirement of GT to obtain quantitative results. They also propose the main lines of future research of this kind of methods.

In [6] a way to design synthetic images and a corresponding GT for evaluating segmentations algorithms is presented. They introduce a general framework and general design considerations. They also present a system for generating synthetic images in shades of gray taking into account theirs design considerations. The behavior of a segmentation method in gray images using thresholding is studied and some remarks are obtained.

In [8] is presented a database containing GT from segmentations produced manually by a group of people from a wide variety of color natural scenes. The authors define two related error measures that quantify the consistency between segmentations of different granularities. This measure permits to make comparisons between segmentations made by people and segmentations made by the computer of the same scene.

In [11] is presented a comparative study of 14 evaluation criteria of supervised image segmentation methods by generating edges. The study was done in two parts: (1) evaluation with synthetic images to characterize the general behavior of the algorithm is complemented by (2) an assessment over a selection of real color images. To get the GT of different peoples the authors mention that theirs propositions were merged, but do not give details of how they accomplished it.

3 Design and synthetic image generation

In [6] three important design considerations for synthetic images are presented: (1) Synthetic images should be appropriate for a quantitative study and should allow objective evaluations of its properties. (2) The synthetic images should reflect the main features of real images, i.e. corruption factors such as noise and blurring, variation of parameters such as size, shape, etc. (3). The system should allow the generation of images with progressive variations of each parameter. In this way the study of the influence of each individual parameter is possible.

In this work we try to extend the concepts and design considerations to RGB true color images presented in [6]. We created synthetic images with figure and background in color and selected the circle as the basic object to create our base image. The circle is appropriate because it is a base model of the objects of interest in various fields such as the detection of white blood cells in microscopic images, and others. It is also more difficult to detect than other objects made of straight lines by methods such as convolution with masks that highlight edges.

There are other techniques for generating synthetic images such as edge-oriented methods [11] or by the introduction of texture in the images [12]. In our case, the segmentation method is region-oriented without using the texture information; the synthetic images were designed with these considerations.

The basic colors selected for both figure and background were based on maintaining constant intensity to 0.5 and saturation to 0.3 and only varying the hue.

Hue was selected as the parameter because its change integrates the three RGB color channels together, making it more difficult to be processed by extending grayscale techniques to each color channel, thus forcing the segmentation algorithms in evaluation to use the color information holistically.

We varied the characteristics of the basic image (circle), such as size, shape, inclination, contrast, and color. A median filter was applied to the entire image (before the noise addition) to remove sharp edges in order to produce a blurring in some measure. Additive Gaussian noise was added to simulate white noise commonly present in real images [6]. The amount of Gaussian noise applied to images is based on the concept of signal to noise ratio (SNR), a high SNR means a recognizable signal with low noise respect to it; respectively a low SNR means high noise respect to the signal. We use as signal the contrast between figure and background: signal = abs (RGB (Figure)-RGB (Background))/3, extending its use in [6] to RGB images. A sample of the set of images can be seen in Fig. 1 where the vertical axis represent changes in size and horizontal axis represent changes in the amount of noise added for SNR = 32 (far left), 8, 4, 2 and 1 (far right).

About the images varying the color contrast, the common hue difference was decreased from 32 degrees to 2, where even to a human observer is hard to find any color difference (see Fig. 2 far right). Figure 2 shows a set of test images varying the color contrast adding a Gaussian noise in steps of SNR equal to two.

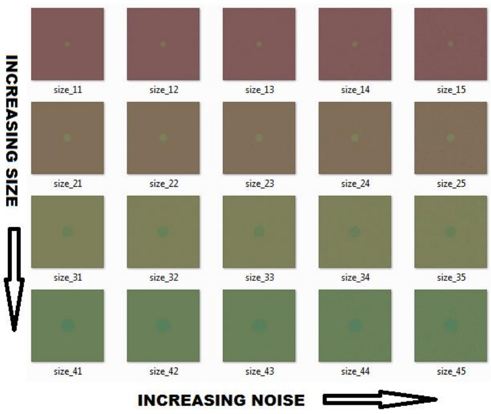


Fig. 1 Set of test images changing in size and signal to noise ratio.

Fig. 2 Set of test images changing the color contrast: 32, 24, 16, 8 and 4 degrees with additive Gaussian noise of SNR = 2

4 Experiments and results

We used the synthetic color images generated with its corresponding GT through the method described above. The color segmentation method requires samples from the operator having, typically from 1 to 9 pixels, which are representative of the desired color target. Samples taken were created trying to select pixels from a zone with a plain color as well as from the edges where the color presents a transition zone (Fig. 3 left).

The evaluation was carried out in the following manner:

1. A database of synthetic color images with its correspondent GT was generated modifying gradually the basic image (in our case a circle) the following parameters: size, shape (changing the eccentricity towards increasingly elongated ellipses) with arbitrary angles, color contrast, varying the color difference between the object and background shrinking almost imperceptible (see Fig. 2 far right), noise varying from SNR of 32 to 1 applied to previous pictures. While generating the synthetic images, the system creates simultaneously its corresponding GT.

2. The semi-automatic color segmentation algorithm [1], [7] was applied to the synthetic images by taking samples from 1 to 9 pixels. Two examples of pixel samples are shown in Fig. 3. In some cases pixels were taken randomly in some other cases in line patterns from the center of the image (as in case of Fig. 3 right) where the operator cannot easily see the color to select because of the applied noise or by the low color contrast (as in Fig. 2 far right).

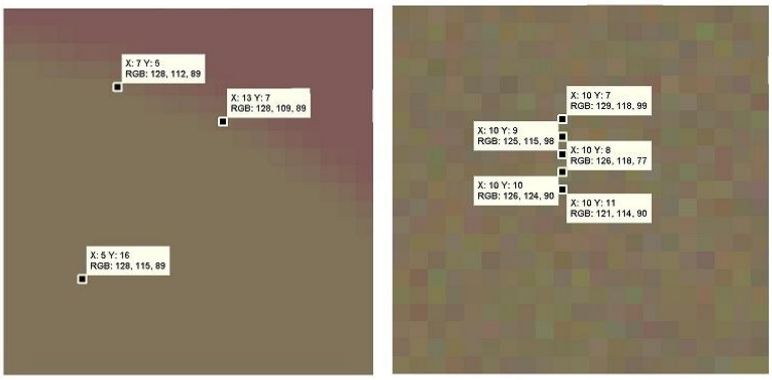


Fig. 3 Examples of pixel samples

3. Any test image was pre-processed; as post-processing a morphological closing filter with a disk of radius of 5 pixels was applied to all images to remove groups of connected pixels in number lower than 30 .

4. We displayed the results of the evaluation using ROC curves due to its ability of visualizing the successes and the failures and their interrelation [10]. They also have been successfully used to assess segmentation algorithms. We used ROC curves for True positives (TP) rate vs. False positives (FP) rate: TP rate = TP/P, FP rate = FP/N

using pixel classification either for object and background [10]. We also use graphs of TP rate, FP rate separated for convenience.

Figure 4 shows the ROC curves for varying size, shape and color contrast. As it can be appreciated, the majority of points (in red) are concentrated in the left corner, showing good performance of the segmentation algorithm in most cases with TP rate around 95% and FP rate lower than 0.1% in most cases. It was expected unvarying results from changes in size and shape because of the algorithm is pixel oriented. The color contrast ROC curve shows good performance of the algorithm even in cases where the color contrast was so low that to an observer is difficult to find any color difference (as in Fig 2 far right).

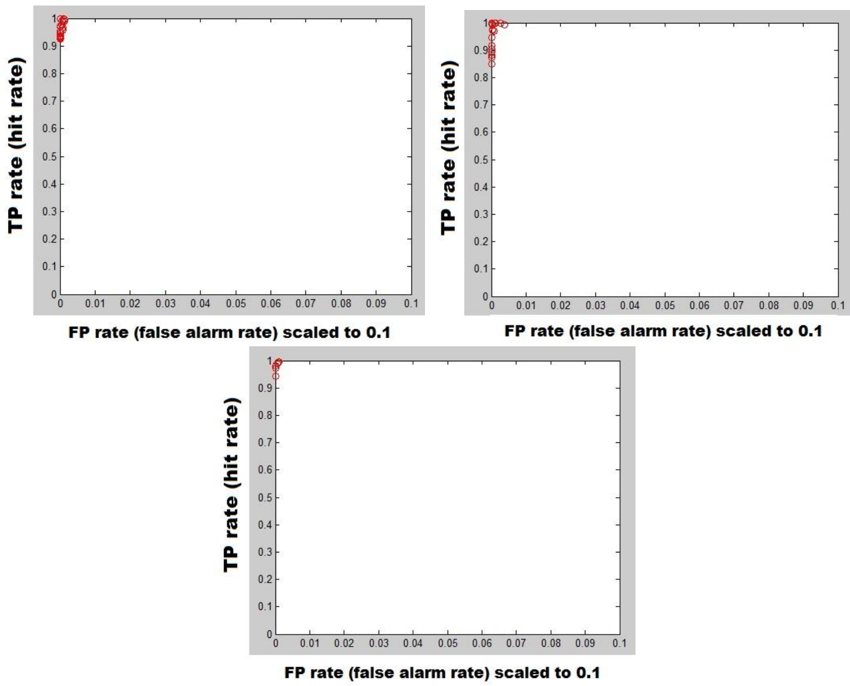


Fig. 4 ROC curves for size (left), shape (center) and color contrast (right)

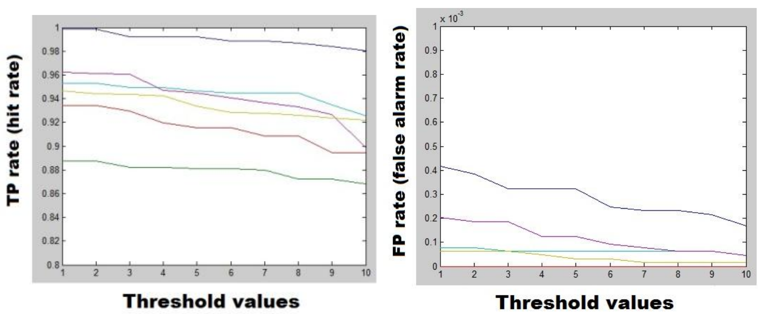


Fig. 5 Graphs for TP rates (left) and FP rates (right) for different thresholds

Figure 5 shows two graphs where it is plotted the behavior of the segmentation of six different images varying in size and shape modifying the threshold applied to the algorithm. The thresholds varied, using Otsu´s threshold [9] at a center value (selected by Otsu), 25 gray levels above and 25 levels below. Abscissa value 1 corresponds to Otsu at –25 gray levels, abscissa value 5 correspond to Otsu´s value, and abscissa value 10 to Otsu´s value at +25 gray levels above. As it can be observed from the graph the TP rate and FP rate values vary around 2% only, showing that the segmentation algorithm is scarcely sensitive to the thresholding value finally chosen.

Figure 6 (left) shows another graph where the TP rates (three graphs without marker) and FP rates (three graphs with marker) are plotted against the number of pixels per sample taken for the initial color sample. The blue line corresponds to the segmentation of different types of objects in images without noise, the red line corresponds to noise with SNR = 2 and the green line to noise with SNR = 1. The magenta line corresponds to FP rate with SNR = 2 that overlaps with FP rate of images with no noise, showing low error rates (near cero) in both cases. The cyan line corresponds to FP rate of noisy images of SNR = 1. It fluctuates showing the lower values near cero for 1, 3, 5 and 10 samples. From analysis of the graph we can point out that 5 pixels per sample is a good choice for most cases, having high TP rate and low PF rate despite noise. We can also observe that in cases without noise, only one pixel per sample is needed having TP rate around 90% and FP rate near cero. Figure 6 (right) shows the ROC curve corresponding to the noisiest cases were the points closer to the top-left corner correspond to the 3, 5 and 10 pixels per sample cases. Then, we can conclude that choosing samples of 5 pixels maximum the algorithm has good performance.

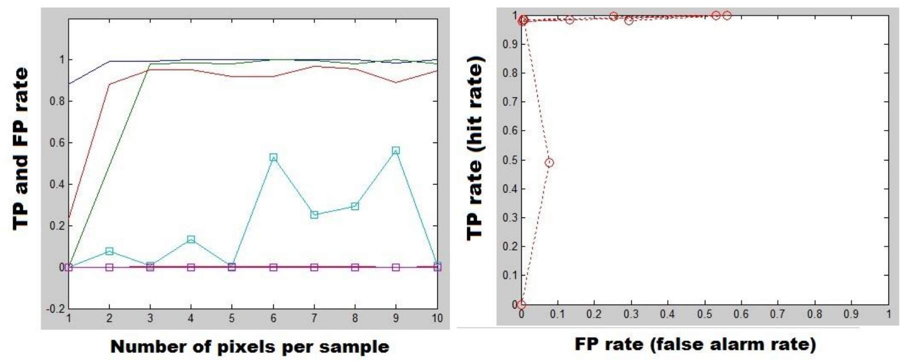


Fig. 6 Graph of TP rates for number of samples (left) and ROC curve for noise level SNR = 1 (right)

5 Conclusion

We presented a quantitative evaluation and characterization of a semi-automatic color image segmentation algorithm by generating synthetic images each with its corresponding ground truth image, with a circular object of interest gradually varying their size, shape, color contrast and amount of additive noise. We calculated the True Positive rate (TP rate) and False Positive rate (FP rate) for every image to obtain ROC curves of the results. This system may be useful for assessing the quality of use of the color information inside the segmentation algorithms in general.

This evaluation system was applied to evaluate the performance of our semi-automatic color segmentation algorithm of color images. In this algorithm color is the only discriminating feature. Behavior was assessed by varying size, shape, and color contrast, amount of additive noise, threshold and number of pixels taken in the sample. After an analysis of its corresponding ROC curves the algorithm shows good performance in most cases with TP rate around 95% and FP rate lower than 0.1%. The ROC curve for color contrast shows good performance of the algorithm even in cases where the color contrast is so low, that for a normal observer is difficult to find any color difference.

We made a characterization study of the algorithm behavior by varying the number of pixels per sample and the threshold value needed for final segmentation. We observed from analysis of the graphs that the segmentation algorithm barely sensitive to the threshold value. Even with changes of +/- 25 gray values above and below from Otsu´s chosen value, the obtained TP rate varied in 2% and FP rate in less than 0.1%. From a similar analysis of the graphs corresponding to the number of pixels samples we can observe that five pixel per sample give fine results of above 90% for the TP rate and less than 0.01% for the FP rate in average in most cases even with a high level of noise of SNR = 1 giving the best results in all cases of study.

Acknowledgements

The authors of this paper wish to thank the Computing Research Center (CIC), Mexico; Research and Postgraduate Secretary (SIP), Mexico; National Polytechnic Institute (IPN), Mexico; and CONACyT, Mexico, for their economic support to this work.

References

1. Alvarado-Cervantes R., Segmentación de patrones lineales topológicamente diferentes, mediante agrupamientos en el espacio de color HSI, M. Sc. Thesis, Center for Computing Research, National Polytechnic Institute, Mexico. (2006).
2. Zhang Y.J., A Survey on Evaluation Methods for Image Segmentation, Pattern Recognition, Vol. 29 No 8 pp. 1335-1346, Elsevier Ltd. (1996).
3. Zhang Y.J., A Summary of Recent Progresses for Segmentation Evaluation. In: Zhang Y.J., Advances in Image and Video Segmentation. USA IRM Press, pp. 423-439 (2007).
4. Zhang Y.J., Gerbrands J.J. On the Design of Test Images for Segmentation Evaluation. In: Signal Processing VI, Theories and Applications pp. 551-554, (1992).
5. Zhang Y.J., A Summary of Recent Progresses for Segmentation Evaluation. In: Zhang Y.J., Advances in Image and Video Segmentation. USA IRM Press, pp. 423-439 (2007).

6. Martin, D., Fowlkes, C., & Tal, D., et al., A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings of the 8th International Conference on Computer Vision (2001), 2, 416-423.
7. Alvarado-Cervantes R., Felipe-Riveron E.M. and Sanchez-Fernandez L.P., Color Image Segmentation by means of a Similarity Function. In: I. Bloch and R.M. Cesar, Jr., Jacobo (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. LNCS Volume 6419 pp. 319-328. Springer, 15th Iberoamerican Congress, CIARP 2010, Sao Paulo, Brazil, 8-11 Nov (2010), ISBN: 978-3-642-16686-0.
8. Martin, D., Fowlkes, C., & Tal, D., et al., A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings of the 8th International Conference on Computer Vision (2001), 2, 416-423.
9. Otsu N., A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man and Cybernetics, 9(1): pp. 62-66. (1979).
10. Fawcett T., An introduction to ROC analysis. Pattern Recognition Letters 27 (2006) pp. 861- 874.
11. Chabrier S., Laurent H., Rosenberg C., Zhang Y.J., Supervised evaluation of Synthetic and Real Contour Segmentation Results, 14th European Signal Processing Conference (EUSIPCO 2006), Fawcett Florence, Italy, September 4-8, (2006).
12. Haindl, M. - Mikes, S., Texture Segmentation Benchmark, 19th ICPR, Tampa 2008, IEEE Press, http://mosaic.utia.cas.cz.

Extracting Structural Features from Manuscript Texts for Forensic Documentoscopy and Graphoscopy Applications

Edgardo M. Felipe-Riverón, Salvador Godoy-Calderón and Edith C. Herrera-Luna

Centro de Investigación en Computación, Instituto Politécnico Nacional, Juan de Dios Bátiz and Miguel Othón de Mendizábal, P. O. 07738, Gustavo A Madero, México

edgardo@cic.ipn.mx, sgodoyc@cic.ipn.mx, edith.hluna@gmail.com

Abstract. This paper presents a new approach for extracting features from a manuscript, as well as a novel approach for modeling the graphoscopic structure of that manuscript at the word and the line levels by using the extracted features. As this new approach is independent of the document's semantics it allows the use of collective decision algorithms for author recognition tasks. Also, this approach represents a hybrid or eclectic paradigm between the texture-related and structure-related modeling approaches. The structural model is explained and the text features to be extracted are analyzed; then a series of test experiments in author identification, using the extracted features, is presented along with a comparison with other similar researches.

Keywords. Feature extraction, structural features, author identification, manuscript text, supervised classification.

1 Introduction

Nowadays, identification and verification are very common tasks that people face every day, not only when extreme security measures are needed, but as part of their daily routines within personal and professional activities. Identification and verification of a person is, nonetheless, a difficult task. That is why both theoretical and applied methods for these tasks are increasingly common, and a wide range of software and hardware tools, commercially available, allow these tasks to be performed with confidence, speed, and with a much smaller margin of error.

Graphoscopic techniques analyze a person's handwriting style and use that information for identification purposes, as well as to determine the specific circumstances under which each document was written. Under such consideration, Graphoscopy is both, a branch of Biometry [1] and a fundamental tool for Documentoscopy [14].

From the view point of Biometry, a person's writing style is a dynamic feature that, in combination with other biometric techniques, can be successfully used for identification purposes [2]. Criminalistics and Criminology research groups have developed a special interest on Documentoscopy techniques, not only for their potential capability to reveal the identity of a manuscript's author, but also for some other psychological traits about the author's health and frame of mind, it may reveal, by analyzing drawing samples, as well as handwriting and signature samples [3]. Other potentially interesting traits include, forfeit detection, multiple authorship identification, and historical edition profiling.

In order to analyze handwriting, two kinds of data acquisition techniques are traditionally used: online and offline. Online techniques acquire data directly from the author using some capturing device that delivers the data in real time and is capable of automatically identifying dynamic changes in the writing style, such as changes in pressure, speed, and inclination. On the other hand, offline techniques acquire data by analyzing texts previously written on some support surface (paper, parchment, leather, etc.). This type of analysis is usually performed by extracting features from digital images of the text, and thus, it is more suited for forensic applications.

The main goal of this work is to propose a particular approach for modeling a manuscript, as well as the particular set of features that need to be extracted from a digital image of the manuscript. Also, we would like to show how the proposed set can be effectively used for forensic Documentoscopy and Graphoscopy applications, with sample results from a previous author identification research. The proposed model effectively constitutes a different and hybrid approach between the classic texture-related and the structure-related approaches to handwritten text's feature extraction.

2 Background

One of the most relevant researches concerning the writing of an individual as a recognizable unique mark of each person was performed by Sargur N. Srihari *et al.* [5, 6]. Srihari's work group created a database with writing samples of approximately 1,500 participants, who were asked to write a letter in English, over a blank sheet of paper [4, 6]. This research identified and analyzed text features from two groups: conventional and computational features. Conventional features are found by the manual analysis of graphoscopic specialists, while computational features come from the analysis of some intermediate media (typically a digital image), and extracted with morphologic operators and/or data transformations like the Fast Fourier Transformation or Wavelets [14]. The set of computational features extracted, was divided in two groups, called macro-features and micro-features. Macro-features are extracted at the whole document level, as well as at the paragraph, row and word levels of analysis. Micro-features, on the other hand, are extracted at the character and stroke levels. For testing purposes, they selected a subset of 11 macro-features, and 3 micro-features, and reported an effectiveness range from 78% to 96% when using those features for author identification. In [7], H.E. Said *et al,* proposed an algorithm for offline data analysis, based on texture features. By using multi-channel filters, Grey Scale Co-occurrence Matrices (GSCMs), and a K-Nearest Neighbor with weighted Euclidian distance criteria as a classifier, they achieved a 96% rate of correct author identification, over a database of 150 writing samples. The main relevance of this research lies on the fact that all the extracted features can be used in independence from the document's content, and its text distribution. E.Zois and V.

Anastassopoulos, proposed in [9], a distinct algorithm that uses horizontal projections and morphologic operators, as well as a texture-related analysis over English and Greek handwritten words. To prove their method to be language independent, they constructed a sample database from 50 authors, each writing 45 different words in English and Greek. In general terms, their method applies a threshold over each word's image, then, a series of morphological thinning and opening operations. A combination of multi-layer perceptron and Bayesian analysis is used to classify samples, and they report a 90% precision in identifying authors. In [1], a recognition method, using some of the features proposed by Srihari *et al.*, was shown, although they were extracted at the character level exclusively. The proposed method starts form digital images, scanned at 300 dpi, where all the characters are manually segmented and saved into a database. After some preprocessing to get each character in a binary form, all GSC [8, 9], and geometric features [10] are automatically extracted, as well as some novel gradient direction related features. At the end of this research, authors report that, using a K-Nearest Neighbor classifier, gradient features better allow them to reach a 100% precision in identification tasks, while the whole set of the same features, along all GSC, and geometric features, only allows them a 90% precision. All their reported experiments were performed over a 1,400 characters database, with 40 different authors. Along the same line of thought, on processing individual characters, is the research performed by V. Pervouchine in [11]. His work focus on the specific analysis of letters 'd', 'y', and 'f', as well as the 'th' combination in English texts. After constructing a custom database, with 15 to 30 different writing samples, from 150 authors, Pervouchine managed to extract a total of 31 features, and selected 13 of them to be essential, 14 to have only partial relevance, and the last 4 to be irrelevant. However, his reported results show an identification performance of only 58%. Finally, a generally accepted top reference in the field is the work of Bansefia *et al.* in [12]. This workgroup used two different databases; the first one, custom-made by them, was called *PSI-Database*, and contained a sample handwritten letter, with 107 words, from 88 French authors. The second database was the relatively famous *IAM* database, which includes samples from approximately 150 authors, and is freely available on [13]. Their proposed processing technique, separates the text in graphemes, based on the analysis of the upper outline of the letters, and they report a 95% precision identifying authors from the *PSI-DataBase*, while only 86% with authors from the *IAM* database.

3 Handwriting characteristics

In order to analyze a handwriting sample, several distinct measurements are required. The measured elements of a person's writing form two disjoint groups, generally called formal elements and structural elements. See Figures 1 and 2.

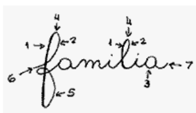


Fig. 1. Formal elements of handwritten text: (1, 2, 4) Crest, (3) Oval, (5) Axis, (6) Initial & (7) Final point.

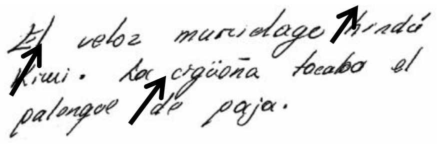


Fig. 2. Example of the structural element called **Slant**. The text in this figure has a right slant.

Formal elements are those characteristics that form letters or words, like strokes, lines, initial and end points, etc. Also, as described in [14], the closed central portion of letters, called ovals, can also be considered as formal elements. The superior portions of letters, called crests, and the axis that guide the lower portion of letters, as well as the punctuation signs like colons, semicolons, etc., are considered formal elements as well. In contrast, structure elements are those characterizing the authors writing style, such as its size, form, direction, and organization. Although both groups of elements have been studied for the last 100 years, there is still no agreement on a unified set of features that should be considered.

The measurement of formal and structural features requires the definition of writing zones which are defined with the aid of two imaginary horizontal lines, tangent to the upper and lower parts of the text, and which delimit three writing zones. The middle zone includes most of the writing strokes, the upper zone includes the letters crests, and finally, the lower zone includes the axis of letters (Figure 3).

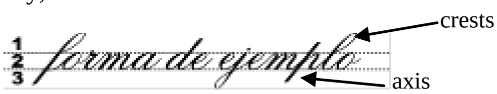


Fig. 3. Writing zones (1) upper zone, (2) middle zone, and (3) lower zone.

The most relevant handwritten text features for this research follows:

Order (Regularity). It refers to the distribution and organization among letters, words, text lines, paragraphs and text margins. **Size (Dimension).** Refers to the breadth of writing, and the amount of space each letter occupies within a word, or a word within a line. The proportion among writing zones, as well as the change in size from lower letters to capital letters is also considered (See [3, 15]). **Proportion.** This can be modeled as a subset of the size measurements, or it can be based on the size ratio of the writing zones, the ratio of the document's margins, or the ratio between the area occupied by the text and the whole area available on the writing media. **Shape.** Measures some traits of the writing, such as if letters are angled, curved, typographical or decorated (Widely in [3, 15, 18]). **Angularity.** This trait can well be a subset of the shape measurements. The text angularity refers to the presence of angles and curves in the writing, the angled end of each character or word, as well as the breadth of the curved strokes. **Direction.** The trajectory followed by the lines or strokes that comprise characters. Some researchers consider the slope of the writing zone, while others prefer the angle of an imaginary line below the text (Refer also to [2, 3]). **Slant. It r**efers to the angle of single words, lines, or paragraphs. The standard is to look for right-slant or left-slant. **Continuity and Linkage.** It is the measure of the strokes that bind two consecutive words in a line of text. **Separation and Cuts**. It measures the blank space between letters, words or text lines. These gaps are also considered as writing strokes.

In order to measure all the above traits, and construct an adequate database for offline analysis, test subjects were asked to write over white Bond paper and with the same writing tool. The complete set of details about the database and the form filled by test subjects can be found in [15].

The digital image of the manuscript is thresholded to get a binary image, and then the horizontal and vertical projections of the binary image are used. For the horizontal projection, the profile of each row of pixels is considered, and the amount of those pixels with a zero value (pixels in black) is used to construct a histogram. The vertical projection follows the exact same procedure, but using the profile of each column of pixels in the image. A detailed description about this can be found in [15, 18].

The rest of the macro-features were extracted by using morphological dilations and erosions, with squared and straight line structure elements previously rotated. Other morphological operators, such as openings, closures and geodetic reconstructions were also used. To review these procedures refer to [15, 19].

4 Methodology

Our general methodology starts with a data acquisition phase, scanning the test forms with a standard color scanner at 300 dpi, and setting up a database with them. Before the preprocessing phase takes place, the characteristic writing zones are calculated for each scanned form. Then, after enhancing the sample images, the handwritten text is segmented and features at the word, line and paragraph are extracted. Once a full supervision sample has been constructed, the specific comparison criteria are selected and each extracted feature is weighted. Some important details about the procedure follow.

During the preprocessing phase, two different methods are used to enhance the digital image of each text. First, a threshold is applied over the green plane, with the Otsu method [16]. Also, the same original green plane is processed using the Khashman and Sekeroglu algorithm [17]. The result from the Otsu method turns out to be very clear; however it removes some important regions from the manuscript. On the other hand, the Khashman and Sederoglu algorithm yields a very noisy image, but the text is much better shown in it. Therefore, a geodetic reconstruction, following the morphological operation of erosion, is used. The image resulting from the Khashman algorithm is used as a mask, and the Otsu image is used as the mark for this reconstruction procedure. As a result, we get a very clean image, without noise at all, but with the handwritten text shown as best as possible (Fig. 4)

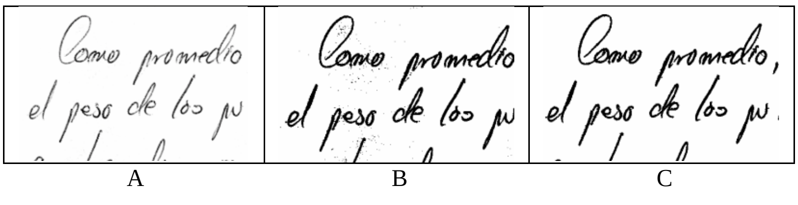


Fig. 4. (A) is the original image, (B) is the result from the Khashman and Sekeroglu algorithm, and (C) is the geodetically reconstructed image.

The exact set of extracted features is the following:

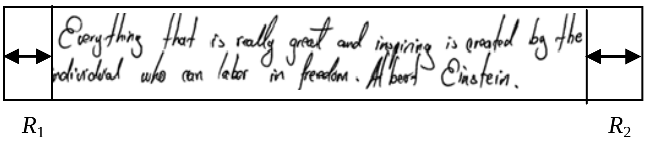
- Margin proportionality features
  1) Average paragraph's distance to the left and right physical margins (R_1, R_2).
  2) Average text line's distance to the left and right physical margins (R_3, R_4).
- Direction and size features
  3) Average distance between successive rows (R_5).
  4) Ratio of the space occupied by a row and its distance from the previous and the following rows (R_6, R_7).
  5) Average space among words from a same row (R_8).
  6) Average row direction (R_9).
- Writing zone's features
  7) Upper zone – Middle zone ratio (R_{10}).
  8) Lower zone – Middle zone ratio (R_{11}).
- Slant features
  9) Average word slant (R_{12}).

5 Classification model

The margin proportionality feature is extracted at the paragraph and line levels, since different indentation formats can be notoriously discriminant under some circumstances. In both cases, the minimum size rectangle, containing the paragraph or row, is cut from the binary image, then, we simply count the number of pixels, form the original image, that were cut, so extracting the first two features R_1 and R_2 (Fig. 5).The same minimum size containing rectangle procedure is applied at the text line level, and the comparison between successive rectangles yields features R_3 and R_4 (Fig. 6).

The first six features are measured directly from the text lines on the manuscript. The next features are calculated as an average of some other features at the word level. Words with crest or axis are clustered into four groups: words with crest, without crest, word with axis, and without axis. For each group features R_{10}, R_{11} and R_{12} are calculated. These features are extracted at the paragraph, line, and word levels. However, since a feature vector represents a text line, features R_1, R_2, and R_5, which are paragraph features, are not included. Also, these three features are equivalent, at the line level, to features R_3, R_4, and R_6.

Then, each pattern represents a line of text in the document, and is composed by 22 features, calculated with the text traits extracted previously. From the form filled by test subjects, 605 text lines, from 30 different authors, were obtained.



R_1 R_2

Fig. 5. Features R_1 and R_2.

R_3 R_4

Fig. 6. Features R_3 and R_4.

5.1 Feature selection

In order to compare patterns, a similarity or dissimilarity function must be used. A weighted syntactic distance is selected for that task. The idea is to allow automatic systems the opportunity to assign a greater weight or relevance to the most typical traits in the class that describes the writing of a person. A complete description of the weighting scheme can be found in [29, 30 and 31].

A collective decision algorithm is used to identify the author of a manuscript by means of the individual classification of the text lines that comprise the manuscript. The descriptive features used to make up the patterns representing such lines include features extracted at the word level, as well as at the line level. When all the text lines have been classified, a final collective decision regarding the author of such text is applied.

6 Experimental results and precision assessment

An *ad hoc* database was created with manuscripts written by 50 test subjects, who wrote three handwritten texts. Each manuscript contains from 5 to 9 text lines, giving closely a total of 600 lines. Text contents were selected arbitrarily from non-technical books with no restrictions on the logic or semantics. Images of those manuscripts were digitally scanned at 300dpi with a conventional scanner and all manuscripts were written with the same black ink pen over white paper.

Three different types of experiments were performed: in experiments of type1 the supervision sample contains the most representative patterns in each class (e.g. those patterns with the maximum average similarity with all others in the same class); in experiments of type2 it contains a random selection from all the patterns in all classes; and finally, in experiments of type3 the supervision sample contains only the least representative patterns from each class (e.g. those with the minimum average similarity with all others in the same class).

Three class representative patterns were selected from each class within each supervision sample, according to the previously described procedure. Experiment results are shown in Table 1.

In the table, the use of a differentiated weighting scheme significantly increases the classification rate, both for text lines and for the whole manuscript. Turns out that this type of weighting allows a more precise characterization of an author's writing style. For experiments of type 2, the higher classification rate is 88.89%.

Table 1. Manuscript line and whole text classification rates for experiments of type 1, 2 and 3.

No.	Text line classification rate (%)	Manuscript classification rate (%)	No.	Text line classification rate (%)	Manuscript classification rate (%)
Experiment type 1			**Experiment type 2**		
1.1	42.28	61.11	1.1	52.76	75.00
1.2	56.10	66.67	1.2	67.72	88.89
1.3	60.16	72.22	1.3	64.57	94.44
1.4	56.10	77.78	1.4	58.26	83.33
1.5	73.17	94.44	1.5	70.08	88.89
1.6	47.97	66.67	**Experiment type 3**		
1.7	61.79	72.22	1.1	46.46	66.67
1.8	60.16	63.89	1.2	65.35	88.89
1.9	51.22	72.22	1.3	62.99	83.33
1.10	63.42	83.33	1.4	58.26	83.33
			1.5	72.44	94.44

A comparative analysis with several state-of-the-art works in manuscripts author identification shows some advantages of the method herein proposed. First, there is the opportunity to increase the classification rate, just by changing the weighting scheme. Second, there is also a chance to change the function for comparing text patterns. Third, although a big number of text line patterns can eventually lead to a deeper analysis of a manuscript, usually a small number of text lines can be enough for achieving high classification rates.

Text independence is generally a difficult problem when automatically analyzing manuscripts. However, results obtained by this research are not significantly different from those obtained by other research groups. Table 2 compares the classification rate of this work with that obtained by several other relevant studies in the field, as presented by [15].

Although each research uses a different sample database, the relative complexity for feature extraction of the database used for this research can be regarded as slightly higher as that from all other works, since those other works usually show a bigger inter-line spacing which simplifies the text segmentation procedure.

Table 2. Comparison with classification rate results in other researches.

	# of Authors	Samples	Text dependency	Classification rate (%)
Said et al. [8]	2x20	25 blocks of text	NO	95
Zois y Anastassopoulos [9]	50	45 samples of the same word	YES	92.48
Marti and Bunke [13]	20	5 samples of the same text	YES	90
A. Bensefia et al. [12]	88	Paragraphs of 3-4 words	YES	93 / 90
This work	**30**	**3 samples of the same text**	**NO**	**94.44**

7 Conclusions

Two clearly distinct methodological phases can be identified in this proposal: the first one deal exclusively with feature extraction, the second one, does characterization and classification. These two phases are not strictly sequential, however. As the first text features are extracted, an adequate similarity/dissimilarity measure[1] is selected, as well as all related comparison criteria. Semantics of each text trait must be carefully analyzed, in order to set an appropriate mixing of features. The weighting scheme turns out to be extremely useful when characterizing an author's handwriting style.

A comparison with previously published works shows that the modeling approach herein proposed yields better results, with the added advantage that the recognition process needs not to be dependent on the semantic contents of the text. The implementation of these improvements may be extremely useful for forensic Documentoscopy and Graphoscopy applications, as well as for more traditional authentication and security systems.

Acknowledgements

The authors of this paper wish to thank the Centro de Investigación en Computación (CIC), Mexico; Research and Postgraduate Secretary (SIP), Mexico, and Instituto Politécnico Nacional (IPN), Mexico, and CONACyT, Mexico, for their economic support to this research.

References

1. Tapiados Mateos, Marino, Sigüenza Pizarro, Juan A. Tecnologías biométricas aplicadas a la seguridad. 1a. México, D.F.: Alfaomega Grupo Editor, 2005. ISBN: 970-15-1128-X.
2. Del Val Latierro, Félix. Grafocrítica. El documento, la escritura y su proyección forense. Madrid, España: Tecnos S.A., 1963.
3. Tesouro De Grosso, Susana. Grafología. Análisis e interpretación científica de la escritura. Buenos Aires, Argentina: Editorial Kier, 2006. ISBN: 950- 17-7011-7.
4. Moreno, G. R. y García, R. Temas de Criminalística. Jornadas sobre Justicia Penal sobre Temas de Derecho Penal, Seguridad Pública y Criminalística. s.l.: Universidad Nacional Autónoma de México, 2005.
5. Srihari Sargur, N. Handwriting identification: research to study validity of individuality of handwriting and develop computer-assisted procedures for comparing handwriting. University of Buffalo, U.S.A.: Center of Excellence for Document Analysis and Recognition, 2001. Tech. Rep. CEDAR-TR-01-1.
6. Srihari, Sargur, N. et al. Individuality of Handwriting. E.U.: Journal of Forensic Sciences, 2002. Vols. 47, No. 4. Paper ID JFS2001227-474.
7. Said, H., Tan, T. y Baker, K. Personal Identification Based on Handwriting. s.l.: Pattern Recognition, 2000. págs. 149 -160. Vol. 33 No. 1.

[1] Although in theory there are many similarity/dissimilarity functions that can be selected, the authors have found that a term-by-term comparison on the patterns (a syntactic distance) usually promotes faster classification processes, as well as clearer interpretations of the obtained results.

8. Said H. E., et al. Writer identification from non-uniformly skewed handwriting images. s.l.: in Proc. 9th British Machine Vision Conference, 1998. págs. 478-487.
9. Zois, E y Anastassopoulos, V. Morphological waveform coding for writer identification. s.l.: Pattern Recognition, 2000. págs. 385 - 398. Vol. 33. No. 3.
10. Srihari, Sargur, N. Recognition of handwritten and machine-printed text for postal address interpretation. s.l.: Pattern Recognition Letters, 1993. págs. 291-302. Vols. 14, No. 4. ISSN: 0167-8655.
11. Pervouchine, Vladimir y Leedham, Graham. Extraction and analysis of forensic document examiner features used for writer identification. s.l.: Pattern Recognition, 2007. págs. 1004-1013. Vol. 40. ISSN: 0031-3203.
12. Bensefia A., Paquet, T. y Heutte, L. A writer identification and verification system. s.l.: Pattern Recognition Letters, 2004. Vol. 26. 2080–2092.
13. Bunke, H. y Marti, U. A full English sentence database for off-line handwriting recognition. [ed.] University of Bern Institute of Informatics and Applied Mathematics. s.l.: Proc. of the 5th Int. Conf. on Document Analysis and Recognition, ICDAR´99, 1999, pp. 705-708. http://www.iam.unibe.ch/fki/databases/iam-handwriting-database/iamhandwriting- database#icdar99.
14. Del Picchia, José y Celso. Tratado de documentoscopía: La falsedad documental. Argentina: Ediciones La Rocca, 1993. ISBN 959-97-1450-4.
15. Guzmán, Carlos Alberto. El peritaje caligráfico. 1a. y 2a. reimp. Buenos Aires: Ediciones La Rocca, 2005. pág. 336. ISBN: 950-9714-59-3.
16. Otsu N., A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man and Cybernetics, 9(1): pp. 62-66. (1979).
17. Khashman, Adnan y Sekeroglu, Boran. A Novel Thresholding Method for Text Separation and Document Enhancement. Near East Univ., Mersin: ICIT 2006. IEEE International Conference on Industrial Technology, 2006. ISBN: 1-4244-0726-5.
18. Herrera-Luna, Edith C., Identificación del autor de un texto manuscrito., Tesis de Maestría en Ciencias de la Computación, Centro de Investigación en Computación, Instituto Politécnico Nacional. México D. F., México.
19. Soille, Pierre. Morphological Image Analysis. Segunda Edición. s.l.: Springer. pág. 391. ISBN: 3-540-42988-3.

VPSNR: Visual Perceptual Full-Reference Image Quality Assessment for JPEG2000

Jaime Moreno, Carlos Correa, Nallely Hernández, Diana Pérez

Superior School of Mechanical and Electrical Engineering
National Polytechnic Institute of Mexico,
IPN Avenue, Lindavista, Mexico City, 07738, Mexico.
XLIM Laboratory, Signal, Image and Communications Department,
University of Poitiers, 86962 Futuroscope, France.
jmorenoe@ipn.mx

Abstract. Estimation of image quality is decisive in the image compression field. This is important in order to minimize the induced error via rate allocation. Traditional full-reference algorithms of image quality try to model how Human Visual System detects visual differences and extracts both information and structure of the image. In this work we propose a quality assessment, which weights the mainstream PSNR (Peak Signal-to-Noise Ratio) by means of a perceptual model (VPSNR). Perceptual image quality is obtained by estimating the rate of energy loss when an image is observed at monotonically increasing distances. Experimental results show that VPSNR is the best-performing algorithm, compared with another eight metrics such as MSSIM, SSIM or VIF, among others, when an image is distorted by a wavelet compression. It has been tested across TID2008 image database.

1 Introduction

Mean Squared Error (MSE) is still the most used quantitative performance metrics and several image quality measures are based on it, being Peak Signal-toNoise Ratio (PSNR) the best example. Wang and Bovik in [1,2] consider that MSE is a poor device to be used in quality assessment systems. Therefore it is important to know what is the MSE and what is wrong with it, in order to propose new metrics that fulfills the properties of human visual system and keeps the favorable features that the MSE has.

In this way, let $f(i,j)$ and $\hat{f}(i,j)$ represent two images being compared and the size of them is the number of intensity samples or pixels. Being $f(i,j)$ the original reference image, which has to be considered with perfect quality, and $\hat{f}(i,j)$ a distorted version of $f(i,j)$, whose quality is being evaluated. Then, the MSE and the PSNR are, respectively, defined as:

© E. A. Santos, J. C. Chimal, L. Cabrera, E. Castillo.
Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 65–79
Paper Recived 21-09-2013 and Acepted 21-10-2013

$$MSE = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[f(i,j) - \hat{f}(i,j)\right]^2 \quad (1)$$

and

$$PSNR = 10\log_{10}\left(\frac{\mathcal{G}_{max}{}^2}{MSE}\right) \quad (2)$$

where G_{max} is the maximum possible intensity value in $f(i,j)$ ($M \times N$ size). Thus for gray-scale images that allocate 8 bits per pixel (bpp) $G_{max} = 2^8 - 1 = 255$. For color images the PSNR is defined as in the Equation 2, whereas the color MSE is the mean among the individual MSE of each component. MSE does not need any positional information in the image, thus pixel arrangement is ordered as a one-dimensional vector.

Both MSE and PSNR are extensively employed in the image processing field, since these metrics have favorable properties, such as:

1. A convenient metrics for the purpose of algorithm optimization. For example in JPEG2000, MSE is used both in Optimal Rate Allocation [3,4] and Region of interest [5,4]. Therefore MSE can find solutions for these kind of problems, when is combined with the instruments of linear algebra, since it is differentiable.
2. By definition MSE is the difference signal between the two images being compared, giving a clear meaning of the overall error signal energy.

However, the MSE has a poor correlation with perceived image quality. An example is shown in Figure 1, where both *Baboon*(a) and *Splash*(b) Images are distorted by means of a JPEG2000 compression with 30dB of PSNR. These noisy images present dramatically different visual qualities. Thereby either MSE or PSNR do not reflect the way that human visual system (HVS) perceives the images, since these measures represent an input image in a pixel domain.

(a) Image *Baboon*

(b) Image *Splash*

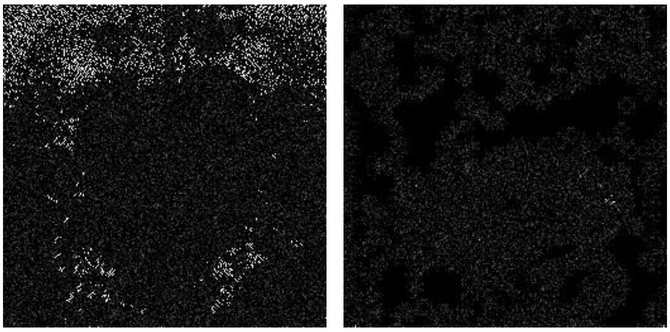
Fig.1. 256×256 patches of Images *Baboon* and *Splash* distorted by means of JPEG2000 both with PSNR=30dB, cropped for visibility.

2 Perceptual peak signal-to-noise ratio

2.1 Introduction

In the referenced image quality issue, there is an original image $f(i,j)$ and a distorted version $\hat{f}(i,j) = \Lambda[f(i,j)]$ that is compared with $f(i,j)$, being Λ a distortion model. The difference between these two images depends on the characteristics of the distortion model Λ. For example, blurring, contrast change, noise, JPEG blocking or wavelet ringing.

In Figure 1, the images *Babbon* and *Splash* are compressed by means of JPEG2000. These two images have the same PSNR=30 dB when compared to their corresponding original image, that is, they have the same numerical degree of distortion (i.e. the same objective image quality PSNR). But, their subjective quality is clearly different, showing the image *Baboon* a better visual quality. Thus, for this example, PSNR and perceptual image quality has a small correlation. On the image *Baboon*, high spatial frequencies are dominant. A modification of these high spatial frequencies by Λ induces a high distortion, resulting a lower PSNR, even if the modification of these high frequencies are not perceived by the HVS. In contrast, on image *Splash*, mid and low frequencies are dominant. Modification of mid and low spatial frequencies also introduces a high distortion, but they are less perceived by the HVS. Therefore, correlation of PSNR against the opinion of an observer is small. Figure 2 shows the diagonal high spatial frequencies of these two images, where they are more high frequencies in image *Baboon*.



(a) (b)

Fig.2. Diagonal spatial orientation of the first wavelet plane, Images *Baboon*(a) and *Splash*(b) distorted by JPEG2000, PSNR=30dB.

If a set of distortions $\hat{f}_k(i,j) = \Lambda_k[f(i,j)]$ are generated and indexed by k (for example, let Λ be a blurring operator), the image quality of $\hat{f}_k(i,j)$ evolves while varying k, being k, for example, the degree of blurring. Hence, the evolution of $\hat{f}_k(i,j)$ depends on the characteristics of the original $f(i,j)$. Thus, when increasing k, if $f(i,j)$ contains many high spatial frequencies the PSNR rapidly decreases, but when low and mid frequencies predominated PSNR decreases slowly.

Similarly, the HVS is a system that induces a distortion on the observed image $f(i,j)$, whose model is predicted by The *C*hromatic *I*nduction *Wa*velet *M*odel (CIWaM) [6].

CIWaM takes an input image I and decomposes it into a set of wavelet planes $\omega_{s,o}$ of different spatial scales s (i.e., spatial frequency ν) and spatial orientations o. It is described as

$$\mathcal{I} = \sum_{s=1}^{n} \sum_{o=v,h,dgl} \omega_{s,o} + c_n \;, \tag{3}$$

where n is the number of wavelet planes, c_n is the residual plane and o is the spatial orientation either *v*ertical, *h*orizontal or *d*ia*g*ona*l*.

The perceptual image I_ρ is recovered by weighting these $\omega_{s,o}$ wavelet coefficients using the *extended Contrast Sensitivity Function* (e-CSF).

Perceptual image I_ρ can be obtained by

$$\mathcal{I}_\rho = \sum_{s=1}^{n} \sum_{o=v,h,dgl} \alpha(\nu, r)\, \omega_{s,o} + c_n \tag{4}$$

where $\alpha(\nu,r)$ is the e-CSF weighting function that tries to reproduce some perceptual properties of the HVS. The term $\alpha(\nu,r)\ \omega_{s,o} \equiv \omega_{s,o;\rho,d}$ can be considered the *perceptual wavelet coefficients* of image I when observed at distance d. For details on the CIWaM and the $\alpha(\nu,r)$ function, see [6].

Hence, CIWaM is considered a HVS particular distortion model $\Lambda \equiv$ CIWaM that generates a perceptual image $\hat{f}_\rho(i,j) \equiv I_\rho$ from an observed image $f(i,j) \equiv$ I, i.e $I_\rho = CIWaM[I]$. Therefore, a set of distortions is defined as $\Lambda_k \equiv CIWaM_d$, being d the observation distance. That is, a set of perceptual images are defined $I_{\rho,d} = CIWaM_d[I]$ which are considered a set of perceptual distortions of image I.

When images $f(i,j)$ and $\hat{f}(i,j)$ are simultaneously observed at distance $\bar{d}$ and this distance is reduced, the differences between them are better perceived. In contrast, if $f(i,j)$ and $\hat{f}(i,j)$ are observed from a far distance human eyes cannot perceive their differences, in consequence, the perceptual image quality of the distorted image is always high. The distance where the observer cannot distinguish any difference between these two images is $\bar{d} = \infty$. In practice, $\bar{d} = D$ where differences are not perceived and range some centimeters from the position of the observer. Consequently, the less distorted $\hat{f}(i,j)$, that is, the highest the image quality of $\hat{f}(i,j)$), the shorter the distance D.

2.2 Methodology

Let $f(i,j)$ and $\hat{f}(i,j) = \Lambda[f(i,j)]$ be an original image and a distortion version of $f(i,j)$, respectively. VPSNR methodology is based on finding a distance D, where there is no perpetual difference between the wavelet energies of the images $f(i,j)$ and $\hat{f}(i,j)$, when an observer watch them at d centimeters of observation distance. So measuring the PSNR of $\hat{f}(i,j)$ at D will yield a fairer perceptual evaluation of its image quality.

VPSNR algorithm is divided in five steps, which is described as follows:

Step 1: Wavelet transformation

Wavelet transform of images $f(i,j)$ and $\hat{f}(i,j)$ is performed using Eq. 3, obtaining the sets $\{\omega_{s,o}\}$ and $\{\hat{\omega}_{s,o}\}$, respectively. The employed analysis filter is the Daubechies 9-tap/7-tap filter (Table 1).

Table 1. 9/7 Analysis Filter.

Analysis Filter		
i	Low-Pass Filter $h_L(i)$	High-Pass Filter $h_H(i)$
0	0.6029490182363579	1.115087052456994
±1	0.2668641184428723	-0.5912717631142470
±2	-0.07822326652898785	-0.05754352622849957
±3	-0.01686411844287495	0.09127176311424948
±4	0.02674875741080976	

Step 2: Distance D

The total energy measure or the *deviation signature*[7,8] $\bar{\varepsilon}$ is the absolute sum of the wavelet coefficient magnitudes, defined as

$$\bar{\varepsilon} = \sum_{n=1}^{N} \sum_{m=1}^{M} |x(m,n)| \tag{5}$$

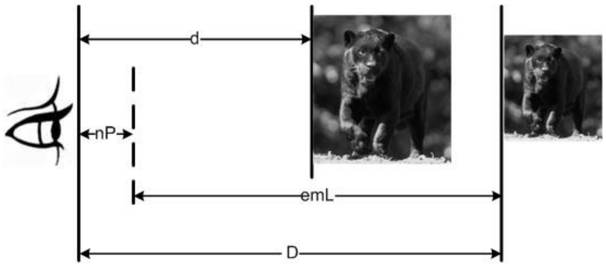
where $x(m,n)$ is the set of wavelet coefficients, whose energy is being calculated, being m and n the indexes of the coefficients. Basing on the traditional definition of a calorie, the units of $\bar{\varepsilon}$ are wavelet calories (wCal) and can also be defined by Eq. 5, since a wCal is the energy needed to increase the absolute magnitude of a wavelet coefficient by one scale.

From wavelet coefficients $\{\omega_{s,o}\}$ and $\{\hat{\omega}_{s,o}\}$ the corresponding perceptual wavelet coefficients $\{\omega_{s,o;\rho,\tilde{d}}\} = \alpha(\nu,r) \cdot \omega_{s,o}$ and $\{\hat{\omega}_{s,o;\rho,\tilde{d}}\} = \alpha(\nu,r) \cdot \hat{\omega}_{s,o}$ are obtained by applying CIWaM with an observation distance $\tilde{d}$. Therefore Equation 6 expresses the relative wavelet energy ratio $\varepsilon R(\tilde{d})$, which compares how different are the energies of the reference and distorted CIWaM perceptual images, namely ε_ρ and ε_ρ respectively, when these images are watched from a given distance $\tilde{d}$.

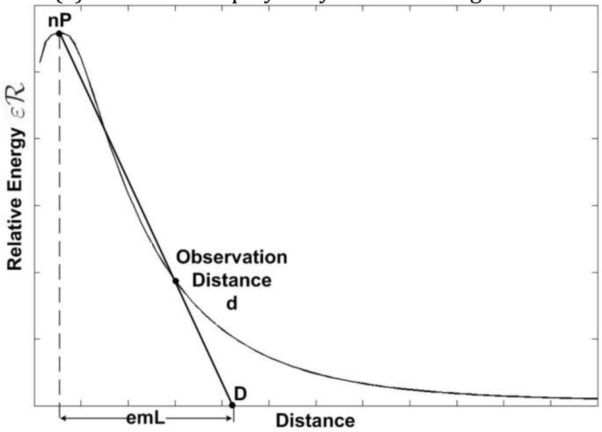
$$\varepsilon\mathcal{R}\left(\tilde{d}\right) = 10 \cdot \left|\log_{10} \frac{\varepsilon_\rho\left(\tilde{d}\right)}{\widehat{\varepsilon}_\rho\left(\tilde{d}\right)}\right| \tag{6}$$

Figure 3(a) shows that distance D is composed by the sum of two distances, nP and εmL. Thereby for the estimation of D, Eq. 7, it is necessary to know the observation distance d besides to figure out the nP and εmL distances. Furthermore Figure 3(b) depicts a chart of εR, which sketches both the behavior of the relative energy when $\tilde{d}$ is varied from 0 to ∞ centimeters and the meaning of the distances D, nP and εmL inside an εR chart.

$$D = nP + \varepsilon mL \tag{7}$$



(a) Distances employed by the VPSNR algorithm.



(b) εR Chart.

Fig.3. Definition of distances D, nP and εmL both graphically (a) and inside an εR Chart (b).

The peak inside an εR chart is nP, which is the distance where the observer is able to better assess the difference between the images $f(i,j)$ and $\hat{f}(i,j)$. From this point nP the observer starts to perceive fewer the differences, until in ∞ these differences disappear, in practice, this point varies from 15 to 25 centimeters. Our metrics is based on finding an approximation of the distance D where the wavelet energies are linearly the same, that is, $\varepsilon R(D) \approx 0$. This is achieved by projecting the points $(nP, \varepsilon R(nP))$ and $(d, \varepsilon R(d))$ to $(D,0)$.

Therefore εmL is the needed length to match the energies from the point where the observer has the best evaluation of the assessed images to D and is described as follows:

$$\varepsilon m\mathcal{L} = \frac{\varepsilon\mathcal{R}\,(n\mathcal{P})}{d\varepsilon\mathcal{R} + \varsigma} \tag{8}$$

where εR(nP) is the relative energy at nP and $d\varepsilon$R is the energy loss rate (wCal/cm or wCal/visual degrees) between (nP,εR(nP)) and (d,εR(d)), namely, the negative slope of the line joining these points, expressed as:

$$d\varepsilon\mathcal{R} = \frac{\varepsilon\mathcal{R}\,(n\mathcal{P}) - \varepsilon\mathcal{R}\,(d)}{d - n\mathcal{P}} \tag{9}$$

When a lossless compression is performed, consequently $f(i,j) = \hat{f}(i,j)$, hence $d\varepsilon$R = 0 and εmL → ∞. In order to numerically avoid it, parameter ς is introduced, which is small enough to not affect the estimation of εmL when $d\varepsilon$R ≠ 0, in our MatLab implementation ς = *realmin*.

Step 3: Perceptual Images

Obtain the perceptual wavelet coefficients $\{\omega_{s,o;\rho,D}\} = \alpha(\nu,r) \cdot \omega_{s,o}$ and $\{\hat{\omega}_{s,o;\rho,D}\} = \alpha(\nu,r) \cdot \hat{\omega}_{s,o}$ at distance D, using Equation 4.

Step 4: Inverse Wavelet Transformation

Perform the Inverse Wavelet Transform of $\{\omega_{s,o;\rho,D}\}$ and $\{\hat{\omega}_{s,o;\rho,D}\}$, obtaining the perceptual images $f_{\rho(i,j),D}$ and $\hat{f}_{\rho(i,j),D}$, respectively. The synthesis filter in Table 2 is an inverse Daubechies 9-tap/7-tap filter.

Table 2. 9/7 Synthesis Filter.

Synthesis Filter		
i	Low-Pass Filter $h_L(i)$	High-Pass Filter $h_H(i)$
0	1.115087052456994	0.6029490182363579
±1	0.5912717631142470	-0.2668641184428723
±2	-0.05754352622849957	-0.07822326652898785
±3	-0.09127176311424948	0.01686411844287495
±4		0.02674875741080976

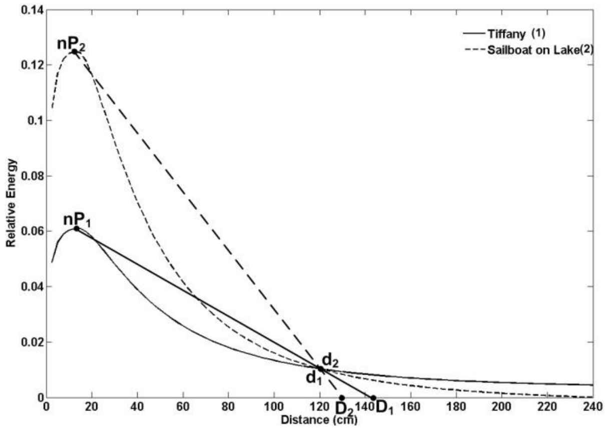
Step 5: PSNR between perceptual images

Calculate the PSNR between perceptual images $f_{\rho(i,j),D}$ and $\hat{f}_{\rho(i,j),D}$ using Eq. 2 in order to obtain the CIWaM weighted PSNR i.e. the VPSNR.

2.3 Discussion

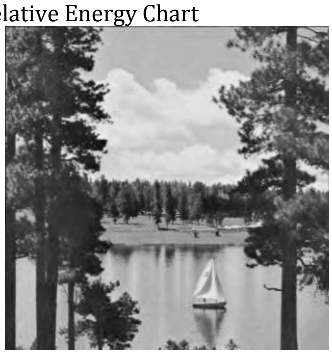
Figures 4(c) *Sailboat on Lake*$_2$ and 5(b) *Splash*$_1$ $D_2 = D_1 = 129cm$, but subjective quality of *Splash*$_1$ is clearly better than the one of *Sailboat on Lake*$_2$. Thus, even when CIWaM versions of *Splash*$_1$ and *Sailboat on Lake*$_2$ are calculate at 129cm, the resultant perceptual images have different objective quality. Hence VPSNR predicts that the error in Figure 5(b) is twice less ($\sim 3dB$) than in Figure 4(c).

That is why overall VPSNR algorithm is the estimation of the objective quality taking into account the set of the interactions of parameters nP, d and D. Figures 5 and 6 show examples when perceptual quality is equal and their respective points (nP,εR(nP)) do not correspond. In Figure 5, there is a difference of $6cm$ between D_1 and D_2, while in Figure 6, there is no difference between distances D_1 and D_2.



(a) Relative Energy Chart

(b) D_1=141.45cm VPSNR=34.82dB

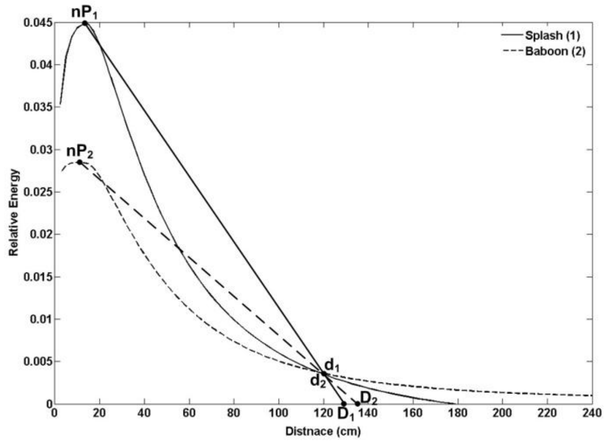


(c) D_2=129.67cm VPSNR=36.77dB

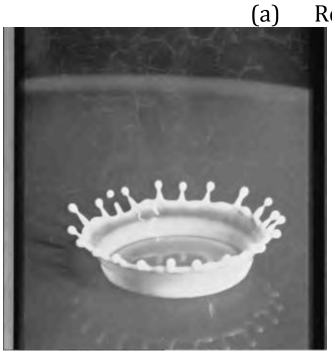
Fig.4. Relative Energy Chart of Images *Tiffany* and *Sailboat on Lake* (a), which are distorted by means of JPEG2000 PSNR=31dB and Observation Distance d=120cm. Perceptual quality VPSNR is equal to 34.82dB for (b) and 36.77dB for (c).

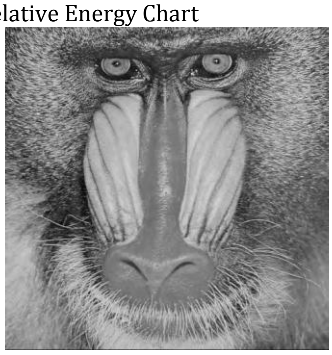
3 Experimental results

In this section, VPSNR performance is assessed by comparing the statistical significance with the psychophysical results obtained by human observers when



(a) Relative Energy Chart





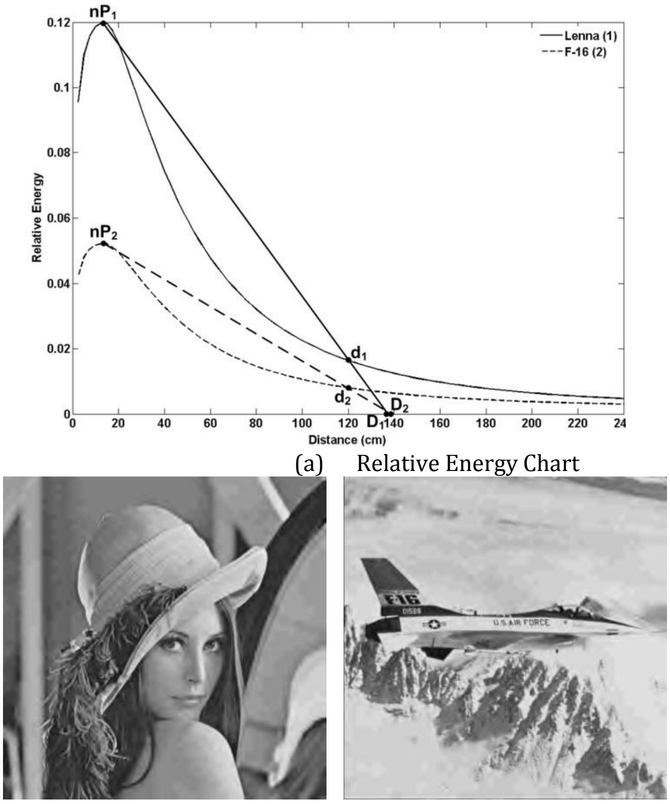
(b) D_1=129.10cm PSNR=35.88dB

(c) D_2=135.89cm PSNR=31.74dB

Fig.5. Relative Energy Chart of Images *Splash* and *Baboon*, which are distorted by means of JPEG2000 VPSNR=39.69dB and Observation Distance d=120cm. Objective quality PSNR is equal to 35.88dB for (b) and 31.74dB for (c).

judging the visual quality of an specific image. These results are expressed in Mean Opinion Scores. In this way, perceived image quality predicted by VPSNR is tested for JPEG2000 distortion across the Tampere Image Database (TID2008) of the Tampere University of Technology, presented by Ponomarenko et.al. in [9,10].

TID2008 Database contains 25 original images (Figure 7), which are distorted by 17 different types of distortions, each distortion has 4 degrees of intensity, that is, 68 versions of each source image. TID2008 also supplies subjective ratings by comparing original and distorted images by 654 observers from Italy, Finland and Ukraine. Thus, for JPEG and JPEG2000 compression distortions, there are 200 (25 images × 2 distortions × 4 distortion degrees) images in the database. MOS is presented as the global rating.



(a) Relative Energy Chart

(b) D_1=137.12cm PSNR=31.00dB (c) D_2=139.49cm PSNR=30.87dB

Fig.6. Relative Energy Chart of Images *Lenna* and *F-16*, which are distorted by means of JPEG2000 VPSNR=34.75dB and Observation Distance d=120cm. Objective quality PSNR is equal to 31.00dB for (b) and 30.87dB for (c).

3.1 Performance measures

Strength of Relationship (SR) is measured by a correlation coefficient. SR means how strong is the tendency of two variables to move in the same (opposite) direction. Pearson Correlation Coefficient (PCC) is the most common measure for predicting SR, when parametric data are used. But in the case of the correlation of non-parametric data the most common indicator is Spearman Rank-Order Correlation Coefficient (SROCC). Results of image quality metrics have no lineal relationship, which is why, it is not convenient to employ PCC, since even PSNR and MSE are the same metrics, PCC calculates different values.

Hence SROCC is a better choice for measuring SR between the opinion of observers and the results of a given metrics. However SROCC is appropriate for

Fig.7. Tested 512 × 384 pixel 24-bit color images, belonging to the Tampere Image Database.

testing a null hypothesis, but when this null hypothesis is rejected is difficult to interpret[11]. In the other hand, Kendall Rank-Order Correlation Coefficient (KROCC) corrects this problem by reflecting SR between compared variables. Furthermore KROCC estimates how similar are two rank-sets against a same object set. Thus, KROCC is interpreted as the probability to rank in the same order taking into account the number of inversions of pairs of objects for transforming one rank into the other[12]. Which is why, VPSNR and the rest of metrics are evaluated using KROCC. One of Limitation of KROCC is located in complexity of the algorithm, which takes more computing time than PCC and SROCC, but KROCC can show us an accurate Strength of Relationship between a metric and the opinion of an human observer.

MSE[13], PSNR[13], SSIM[14], MSSIM[15], VIF[16], VIFP[14], IFC[17] and WSNR[18] are compared against the performance of VPSNR for JPEG2000 compression distortion. I chose for evaluating these assessments the implementation provided in [19], since it is based on the parameters proposed by the author of each indicator.

VPSNR is implemented assuming the following features:

- Observation Distance, d=8H, where H is the height of a 512 × 512 image.
- 19” LCD monitor with horizontal resolution of 1280 pixels and 1024 pixels of vertical resolution.
- Gamma correction, $\gamma = 2.2$
- Wavelet Transform, set of wavelet planes ω with $n = 3$, Eq. 3.

3.2 Overall performance

Table 3 shows the performance of VPSNR and the other eight image quality assessments across the set of images from TID2008 image database employing KROCC for testing the distortion produced by a JPEG2000 compression.

Table 3. KROCC of VPSNR and other quality assessment algorithms on multiple image databases using JPEG2000 distortion. Bold and italicized entries represent the best and the second-best performers in the database, respectively. The last column shows the KROCC average of all image databases.

Metrics	TID2008 Image Database
Images	100
MSE	0.6382
PSNR	0.6382
SSIM	0.8573
MSSIM	*0.8656*
VIF	0.8515
VIFP	0.8215
IFC	0.7905
WSNR	0.8152
VPSNR	**0.8718**

Thus, for JPEG2000 compression distortion, VPSNR is also the best metrics for each database. VPSNR gets its better results when correlation is 0.8718 for a corpus of 100 images of the TID2008 database. For this distortion, MSSIM is the second best indicator. Furthermore VPSNR improves 0.2336 the perceptual functioning of PSNR when this metrics compares perceptual images in a dynamic way.

4 Conclusion

In this work, I presented a new metrics of full-reference image quality based on perceptual weighting of PSNR by means of a perceptual model. The VPSNR metrics is based on the measurement the objective quality of perceptual images predicted by CIWaM at *D*.

VPSNR was tested in TID2008 image database, over viewing distances proposed in this image database. Results show that VPSNR significantly increases the correlation of PSNR with perceived image quality, maintaining its advantageous features, in addition to be the best-ranked image quality gauge in overall performance, in comparison to a diversity of existing metrics. Accuracy of Second best-performing algorithm, MSSIM, is 1.5% lower than VPSNR forJPEG2000 distortion. While when CIWaM weights PSNR correlation of predicting subjective ratings either of PSNR or MSE improves the results by 23.36% for the same kind of distortions, on the average.

VPSNR is mainly developed for optimizing the perceptual error under the constraint of a limited bit-budget, but it contains another properties that can be used for quantizing, since the CIWaM algorithm calculates one value of an extended contrast sensitivity function by pixel. In this way, it is possible to quantize a particular pixel while an algorithm of bit allocation is working, incorporating into embedded compression schemes such as EZW[20], SPIHT[21], JPEG2000[22] or H*i*-SET[23].

Acknowledgment

This work is supported by National Polytechnic Institute of Mexico by means of Project No. 20131312 granted by the Academic Secretary, National Council of Science and Technology of Mexico by means of Project No. 204151/2013, Coimbra Group Scholarship Programme granted by University of Poitiers and Region of Poitou-Charentes, France.

References

1. Z. Wang and A. Bovik, Mean squared error: Love it or leave it? a new look at signal fidelity measures, Signal Processing Magazine, IEEE, vol. 26, no. 1, pp. 98 –117, jan. (2009).
2. Z. Wang and A. C. Bovik, Modern Image Quality Assessment, 1st ed. Morgan & Claypool Publishers: Synthesis Lectures on Image, Video, & Multimedia Processing, February (2006).
3. F. Auli-Llinas and J. Serra-Sagrista, Low complexity JPEG2000 rate control through reverse subband scanning order and coding passes concatenation, IEEE Signal Processing Letters, vol. 14, no. 4, pp. 251 –254, april (2007).
4. D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, ser. ISBN: 0-7923-7519-X. Kluwer Academic Publishers, 2002.
5. J. Bartrina-Rapesta, F. Auli-Llinas, J. Serra-Sagrista, and J. Monteagudo-Pereira, JPEG2000 Arbitrary ROI coding through rate-distortion optimization techniques, in Data Compression Conference, pp. 292 –301. (2008).
6. X. Otazu, C. Párraga, and M. Vanrell, Toward a unified chromatic induction model, Journal of Vision, vol. 10(12), no. 6, (2010).
7. G. van de Wouwer, P. Scheunders, and D. van Dyck, Statistical texture characterization from discrete wavelet representations, IEEE Transactions on Image Processing, vol. 8, no. 4, pp. 592 –598, Apr. (1999).

8. B. A. Wilson and M. A. Bayoumi, A computational kernel for fast and efficient compressed-domain calculations of wavelet subband energies, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 50, no. 7, pp. 389 – 392, July (2003).
9. N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, TID2008 - a database for evaluation of full-reference visual quality assessment metrics, Advances of Modern Radioelectronics, vol. 10, pp. 30–45, (2009).
10. N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, Metrics performance comparison for color image database, Fourth international workshop on video processing and quality metrics for consumer electronics, p. 6 p., (2009).
11. M. Hollander and D. Wolfe, Non-parametric Statistical Methods, 2nd ed. Wiley, (1999).
12. H. Abdi, Kendall rank correlation., N. Salkind, Ed. Encyclopedia of Measurement and Statistics. Thousand Oaks (CA), (2007).
13. Q. Huynh-Thu and M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electronics Letters, vol. 44, no. 13, pp. 800–801, (2008).
14. H. Sheikh and A. Bovik, Image information and visual quality, IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430 –444, feb. (2006).
15. Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment, in Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers., vol. 2, pp. 1398 – 1402, (2003).
16. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600 –612, (2004).
17. R. Sheikh, A. Bovik, and G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Transactions on Image Processing, vol. 14, pp. 2117–2128, (2005).
18. T. Mitsa and K. Varkur, Evaluation of contrast sensitivity functions for formulation of quality measures incorporated in halftoning algorithms, IEEE International Conference on Acustics, Speech and Signal Processing, vol. 5, pp. 301–304, (1993).
19. C. U. V. C. Laboratory: MeTriX MuX visual quality assessment package, available at http : //foulard.ece.cornell.edu/gaubatz/metrix mux/. Cornell University Visual Communications Laboratory. (2010).
20. J. Shapiro, Embedded image coding using Zerotrees of wavelet coefficients, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 41, no. 12, pp. 3445 – 3462, (1993).
21. A. Said and W. Pearlman, A new, fast, and efficient image codec based on Set Partitioning In Hierarchical Trees, IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 3, pp. 243 – 250, (1996).
22. A. Skodras, C. Christopoulos, and T. Ebrahimi, The JPEG 2000 still image compression standard, IEEE Signal Processing Magazine, vol. 18, no. 5, pp. 36–58, (2001).
23. J. Moreno and X. Otazu, "Image coder based on Hilbert Scaning of Embedded quadTrees," IEEE Data Compression Conference, p. 470, (2011).

Agrupamiento de Datos Linealmente Separables Mediante un Algoritmo Genético Celular

Migdaled López-Juárez, Ricardo Barrón Fernández, Salvador Godoy-Calderón

Centro de Investigación en Computación del Instituto Politécnico Nacional, México
migdaled.lopez@gmail.com, barron2131@gmail.com,
sgodoyc@gmail.com.

Resumen. El uso de algoritmos genéticos es amplio, ya que se permite encontrar soluciones óptimas con una función aptitud que así lo permita, tal es el caso de el agrupamiento particional de objetos, considerando particularmente aquellos que son linealmente separables. En este trabajo se propone un algoritmo genético celular para encontrar el número de particiones óptimo considerando el valor DB (índice que mide el agrupamiento de datos realizado por un algoritmo, introducido por David L. Davies y Donald W. Bouldin) como función de aptitud.

Palabras clave: Algoritmos genéticos, algoritmos genéticos celulares, particionamiento.

1 Introducción

Los algoritmos genéticos (GA) son métodos de búsqueda para problemas de optimización y están inspirados en los principios de selección natural de Charles Darwin [1]. Estos algoritmos tienen como objetivo evolucionar un conjunto de posibles soluciones (población) a un problema, hacia un óptimo global [2], permitiendo ser aplicados en una amplia gama de problemas. Tal es el caso del agrupamiento particional de datos, el cual forma parte de una de las técnicas de aprendizaje no supervisado (clustering). Dicho agrupamiento puede emplear el criterio de selección de centroides u holotipos que permitan particionar los datos [2]. Generalmente es necesario validar que tan adecuadas son las particiones creadas, es decir evaluar si es mejor tener 2 particiones que 3 ó 5, es por ello que existen índices de validación. Estos índices toman en cuenta los criterios de compactación y separación entre las particiones [3].

Como se ha mencionado, los GA trabajan con una población conformada por las posibles soluciones a un problema, en cuyo caso es necesario codificar la información de cada individuo (cromosoma) en números reales, números enteros, etc. A cada uno de los cromosomas se le aplicarán los operadores genéticos de selección, cruza (recombinación) y mutación con el objetivo de encontrar un individuo que represente una solución óptima global.

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 79–88
Paper Recived 22-09-2013 and Acepted 21-10-2013

En este trabajo se propone el uso de una familia de algoritmos que forman parte de los GA: algoritmos celulares genéticos (cGA). En los cGA cada individuo obteniene información de los individuos vecinos (vecindario) durante la selección y recombinación [7].

2 Revisión del estado del arte

El trabajo de Sanghamitra Bandyopadhyay y Ujjwal Maulik [4] propone el agrupamiento de datos de manera tal que, no necesita un número de particiones o bien centroides preestablecidos. De hecho, la población está conformada por individuos que no tienen la misma longitud (cromosomas de longitud variable). El cromosoma lo conforman números reales los cuales representan las coordenadas de los centroides en un espacio n-dimensional. El índice I desempeña el papel de función de aptitud.

Ge Xiufeng y Xing Changzheng [5], trabajan de igual forma con cromosomas de longitud variable codificados en números reales, tal como en [4]. Lo más sobresaliente de este trabajo es que dividen a la población en un número arbitrario de subpoblaciones llamadas islas, y éstas evolucionan de manera independiente. El índice DB es usado como función de aptitud [2,3].

De igual forma Venkatesh Katari et al. [6] trabajo con cromosomas de longitud variable como [4,5] pero ahora los cromosomas se encuentran codificados en números enteros positivos representando los niveles de RGB (por sus siglas en inglés, Red Green Blue) en una imagen digital. La función de aptitud es de nueva cuenta el índice DB.

3 Preliminares

3.1 Agrupamiento por particionamiento

El agrupamiento (clustering) de patrones es una técnica de clasificación no supervisada de patrones en la cual se particionan los datos en K regiones. El criterio empleado para dicho particionamiento es la medida de distancia (o bien, similitud) con respecto a un objeto arbitrario (holotipo o centroide).

Es necesario proporcionar el número K de grupos para calcular los holotipos o centroides [8], para este trabajo no es necesario indicar el número de particiones. Esto es, el algoritmo se encargará de evolucionar los holotipos para poder particionar los datos de manera óptima.

3.2 Algoritmos celulares genéticos (cGA)

Los cGA como se ha mencionado, forman parte de una familia de los GA. La población está usualmente estructurada en una rejilla (o malla) n-dimensional de individuos. Este modelo simula la evolución natural desde el punto de vista del individuo, los individuos sólo realizan la selección y recombinación con sus vecinos. Con base en los experimentos realizados [7] se observa que la lenta difusión de soluciones inducida en la población con los vecindarios solapados permite la exploración (diversificación), mientras que la explotación (intensificación) se logra con la interacción de cada uno de los vecindarios (ver figura 1).

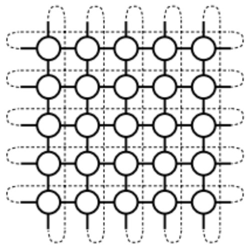


Fig. 1. Malla toroidal de 2 dimensiones donde se representa la estructura de los individuos

4 Propuesta

A continuación se resumen los pasos principales que conforman el algoritmo propuesto (ver tabla 1).

Tabla 1. Algoritmo genético celular propuesto.

1. Inicializar población
2. Evaluar la población inicial
3. Verificar si la condición de paro se cumple
   (a) Finalizar el algoritmo en caso de que se cumpla
   (b) Continuar con el paso 4
4. Determinar el vecindario de cada individuo
5. Seleccionar la pareja de recombinación para cada individuo
6. Recombinar a cada individuo con la pareja previamente seleccionada para generar un nuevo individuo
7. Mutar a cada uno de los nuevos individuos generados
8. Reemplazar cada uno de los individuos, sólo si el nuevo individuo que ha generado es mejor que él
9. Regresar a paso 2

En las siguientes sub secciones se mencionan más a detalle las etapas que intervienen para llevar a cabo la ejecución del algoritmo propuesto.

4.1 Estructura de la población

En este caso se utilizan rejillas de *nxn*, donde *n* es el número máximo de particiones que se desean probar. Para obtener el valor de *n* es necesario considerar el número de datos que se tienen, obtener su raíz cuadrada, finalmente el número cuadrado superior cercano a la raíz cuadrada será el resultado.

Es decir, en caso de tener 30 datos, *n* será igual a 9.

La política empleada para formar al vecindario es la selección de los 4 vecinos más cercanos, considerando que el individuo se encuentre en las coordenadas *i,j* dentro de la malla los vecinos serán: *(i,j-1), (i+1,j),(i-1,j),(i,j-1).* Este tipo de vecindario fue propuesto por Von Neumann en la década de los 50's.

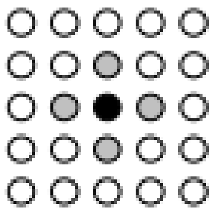


Fig. 2. Selección de vecinos para cada individuo

4.2 Codificación de los individuos

La población que conforma nuestro espacio de búsqueda son coordenadas de números reales (holotipo) para formar un determinado número de particiones.

El cromosoma que se representa para cada individuo es de longitud variable, la longitud depende del número de particiones que se desean formar y la dimensión de los holotipos (R^n), como se muestra en la tabla 2.

$(z_{11}, z_{21}, \ldots, z_{i1})$	$(z_{12}, z_{22}, \ldots, z_{i2})$	...	$(z_{i1}, z_{i2}, \ldots, z_{ij})$

Tabla 2. Codificación de los individuos

4.3 Operadores genéticos

Para llevar a cabo la recombinación entre individuos es necesaria una previa selección de aquellos individuos (padres) que generarán una nueva descendencia. En este caso, tenemos la a un individuo dentro de la población como primer padre (P1) y tomando en cuenta a los individuos más cercanos a él (vecinos), se selecciona al mejor (P2) por medio del método de ruleta, donde básicamente la selección es proporcional a la función de desempeño [1].

La recombinación se logra tomando como base a P1, posteriormente seleccionar de un punto de cruza para poder emplear información de P2. Para este caso nos referimos a coordenadas de holotipos. La cantidad de coordenadas a ser intercambiadas queda en función de una probabilidad de recombinación.

Durante la mutación el nuevo individuo que se formó con P1 y P2 será alterado por incrementos aleatorios.

$$x_{t+1} = x_t(1 \pm 2\delta) \quad (1)$$

Donde,

δ: un número en el rango [0, 1] generado con una distribución uniforme
z: indica el valor de una de las coordenadas del holotipo
t: indica el número de generación

El reemplazo de z_t por z_{t+1} sólo en el caso que z_{t+1} tenga un mejor valor de aptitud.

4.4 Evaluación de los individuos

El cálculo de aptitud de los individuos está basado en el índice DB. Este índice fue seleccionado dado que responde adecuadamente a la evaluación de los agrupamientos en los datos que se manejan (hiperesféricos)

$$\text{Disp Ci} = \sqrt{\frac{1}{|\text{Ci}|} \sum_{\text{oi,oj} \in \text{Ci}} \|\text{oi} - \text{oj}\|} \quad (2)$$

$$\text{DB} = \frac{1}{\text{k}} \sum_{\text{i}=1}^{\text{k}} \max_{\text{j}=1\ldots\text{k},\text{i}\neq\text{j}} \left\{ \frac{\text{Disp(Ci)} + \text{Disp(Cj)}}{\text{d(Ci, Cj)}} \right\} \quad (3)$$

Donde,

i, j: indican el índice de los holotipos que conforman la partición
$Disp\ Ci$: indican la dispersión de los elementos de una partición
k: número de particiones

4.5 Condición de paro

La condición de paro establecida es llegar a un máximo de w generaciones, este dato es arbitrario.

5 Experimentos

Para poner en práctica el algoritmo genético celular propuesto en la sección 4, se implementó el código en C empleando gcc 4.7.1 sobre una plataforma OpenSuse.

6 Experimentos

Para poner en práctica el algoritmo genético celular propuesto en la sección 4, se implementó el código en C empleando gcc 4.7.1 sobre una plataforma OpenSuse Linux 12.2. Se emplearon dos conjuntos de datos. Uno de los conjuntos es sintético y está constituido por tres clases linealmente separables entre ellas. El otro conjunto es real [10] y cuenta con 3 clases, sólo una de ellas es linealmente separable de las otras 2 (estas dos últimas no son linealmente separables una de la otra.)

El objetivo de los experimentos es validar que se pueda obtener el número óptimo de clases para cada uno de los conjuntos de datos, por lo que se toma en consideración la frecuencia relativa en las pruebas ejecutadas para dicho número. El número de ejecuciones para cada experimento es 30.

6.1 Datos sintéticos

Se empleó un conjunto de 300 datos hiperesféricos en R^2 con 100 puntos en cada clase, el conjunto de datos se puede observar en la Figura 2. En la Tabla 3 se muestran los parámetros empleados en la ejecución.

Tabla 3. Parámetros para la ejecución del experimento con 300 datos hiperesféricos en R^2.

Probabilidad de cruza: 0.65
Probabilidad de mutación: 0.01
Número límite de generaciones: 25

A continuación, se muestra los resultados obtenidos con la ejecución del algoritmo, el número de clases que se encuentra como el óptimo y la frecuencia relativa. El resultado es satisfactorio, dado que el conjunto es de tres clases.

Tabla 4. Resultados con un conjunto 300 datos hiperesféricos.

Número óptimo de clases	Frecuencia relativa (%)
3	80%
2	20%

En la Figura 3 se observa el valor promedio de aptitud (fitness) obtenido, se puede observar que la difusión del óptimo local dentro de cada vecindario permite llegar a un óptimo global en 25 generaciones.

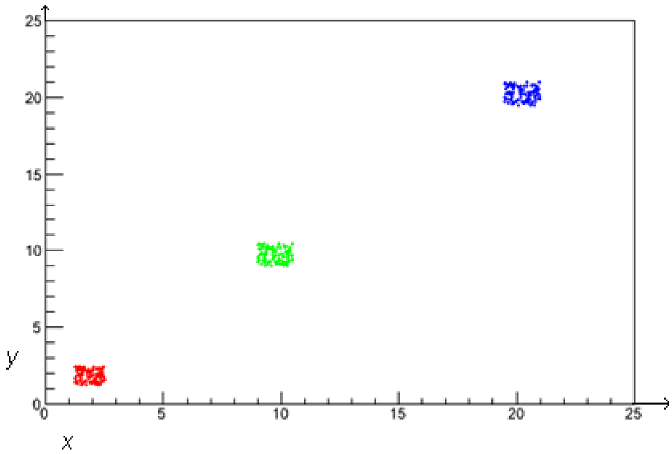


Figura 1. Datos en espacio R^2

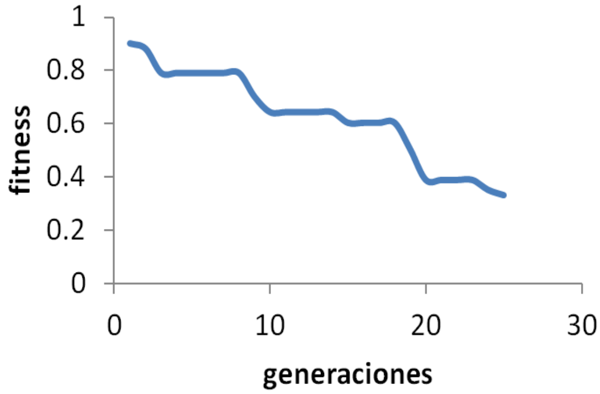


Figura 2. Gráfica de convergencia para 300 datos en R^2.

6.2 Datos reales

Se empleó el conjunto Iris [10] en R^4, el conjunto de datos se puede observar en la Figura 4. En la Tabla 5 se muestran los parámetros empleados en la ejecución.

Tabla 5. Parámetros para la ejecución del experimento con el conjunto Iris.

Probabilidad de cruza: 0.65
Probabilidad de mutación: 0.01
Número límite de generaciones: 45

A continuación, se muestra los resultados obtenidos con la ejecución del algoritmo, el número de clases que se encuentra como el óptimo y la frecuencia relativa del evento. El resultado es aceptable en comparación con [4,11-12] pues encuentran que el número de clases detectadas son dos.

Tabla 6. Resultados con el conjunto Iris Data.

Número óptimo de clases	Frecuencia relativa (%)
2	60%
3	40%

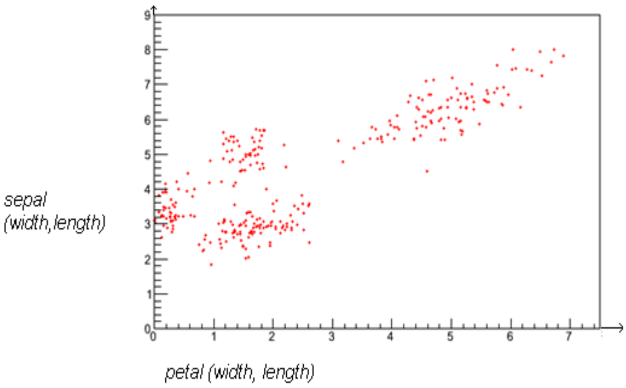


Figura 3. Conjunto Iris Data, en el eje de las abscisas se consideran las medidas de los pétalos y en el eje de las ordenadas los sépalos.

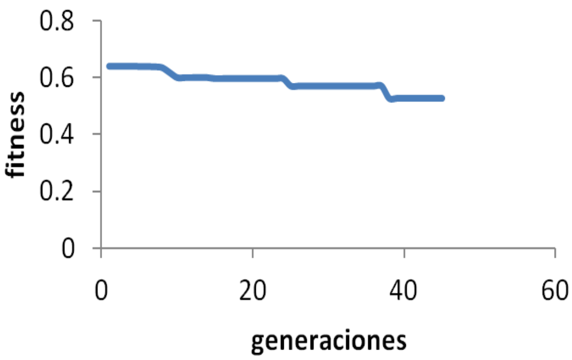


Figura 4. Gráfica de convergencia para Iris Data.

7 Conclusiones

Luego de implementar el algoritmo genético celular y realizar pruebas, se obtiene un resultado satisfactorio, dato que es posible encontrar un número óptimo de clases empleando el índice DB para medir la aptitud (fitness) de cada individuo. Para el conjunto de datos sintético se encuentran tres, y para el conjunto real dos.

Así como también es importante señalar que, en menos de 50 generaciones es posible determinar un valor óptimo para los dos conjuntos de datos probados.

Agradecimientos

Agradecemos el apoyo otorgado por el Instituto Politécnico Nacional a través del proyecto 20131210.

Referencias

1. Holland, J.H. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor, University of Michigan Press.
2. Naldi, MuriloCoelho and Carvalho, André C.P.L.F. and Campell, Ricardo José GabrielliBarreto and Hruschka, Eduardo Raul. (2008). Soft Computing for Knowledge Discovery and Data Mining, Genetic Clustering for Data Mining. Springer US.
3. Halkidi, Maria and Batistakis, Yannis and Vazirgiannis, Michalis. (2001). On Clustering Validation Techniques. Kluwer Academic Publishers.
4. Bandyopadhyay, S., Maulik, U., (2001), Nonparametric genetic clustering: Comparison of validity indices. Systems, Man and Cybernetics, Part C, IEEE Transactions on: Applications and Reviews.
5. Ge Xiufeng, Xing Changzheng. (2010). K-means Multiple Clustering Research Based on Pseudo Parallel Genetic Algorithm. Information Technology and Applications (IFITA), 2010 International Forum on.
6. Venkatesh Katari, Suresh Ch, Ra Satapathy, Member Ieee, Jvr Murthy, Pvgd Prasad Reddy. (2007). Hybridized Improved Genetic Algorithm with Variable Length Chromosome for Image Clustering Abstract. IJCSNS International Journal of Computer Science and Network Security.
7. Alba, Enrique and Dorronsoro, Bernabè. (2008). Cellular Genetic Algorithms. Introduction to Cellular Genetic Algorithms. Universidad de Málaga. Springer US.
8. Xiu Riu, Wunsch II Donald C. (2009). Clustering. IEEE Press Series on Computational Intelligence.
9. Niloy Gangul, Biplab K Sikdar, Andreas Deutsch, Geoffrey Canright, P Pal Chaudhuri. (2001). A Survey on Cellular Automata.
10. Iris Dataset, http://archive.ics.uci.edu/ml/datasets/Iris.
11. J. C. Bezdek and N. R. Pal. (1998). Some new indexes of cluster validity. IEEE Trans. Syst., Man, Cybern.
12. R. Kothari and D. Pitts. (1999). On finding the number of clusters. Pattern Recognit. Lett.

Análisis Comparativo de Algoritmos Genéticos Aplicados a Calendarización de Trabajos en un Grid Computacional

Victor Hugo Yaurima Basaldúa[1], Andrei Tchernykh[2], Moisés Torres Martínez[3]

[1] Universidad Estatal de Sonora, San Luis R.C., México, [2] Centro de Investigación CICESE, Ensenada, México, [3] Universidad de Guadalajara, Guadalajara, México,
victor.yaurima@ues.mx, chernykh@cicese.mx, moises.torres@redudg.udg.mx

Resumen. Este artículo aborda la calendarización de trabajos paralelos en un Grid jerárquico con dos etapas. En esta configuración uno de los grandes retos es asignar las tareas de manera que permita un uso eficiente de los recursos, al mismo tiempo que satisface otros criterios. En general, los criterios de optimización a menudo están en conflicto. Para resolver este problema, proponemos un algoritmo genético bi-objetivo y presentamos un estudio experimental de seis operadores de cruzamientos, y tres operadores de mutación. Se determinan los parámetros más influyentes a través de un análisis estadístico de varianza multifactorial y comparamos nuestra propuesta con cinco estrategias de asignación conocidas en la literatura.

Palabras clave: Algoritmos genéticos, grid, calendarización.

1 Introducción

En este artículo presentamos un trabajo experimental de calendarización en un Grid computacional jerárquico de dos etapas [1, 2]. Uno de los grandes retos es obtener una calendarización que permita un uso eficiente de los recursos, al mismo tiempo que satisface otros criterios. Los criterios de optimización a menudo están en conflicto. Por ejemplo, los proveedores de los recursos y los usuarios tienen diferentes objetivos: los proveedores buscan una alta utilización de sus recursos, mientras que los usuarios están interesados en una rápida respuesta. En este trabajo se consideran ambos objetivos utilizando el método de agregación [3]. Para la primera etapa se desarrolla un algoritmo genético bi-objetivo para seleccionar recursos computacionales. Se examina el desempeño de un Grid computacional basado en datos reales de cinco días de actividad y se presenta un análisis comparativo de seis operadores de cruzamiento y tres operadores de mutación. Para calibrar el algoritmo genético se aplica un análisis de varianza multifactorial. El algoritmo genético calibrado es comparado con cinco estrategias conocidas.

Después de abordar el modelo del problema en la Sección 2, describimos en detalle el algoritmo genético en la Sección 3. En la Sección 4 calibramos el algoritmo genéti-

© E. A. Santos, J. C. Chimal, L. Cabrera, E. Castillo.
Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 89–99
Paper Recived 22-09-2013 and Acepted 21-10-2013

co y presentamos la evaluación comparativa en la Sección 5. Finalmente concluimos en la Sección 6.

2 Modelo

Nos enfocamos en un modelo de calendarización fuera de línea: n trabajos paralelos $J_1, J_2, \dots, J_n$ deben ser calendarizados en m máquinas paralelas (sitios) $N_1, N_2, \dots, N_m$. Sea m_i el número de procesadores idénticos de la máquina N_i. Se asume que las máquinas están enlistadas en orden descendente de acuerdo al número de procesadores, esto es $m_1 \leq m_2 \leq \cdots \leq m_m$. Cada trabajo J_i se describe por una tupla $(size_j, p_j, p'_j)$: donde $1 \leq size_j \leq m_m$ representa el número de procesadores requeridos, también llamado grado de paralelismo, tiempo de ejecución p_j y tiempo de ejecución estimado p'_j. Todos los trabajos están disponibles antes de iniciar el proceso de calendarización. El tiempo de procesamiento del trabajo es desconocido hasta que el trabajo ha completado su ejecución. El tiempo de ejecución estimado p'_j es proporcionado por el usuario. Una máquina debe ejecutar un trabajo dedicando exactamente $size_j$ procesadores por un periodo ininterrumpido de tiempo p_j. Como no se permiten ejecuciones multi-sitios, un trabajo J_j puede ejecutarse solo en una máquina N_i si $size_j \leq m_j$.

Dos criterios son considerados: La terminación del calendario (*makespan*): $C_{max} = max(C_i)$, $i = 1,2,3, \dots, N_m$, donde C_i es el tiempo máximo de terminación de un trabajo en la máquina N_i, y el promedio de finalización (*mean turnaround time*): $TA = \frac{1}{n}\sum_{j=1}^{n} c_j$, donde c_j es el tiempo de terminación del trabajo J_j.

De acuerdo a la notación de tres campos (α|β|γ) [4], nuestro problema de calendarización es caracterizado como $GP_m \left| size_j, p_j, p'_j \right| OWA$, donde GP_m es el modelo de Grid de m maquinas con procesadores idénticos, OWA es un criterio de optimización que concatena dos criterios por el método de Agregación ($OWA = w_1 C_{max} + w_2 TA$), y w_i es el peso asignado a cada criterio. Para el problema en la segunda etapa usamos la estrategia Easy Backfilling [5].

2.1 Trabajos relacionados

Los algoritmos de calendarización para modelos de Grid en dos etapas pueden ser divididos en calendarización global y calendarización local [6]. En la primera etapa seleccionamos una máquina para cada trabajo usando un algoritmo genético. En la segunda etapa, usamos una estrategia de ejecución local para los trabajos recibidos.

La administración de recursos de un Grid es influenciada por múltiples objetivos y pueden requerir apoyo en la toma de decisiones con múltiples criterios. En [7], se considera la calendarización de tareas en recursos tomando en cuenta dos criterios: tiempo máximo de completar una tarea y tiempo promedio de completar las tareas. En [8], se presenta una calendarización de tareas basada en negociación de recursos, reservación de antemano y preferencias de usuario. Para ayudar a escoger la mejor estrategia en [9], se desarrolla un análisis de diferentes métricas de acuerdo a la metodología de sus degradaciones. El objetivo es encontrar una estrategia que se comporte con buenos resultados en todos los casos de estudio, considerando cargas de trabajo y configuraciones de Grid diferentes.

La metodología para una decisión multi-criterio puede basarse en la optimalidad de Pareto, sin embargo es muy difícil alcanzar soluciones en un tiempo razonable usando esta opción. En [3], usan el método de agregación para modelar las preferencias de los actores involucrados. En [10] se presentan estrategias de asignación a diferentes centros de cómputo y proponen un modelo de calendarización multi-criterio. Ellos concluyen que tal calendarización puede ser desarrollada eficientemente usando algoritmos genéticos. En [11], se aplican algoritmos genéticos con dos objetivos usando el método de agregación, comparando cinco operadores de cruzamiento.

3 Algoritmo genético

Los algoritmos genéticos son una técnica conocida, usada para encontrar soluciones a problemas de optimización combinatoria. Las soluciones candidatas son codificadas como cromosomas, también llamadas genomas o individuos. En nuestro caso, cada individuo o solución es codificada en una matriz $n \cdot m$. El número -1 en una celda significa que no existe un trabajo en esa posición de la cola local.

La fila $i = 0, \dots, m-1$ representa la cola local en la máquina N_i. El conjunto de máquinas disponibles para el trabajo J_j son las máquinas con índices $\{f_j \dots m\}$, donde f_j es el índice i más pequeño tal que $m_i \geq size_j$.

Una población está formada por un conjunto de individuos que estarán modificándose para mejorar su aptitud. Un valor de aptitud es la medida de la calidad de la solución de acuerdo al criterio de optimización usado. En este caso usamos dos criterios (Sección 2). Para que los dos criterios se evalúen simultáneamente, usamos el método de agregación: promedio ponderado ordenado (*Ordered Weighted Averaging* - *OWA*) [3]. Se contemplan valores o pesos que representan la relativa importancia de cada criterio:

$$OWA(x_1, x_2, \dots, x_n) = \sum_{c=1}^{k} w_c s(x)_{\sigma(c)} \quad (1)$$

Donde w_c es el peso, $c = 1, \dots, k$, x_c es un valor asociado con la satisfacción del criterio c. Se realiza la permutación de valores: $s(x)_{\sigma(1)} \leq s(x)_{\sigma(2)} \leq \dots s(x)_{\sigma(k)}$. Los pesos (w_c) son positivos y $\sum_{c=1}^{k} w_c = 1$. El objetivo es encontrar un esquema de pesos que provea el mejor valor medio de acuerdo a la conveniencia de los interesados y el más alto valor posible para el peor caso. Para alcanzar esto el peso w_1 debe ser relativamente grande, mientras que el peso w_k debe ser pequeño, k denota el número de criterios. Los pesos restantes son decrementados en valor desde w_1 hasta w_k de acuerdo a:

$$w_c = \begin{cases} 3/2k, & c = 1 \\ (3k - 2c - 1)/2n(k - 1), & 1 < c \leq k \end{cases} \quad (2)$$

El valor mínimo de OWA corresponde a la mejor aptitud.

Se aplican tres operadores genéticos: selección, cruzamiento y mutación. La selección identifica los individuos que van a cruzarse para producir la siguiente generación. Utilizamos la selección por torneo binario, donde dos individuos son elegidos aleatoriamente de la población, el que tenga mejor aptitud, gana. Este proceso se repite dos veces con el fin de seleccionar dos padres. Los individuos se desarrollan hasta que el criterio de paro se cumple. En nuestro caso, el algoritmo se detiene si la aptitud del mejor individuo encontrado no mejora en 10 generaciones.

3.1 Operadores de cruzamiento

El operador de cruzamiento se aplica bajo cierta probabilidad (P_c). En este trabajo son considerados seis operadores que a continuación se describen.

One Segment Crossover for Matrix (OSXM). Está basado en el operador de cruzamiento OSX - *One Segment Crossover* [14]. En este operador, se seleccionan aleatoriamente dos puntos en la matriz, S1 y S2 desde 0 hasta el máximo índice usado. El hijo hereda las columnas del padre 1 desde la posición 0 hasta S1. Hereda las columnas del padre 2 desde S1 hasta S2, considerando solamente aquellos elementos que no han sido copiados del padre 1. Finalmente el hijo hereda el resto de los elementos del padre 1.

Two Point Crossover for Matrix (TPM). Está basado en el operador de cruzamiento *Two Point Crossover* [15]. En este operador, dos posiciones de la matriz son elegidas aleatoriamente *S1* y *S2*. Las columnas desde la posición 0 hasta *S1* y desde *S2* hasta el final son copiados del padre 1. El resto de los elementos son copiados del padre 2.

Order Based Crossover for Matrix (OBXM). Está basado en el operador de cruzamiento *OBX - Order Based Crossover* [12]. Se usa una máscara binaria generada aleatoria e uniformemente de acuerdo al número de columnas en la matriz. Los valores de la máscara binaria iguales a uno indican que sus correspondientes columnas son copiadas al hijo del padre 1. El resto de los elementos son copiados del padre 2 (Fig. 1).

Precedence Preservative Crossover for Matrix (PPXM). Está basado en el operador de cruzamiento *PPX - Precedence Preservative Crossover* [13]. Se usa una máscara binaria generada aleatoria e uniformemente de acuerdo al número de columnas en la matriz. La columna cuyo valor corresponde a la máscara igual a uno, es copiada al hijo del padre 1. La siguiente columna si la máscara es igual a cero, es copiada del padre 2, de lo contrario es copiada del padre 1, así sucesivamente son copiados los elementos. Se toma en cuenta el orden de izquierda a derecha en cada iteración y que los elementos no hayan sido copiados al hijo de alguno de los padres.

Order Segment Crossover for Matrix with Setup (OSXMS). Está basado en el operador de cruzamiento *OSX-Order Segment Crossover* [11, 14]. Similar al OSXM, pero los elementos son ordenados ascendentemente de acuerdo al número de procesadores requeridos antes de ser copiados al hijo, tomando en cuenta que la capacidad de la máquina en cuanto a número de procesadores sea suficiente para ejecutar el trabajo.

Order Based Crossover for Matrix with Setup (OBXMS). Está basado en el operador de cruzamiento OBX - *Order Based Crossover* [12]. Se usa una máscara binaria generada aleatoria e uniformemente de acuerdo al número de columnas en la matriz. Los valores de la máscara binaria iguales a uno indican que sus correspondientes columnas son copiadas al hijo del padre 1 tomando en cuenta un orden ascendente de los trabajos de acuerdo al número de procesadores requeridos. El resto de los elementos son copiados del padre 2.

Máscara

1	0	0	0	1	1	0	1

	0	1	2	3	4	5	6	7
0	21	18	-1	-1	-1	-1	-1	-1
1	32	27	25	-1	-1	-1	-1	-1
2	24	6	20	-1	-1	-1	-1	-1
3	45	12	44	46	-1	-1	-1	-1
4	48	13	22	47	-1	-1	-1	-1
5	38	16	37	40	14	-1	-1	-1
6	26	2	11	49	39	-1	-1	-1
7	43	15	3	36	8	-1	-1	-1
8	9	31	42	35	41	23	-1	-1
9	34	29	10	28	33	0	-1	-1
10	17	4	1	30	19	5	7	-1

Padre 1

Máscara

1	0	0	0	1	1	0	1

	0	1	2	3	4	5	6	7
0	22	47	48	-1	-1	-1	-1	-1
1	23	24	20	18	-1	-1	-1	-1
2	44	32	21	49	-1	-1	-1	-1
3	46	38	25	27	-1	-1	-1	-1
4	17	8	6	45	40	-1	-1	-1
5	13	16	5	34	-1	-1	-1	-1
6	12	15	35	1	-1	-1	-1	-1
7	0	26	9	2	10	-1	-1	-1
8	7	43	14	3	31	-1	-1	-1
9	11	37	41	36	29	19	30	-1
10	28	42	33	39	4	-1	-1	-1

Padre 2

Máscara

1	0	0	0	1	1	0	1

	0	1	2	3	4	5	6	7
0	21	22	47	-1	-1	-1	-1	-1
1	32	20	18	-1	-1	-1	-1	-1
2	24	44	49	-1	-1	-1	-1	-1
3	45	46	25	27	-1	-1	-1	-1
4	48	6	40	-1	-1	-1	-1	-1
5	38	14	13	16	-1	-1	-1	-1
6	26	39	12	15	35	1	-1	-1
7	43	8	2	10	-1	-1	-1	-1
8	9	41	23	7	31	3	-1	-1
9	34	33	0	11	37	36	29	30
10	17	19	5	28	42	4	-1	-1

Hijo

Fig. 1. Order Based Crossover for Matrix (*OBXM*)

3.2 Operadores de mutación

Los operadores de mutación producen pequeños cambios en los individuos de acuerdo a una probabilidad P_m. Este operador ayuda a prevenir caer en óptimos locales y a extender el espacio de búsqueda del algoritmo. Se consideran tres operadores de mutación adaptados para una matriz de dos dimensiones: (1) *Queue_Insert***.** Dos posiciones *S1* y *S2* son seleccionadas aleatoriamente. Los elementos de la columna *S2* son insertados en la columna *S1*, recorriendo el resto de los elementos. (2) *Queue_Swap* selecciona aleatoriamente dos columnas en la matriz e intercambia sus elementos; (3) *Queue_Switch*, selecciona una columna aleatoriamente e intercambia sus elementos con la columna adyacente.

4 Calibración del algoritmo genético

4.1 Carga de trabajo

Dos aspectos fundamentales deben abordarse para configurar un entorno de simulación para la evaluación del desempeño. Por un lado, se necesitan registros representativos de carga de trabajo para producir resultados confiables. Por otro lado, un buen entorno de prueba debe ser configurado para obtener resultados reproducibles y comparables. Se consideran cuatro registros de PWA (Parallel Workloads Archive) [16] correspondientes a: Cornell Theory Center, High Performance Computing Center North, Swedish Royal Institute of Technology y Los Alamos National Lab; y un registro de GWA (Grid Workloads Archive) [17] correspondiente a: Advanced School for Computing and Imaging. Se tomaron los registros correspondientes a cinco días.

4.2 Calibración de parámetros

Se usa una adaptación del método de diseño de experimentos propuesto por [18], donde se consideran los siguientes pasos: (a) se ejecuta cada carga de trabajo con todas las posibles combinaciones de parámetros; (b) se obtiene la mejor solución; (c) se calcula la diferencia relativa de cada algoritmo sobre la mejor solución (d) Se aplica el Análisis de Varianza Multifactorial (ANOVA) para encontrar el parámetro que más influye en la solución y para seleccionar el conjunto de los mejores valores de cada parámetro que constituirá el algoritmo adecuado para el problema abordado. La Tabla 1 muestra los parámetros usados para la calibración.

Tabla 1. Parámetros de calibración

Parámetros	Niveles
Operadores de cruzamiento:	OSXM, TPM, OBXM, PPXM, OSXMS, OBXMS
Operadores de mutación:	Queue_Insert, Queue_Swap, Queue_Switch
Probabilidades:	Cruzamiento: 0.9, Mutación: 0.01
Población:	100 individuos.
Número de trabajos por individuo:	Día 1: 1375; Día 2: 646; Día 3: 564; Día 4: 1041; Día 5: 1083.
Operador de selección:	Torneo binario.
Tamaño de sitios:	4, 4, 4, 8, 8, 8, 16, 16, 32, 32
Criterio de paro:	Si la aptitud no mejora en 10 generaciones.

Se consideraron 18 diferentes algoritmos. Se realizaron 30 ejecuciones para la carga de trabajo en cada combinación, en total 18 x 30 = 540 experimentos. El desempeño de cada algoritmo es calculado como el porcentaje del incremento relativo sobre la mejor solución obtenida (*IRMS*), calculado con la siguiente fórmula: $IRMS = (Heu_{sol} - Best_{sol})/Best_{sol} \cdot 100.$ Donde Heu_{sol} es el valor de la función objetivo obtenido por el algoritmo considerado y $Best_{sol}$ es el mejor valor obtenido durante la ejecución de todas las posibles combinaciones de los parámetros.

4.3 Análisis de varianza.

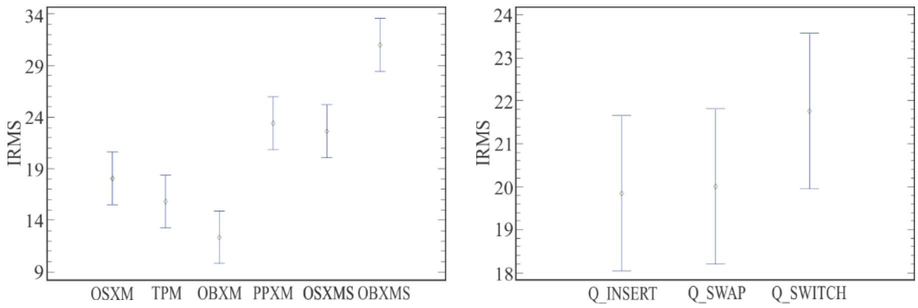
El análisis de varianza se aplica para evaluar la diferencia estadística entre los resultados experimentales y observar el efecto de los parámetros sobre la calidad de los resultados. Se usa para determinar los factores que tienen un efecto significativo y saber cuáles son los factores más importantes (Tabla 2).

Tabla 2. Análisis de varianza para IRMS. 5 días.

		(a) *Cmax*		(b) *TA*	
Días	**Factores**	**Razón-F**	**Valor-P**	**Razón-F**	**Valor-P**
Día 1	Cruzamiento	16.43	**0.0002**	11.76	**0.0006**
	Mutación	0.88	0.4452	2.23	0.1580
Día 2	Cruzamiento	32.25	**0.0000**	7.26	**0.0041**
	Mutación	1.64	0.2416	0.77	0.4874
Día 3	Cruzamiento	23.70	**0.0000**	11.00	**0.0008**
	Mutación	0.82	0.4681	0.99	0.4056
Día 4	Cruzamiento	26.49	**0.0000**	7.50	**0.0036**
	Mutación	0.35	0.7104	0.90	0.4372
Día 5	Cruzamiento	51.64	**0.0000**	13.34	**0.0004**
	Mutación	1.48	0.2742	0.61	0.5628

Los resultados de ANOVA descomponen la variabilidad de IRMS en contribuciones debidas a varios factores. Puesto que se ha escogido la suma de cuadrados Tipo III (por omisión), la contribución de cada factor se mide eliminando los efectos de los demás factores.

La Fig. 2 (a) muestra los resultados obtenidos en el primer día para el factor de operador de cruzamiento. El eje vertical es el valor del *IRMS.* Podemos ver que el operador de cruzamiento *OBXM* es mejor. La Fig. 2 (b) muestra los resultados para los operadores de mutación. El operador *Queue_Insert* resultó ser ligeramente mejor. Este comportamiento es similar en los cinco días analizados. Con el criterio *TA* los operadores de mutación tienen el mismo comportamiento y respecto al operador de cruzamiento *OSXM* es mejor.



(a) Operadores de cruzamiento. Cmax

(b) Operadores de mutación. Cmax

Fig. 2. Resultados Día 1

Hasta este punto, hemos calibrado el algoritmo genético, obteniendo como operador de cruzamiento: *OBXM* para el criterio *Cmax* y *OSXM* para el criterio *TA*. Siendo que en la siguiente sección de comparación usaremos el criterio *Cmax*, seleccionamos el operador *OBXM.* Como operador de mutación proponemos el mejor en ambos casos: *Queue_Insert*. Usamos 0.9 como probabilidad de cruzamiento de acuerdo a [11]. A este algoritmo calibrado, llamaremos *Calib_GA*.

5 Evaluación comparativa

Después de obtener la calibración del algoritmo genético, en esta sección procedemos a comparar *Calib_GA* con 5 estrategias que se presentaron en [9] (Tabla 3).

Tabla 3. Estrategias de asignación

Estrategia	Descripción
MIN LB	Asigna el trabajo j al sitio con los menores recursos consumidos por procesador. $min_{i=1..m}\left\{\sum_{k=1}^{l_i}\frac{size_k \cdot p_k^j}{m_i}\right\}$
MIN CT	Asigna el trabajo j al sitio con el menor tiempo de finalización del Grid $min\{C_{max}^i\}$,
MIN WT	Asigna el trabajo j al sitio con el menor promedio de tiempo de espera de los trabajos. $min_{i=1..m}\left\{\sum_{g_k=i}\frac{t_w^k}{n_i}\right\}$
MIN ST	Asigna el trabajo j al sitio que pueda iniciar más temprano su ejecución. $min_{\ i=1..m}\{s_j^i\}$
MIN TA	Asigna el trabajo j al sitio con el menor tiempo de permanencia de los trabajos $min_{i=1..m}\left\{\sum_{g_k=i}\frac{c_w^k}{n_i}\right\}$

Del total de trabajos hemos tomado la cantidad correspondiente al registro de 5 días: 1375, 646, 564, 1041, 1083 trabajos por día. Hemos realizado 30 ejecuciones y tomado el promedio de éstas para efecto de la comparación.

El desempeño de los algoritmos es calculado usando el *IRMS*. Para todos los casos el algoritmo Calib_GA resultó mejor. El *IRMS* de este algoritmo está en el rango 2.7% - 12.7% en promedio. Mientras que los otros algoritmos tienen una distancia de 9.5% - 118.7% respecto a la mejor solución. La Tabla 4 muestra los promedios del *IRMS* respecto a los 30 experimentos para cada día. El algoritmo Calib_GA se desempeñó aproximadamente cuatro veces mejor con un promedio de 9.3 comparado con las otras estrategias que resultaron en el rango de 30.3 a 69.8.

Tabla 4. Promedio del *IRMS*

Estrategia	Día 1	Día 2	Día 3	Día 4	Día 5	Prom.
MIN LB	49.8	25.6	85.4	22.8	33.7	43.4
MIN CT	54.8	30.0	24.1	33.3	9.5	30.3
MIN WT	52.5	24.2	54.6	35.3	36.9	40.7
MIN ST	77.7	25.7	52.4	28.2	11.2	39.0
MIN TA	93.5	44.9	62.1	118.7	29.9	69.8
Calib_GA	11.9	12.7	4.8	14.3	2.7	9.3

La Tabla 5 muestra los tiempos de ejecución. Las estrategias clásicas tuvieron un tiempo de ejecución de menos de 3 segundos, mientras que el algoritmo Calib_GA consumió un tiempo mayor, con un promedio de 644 segundos (10.7 minutos).

Tabla 5. Tiempo de ejecución en segundos.

Estrategia	Día 1	Día 2	Día 3	Día 4	Día 5
MIN LB	1	< 1	< 1	< 1	< 1
MIN CT	< 1	1	1	1	1
MIN WT	< 1	1	2	2	2
MIN ST	< 1	1	1	1	2
MIN TA	< 1	1	1	3	1
Calib_GA	942	276	218	495	1290

6 Conclusiones

Se realizó un estudio comparativo de un algoritmo genético y estrategias conocidas, enfocados a optimizar la ejecución de trabajos computacionales en un grid jerárquico con dos etapas, considerando una carga de centros de cómputo reales. Primeramente se calibró el algoritmo genético. La calibración fue realizada comparando seis operadores de cruzamiento y tres operadores de mutación. Así, se analizaron 18 diferentes algoritmos genéticos con 30 ejecuciones para cada uno. En total, se evaluaron 540 experimentos. Usamos el método de agregación de criterios para considerar dos objetivos: C_{max} y *TA*. Para evaluar la diferencia estadística entre los resultados experimentales y observar el impacto de diferentes parámetros sobre la calidad del resultado, se aplicó la técnica de ANOVA. De los seis operadores de cruzamiento comparados, se seleccionó el operador *OBXM*. Después, se comparó con cinco estrategias conocidas en la literatura. Los resultados muestran que nuestro algoritmo es 480% mejor, sin embargo, consumió más de 10 minutos en promedio en comparación con menos de 3 segundos de las otras estrategias. Concluimos que para los casos donde la calendarización en ambientes dinámicos es administrada por lotes, la calendarización realizada con algoritmos genéticos puede ser muy eficiente. Los resultados obtenidos pueden servir como punto de partida para algoritmos de calendarización que pueden ser implementados en un ambiente de Grid computacional real donde es posible agrupar los trabajos en lotes.

Referencias

1. Hirales-Carbajal, A., Tchernykh, A., Yahyapour, R., Röblitz, T., Ramírez-Alcaraz, J.M., González-García, J.L.: Multiple Parallel Workflow Scheduling Strategies on a Grid. Journal of Grid Computing, 10(2), 325-346 (2012).
2. Quezada-Pina, A., Tchernykh, A., González-García, J.L., Hirales-Carbajal, A., Ramírez-Alcaraz, J.M., Schwiegelshohn, U., Yahyapour, R., Miranda-López, V.: Adaptive parallel job scheduling with resource admissible allocation on two-level hierarchical grids. Future Generation Computer Systems. 28(7), 965-976 (2012).
3. Kurowski, K., Nabrzyski, J., Oleksiak, A. and Węglarz, J.: A multicriteria approach to two-level hierarchy scheduling in Grids. Journal of Scheduling, 11(5), 371-379 (2008).

4. Graham, R.L., Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G.: Optimization and approximation in deterministic sequencing and scheduling: a survey. In: Hammer, P.L., Johnson, E.L., Korte, B.H. (eds.) Annals of Discrete Mathematics 5. Discrete Optimization II, North-Holland, 287–326 (1979).
5. Lifka, D. A.: The ANL/IBM SP Scheduling System. In IPPS '95: Proceedings of the Workshop on Job Scheduling Strategies for Parallel Processing, London, UK: Springer-Verlag, 295-303 (1995).
6. Tchernykh A., Schwiegelsohn U., Yahyapour R., Kuzjurin N.: Online Hierarchical Job Scheduling on Grids with Admissible Allocation. Journal of Scheduling, Springer-Verlag, 13(5), 545-552 (2010).
7. Dutot, P., Eyraud, L., Mounie, G., Trystram, D.: Models for scheduling on large scale platforms: which policy for which application? 18th International Symposium on Parallel and Distributed Processing, pp.172 (2004).
8. Siddiqui, M., Villazon, A., Fahringer, T.: Grid Capacity Planning with Negotiation-based Advance Reservation for Optimized QoS. In: Proceedings of the 2006 ACM/IEEE conference on Supercomputing (SC '06). ACM (2006).
9. Ramírez-Alcaraz J.M, Tchernykh A., Yahyapour R., Schwiegelshohn U, Quezada-Pina A., González-García, J.-L., Hirales-Carbajal A.: Job Allocation Strategies with User Run Time Estimates for Online Scheduling in Hierarchical Grids. J. Grid Computing, Springer, 9, 95-116 (2011).
10. Lorpunmanee, S., Noor, M., Hanan, A., Srinoy, S.: Genetic algorithm in Grid scheduling with multiple objectives. Proceedings of the 5th WSEAS int. Conf. on Artificial Intelligence. Knowledge Engineering and Data Bases. Madrid, Spain, 429-435 (2006).
11. Yaurima-Basaldúa, V. H, Tchernykh, A., Castro-García, Y., Villagómez-Ramos, V.M., Burtseva, L.: Genetic algorithm calibration for two objective scheduling parallel jobs on hierarchical Grids. In Parallel Processing and Applied Mathematics, Wyrzykowski et al. (Eds.). LNCS, Springer-Verlag, 7204, 61-70 (2012).
12. Gen, M., Cheng, R.: Genetic algorithms & engineering optimization. John Wiley & Sons, New York, 512 pp. (1997).
13. Bierwirth, C., Mattfeld, D. and Kopfer, H.: On permutation representations for scheduling problems. En: Voigt, H.-M., W. Ebeling, I. Rechenberg, y H.-P. Schwefel, editores, Parallel Problem Solving from Nature - PPSN IV, Berlin, Germany, LNCS. Springer, 1141, 310-318 (1996).
14. Guinet, A., Solomon, M.: Scheduling Hybrid Flowshops to Minimize Maximum Tardiness or Maximum Completion Time, Int. J. Production Research, 34(6), 1643-1654 (1996).
15. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolutions Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996).
16. PWA. Parallel Workloads Archive, http://www.cs.huji.ac.il/labs/parallel/workload/
17. GWA. Grid Workloads Archive, http://gwa.ewi.tudelft.nl/pmwiki/
18. Ruiz, R., Maroto, C.: A genetic algorithm for hybrid flowshops with sequence dependent setup times and machine eligibility. European Journal of Operational Research, 169, 781-800 (2006).

Communications & Computer Networks

Design of Pulse Oximeter with WiFi Connectivity

Victor Ochoa, Fernando Vazquez, Martha Jimenez

Computer Academy, Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas, Instituto Politécnico Nacional, Av. Te #950, Col. Granjas México, C.P. 08400, Del. Iztacalco, Distrito Federal, México
{vochoag,fvazquezt,majimenez}@ipn.mx

Abstract. Given that health is so relevant for global productivity and competitiveness, and that the TICs play an important role in all of the productivity factors, this work makes use of the TICs in health matters with a proposal to use a WiFi oximeter. This article will describe the operating principles of a Pulse Oximeter (PO), which is an opto-electronic non-invasive medical instrument capable of measuring changes in HR and SpO2 at the fingertip; its design, and its validation against the three existing devices. Variables (SpO2%, Ppm, Temperature) were compared, and the WiFi oximeter had a greater impact because of its performance in responses; in addition, the WiFi oximeter incorporates the Temperature variable and the WiFi connectivity, for which we can conclude that implementing this device, will contribute to global competitiveness.

Keywords. Pulse oximetry, optical sensor, heart rate level, wifi protocol.

1 Introduction

Health is very important to be a competitive person in a world like ours, so this paper is related about a biodevice (Oximeter WiFi) for health. Pohjola, Venturini (2009) say that ICT investments have a positive impact on economic growth of GDP, also Edwards (2001) said that areas need large investments in ICT research, development, education, infrastructure and health to generate economic growth. Baily, Katz and West (2011) suggest the investment of ICT innovation as a key factor in the economy, so we can conclude that the oximeter is a technology that can help economic growth regardless of health satisfier give the population [2, 3, 4, 5].

The pulse oximeter has become a vital NICU instrument [15], [16] and may have been adopted as a standard [17]. Various studies have concluded that with better technology, pulse oximeters now provide highly accurate measurements of oxygenation[17-18]. Bierman demonstrated that with other factors being equal, pulse oximetry significantly reduced the need for arterial blood gas collection [17], [19-20]. Zengel examined the effects of subareolar isosulfan blue injection on pulse oximeter (SpO_2) readings and concluded that Time to peak SpO_2 fall, and the recovery period, are delayed in the subareolar technique [21].

On the another hand Rodríguez, Garrido, Martínez, & García (2013) presented a paper was related with the accuracy of pulse oximeters. It included a brief introduc-

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 101–111
Paper Recived 21-06-2013 and Acepted 21-10-2013

tion to the pulse oximetry operation principles, calibration procedure, and discusses main aspects related with the accuracy of measurements and said that the magnitudes of the errors due variations of the wavelengths of the LEDs used were highlighted together with the risks that those errors produced to the patients [1]. And we take this very serious for the design process.

Hülsbusch, et. al (2010) studied about cardiovascular diseases of irregularities in the human cardiovascular system was developed, which was based on a micro-optic in-ear sensor. The resulting signal was then transferred wirelessly to a personal digital assistant (PDA) smart phone or PC where the heart beat, oxygen saturation (SpO_2), breathing frequency and slower perfusion rhythms could be calculated. This contribution introduced the system concept of the monitoring [13].

Besides, the oxygen is vital to the functioning of each cell in the human body. Without oxygen for a prolonged amount of time, cells will die. Thus, oxygen delivery to cells is an important indicator of a patient health. Several methods have been developed to analyze oxygen delivery. Pulse oximetry is a common, noninvasive method used in clinical environments [6].

Blood red cells contain a protein called hemoglobin. Red cells with oxygenated hemoglobin circulate in the blood through the whole body, irrigating tissues. When blood gets in contact with a cell, the red cells hemoglobin releases oxygen and becomes Deoxyhemoglobin (Hb) (deoxygenated hemoglobin) [7].

More over pulse oximetry is the non-invasive measurement of the oxygen saturation (SpO_2). And pulse oximetry systems are based on two principles related to the characteristic of blood flow rate in the context of the oxy-hemoglobin and deoxy-hemoglobin status. Both oxy-hemoglobin and deoxy-hemoglobin are different in their absorption of red (660 nm to 750 nm) and infrared light (850nm-1000nm), and because the volume of the arterial blood in tissue changes as the pulse changes. With each heartbeat, the volume of the arteries becomes larger before the blood enters the capillaries. This change makes possible for the oximetry system to differentiate the arterial blood from all other absorbing substances [8, 9].

When light is emitted into the body tissue, some light will be absorbed by the skin, bones and muscle tissue. This represents the static direct current (DC) component of the signal received at the photo detector receiver. The pulsatile flow in arteries and arterioles during diastole and systole will create some variation in light intensity. This will produce the alternating current (AC) part of the signal [9]. At this point the absorption that occurs is known as the Beer-Lambert Law. Both AC and DC components are shown in figure 1. [10].

As mentioned above, we agree with Shafique, Kyriacou, & Pal (2012), who investigated about Photoplethysmography (PPG), which is a technique widely used to monitor volumetric blood changes induced by cardiac pulsations and pulse oximetry uses the technique of PPG to estimate arterial oxygen saturation values (SpO_2) [14].

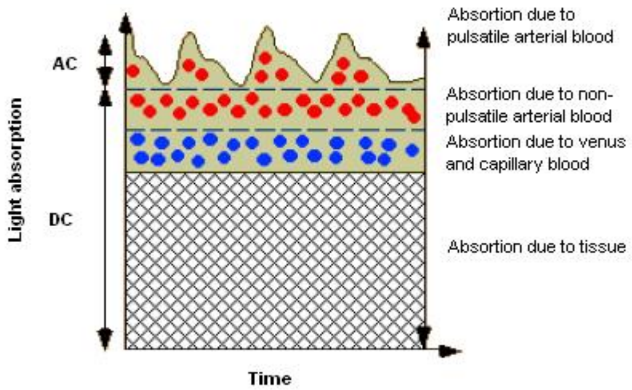


Fig. 1. Light absorption through living tissue.

2 Descriptive and methodological section

This research was conducted as an experiment in Mexico, at the "Hospital de Ortopedia y Traumatología - Dr. Victorio De La Fuente Narváez" in Mexico City, in the year 2013, with a sample of 32 patients. This experiment consisted of taking samples of variables 1) SpO_2%, 2) Ppm and 3) The temperature, which are shown in Figure 8 and are described in Table 1.

Table 1. Study variables.

Variable	Description
SpO_2%	Oxygen level (0 to 100%)
Ppm	Pulses per minute (0 to 200ppm)
Temperature	Body Temperature (0 to 100ºC)

The research instrument consisted of a form where measures of values resulting from the three variables SpO_2%, ppm, temperature were written down. This instrument was applied to 32 patients. Three treatments were also used, i.e., oximeters: Nonin, Mazimo, and oximeter WiFi.

An intensive care monitor alarm has been a major burden on both nurses and patients. Between 44 and 63 % of alarms are caused by pulse oximeters, with 94 % of these being non-significant [22-24]. Any technique for measuring pulse oximeter saturation (SpO_2) has been developed this technique uses mathematical manipulation of the pulse oximeter is red and infrared light absorbance to identify and subtract the noise components associated with these signals [25]. Theoretically the pulse oximeter analyzes the light absorption of two wavelengths from the pulsatile-added volume of oxygenated arterial blood ($AC_{red\ light}/DC_{infrared\ light}$) and calculates the absorption ratio "R" using the following equation 1.

$$R = \frac{AC_{660}/DC_{660}}{AC_{940}/DC_{940}} \tag{1}$$

SpO_2 is taken out from a table stored on the memory calculated with empirical formulas. A ratio of 1 represents a SpO_2 of 85%, a ratio of 0.4 represents SpO_2 of 100 %, and a ratio of 3.4 represents SpO_2 of 0 %. For more reliability, the table must be based on experimental measurements of healthy patients.

Another way to calculate SpO_2 is taking the AC component only of the signal and determines its ratio by using the following equation 2. SpO_2 is the value of "R" X 100.

$$R = \frac{Log_{10}/(I_{AC_{660}})}{Log_{10}/(I_{AC_{940}})} X100 \tag{2}$$

Where:

I_{ac} = Light intensity at 1 (660nm) or 2 (940nm), where only the AC level is present.

R= Absorption ratio of light.

The system will consist of five parts; sensor, amplifier, processing, LCD display and WiFi communication protocol, as shown in figure 2.

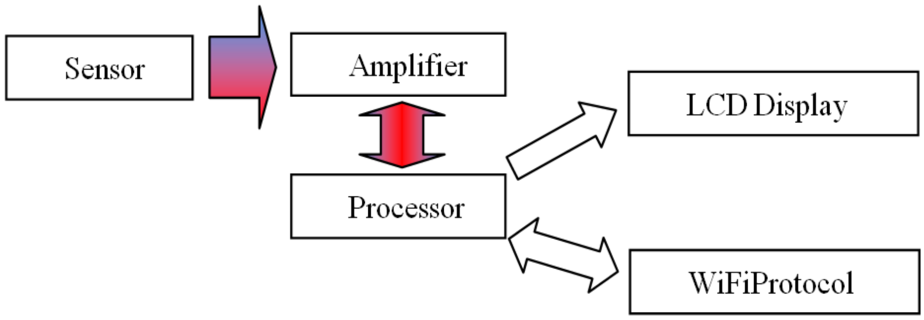


Fig. 2. Block diagram showing the flow of operation for the Pulse Oxymetry WiFi System.

2.1 Sensor of pulse oximetry

The designed Pulse Oximeter WiFi system will consist of a probe (sensor); the sensing probe consists of two LEDs, and a photo-detector. The two LEDs used in the sensor part are the red and infrared (refer to figure 3). The signal collected from the photo-detector. To perform our tests, we used the finger. The detectors must be highly sensitive and be able to measure the weak emission through to the tissues.

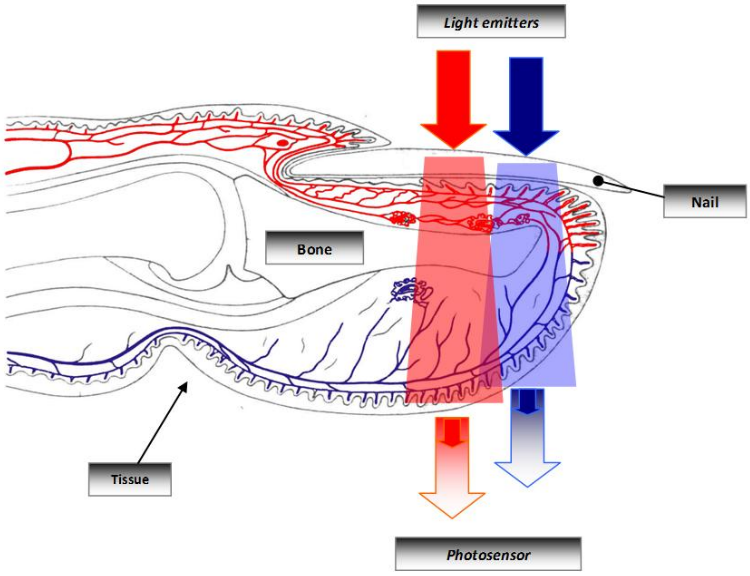


Fig. 3. Sensor orientation for light transmittance in the designed pulse oximeter.

2.2 Acquiring the signal

The optical receiver element is the photodiode. For the purpose of acquiring the signal, we have to amplify and filter, the signal that amplification from the photocurrent to be transformed into moderate voltage output impedance is then taken to a bandpass filter section designed to operate at frequencies of 0.15 Hz to 7.5 Hz mainly intended to eliminate DC component and high frequency noise. [12] See Figure 4.

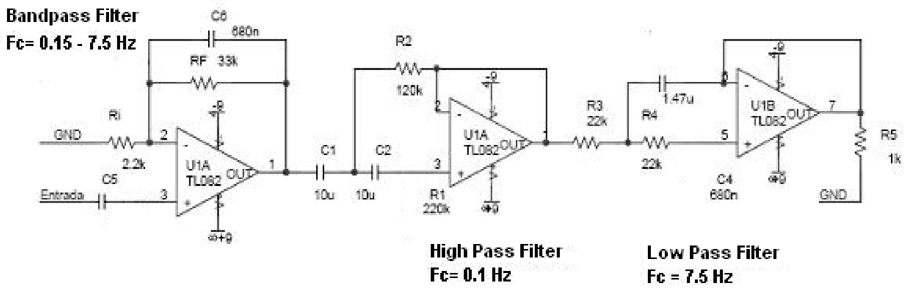


Fig. 4. Filtering and Amplifying Circuits.

2.3 Processing pulse oximetry signal

The output signal from the sensors (filtered and amplified) will be further supplied to a Programmable Interface Controller which will be converted from analogue signal into digital signal through the built-in Analogue to Digital Converter of 12bit. However, this converting process will require a C programming software and C18 compilation process to generate the Hexadecimal “.hex” file. An example of line code in C18 to calculate de SPO_2%, applying equation 2:

$$\text{calculoSpo2} = (((\log((1 / \text{red}))) / (\log((1 / \text{infrared}))))*100) \tag{3}$$

2.4 Display of pulse oximetry

For the device to be user friendly, the measured values are shown; the output produced by the PO will be displayed via a Liquid Crystal Display (LCD) screen. The organization chacters are as shown in figure 5.

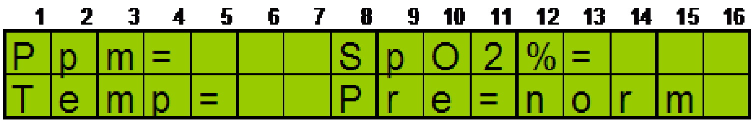


Fig. 5. Display test data processed as the Table 1.

2.5 The WiFi protocol and communication

To implement the WiFi protocol, we use the Microchip TCP/IP Stack because it is a suite of programs that provides services to standard TCP/IP-based applications (HTTP Server, Mail Client, etc.). The software stack has an integrated driver that implements the API that is used in the modules for command and control, and for management and data packet traffic [11].

When the device has the final results (SpO_2% and the HR), we use a micro embedded WiFi card to communicate the microcontroller with the most nearest access point (AP) to enable the Pulse Oximeter WiFi system to be reachable for other devices like laptops, computers, Smartphones connected to the same AP. See figure 6.

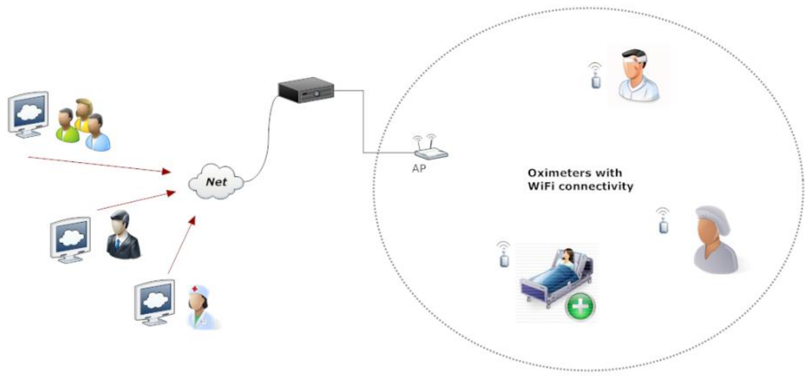


Fig. 6. Oximeters with WiFi connectivity over the network.

2.6 Implementation

Since, the pediatricians used a Nellcor pulse oximeter to measure SpO_2. The prolonged measurements took into account the pulsatile nature of arterial flow, as well as the highest SpO_2 value recorded and its stability, their results below < 95% were verified by a second measurement, and only the highest value was recorded [26].

As a result of, the first implementation was made at the “Hospital de Ortopedia y Traumatología - Dr. Victorio De La Fuente Narváez” in Mexico City, with optimum results at the moment the Pulse Oximeter was tested [8]. All tests were reviewed by specialist doctors under the ISO 9919:2005, in which ISO defines the procedure to prove the Oximeters. [14] figure 7 shows the tests.

Fig. 7. Oximeters test; Mazimo, Nonin and Oximeter WiFi in a patient.

3 Results and discussion

The results obtained from the three treatments (Nonin, Mazimo and Oximeter WiFi), are quite similar as to the minimum, maximum and average measures. Regarding temperature, only the Oximeter WIFI treatment shows the temperature by using descriptive Statistics in table 2.

Table 2. Results from the Three Treatments.

	Nonin			**Mazimo**			**Oximeter WiFi**		
	SpO_2%	PPM	Tem.	SpO_2%	PPM	Temp.	SpO_2%	PPM	Temperature
Min.	87.00	70.00		86.00	69.00		86.00	70.00	29.00
Max.	99.00	94.00		100.00	93.00		100.00	94.00	32.00
Average	94.16	82.84		93.72	82.78		93.78	83.00	30.38

In the figure 8 shows visually the values obtained for the SpO_2% variable, which gives reliability on the results by using the WiFi proposed treatment.

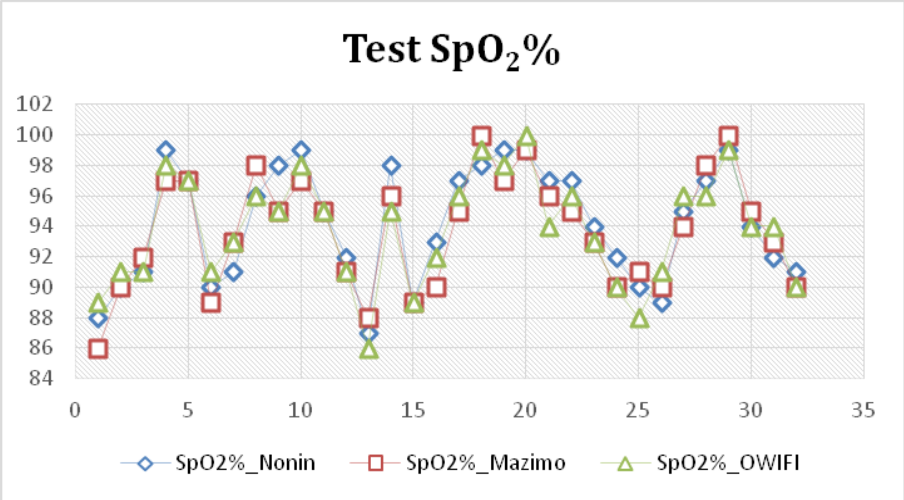


Fig. 8. Test Results SpO_2%.

Discussion: Many authors have worked with pulse oximeter. However in connection with the figure 8, the treatments are quite similar and the proposal (Oximeter WiFi) has good results. We show the same result with a plus (Temperature). So we think that it is important to help the health, and we agree with authors Pohjola (2000), Venturini (2009), in that the treatment will be a call for health, which will promote an economic growth [4].

4 Conclusions at the moment

The preliminary conclusions are: First, from the electronic point of view, the necessary research and tests were carried out to join the project and bring it to a first phase of its construction. Second, notwithstanding that results were successful in its implementation, doctors made it clear that further testing in a more specialized area was needed. And of course, there will be a more advanced version of this oximeter. This was because doctors asked if the oximeter could autosave the results into a system or database and the answer was yes, this gives the opportunity to continue the development of a second phase of this oximeter. The following features are emphasized: scalable technology, on-line monitoring, provides connectivity and networking, will provide more timely and easy monitoring, use of Standard ISO 9919:2005.

Acknowledgements

The authors would like to take this opportunity to express his heartfelt appreciation and thanks to the Instituto Politecnico Nacional and the UPIICSA, IMSS especially to the Hospital de Ortopedia y Traumatología - Dr. Victorio De La Fuente Narváez for their support, which made it possible for the authors to produce this paper.

References

1. Rodríguez, J., Garrido, H., Martínez, A., & García, M.: Exactitud y Errores de la Oximetría de Pulso. IFMBE Proceedings (33), 758–761, (2013).
2. Baily, M., Katz, B., & West, D.: Building a Long-Term Strategy for Growth through Innovation. Brookings, (2012).
3. Edwards, S.: Information Technology and Economic Growth in the Emerging Economies. 1-22, (2001).
4. Pohjola, M.: Information technology and economic growth: a cross-country analysis. UNU Word Institute for Development Economics Research, Working Papers No. 173, pp 1, (2000).
5. Venturini, F.: The long-run impact of ICT. Empirical Economics, 37(3), 497–515, (2009).
6. John G.: Design of Pulse Oximeters. En: John G II, USA: Taylor & Francis; pages. 12-25, (1997).
7. Santiago L.: Pulse Oximeter Fundamentals and Design, Freescale Semiconductor Document Number: AN4327, (2012).
8. Hoff, D. Zhang, R. Stalter, T. and Carlson, M.: Pulse Oximetry. Undergraduate Thesis Electrical and Computer Engineering, North Carolina State University, USA, (2003).
9. Oximetery.org: Principles of Pulse Oximetry Technology. Retrieved on: 16th of June 2013 URL: (http://www.oximeter.org/pulseox/principles.htm).
10. The analysis of transesophageal oxygen saturation photoplethysmography from different signal sources. Retrieved on: 16th of June 2013 URL: (http://www.ncbi.nlm.nih.gov/pubmed/23475176).
11. Nilesh R: The Microchip TCP/IP Stack, Microchip Document Number: AN833, (2002).
12. J. T. B. Moyle, Pulse Oximetry, 1st ed. London, U.K.: BMJ, (1994).
13. Hülsbusch, M., Blazek, V., Herzog, M., Vogel, S., Wartzek, T., Starke, D., &Hennig, T. (2010). Development of a miniaturized in-ear pulse oximeter for long term monitoring of risk patients. En O. Dössel, & W. C. Schlegel, World Congress on Medical Physics and Biomedical Engineering (Vol. 25, pags. 779-781). springer. doi:10.1007/978-3-642-03882-2_207.
14. ISO 9919:2005 Medical electrical equipment - Particular requirements for the basic safety and essential performance of pulse oximeter equipment for medical use. URL: (http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=33780). Retrieved on: 16th of June 2013.
15. Cresi, F., Pelle, E., Calabrese, R., Costa, L., Farinasso, D., &Silvestro, L.: Perfusion index variations in clinically and hemodynamically stable preterm newborns in the first week of life. Italian journal of pediatrics, 36(1), 1-5, (2010).
16. Salyer, J. W.: Neonatal and pediatric pulse oximetry. 48(4):386-396, Respir Care 2003.

17. Shah, A., & Shelley, K. H.: Is pulse oximetry an essential tool or just another distraction? The role of the pulse oximeter in modern anesthesia care. Journal of clinical monitoring and computing, 1-8, (2013).
18. Anesth, J.: Pulse oximetry: its invention, theory, and future. 17(4):259-66, (2003).
19. Pedersen T, Moller A, Pedersen B. Pulse oximetry for perioperative monitoring: systemic review of randomized, controlled trials. Anesth Analg. 96(2):426, (2003).
20. Bierman, M, Stein, K, Snyder J. Pulse oximetry in the postoperative care of cardiac surgical patients. A randomized controlledtrial. Chest. 102(5):1367–70, (1992).
21. Zengel, B., Yararbas, U., Bingolballi, O., &Denecli, A. G.: The Effect of Subareolar Isosulfan Blue Injection on Pulse Oximeter Readings. Indian Journal of Surgery, 1-5, (2012).
22. Sabar, R., Zmora, E.: Nurses' response to alarms from monitoring systems in NICU. Pediatr Res 41: 174A, (1997).
23. Lawless, S. T.: Crying wolf: false alarms in a pediatric intensive care unit. Crit Care Med 22:981-985, (1994).
24. Bohnhorst, B., & Poets, C. F.: Major reduction in alarm frequency with a new pulse oximeter. Intensivecare medicine, 24(3), 277-278, (1998).
25. Dumas, C., Wahr, J. A., Tremper, K. K.: Clinical evaluation of a prototype motion artifact resistant pulse oximeter in the recovery room. Anesth Analg 83:269, (1996).
26. Corrard, F., de La Rocque, F., Martin, E., Wollner, C., Elbez, A., Koskas, M., & Cohen, R.: Food intake during the previous 24 h as a percentage of usual in take: a marker of hypoxia in infants with bronchiolitis: an observational, prospective, multicenter study. BMC pediatrics, (2013).

“Glover Slider”: Sistema Híbrido de Edición y Manipulación de Diapositivas

Yaxkin Flores Mendoza, Axel E. Moreno Cervantes, Gustavo A. Sanvicente Solís

Escuela Superior de Cómputo-IPN, Av. Juan de Dios Bátiz s/n esquina Miguel Othón de Mendizabal. Unidad Profesional Adolfo López Mateos. Col. Lindavista C.P. 07738, México, D.F.

floresya@hotmail.com, axelernesto@gmail.com,
sasg09@hotmail.com

Resumen. Hoy en día la mayoría de las personas requieren en sus actividades diarias realizar presentaciones de diversos temas, por ejemplo, un científico que presenta sus descubrimientos en un congreso o un profesor con sus lecciones diarias. Solo en México el 39.75% de la población usa para ello la computadora. “Glover Slider” es un sistema híbrido compuesto por una aplicación para desarrollar y editar presentaciones y por un guante electrónico e inalámbrico que controla dichas presentaciones solo durante su ponencia. Este prototipo busca aumentar la fluidez y dinamismo durante el proceso de creación y presentación de diapositivas.

Palabras clave: Bluetooth, accelerometer, hybrid system, motion detection, data acquisition.

1 Introducción

En México actualmente hay un total de 112, 322,757 habitantes de los cuales el 39.75% son económicamente activos [1], muchas de las ocupaciones que los mexicanos desempeñan implican hablar en público, es decir que varios de ellos tienen o han tenido que fungir como expositores en algún momento de su vida, como por ejemplo:

- Un profesor presentando un tema de importancia dentro de su asignatura.
- Alumnos exponiendo una investigación.
- Un ejecutivo mostrando planes de acción.
- Congresistas.

Haciendo referencia a lo anterior, para los expositores es importante retener la atención de su audiencia, es por ello que como se menciona en el libro “Claves para hablar en público” del autor Max Atkinson [2] la mayoría de los expositores tienen problemas con las pausas que hacen al hablar provocadas muchas veces por dificulta-

© E. A. Santos, J. C. Chimal, L. Cabrera, E. Castillo.
Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 111–120
Paper Recived 15-07-2013 and Acepted 21-10-2013

des con la manipulación de las presentaciones en las que se apoyan, por lo que recurren principalmente a apuntadores láser para señalar la información que desean resaltar, o bien, a la ayuda de un asistente encargado del cambio de diapositivas, dichas acciones resultan ser distractoras para el espectador ya que según el autor "…Las pausas pueden alterar el impacto de las frases…".

Para atenuar la problemática antes planteada se propone desarrollar un sistema híbrido que permita el desarrollo de presentaciones electrónicas mediante un módulo de edición y otro de presentación, el cual estará complementado con el uso de un dispositivo inalámbrico que le permitirá al usuario manipular el flujo de la presentación, realizar acercamientos, alejamientos, subrayar y apuntar texto por medio de comandos previamente establecidos, además de que dada la simplicidad de los dichos comandos, es decir, a la naturaleza con la que realizamos estos movimientos con nuestras manos, el guante permitirá bloquear y desbloquear con un push-button el sensado de movimientos evitando así errores.

2 Estado del arte

Como se puede apreciar en la Tabla 1, a pesar de que ya existen en el mercado dispositivos similares al que aquí se presenta no están enfocados para la edición y presentación de aplicaciones, por lo que no cuentan con comandos para la manipulación de las mismas.

Tabla 1. Estado del arte con dispositivos similares.

Sistemas / Características	PlayStation Move [3]	HP Swing	Microsoft Kinect	Presentador Laser Perfect Choice	WiiMote	TT N°. 2009-0146 "Free Mouse"	Propuesta de Solución
Seguimiento de Posición	Limitado	Si	Limitado	No			
Seguimiento de Rotación							
Capacidad de Puntero	Si	No	Si	No	Si	Si	Si
Botones	Si (9 Botones)	Si (5 Botones)	No	Si (9 Botones)	Si (11 Botones)	Si (1 flexómetro)	Si (4 flexómetros)
Proporciona información							
Cámara	Si	No	Si	No	No	No	No
Captura de imagen	Si	No	Si	No	No	No	No
Reconocimiento de voz	Si	No	Si	No	No	No	No
Wireless	Si	Si	Si	Si	Si	Si	Si
Manipulación de Presentaciones	No	No	No	Si	No	No	Si
Acercamiento de Imágenes	No	No	No	No	No	No	Si
Subrayar sobre presentaciones	No	No	No	No	No	No	Si
Documentación útil para el proyecto	No	No	Limitada	No	Limitada	Si	No Aplica
Cronómetro	No	No	No	Si	No	No	
Precio	$1,900.00	$29,990.00	$2,299.00	$500.00	$340.00	No Aplica	$1,311.00

Nota: El precio de la propuesta de solución solo contempla los componentes electrónicos correspondientes al guante. Se debe considerar aparte la batería de 9 volts como fuente de alimentación.

3 Metodología usada

Para el desarrollo del Sistema Híbrido "Glover Slider" se utilizó la metodología OpenUP [9] que es un modelo de desarrollo de software y parte del Framework de modelo de proceso de Eclipse. Se decidió ocupar esta metodología por los beneficios que ofrece al tipo de proyecto planteado, de los cuales los más importantes se muestran a continuación:

- Es apropiado para proyectos pequeños y de bajos recursos incrementando las posibilidades de éxito.
- Permite detectar errores tempranos a través de un ciclo iterativo.
- Evita la elaboración de documentación, diagramas e iteraciones innecesarios requeridos en la metodología RUP.
- Tiene un enfoque centrado al cliente y con iteraciones cortas debido a que es una metodología ágil.

4 Desarrollo

Se realizó un sistema híbrido conformado por un guante que permite a su usuario manipular diapositivas con el simple movimiento de los dedos de la mano derecha y de una aplicación capaz de crear y editar presentaciones. El sistema está conformado por dos módulos: el modulo del periférico y el modulo de la aplicación, (ver Figura 1).

El desarrollo de la aplicación del sistema híbrido "Glover Slider" se dividió en dos sub-módulos los cuales se muestran a continuación:

- Edición
- Presentación

Dentro del módulo de edición se tiene la opción de personalizar cada diapositiva mediante la inserción de tablas, imágenes, sonido, vídeo y texto así como animaciones y transiciones. El módulo de presentación es el encargado de mostrar el trabajo previamente creado así como las transiciones y animaciones de cada componente, además de ser este el único modo que permitirá el uso del dispositivo inalámbrico para la manipulación de dichas diapositivas. Los requerimientos mínimos necesarios para el correcto funcionamiento de la aplicación se muestran en la Tabla 2.

Para el desarrollo del periférico fue necesario elaborar un circuito que permitiera la comunicación con la computadora mediante Bluetooth® y que a la vez realizara el sensado tanto de las resistencias variables (Flex Bend Sensor) como del acelerómetro que lo componen, para ello se determinó el uso de ciertos componentes electrónicos que se pueden observar en la Tabla 2.

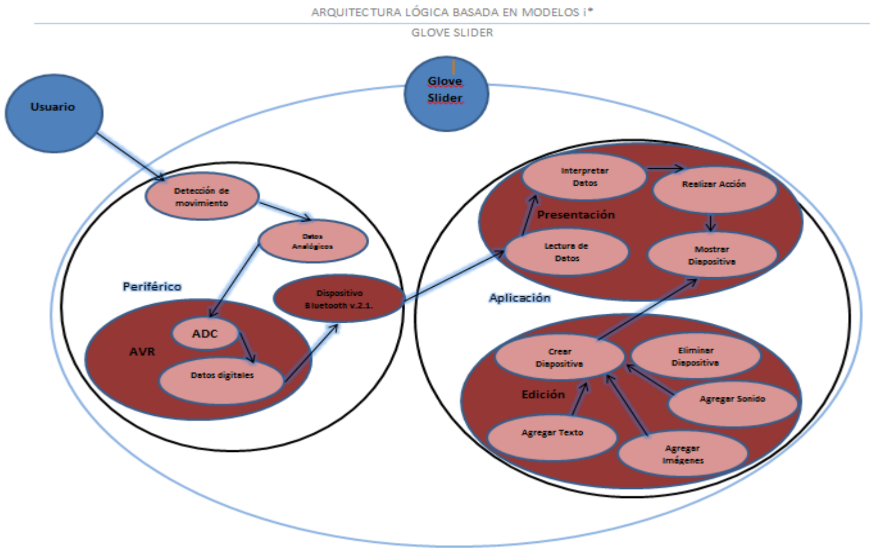


Fig. 1. Arquitectura Lógica basada en modelos i* [10].

Tabla 2. Componentes y requerimientos del sistema.

PERIFÉRICO	APLICACIÓN
Bluetooth® RN-42	Bluetooth® 2.0/2.1 Wireless Technology
Sensor de flexión FSL-0055-253-ST	Adaptador Bluetooth® 2.0/2.1 para PC
Acelerómetro MMA7341L	Procesador Intel Pentium IV ó Celeron D @ 2.0 GHz
AVR ATMEGA 8	2GB en RAM DDR2
Resistencias	Disco duro de 5400 RPM con 30GB disponibles
Regulador de 5 volts	Monitor capaz de soportar una resolución de 1024×768 o mayor.
Push-button	Mouse
Zócalos de 28 pines	Teclado
Conector para pila de nueve volts	Tarjeta de video compatible con DirectX 9 capaz de soportar una resolución de 1024 o mayor, compatible con pixel shader 3.0 y al menos 128MB de memoria en gráficos.
Placa fenólica de 5×10.	Sistema Operativo Multiusuario Windows XP Service Pack 3
Fuente, pilas, eliminador	Sistema Operativo Multiusuario Windows 7 de 32 y 64 bits (recomendado)

Programador de AVR USB compatible con ATMEGA 8	Microsoft Visual Studio Ultimate 2010
Headers (nueve para el acelerómetro y seis para el Bluetooth)	Reproductor de Windows Media
Sistema Operativo Multiusuario Windows XP Service Pack 3	Microsoft Expression Blend 4
Windows 7 Basic de 64 bits	Microsoft Office 2010
Microsoft Visual C# 2010 Express.	Visual Paradigm Suite 4.1
Hyperterminal Private Edition 7.0	
Virtual Serial Port Driver 7.1(Build 7.1.289) de Eltima Software	
Proteus 7.7 Professional	
BASCOM AVR 2.0.7.1	

5 Resultados experimentales

Para obtener los siguientes resultados se realizaron pruebas en un laboratorio simulando el caso idóneo así como en un caso práctico durante una ponencia. Para la manipulación del guante se ocupan los comandos descritos en la Tabla 3.

Tabla 3. Comandos de Manipulación.

Comando	Acción	Comando	Acción
	Trazo		Eliminar Trazo
	Siguiente Diapositiva		Zoom In
	Anterior Diapositiva		Zoom Out

Las pruebas de laboratorio consistieron en la realización de una presentación que contenía: animaciones, transiciones, tablas, vídeo e imágenes, posterior a esto se conectó el guante bajo circunstancias ideales evitando interferencias de otros dispositivos. Una vez realizada la conexión entre el dispositivo y la aplicación se procedió a realizar cada una de las acciones previamente descritas (ver tabla2).

En cuanto a las pruebas prácticas se realizó una presentación haciendo uso del sistema “Glover Slider” durante la clase de “Liderazgo” a cargo del profesor Jorge Ferrer Tenorio, dicha prueba tenía como objetivo probar si el uso del guante y la aplicación cumplen con los objetivos del proyecto. Los resultados obtenidos durante las pruebas antes mencionadas se muestran a continuación.

4.1 Caso: zoom in

Laboratorio	Clase de Liderazgo
	

Fig. 2. Pruebas Zoom In.

En la Figura 2, del lado izquierdo se puede observar la prueba en el laboratorio con la acción Zoom In que se realiza con la flexión del dedo medio y el pulgar de la mano derecha. Del lado derecho se puede observar la prueba realizada en la clase de liderazgo donde se puede notar que no es necesario estar enfrente de la computadora para poder ser detectada la acción en cuestión.

4.2 Caso: trazar

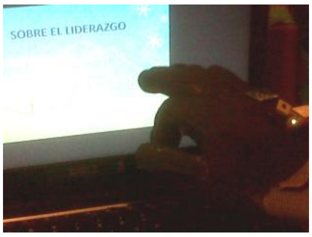
Laboratorio	Clase de Liderazgo
	

Fig. 3. Prueba trazar en la diapositiva, en laboratorio y clase de liderazgo

En la Figura 3, podemos observar tanto el caso idóneo como el caso práctico donde se probo la acción de trazar, dicha acción se realiza flexionando el dedo índice de la mano derecha en un ángulo igual o mayor a los 90°. En ambos casos podemos notar el trazo realizado y que no es necesario estar en cierta posición para poder realizarlo.

5.3 Pruebas de respuesta

Para evaluar la respuesta del dispositivo en cuanto a los comandos se realizaron diez intentos de cada acción a una distancia promedio de 10m obteniendo los resultados mostrados en la Tabla 4.

Tabla 4. Resultados de Pruebas.

COMANDO	Comandos Interpretados Correctamente	Comandos Interpretados Incorrectamente
Trazar	9	1
Borrar Trazo	10	0
Siguiente Diapositiva	10	0
Zoom in	8	2
Anterior Diapositiva	10	0
Zoom Out	8	2
Bloqueo	7	3

6 Conclusiones y trabajos futuros

"Glover Slider" es un sistema híbrido que está conformado por un módulo de hardware integrado por un guante que envía datos vía Bluetooth® a la computadora y un modulo de software el cual es una aplicación que le ofrece a los usuarios una opción más dinámica, vistosa y funcional para la realización de sus presentaciones electrónicas. Con base en lo anterior se logró desarrollar una aplicación que permitiera crear, editar y visualizar presentaciones conformadas por diapositivas con la extensión .gsp. Además de haberse logrado establecer una comunicación inalámbrica mediante un puerto serial virtual que recibe los datos del dispositivo inalámbrico encargado de sensar los movimientos realizados por la mano derecha para determinar si se está realizando algún comando válido para la manipulación de las diapositivas durante su presentación.

6.1 Trabajos futuros

Este Sistema Híbrido podrá ser extendido mediante el desarrollo de complementos para la aplicación así como de mejoras en cuanto al guante que permitan una mejor manipulación de las diapositivas. A continuación se enlistan algunas de estos puntos para el caso de la aplicación:

- Añadir compatibilidad con otras extensiones.
- Incrementar el número de herramientas destinadas a la edición y personalización
- Mayor grado de personalización en las tablas.
-Permitir la inserción de figuras geométricas regulares e irregulares.

de los elementos.

- Incrementar el número de efectos y transiciones ente diapositivas y elementos.
- Incremento en el número de temas de fondo y permitiendo la creación de estos.

Mientras que en el caso del guante se podrán incorporar las siguientes características:

- Compatibilidad al guante, es decir, que pueda manipular presentaciones de otras aplicaciones.
- Rediseño de la línea del guante (para considerar diferentes tallas, terminados, etc.).
- Incrementar el número de comandos definidos.
- Automatización en la configuración del guante.
- Permitir la conexión de varios guantes.
- Actualización de los componentes del guante.
- Diseño de guante para zurdos.

Agradecimientos

Se agradece a la Escuela Superior de Cómputo del Instituto Politécnico Nacional por las facilidades otorgadas para el desarrollo del sistema “Smartway”.

Referencias

1. H. Sossa, P. Rayón y J. Figueroa. Arquitectura basada en redes neuronales para el reconocimiento de rostros. Soluciones Avanzadas. Año 7, No. 63. (1998).
2. P. Rayón and H. Sossa. A procedure to select the vigilance threshold for the ART2 for supervised and unsupervised training. LNAI 1793, pp. 389-400, Springer Verlag, (2000).
3. PLAYSTATION®MOVE http://vive.playstation.com/?gclid=COrG08OXwKYCFQlPgwodmQahHQ#/move [Consultado: 2011-10-14]
4. HP®SWING http://www.zdnet.com/blog/gadgetreviews/hp-swing-motion-controller-launched-in-india/14131 [Consultado: 2011-10-10]
5. MICROSOFT®KINECT http://www.xbox.com/es-mx/kinect [Consultado: 2011-10-09]
6. Perfect Choice. Disponible en: http://www.perfectchoice.com.mx/PerfectChoice/Master.aspx?sWebSiteCode=Perfect, [Consultado: 2011-10-14]

7. Nintendo®. Fecha de Consulta: 10 de Agosto del 2011. Disponible en: http://www.nintendo.com/wii&IP=148.204.56.80&CAT=GGAMES&USER=DEFAULT&CE=0 [Consultado: 2011-10-09]
8. Mouse Inalámbrico con Acelerómetros "Free Mouse", Trabajo Terminal N°.2009-0146, M. en C. Axel Ernesto Moreno Cervantes, Dr. Miguel Ángel Alemán Arce, (2010).
9. Open UP. Fecha de Consulta: 17 de Septiembre del 2011. Disponible en: http://epf.eclipse.org/wikis/openup/ [Consultado: 2011-10-14]
10. Carvallo, J. P.: Descubriendo la Arquitectura de Sistemas software Híbridos: Un enfoque basado en modelos i*. España. Unidad Politécnica de Cataluya, (2009).

GateWatcher: Android Phone Based Video Surveillance System Using App Inventor

Sergio Sandoval-Reyes, Juan Carlos Castañeda-Martínez

Centro de Investigación en Computación, Instituto Politécnico Nacional, México
Sergio.Sandoval9@gmail.com; jc.castmart@gmail.com

Abstract. Video surveillance is an application that employs video cameras for the purpose of observing an area to monitor traffic, streets, facilities, and for home security. Video cameras can send video and audio via a wired or wireless network and the Internet, to watch live and recorded video from different devices, for example, cell phones. This paper describes GateWatcher an Android Phone based video surveillance system in which a video-doorbell is implemented using a cell phone and App Inventor. When a visitor rings the phone-based doorbell, you receive a sms message notification on your phone. You can accept this notification to watch the visitor returning another sms to the doorbell, which triggers its video-camera to start a Bambuser audio-video transmission to the Bambuser website via a Wi-Fi or a 3G network, which is accessed automatically from your phone to see and hear the visitor, but he/she cannot hear or see you. The application also allows communicating with the visitor through chat instant messages.

Keywords. Remote video surveillance, android, app inventor, bambuser.

1 Introduction

Video surveillance is an application that employs video cameras for the purpose of observing areas to monitor traffic, streets, facilities, and for home security. Video cameras can send audio and video via a wired or wireless network and the Internet, to watch remotely live and recorded video content from different devices. The traditional video surveillance systems like Closed Circuit Television (CCTV) or the PC based video systems require large and expensive equipment and do not provide mobility to monitor the premises remotely. Wireless networks such as Wi-Fi and 3G not only provide ubiquity but also large wireless bandwidth which make them possible to develop video applications for mobile phones such as an Android based video surveillance system.

Although there are many ways to use a smartphone to develop a surveillance system, most of them requires the knowledge of a classical computer programming language such as C, C++, Java, Python, etc. Lately however, visual programming languages such as Scratch [1], and App Inventor for Android [2], can also be used to interact with phones without the need of knowing the classical programming languages.

© E. A. Santos, J. C. Chimal, L. Cabrera, E. Castillo.
Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 121–131
Paper Recived 02-08-2013 and Acepted 21-10-2013

This paper describes GateWatcher an Android Phone based video surveillance system in which a video-doorbell is implemented using a cell phone and App Inventor. When a visitor rings the phone-based doorbell, you receive a sms message notification on your phone. You can accept this notification to watch the visitor returning another sms to the doorbell, which triggers its video-camera to start a Bambuser audio-video transmission to the Bambuser website via a Wi-Fi or 3G network, which is accessed automatically from your phone to see and hear the visitor, but he/she cannot hear or see you. The application also allows communicating with the visitor through chat instant messages.

The remainder of this paper is organized as follows: Section 2 presents a summary of works related to mobile phone based surveillance systems. Section 3 describes the design of GateWatcher using App Inventor. Section 4 shows the results of the implementation. Our conclusions are presented in Section 5. Finally, Section 6 outlines the future work.

2 Related work

Several research works on surveillance systems have been proposed. Some of the most important are described in the following.

2.1 Image based surveillance systems

In [3] a server with an embedded webcam capture images continuously of the desired location and stores them. Then a client login to the URL of the server can fetch the images from his phone using the HTTP protocol (Figure1). Although simple, the system only manages images but not video nor audio. A variant of this scheme [4] employs IP cameras, storing the images in the cloud and accessing it through a peer-to-peer network.

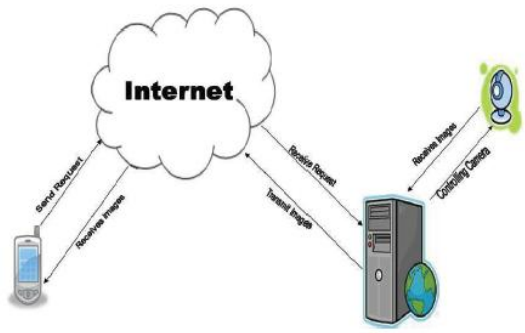


Fig. 1. An image based surveillance system.

In [5] the capture of the images is accompanied with an e-mail or sms text message alert to inform the presence of an intruder, while that in [6] the surveillance system adds the transmission of live or recorded video to the smartphone.

2.2 Video surveillance systems based on Wi-Fi and 3G wireless networks

Wi-Fi and 3G wireless networks feature faster data rates and higher stability of data links, allowing to develop video surveillance systems for mobile phones. In [7] a surveillance system is developed using a Wi-Fi network for transmitting data from each of the video data nodes to a video control server (Figure 2). This video server which has 3G Internet access, transmits the video data to a web site, from which the video data is stored in a video database and published to be accessed from 3G smartphones. A similar system [8] employs a WIMAX mobile network and SIP phones, but the system transmits live video.

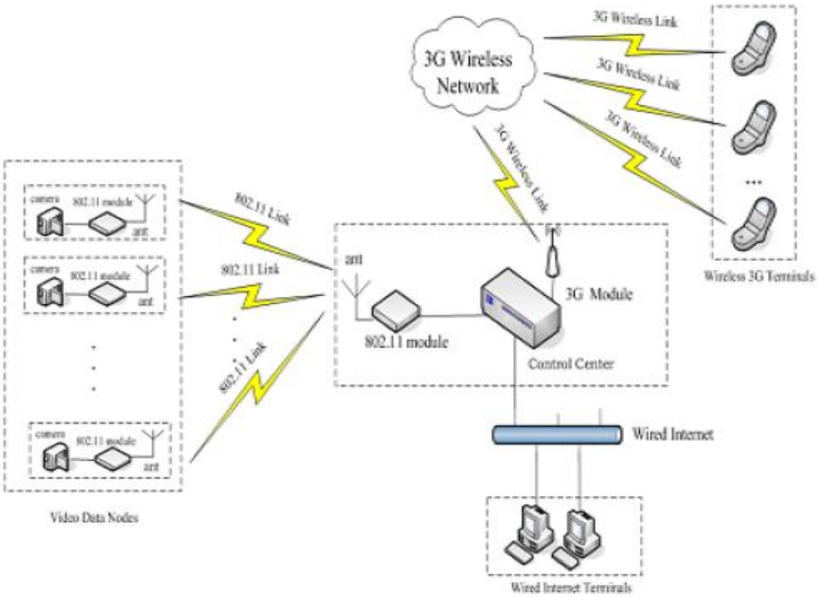


Fig. 2. A Wi-Fi and 3G based surveillance system.

2.3 Doorbell video surveillance systems

Currently there are several commercial video surveillance systems in the market focused specially to monitor visitors at the door of a home or premise. One of them is the MOBOTIX T24 [9], which is an IP door station based on the video telephony standard VoIP/SIP (Figure 3). When a visitor rings the doorbell, a connection is established with an IP video phone or a standard computer via the network to conduct a video conversation from anywhere with the visitor at the door. DoorBot a similar product [10], is a Wi-Fi based doorbell that streams live video and audio of the front door directly to a smartphone or tablet (iOs and Android).

Fig. 3. The MOTOBOX VoIP/SIP door station.

CellNock is a Wi-Fi based video doorbell [11]. When a visitor pushes the doorbell a call alert is sent to a smartphone (Figure 4). The called party can accept or reject the call. Accepting the call allows to communicate with the visitor like a video conference call. If the call is rejected, the visitor has the option to record a video message which can be retrieved at convenience.

Fig. 4. CellNock doorbell.

2.4 App inventor for android (AIA)

AIA is a visual blocks language provided originally by Google and from January 2012, by the MIT Center for Mobile Learning. AIA graphical interface is very similar to the Scratch programming language, allowing users to drag-and-drop visual objects to create applications that run on many mobile phones with the Android OS. [12]. AIA has two main windows: a component designer for building the user interface and a blocks editor for defining the application behavior. Applications can be tested directly on the phone or an emulator (Figure 5).

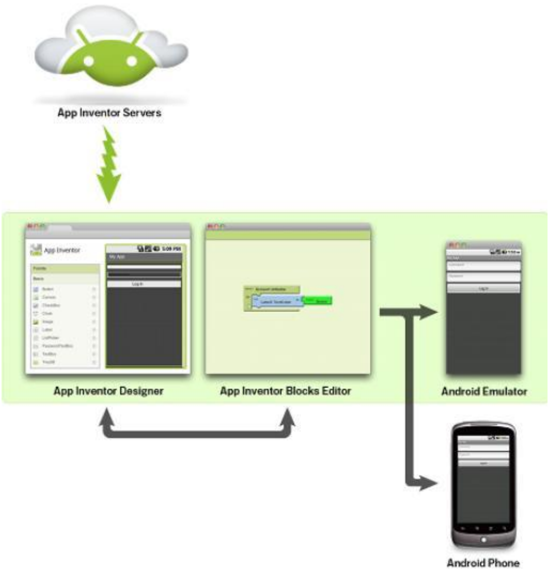


Fig. 5. AIA Component Designer, Blocks Editor, Emulator and Phone.

After reviewing the above research and commercial works and taking advantage of the video camera and communication capabilities of Android phones, in the following sections we describe the design of GateWatcher an Android Phone based video surveillance system in which a video-doorbell is implemented using a cell phone and App Inventor. When a visitor rings the phone-based doorbell, a sms message notification is sent to a second phone. The receiver can accept this notification to watch the visitor returning another sms to the doorbell, which triggers its video-camera to start a Bambuser audio-video transmission to the Bambuser web site via a Wi-Fi or a 3G network. This web site is accessed automatically from the second phone to see and hear the visitor. The application also allows communicating with the visitor through chat instant messages.

3 Gatewatcher design

The app uses touch screen buttons, microphone and speaker to emulate the doorbell, thus as the video camera and sms notification capability of an Android phone (Figure 6). The phone uses also Bambuser a free live video broadcast service app [13], for streaming live video from mobile phones or PC-webcams to Internet via a Wi-Fi, or a 3G wireless network to the Bambuser web site, which can be accessed from a second mobile phone or PC with internet connection. The next section describes the design of the video doorbell using App Inventor.

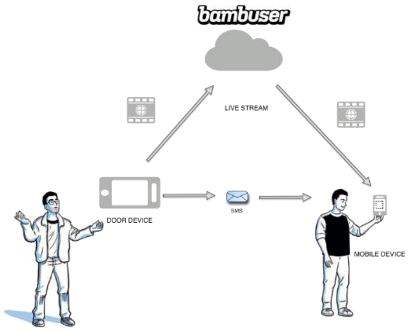


Fig. 6. GateWatcher architecture.

3.1 Building the gatewatcher GUI with the component designer, part I

The Component Designer shown on Figure 7 is the tool for designing the app interface. The left side palette has two types of components: 1) Visible like button, image, label, etc., and 2) Non-Visible like sound, texting, and ActivityStarter. We drag components from this palette to the viewer to specify the way the phone´s screen will appear when the application run, for saving data persistently, and for talking to the web. As a component is dragged into an app, its name and type appears in the list of the Components window. Components can be renamed, deleted, and another media added. When a component is selected, its properties that appear in the Properties window, can be modified. The visible and non visible components used are shown in Table 1.

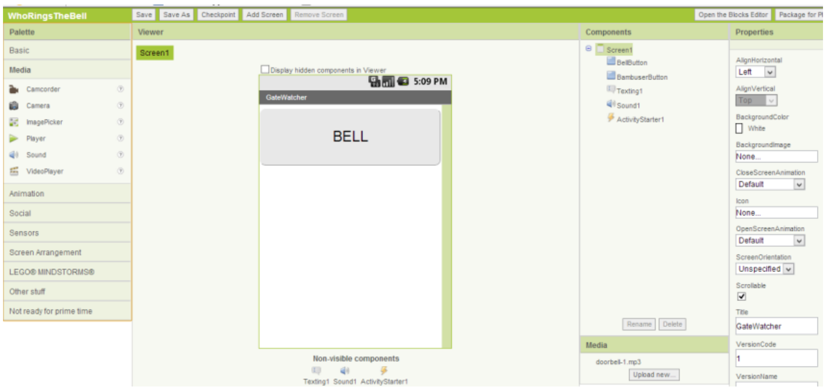


Fig. 7. AIA Component Designer for the GateWatcher application.

BellButton starts the application, produces a doorbell sound and sends a sms alert notification to the called party using the Texting1 component. The *BambuserButton*, when touched, displays a message and launches the Bambuser app using the ActivityStarter1 component (whose values were obtained using the Log Viewer app [14]).

Table 1. Visible and non visible components for the GateWatcher application.

Component type	Palette group	Component named as:	Purpose
HorizontalArrangement	Screen Arrangement	HorizontalArrangement1	A visual container
Button	Basic	BellButton	To ring the bell
Button	Basic	BambuserButton	To launch the Bambusser app
Sound	Media	BellSound	To sound the bell
Texting	Social	Texting1	To send and receive sms alerts
ActivityStarter	Other stuff	ActivityStarter1	To launch the Bambuser app

3.2 Programming the behavior of gatewatcher with the blocks editor

The behavior of the application is defined in the Blocks Editor. This editor is launched clicking the “Open The Blocks Editor” button in the Component Designer. This Editor has two palettes from which blocks are dragged, the Built-in palette and the My Blocks palette (Figure 8).

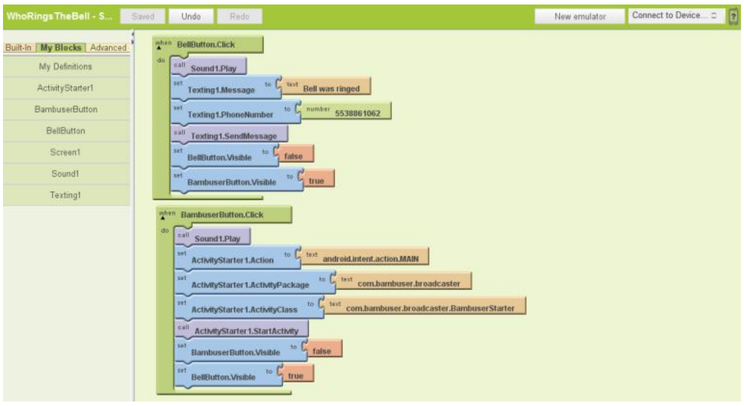


Fig. 8. The Blocks Editor with the My Blocks palette of GateWatcher.

The Built-in palette contains built-in blocks for standard programming control and functionality, for text and list manipulation, and mathematical, logical and control operators (Figure 8). The My Blocks palette contains blocks representing the components of the applications that were added in the Component Designer. In a very similar way like Scratch, the App Inventor blocks language provide visual cues to ease the development tasks, and only some blocks lock in place, reducing the possibility of errors (Figure 8). The application behavior is directly defined through a set of event-handlers (e.g., “when event *BellButton.Click* occurs, do *callTexting1.SendMessage*”). Live testing can be performed with a Wi-Fi connected Android phone (using the MIT AICompanion application), clicking the *Connect to Device* button located in the upper right side. Once the application is tested, it can be deployed by packing it into an Android apk application by clicking the upper right *Package for Phone* button in the Component Designer (Figure 7).

3.3 Building the gatewatcher GUI with the component designer, part II

The second part of the GateWatcher application corresponds to the receiver. Here we want to receive an alert in our mobile phone when somebody knocks at the door of lets say, our home, and to decide if we want to do something about it, like acknowledging the alert notification sending an sms to the sender and accessing automatically the Bambuser web site to watch and hear the visitor and optionally to interact with he/she. In order to do this we used an application reported in [15], but modifying it with the ActivityStarter component used before in section 3.2 to access the Bambuser web site. The user interface and the blocks editor of this second part is shown in figure 9.

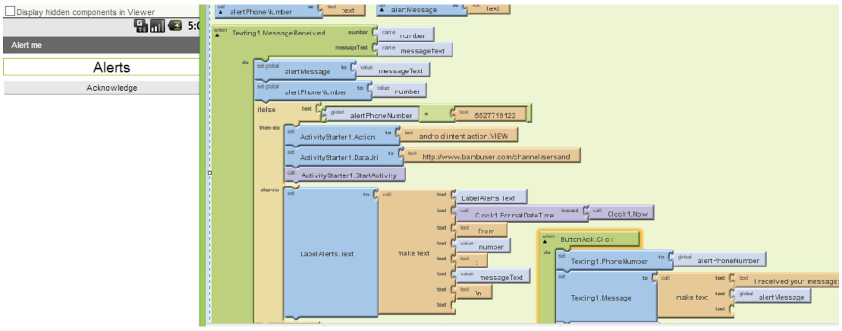


Fig. 9. GateWatcher receiver alert notification and access to the Bambuser web site.

4 Experiments and results

For the experiments we use as the video doorbell a Lanix Ilium S400 Android 4.0.2 phone with Internet connection via Wi-Fi, and the *Bambuser* live video broadcast application installed, thus as the *Log Viewer* app. This logger app was very useful to set the properties of the Activity Starter component to launch the Bambuser app. These properties were: the action name, the activity package name, the activity class name, thus as the Data URI of the Bambuser app as is shown in [16]. For the reception part, a LG P500 Android 2.2 phone with Internet connection via a 3G wireless network was used. The sms text messages used were standard. Google has a free sms service called *Google Voice*. Unfortunately, this service only works at the USA and

Canada. Figure 10 shows the results of the GateWatcher implementation. The left screen captures show the user interface with buttons to send the sms alerts and to launch the Bambuser live video streaming via Wi-Fi. The screen captures toward the right show the reception user interface, the sms alert notifications, and the access to the Bambuser web site to watch and hear the visitor at the door. The Bambuser web site allows the called party to communicate with the visitor through instant messages, but the visitor does not need to type anything, just talk (Figure 11). In this way the visitor cannot hear or see the recipient party.

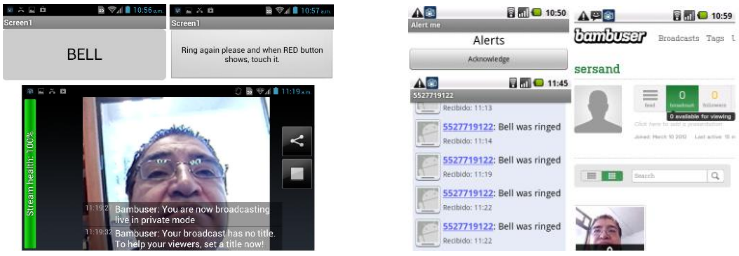


Fig. 10. Left: GateWatcher live video transmission; Right: Alerts & live video reception.

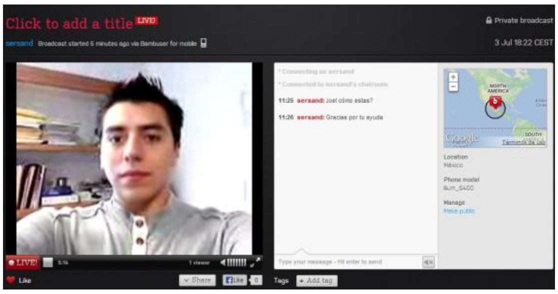


Fig. 11. GateWatcher: communicating with the visitor via Bambuser´s instant messages.

5 Conclusions

Surveillance systems are very important in our days for security reasons. In this work we design GateWatcher a video doorbell surveillance system using an Android phone with Wi-Fi internet connection, to monitor remotely the presence of a visitor when rings the bell of a home. The audio and video of the visitor was sent to the Bambuser web site from which can be can be accessed in real time using a second Android phone with Internet connection. The user interfaces of GateWatcher was eased using App Inventor with its intuitive drag-and-drop graphical interface, and its sms texting messages and Activity Starter components. This paper showed that it is not very difficult the implementation of a smart surveillance system with live video streaming using Android smartphones with sms alert message notifications.

6 Future work

This application may still be expanded beyond the actual state, making it more useful by including a log of doorbell ringing events, thus as the option that allow the visitor to leave a recorded video message that can be retrieved at convenience. The transmission of alert notifications by email or social networks would be convenient.

Acknowledgements

This work was supported by SIP-IPN project 20131854.

References

1. Scratch imagine program share, Scratch: a programming language for everyone. http://scratch.mit.edu , (2012).
2. Wikipedia, Google App Inventor, http://en.wikipedia.org/wiki/Google_App_Inventor.
3. Badrun Nahar and Md. Liakot Ali, Development of Mobile Phone Based Surveillance System, Proceedings of the 13th International Conference on Computer and Information Technology, pp 23-25, Dhaka, Bangladesh, (2010).
4. D. Shiny Irene and R. Dhanalakshmi, Video Surveillance System and Content Sharing Between PC And Mobile Using Android, http://ieeexplore.ieee.org, paper 6508392, (2013).
5. Rashmi, R., Latha, B.: Video Surveillance System And Facility To Access Pc From Remote Areas Using Smart Phone, http://ieeexplore.ieee.org, paper 06508393, (2012).
6. Estevez Ayres, I., Garcia Valle, M., Basanta Val P., Fernandez Pacheco, I.: Using Android Smartphones in a Service-Oriented Video Surveillance System, http://ieeexplore.ieee.org, paper 0572290, (2011).
7. Wang Liwei, Yan Shi and Xu Yiqiu, A Wireless Video Surveillance System based on 3G Network, 2nd Conf. Environmental Science & Information Application Technology, (2010).
8. LUO Yong-mei, WEI Bo, and JIN Zhi-gang, SIP-based WIMAX Wireless Video Surveillance System, 2012 International Conference on Computer Science and Service System.
9. Mobotix T24, “T24, Hemispheric VoIP/SIP video door station”, http://www.mobotix.com/eng_US/content/view/full/78917, (2013).
10. DoorBot, The Wi-Fi enabled Smart doorbell, https//secure.christiestreet.com, may 2013.
11. CellNock, Peerhole Cell Phone, http://www.cellnock.com, (2013).
12. MIT Center for Mobile Learning, App Inventor for Android. http://appinventor.mit.edu.
13. Bambuser Interactive live video broadcast service from mobile phones and web cameras to the Internet, http://www.bambuser.com, (2013).
14. Log Viewer, Real-time system log of Android, https://play.google.com/store/apps/details?id=jp.itplus.android.log.viewer, (2013).
15. MIT App Inventor Tutorials, AlertMe: Texting in the background when app is not running, http://appinventor.mit.edu/explore/content/alertme.html.
16. MIT App Inventor Explore, Using the Activity Starter, http://appinventor.mit.edu/explore/content/Using-activity-starter.html, (2013).

Análisis del Comportamiento del Tránsito Vehicular con Base en el Sensado de Dispositivos Móviles

E. A. Moedano Cardiel, M.J Torres Ruiz, M. A. Moreno Ibarra

Centro de Investigación en Computación, Instituto Politécnico Nacional, México
edwinmoedano@gmail.com, mtorres@cic.ipn.mx,
marcomoreno@cic.ipn.mx

Resumen. Durante los últimos años el sensado de las actividades humanas con dispositivos móviles ha despertado un gran interés para diferentes disciplinas, ya que permiten cuantificar diferentes fenómenos con sus sensores. Uno es el GPS que identifica la posición, generando datos que pueden permitir el análisis del flujo en una red vehicular. Por mucho tiempo el tránsito se ha convertido en un problema que afecta a las grandes ciudades del mundo y crece precipitadamente por su naturaleza. Existen factores que impiden proporcionar una estimación de las condiciones del tránsito, tales como la falta de infraestructura del sensado de vehículos, deficientes sistemas de comunicación entre otras. En este trabajo, el problema que se plantea resolver es estimar las condiciones de una red vehicular con respecto al tránsito. Se propone una metodología de clasificación que consiste en calificar las vialidades con base en el sensado del flujo vehicular con dispositivos móviles y sus caracteristicas fisicas.

Palabras Clave: Tránsito, sensado vehicular, red vehicular, flujo vehicular, clasificación, estimación, dispositivos móviles, GPS.

1 Introducción

Hoy en día se han propuesto diversas técnicas para el sensado y análisis de las actividades humanas en el entorno geoespacial [1] [4].Los GPS y otros sensores para realizar técnicas de rastreo han recibido gran atención, debido al crecimiento de las aplicaciones con un enfoque geográfico y análisis espacial, orientadas hacia diferentes objetivos y enfoques [2]. En [8] [7] [6] se utilizan dispositivos móviles con GPS para el sensado y la estimación del tránsito, el problema para la estimación es el análisis de diferentes variables urbanas, temporales, climáticas y sociales por mencionar algunas. Se menciona en [8] que la eficiencia de las metodologías implementadas es baja ya sea por su poca precisión, lo impráctico que resulta por la falta de datos o por no contemplar el impacto generan las intersecciones por ejemplo los semáforos.

El sensado con dispositivos móviles presenta diferentes problemas comentadas en [2] son principalmente la capacidad de operación de los dispositivos móviles (almacenamiento, batería, tiempo de calibración, precisión) y la zona a sensar como ciudades, carreteras, entre otras.

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 131–138
Paper Recived 22-09-2013 and Acepted 21-10-2013

Existen diferentes metodologías para el sensado colectivo de una red vehicular y depende de la infraestructura de la zona donde se pretenda implementar. Las redes ad-hoc vehiculares (VANETs) se han retomado por los avances en las comunicaciones[3], para estas redes se desarrollaron diferentes esquemas de comunicaciones para colectar datos, procesar y acceder a la información de manera local o cooperativa, como son vehículo a vehículo (V2V) y con la introducción a las redes celulares (GSM, CDMA, EDGE, 3G) se plantean sistemas distribuidos con conectividad a Internet.

Los sensores GPS día con día son más precisos y esto provoca cierta preocupación para el almacenamiento y procesamiento de los datos [3] [4]. Los esquemas mixtos de conexión (Wi-Fi, Celular, DSCR, WLAN) han sido adoptados para solucionar esos problemas, porque dan poder a las VANETs para capturar datos, concentrarlos en algún servidor y compartir la información entre otros dispositivos.

En este trabajo se plantea resolver la problemática de estimar las condiciones del tránsito que pueda existir en una red vehicular. Se propone una metodología para clasificar vialidades con respecto al flujo vehicular, la clasificación se emplea utilizando una evaluación personalizada de vialidades (EPV) con base en el sensado del flujo vehicular con dispositivos móviles.

El resto del artículo está organizado como sigue: La sección 2 menciona los antecedentes en cada uno de los temas en que se involucra el problema a resolver, en la sección 3 se encuentra la metodología propuesta, en la sección 4 se presenta los detalles de la implementación y en la sección 5 los resultados preliminares y conclusiones.

2 Antecedentes

En esta sección examinamos los conceptos involucrados en la metodología y algunos trabajos previos.

2.1 Monitoreo del censado vehicular

El monitoreo del transito no solo se limita a obtener la posición, la trayectoria o movilidad de los vehículos, sino generar el análisis de las interacciones de los usuarios con su entorno para evaluar el comportamiento de ambos y posteriormente tomar decisiones. Existen varios trabajos enfocados a la movilidad y la estimación del flujo vehicular como en [7] y TruTraffic [3] los cuales recolectan datos usando redes 2/3G para entregar a los usuarios información reciente del estado del tránsito, con esto se generan rutas dinámicas y notificaciones sobre la saturación de alguna vialidad. Otro trabajo es Google Traffic que genera rutas con recursos como difusión de reportes de radio, noticias, entre otras. Sin embargo la limitante es que esto solo aplica para avenidas principales o carreteras. En [2] se menciona que para sus experimentos la frecuencia de sensado fue una ventana de dos a cinco segundos, almacenando los datos en la memoria del dispositivo para posteriormente realizar el procesamiento de los datos. En [6] se recolectaron los datos en algunas avenidas principales tomando la velocidad promedio en ventas de 15 minutos, después se realiza interpolación lineal y extrapolación para las vialidades donde no fue posible recolectar datos y establecer una aproximación.

2.2 Datos históricos

Una de las aplicaciones de los Sistemas de Información Geográfica (GIS) es la generación de rutas [4]. De acuerdo con [5], el problema está dividido en dos categorías, enrutamiento estático y dinámico. Se dice que el problema es estático si los datos de entrada no dependen estrictamente del tiempo, dentro de estas dos clasificaciones se subdivide en determinísticos y estocásticos, el problema es determinístico si conocemos todos los datos al momento de generar una ruta, esta clasificación nos permite determinar que este trabajo se realiza enrutamiento dinámico.

Para realizar el enrutamiento se deben seguir criterios y políticas que nos permitan generar rutas. Existen algunos trabajos donde se enfocan en el uso de datos en tiempo real y el análisis de datos previamente conocidos o históricos como los mencionados en [5] y [6], este tipo de trabajos se proponen para resolver el ruteo en redes dinámicas estocásticas, donde el flujo vehicular varía con el tiempo.

Las ventajas analizadas del uso de datos históricos combinados con el análisis de datos en tiempo real conforme a [6], se enfocan en disminuir los costos e incrementar la productividad de las flotillas, donde demuestran disminuir en horas picos un 37% el costo total por las mañanas y un 47% por las tardes, así como la reducción del uso de las unidades en un 28% por las mañanas y un 58% por las tardes, lo que nos indica que este tipo de metodologías pueden ser de gran beneficio para optimizar la carga de una red vehicular y distribuir de manera adecuada el flujo dentro de la red.

3 Metodología

La Metodología estima el tránsito en una red vehicular urbana y se divide en dos etapas, la etapa de sensado y la etapa de clasificación de rutas como se ilustra en la Figura 1. En esta sección detallaremos ambas etapas.

3.1 Etapa de sensado.

Para esta etapa primero normalizamos las vialidades asignándoles un valor denominado tiempo de recorrido óptimo (TRO) que representa una relación en segundos entre la longitud de la vialidad y la velocidad máxima permitida. Posteriormente se recolecta la ubicación de los dispositivos móviles con una granularidad de 15 mestros y se evaluan mediante a un proceso denominado *evaluación personalizada de vialidades* (EPV) para determinar el nivel de flujo vehicular. La EPV se aplica al determinar en una vialidad el *tiempo de arribo* (TA) en el que un conductor ingresa a la vialidad y el *tiempo de partida* (TP) en el que es avandonada la vialidad por completo, la diferencia nos da como resultado el tiempo de recorrido de vialidad (TRV) la cual nos permite clasificar una vialidad como en la Figura 2.

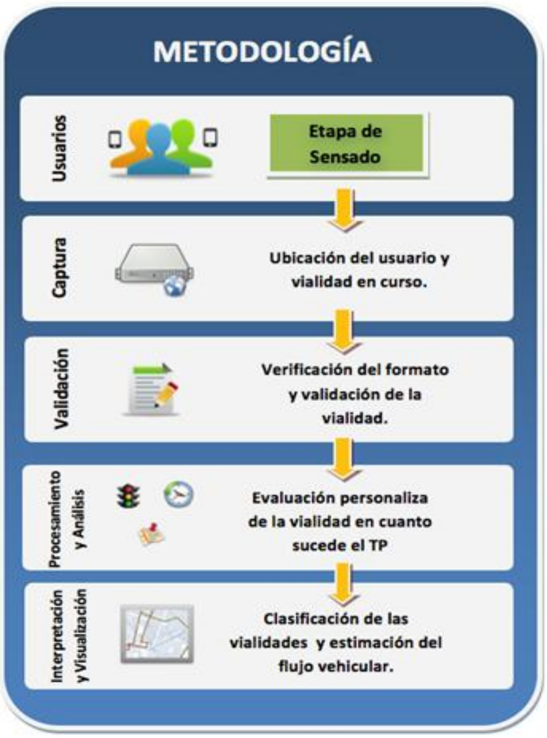


Fig. 1. Fases de la Metodología.

$$TRV < TRO \Rightarrow Flujo\ Adecuado$$
$$TRV \simeq TRO \Rightarrow Flujo\ Moderado$$
$$TRV > TRO \Rightarrow Congestión$$

Fig. 2. Clasificación de las vialidades.

La información procesada posteriormente es almacenada en la *base de datos histórica* (BHD) y genera un impacto en el valor del TRO de las vialidades ya que este valor reperesenta las condiciones ideales de flujo vehicular sin tomar en cuenta semaforos, topes, desperfectos viales, entre otros. Por lo que es necesario aproximar el valor del TRO a los datos sensados, para esto se utiliza una técnica llamada promedio ponderado exponencial y se expresa como en la Ecuación 1 Donde TRO_t es el tiempo promedio en un intervalo t, el valor de α que utilizamos es de 0.125 el cual es sugerido para esta tecnica en [9] y promedia el impacto entre el TRO y el TRV.

$$TRO_t = (1 - \alpha)TRO_{t-1} + \alpha TRV, t > 0 \tag{1}$$

4 Resultados

4.1 Implementación de la aplicación móvil

Se implementó una aplicación móvil que recolecta la posición de un vehículo de forma continua, para su implementación se utilizó *Titanium Studio* [11] debido a sus características multiplataforma, para el almacenamiento y procesamiento de los datos se implementó un servidor web con Apache Http Server [14], la aplicación web que procesa las consultas de los clientes está desarrollada con CakePHP [15] y los datos son gestionados con PostgreSQL [12] utilizando su extensión para datos geográficos PostGIS [13]. La comunicación cliente/servidor es con peticiones HTTP con el método POST incorporando un modulo personalizado de seguridad AES a 256 bits. En la base de datos fue cargado un shapefile de vialidades del Distrito Federaly se utilizarón datos del Instituto Nacional de Estadística y Geografía (INEGI).

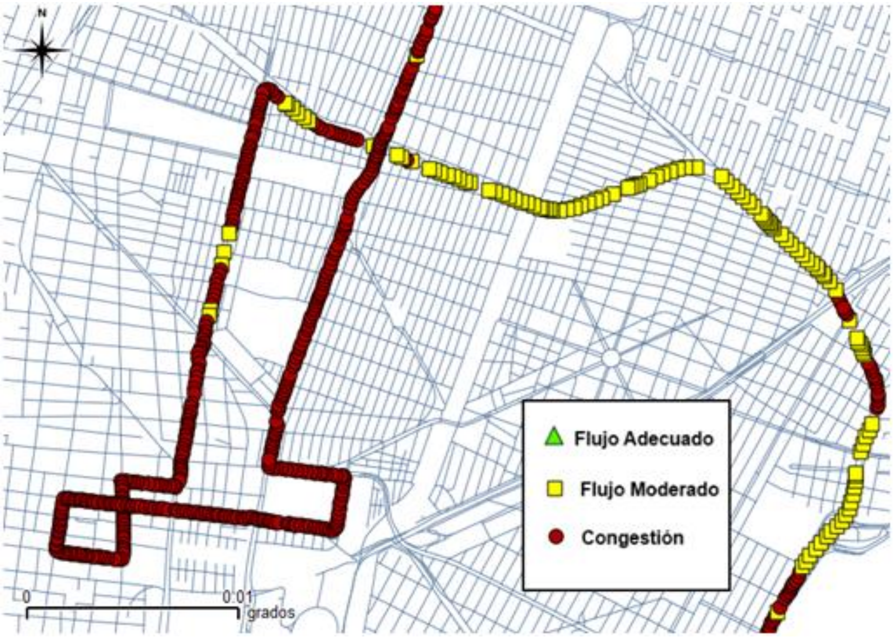


Fig. 3. Clasificación de las vialidades utilizando la *EPV*.

4.2 Resultados del experimento en la etapa de sensado

Dentro de los experimentos realizados se destaca el recorrido que se llevo a cabo en la zona centro de la ciudad de México Eje 1 Nte. La clasificación realizada en la Figura 3 obtuvo resultados más completos que las implementadas en [7] y [10] porque en ellos se obtiene la velocidad promedio de la vialidad completa y no por tramos, se enfocan en clasificar con respecto a la velocidad de los automovilistas y a un umbral de velocidad fijo dando a todas las vialidades sin importar sus caracteristicas fisicas el mismo umbral. Ésta metodología propone un análisis mas detallado por secciones dado que cada tramo que compone a una vialidad tiene diferentes caracteristicas fisicas dando una mejor caracterización de la red vehicular, ademas la cantidad de flujo vehicular de una vialidad depende de muchos factores como las condiciones de la vialidad, semaforos, el horario, clima, entre otros. En vez de tener umbrales de velocidad fijos se propone que los conductores califiquen las vialidades conforme al tiempo que les tomo recorrerlas.

5 Conclusiones

Los resultados obtenidos sugieren que podriamos estimar la cantidad de flujo vehicular conforme a la *Evaluación Personalizada de Vialidades* y la clasificación de las vialidades para las vialidades donde no fue posible recolectar datos mediante la etapa de sensado, el tipo de análisis que requerimos para calcular el *tiempo de recorrido de vialidad* es adecuado ya que no se requiere demasiada precisión del movimiento por las velocidades de conducción en las ciudades y el sensado por granularidad es bastante eficiente para economizar los recursos de los dispositivos moviles. Los problemas presentados para esta etapa fueron la comunicación entre cliente/servidor, el empaquetamiento de los puntos sensados para evitar la saturación del servidor y que las peticiones fueran consideradas como un DDoS attack, esto se resolvió con la frecuencia de recolección y la granularidad. Las ventajas de la metodología propuesta es lo detallado que resulta la clasificacion de las vialidades y como podemos personalizar para cada vialidad su situacion y tratarla de forma distinta a las demas, otras posibles ventajas son la rápida adopción de los usuarios de dispositivos móviles ya que la metodologia está pensada para consumir una cantidad de recursos minima y lo económico que resulta la recolección de datos voluntarios como se menciona [1].

Agradecimientos

Los autores de este trabajo agradecen al Instituto Politécnico Nacional por el soporte proporcionado para la realización de este trabajo.

Referencias

1. Goodchild, M. F.: Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211-221. (2007).
2. S. Van der Spek, J. Van Schaick, P. de Bois, R. de Haan,Sensing Human Activity: GPS Tracking, (2009).
3. Lee, U., & Gerla, M.: A survey of urban vehicular sensing platforms. Computer Networks, 54(4), 527-544. (2010).
4. Goodchild, M. F.: Geographic information systems and science: today and tomorrow. Annals of GIS, 15(1), 3-9. (2009).

5. Ghiani, G., Guerriero, F., Laporte, G., & Musmanno, R.: Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies. European Journal of Operational Research, 151(1), 1-11. (2003).
6. Kim, S., Lewis, M. E., & White III, C. C. Optimal vehicle routing with real-time traffic information. Intelligent Transportation Systems, IEEE Transactions on, 6(2), 178-188. (2005).
7. Qiankun, Z., Qingjie, K., Yingjie, X., & Yuncai, L.: An improved method for estimating urban traffic state via probe vehicle tracking. In Control Conference (CCC), 2011 30th Chinese (pp. 5586-5590). IEEE. (2011).
8. Leontiadis, I., & Mascolo, C.: Geopps: Geographical opportunistic routing for vehicular networks. In World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a (pp. 1-6). IEEE. (2007).
9. Paxson, V., & Allman, M.: Computing TCP's retransmission timer (pp. 1-6). RFc 2988, November. (2000).
10. Pattara-Atikom, W., Pongpaibool, P., & Thajchayapong, S.: Estimating road traffic congestion using vehicle velocity. In ITS Telecommunications Proceedings, 2006 6th International Conference on (pp. 1001-1004). IEEE. (2006).
11. Titanium Studio. http://www.appcelerator.com. [Consultado el 9 de Septiembre del 2013].
12. PostgreSQL. http://www.postgresql.org.es. [Consultado el 9 de Septiembre del 2013].
13. PostGIS.http://postgis.net. [Consultado el 9 de Septiembre del 2013].
14. Apache HTTP Server, http://httpd.apache.org. [Consultado el 9 de Septiembre del 2013].
15. CakePHP. http://cakephp.org. [Consultado el 9 de Septiembre del 2013].

Modeling Work Teams Trough Signed Graphs

Guillermo De Ita Luna, Yolanda Moyao Martínez, Luis Carlos Altamirano Robles

Faculty of Computer Sciences, Universidad Autónoma de Puebla
deita@cs.buap.mx, *ymoyao@cs.buap.mx*, *altamirano@cs.buap.mx*

Abstract. We present a method to count the different ways to form work teams in large corporate companies for projects with specific requirements. The teams are formed by employees and according to the ability of the employees for satisfying the requirements of the project. We model this problem through signed graphs, where each node represents an employee and an edge defines a restriction between the employees. The restrictions of a project is translated as a Boolean Formula Σ in two conjunctive form $(2-CF)$.
The logical one values in a satisfied assignment of Σ determine what employees can be part of an adequate team, while the logical zero values in a satisfied assignments of Σ indicate the employees that they have not to be part of the team.
In this model, $SAT(\Sigma)$ contains the set of different teams that can be formed to develop effectively the project. And $\#SAT(\Sigma)$ will show us how many different teams can be formed to develop the project. In order to compute $\#SAT(\Sigma)$, a signed extended graph is formed and recurrence equations are applied. We show for what class of signed graphs the computation of $\#SAT(\Sigma)$ can be done in polynomial time.
Keywords: Signed graphs, #SAT Problems, Counting Models.

1. Introduction

Signed graphs (graphs whose edges are designated positive or negative). This class of graphs have been very useful for modeling different class of problems in the social sciences [3, 5, 2]. For example, signed graphs are used for modeling the interaction among a group of persons and the type of relationship between certain pair of individuals of the group.

We are interested in computing the different work teams that can be formed to develop certain kind of projects in large corporative companies. This problem can be modeled through signed graphs, considering each node of the graph as an employee and forming edges to join nodes that represent restriction between those employees. A team's leader of the project defines the restrictions of the project, and such constraints form a Boolean Formula Σ in $2-CF$ (conjunction of binary or unary clauses).

Therefore, if we can find $SAT(\Sigma)$ (the satisfy assignments of Σ), we will obtain the adequate teams to develop the project effectively. And #SAT(Σ)

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 139–150
Paper Recived 22-09-2013 and Acepted 21-10-2013

(the number of models in Σ) shows how many different teams can be formed to develop the project.

The SAT problem is a classic NP-complete problem [1], while #SAT is relevant in the following issues: for estimating the degree of reliability in a communication network, for computing degree of belief in propositional theories, for the generation of explanations to propositional queries, in Bayesian inference, in a truth maintenance systems and for repairing inconsistent databases [6, 8]. Those previous problems come from several AI applications such as planning, expert systems, reasoning, etc.

$\#SAT$ is as difficult as the SAT problem, but even when SAT can be solved in polynomial time, it is not known an efficient computational method for $\#SAT$. For example, the 2-SAT problem (SAT limited to consider (≤ 2)-CF's) can be solved in linear time. However, the corresponding counting problem #2-SAT is #P-complete.

The aim is to develop a method to count the different teams that can be formed to develop different projects through to count the number of models in the Boolean formula associated to the requirements of the project.

This kind of automatization, will help the companies to save money and time, by making the team work selection process more efficient and dynamic.

2. Notation and Preliminaries

A signed graph Γ (also called sigraph) is an ordered pair $\Gamma = (G, \sigma)$ where $G = (V, E)$ is a graph called the underlying graph of Γ and $\sigma : E \to \{+, -\}$ is a function called a *signature* or *signing*. $E^+(\Gamma)$ denotes the set of edges from E that are mapped by σ to '+', and $E^-(\Gamma)$ denotes the set of edges from E that are mapped by σ to '-'.

The elements of $E^+(\Gamma)$ are called positive edges and those of $E^-(\Gamma)$ are called negative edges of Γ. A signed graph is *all-positive* (respectively, *all-negative*) if all of its edges are positive (negative); further, it is said to be *homogeneous* if it is either all-positive or all-negative, and *heterogeneous* otherwise.

Signed graphs have been very useful for modeling interactions among a group of persons and for representing the type of relationship between certain pair of individuals of the group. Special focus is on the social inequalities. What is it about such characteristics as sex, race, occupation, education, and so on that leads to inequalities in social interactions? [2].

Given a set of de n Boolean variables, $X = \{x_1, x_2, ..., x_n\}$. It's called a *literal* to any variable x or the negation $\overline{x}$ of it. We use $v(l)$ to indicate the variable involved by the literal l.

The disjunction of different literals is called a *clause*. For $k \in \mathbb{N}$, a $k-clause$ is a *clause* consisting of exactly k $literals$. A variable $x \in X$ appears in a clause c if x or $\overline{x}$ is an element of c. Let $v(c) = \{x \in X : x\ appears\ in\ c\}$. A conjunctive form ($CF$) is a conjunction of *clauses*. A $k - CF$ is a CF containing only $k - clausulas$ and, $(\leq k) - CF$ denotes a CF containing *clauses* with at most k literals.

Let Σ be a 2-CF, then an assignment s for Σ is a function $s : v(\Sigma) \to \{0, 1\}$. An assignment can also be considered as a set of non-complementary pairs of literals. If $l \in s$, being s an assignment, then s makes l $true$ and makes $\bar{l}$ $false$. A $clausula\ c\ is\ satisfied$ if and only if $c \cap s \neq \emptyset$, and if for all $l \in c, \bar{l} \in s$ then s falsifies c.

A CF Σ is satisfied by an assignment s if each *clausula* in Σ is satisfied by s and Σ is contradicted if it is not satisfied. s is a model of Σ if s is a satisfied assignment of Σ.

Let $SAT(\Sigma)$ be the set of models than Σ has over $v(\Sigma)$. Σ is a contradiction or unsatisfiable if $SAT(\Sigma) = \emptyset$. Let $\mu_{v(\Sigma)}(\Sigma) = |SAT(\Sigma)|$, be the cardinality of $SAT(\Sigma)$. Given Σ a CF, the SAT problem consists in determining if Σ has a model. The $\#SAT$ consists of counting the number of models of F defined over $v(\Sigma)$. We will also denote $\mu_{v(\Sigma)}(\Sigma)$ by $\#SAT(\Sigma)$.

Given a signed graph G_Σ, we obtain the associated Boolean formula Σ. Σ can be expressed as a two Conjunctive Form (2-CF) in the following way. Let $G_\Sigma = ((V, E), \sigma)$ be the signed graph, then $v(\Sigma) = V$ and for all positive edge $x \overset{+}{-} y$ in G_Σ the clause (x, y) is part of Σ, while for a negative edge $x \overset{-}{-} y$ in G_Σ the clause $(\overline{v(x)}, \overline{v(y)})$ is part of Σ. This means that the vertices of G_Σ are the variables of Σ, and each signed edge clause in E there is a clause in Σ.

2.1. Extending Signed Graphs

We consider the following approach which represents a very important application of signed graphs:

Corporate companies working based on projects must dynamically form work teams to solve the projects requirements. These companies have a finite set of "n" employees available, which we denote as: X=$\{x_1$, ..., $x_n\}$.

Project managers propose the staff selection to form a work team, based on restrictions between pairs of employees, such restrictions are formed according to the capabilities of employees and the requirements that must be accomplished in the project. For example, a restriction could be where both employees have the same abilities, another example could be where no employee covers some job profile for a certain type of project and this difficult or delay the develop of the project that might be accomplis with another employee that match the required profile.

The project manager defines the restrictions (capabilities or requirements)for each project with a pair of employees, where: $i \neq j = 1, 2, ..., n$. The clause is the way to model the restrictions on the employee x_i and the employee x_j. We identified four cases that model the relation between a pair of employees.

Case 1: (x_i). It indicates that it is mandatory for the employee x_i to be part of the project.

Case 2: $(x_i \vee x_j)$. The clause is unsatisfiable when $x_i = 0$ and $x_j = 0$, indicating that either one employees x_i or x_j must participate in the team, or even both can be on the same team.

Case 3: $(x_i \vee \sim x_j)$. The clause is unsatisfiable when $x_i = 0 \wedge x_j = 1$. It indicates that if x_j is on the team, then x_i should't be on the same team, because they block themselves, in the same way if x_i is in the team then x_j must not be in it.
Case 4: $(\sim x_i \vee \sim x_j)$. The clause is unsatisfiable if $x_i = 1 \wedge x_j = 1$ indicating that either one x_i or x_j employees should not participate in the team, even may not be both. But we must avoid having both in the same team.

This kind of restrictions are joined in a formula Σ in $2-FC$ (conjunction of binary or unary clauses where each variable appears twice at most), and also forms a constraint graph of the formula G_Σ.

G_Σ shows graphically how work teams can be formed to develop a project. Thus, an assignment that satisfies Σ indicates a way of creating the proper work team that will handle a certain project. This means that the variables that take value 1, represents the employee that will form the team, and the value 0 represents the employee that should not be on the team.

Therefore $SAT(\Sigma)$ contains all the possible teams that can be formed to implement the project effectively and $\#SAT(\Sigma)$ will indicate us how many different teams can be formed for the project Σ.

The signed graphs are very useful to represent some kind of relationships among individuals. In this case, each employed is a node of the graph and there is an arc between $\{x, y\}$ if x is in some relation to y. Many of the relationships of interest have natural opposites, for example: likes/dislikes, associates with/avoids, and so on [2]. For this, two different signs $\{+, -\}$ are associated with each edge of the graph.

For example, if we consider that x likes y, then a positive edge is denoted between x and y, but we do not know what about y with x. y might dislike x or maybe like him. In order to be more precise in the kind of relationships between two individuals is better to consider two signs in each edge, one sign associated with each point end of the edge.

Furthermore, in order to consider general Boolean formulas in 2-CF, we consider an extension of the concept of signed graphs. Instead to consider just one sign (+ or -) associated with each edge of a signed graph $G = ((V, E), \sigma)$, we consider here that all edge in E has associated a pair of signs. Then the signed function σ has now the type $\sigma :\rightarrow \{(+,+), (+,-), (-,+), (-,-)\}$ which gives a pair of signs to each edge of G.

In this way, all binary clause $\{x, y\}$ can be represented by a signed edge in the graph in a natural way without importance of the signs associated to the variables in the clause. For example the clause $\{\overline{x}, y\}$ will be represented as the signed edge $x \overset{-\,+}{---} y$ and in this case, - is called the adjacent sign of x and + the adjacent sign of y in the edge $x \overset{-\,+}{----} y$.

Add more, we can consider those extended signed graphs as work team $\Sigma = (G, \sigma)$ where $G = (V, E)$ is the underlying graph of Σ and $\sigma : E \rightarrow \{(+,+), (+,-), (-,+), (-,-)\}$ is the signature function. That means that all edge in the graph of work team is signed in its endpoints.

We are interested here, in count the different assignments associated to the nodes of the work team in such a way that all edge in the work team will be satisfied by at least one of its two possible signs of its endpoints.

For this, we start analyzing the way to build satisfy assignment in the work team considering first the most simple topologies associated with the underlying graph of a work team.

3. #SAT Solution Techniques

3.1. If G_Σ is a Linear Path

Let us consider that the graph of the work team, $G_\Sigma = (V, E)$ **is a linear path**. Let us write down its associated formula Σ, without a loss of generality (ordering the clauses and its literals, if it were necessary), as: $\Sigma = \{c_1, ..., c_m\} = \Big\{\{x_1^{\epsilon_1}, x_2^{\delta_1}\}, \{x_2^{\epsilon_2}, x_3^{\delta_2}\}, \ldots, \{x_{m-1}^{\epsilon_m}, x_m^{\delta_m}\}\Big\}$, where $|\upsilon(c_i) \cap \upsilon(c_{i+1})| = 1, i \in [\![m-1]\!]$, and $\delta_i, \epsilon_i \in \{0, 1\},\ i = 1, ..., m$.

In order to compute the number of signed paths in such graph of the work team, we start with a pair of values (α_i, β_i) associated with each node i of the graph. We call the charge of node i to the pair (α_i, β_i). The value (α_i) indicates the number of times that node i takes the positive value and (β_i) indicate the number of times that node i takes the negative value.

Let f_i be a family of clauses of Σ built as follows: $f_0 = \emptyset, f_i = \{c_j\}_{j \le i}$, $i \in [\![m]\!]$. Note that $f_i \subset f_{i+1}$, $i \in [\![m-1]\!]$. Let $SAT(f_i) = \{s : s\ satisfies\ f_i\}$, $A_i = \{s \in SAT(f_i) : x_i \in s\}$, $B_i = \{s \in SAT(f_i) : \overline{x}_i \in s\}$. Let $\alpha_i = |A_i|$; $\beta_i = |B_i|$ and $\mu_i = |SAT(f_i)| = \alpha_i + \beta_i$. From the total number of models in $\mu_i, i \in [\![m]\!]$, there are α_i of which x_i takes the logical value 'true' and β_i models where x_i takes the logical value 'false'.

For example, $c_1 = (x_1^{\epsilon_1}, x_2^{\delta_1})$, $f_1 = \{c_1\}$, and $(\alpha_1, \beta_1) = (1, 1)$ since x_1 can take one logical value 'true' and one logical value 'false' and with whichever of those values satisfies the subformula f_0 while $SAT(f_1) = \{x_1^{\epsilon_1} x_2^{\delta_1}, x_1^{1-\epsilon_1} x_2^{\delta_1}, x_1^{\epsilon_1} x_2^{1-\delta_1}\}$, and then $(\alpha_2, \beta_2) = (2, 1)$ if δ_1 were 1 or rather $(\alpha_2, \beta_2) = (1, 2)$ if δ_1 were 0.

In general, we compute the values for (α_i, β_i) associated to each node x_i, $i = 2, .., m$, according to the signs (ϵ_i, δ_i) of the literals in the clause c_i, by the next recurrence equation:

$$(\alpha_i, \beta_i) = \begin{cases} (\beta_{i-1} & , \alpha_{i-1} + \beta_{i-1})\ if(\epsilon_i, \delta_i) = (-, -) \\ (\alpha_{i-1} + \beta_{i-1}, & \beta_{i-1})\ if(\epsilon_i, \delta_i) = (-, +) \\ (\alpha_{i-1} & , \alpha_{i-1} + \beta_{i-1})\ if(\epsilon_i, \delta_i) = (+, -) \\ (\alpha_{i-1} + \beta_{i-1}, & \alpha_{i-1})\ if(\epsilon_i, \delta_i) = (+, +) \end{cases} \quad (1)$$

We denote with $' \rightarrow '$ the application of one of the four rules of the recurrence (1), so, the expression $(2, 3) \rightarrow (5, 2)$ denotes the application of one of the rules (in this case, the rule 4), over the pair $(\alpha_{i-1}, \beta_{i-1}) = (2, 3)$ in order to obtain $(\alpha_i, \beta_i) = (\alpha_{i-1} + \beta_{i-1}, \alpha_{i-1}) = (5, 3)$.

In recurrence (1) the (ϵ_i, δ_i) represents the signs associated with the edge joining a child node joining with the father node. These recurrence equations allow to carry the count to indicate the different ways to conform work teams for the project. In fact, the sum $\alpha_i + \beta_i$ obtained from the root node of the graph, indicate the total number of Boolean formula models associated with the graph and it is as well as the number of different ways to conform work teams for the project.

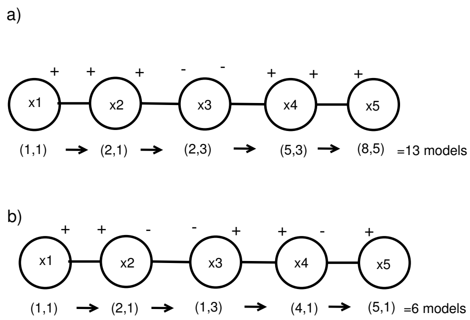


Fig. 1. a) Counting models over a positive path. b) Counting models over a general path.

Example 1 *Let* $\Sigma = \{(x_1, x_2), (x_2, \overline{x}_3), (\overline{x}_3, x_4), \ (x_4, x_5)\}$ *be a path, the series* $(\alpha_i, \beta_i), i \in [\![5]\!]$*, is computed according to the signs of each clause, as it is illustrated if the figure (1a). A similar path with 5 nodes but with different signs in the edges is shown in figure (1b).*

If Σ is a path , we apply (1) in order to compute $\mu(\Sigma)$. The procedure has a linear time complexity over the number of variables of Σ, since (1) is applied while we are traversing the chain, from the initial node y_0 to the final node y_m.

3.2. If G_Σ is a Tree

The charge of the nodes (α_i, β_i) for $i = 1...n$, are calculated as the nodes are visited applying a post-order traversals. The first pair of values (α_i, β_i) for the terminal nodes of the network (tree leaves) are initialized with the value $(1, 1)$. So, to traverse the tree in post-order, we start by the left tree first, followed by the right tree, and then finally the root node. At each visit of a child node $i - 1$ to a father node i, the new values (α_i, β_i) are computed for the father node as shown in the algorithm 1.

This procedure returns the number of signed paths as the ways to conform the work teams for the project in time $O(n+m)$ which is the necessary time for traversing G_Σ in depth-first [7]

Input: A_Σ the tree defined by the depth-search over connectivity graph of the electrical network G_Σ
Output: The number of signed paths of an electrical network.
Procedure: Traversing A_Σ in depth-fist, and when a node $v \in A_\Sigma$ is left (all of its edges have been processed) assign:
1. $(\alpha_v, \beta_v) = (1, 1)$, If v is a leaf node in A_Σ
2. If v is a father node with a list of child nodes associated, i.e., $u_1, u_1, ...u_k$, are the child nodes of v, then as we have already visited all the child nodes, then each pair $(\alpha_{u_j}, \beta_{u_j})$ $j = 1, ..., k$ has been defined based on (1), $(\alpha_{v_i}, \beta_{v_i})$ is obtained by apply (2) over $(\alpha_{i-1}, \beta_{i-1}) = (\alpha_{u_j}, \beta_{u_j})$. This step is iterated until computes all the values $(\alpha_{v_j}, \beta_{v_j})$, $j = 1, , k$. And finally, let $\alpha_v = \prod_k^{j=1} \alpha_{v_j}$ y $\beta_v = \prod_k^{j=1} \beta_{v_j}$
3. If v is the root node of $A\Sigma$ the return $(\alpha_v + \beta_v)$

Algorithm 1: Algorithm Count_Models (A_Σ)

3.3. If G_Σ have a Single Cycle or Parallel Edges

The method now proposed is based on the recurrence equations in Table 1 and set theory.

Parallel Edges The parallel edges are those that join two nodes more than once, to be exact up to 4 times. In this way can be created only 4 cases of edges $\{ \xrightarrow{++}, \xrightarrow{+-}, \xrightarrow{-+}, \xrightarrow{--}\}$, a fifth case would be redundant of any of the above. Hence the idea to combine the recurrence equations with set theory.

The method consists of:
1.We must know which is the source node and destination node, so we choose the direction of the edges.
2. Values are taken Greek letters source node (α_i, β_i) to obtain the final result of the values of the destination node $(\alpha_{i+1}, \beta_{i+1})$. Each value of the destination node is a unitary set. If there is a sum, then they will join Greek letters and form a set of two elements.
3. For each ordered pair value of their corresponding sets intersect. The resulting set will be the value that the destination node. If the intersection is empty set then the value is zero.
4. In case of quadruple parallel edges automatically eliminates the possibility of finding a model or mapping that satisfies the Boolean formula.

Example 2 *Let* $\Sigma = \{(x_1, x_2), (x_1, x_4), (x_1, x_8), (\overline{x}_1, \overline{x}_8), (x_1, \overline{x}_8), (\overline{x}_1, x_8), (x_2, \overline{x}_3), (\overline{x}_2, x_3), (x_4, x_5), (x_4, x_6), (x_6, x_7), (x_6, \overline{x}_7), (\overline{x}_6, x_7)\}$

Step 1. In this example, Σ contains quadruple parallel edges which we give the result that there is no satisfiable assignment for it. We note that the leaf nodes should take values from $(1, 1)$. To resolve double parallels edges that join x_3, x_2, we apply the recurrence (1) where in this case, we have the cases of edges ($\xrightarrow{+-}$, $\xrightarrow{-+}$). In Table 1, we show the step 3, with the help of recurrence (1) for the cases 2 and 3.

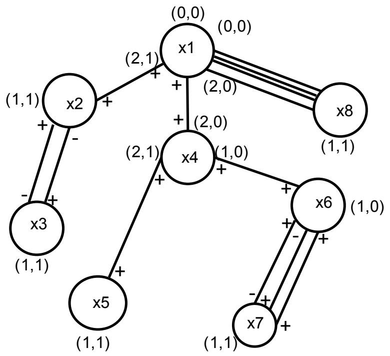


Fig. 2. Graph with parallel edges from a formula in 2-CF.

Table 1. Recurrence Equations expressed in sets

3. (+)⟶(-)	$\alpha_{i+1} = \{\alpha_i\}$ $\beta_{i+1} = \{\alpha_i, \beta_i\}$
2. (-)⟶(+)	$\alpha_{i+1} = \{\alpha_i, \beta_i\}$ $\beta_{i+1} = \{\beta_i\}$

In Table 2, we show the step 4, intersections are performed of $(\alpha_{i+1}, \beta_{i+1})$.

Table 2. Recurrence Equations expressed in sets of triple edges

(x_3, x_2)	$\alpha_{i+1} = \{\alpha_i\} \bigcap \{\alpha_i, \beta_i\} = \{\alpha_i\}$ $\beta_{i+1} = \{\alpha_i, \beta_i\} \bigcap \{\beta_i\} = \{\beta_i\}$

Table 3. Intersection of the sets of $(\alpha_{i+1}, \beta_{i+1})$

1. (+)⟶(+)	$\alpha_{i+1} = \{\alpha_i, \beta_i\}$ $\beta_{i+1} = \{\alpha_i\}$
2. (+)⟶(-)	$\alpha_{i+1} = \{\alpha_i\}$ $\beta_{i+1} = \{\alpha_i, \beta_i\}$
3. (-)⟶(+)	$\alpha_{i+1} = \{\alpha_i, \beta_i\}$ $\beta_{i+1} = \{\beta_i\}$

In Table 2, we show the step 4, intersections are performed of $(\alpha_{i+1}, \beta_{i+1})$.

In Table 4, we obtain an empty set for β_{i+1}. In quadruple edges of (x_8, x_1), the end result for x_1 is $(0, 0)$, to prove this only sets must intersect all four cases, this is shown in Table 5.

Table 4. Intersection of the sets of $(\alpha_{i+1}, \beta_{i+1})$ with triple edges

(x_7, x_6)	$\alpha_{i+1} = \{\alpha_i, \beta_i\} \bigcap \{\alpha_i\} \bigcap \{\alpha_i, \beta_i\} = \{\alpha_i\}$ $\beta_{i+1} = \{\alpha_i\} \bigcap \{\alpha_i, \beta_i\} \bigcap \{\beta_i\} = \emptyset$

Table 5. Intersection of the sets of $(\alpha_{i+1}, \beta_{i+1})$ whit quadruple edges

(x_8, x_1)	$\alpha_{i+1} = \{\alpha_i, \beta_i\} \bigcap \{\alpha_i\} \bigcap \{\alpha_i, \beta_i\} \bigcap \{\beta_i\} = \emptyset$ $\beta_{i+1} = \{\alpha_i\} \bigcap \{\alpha_i, \beta_i\} \bigcap \{\beta_i\} \bigcap \{\alpha_i, \beta_i\} = \emptyset$

Simple Cycles So far we have only count models of Boolean formulas that are converted to tree-like graphs. But what happens if there is a simple cycle? This mystery will be solved below.

The cycles are formed by back edges [4], which is why we subtract the value of the path to that node, the negative value of the ordered pair that makes the back edge unsatisfiable.

The method consists of:

1. Make to traverse in a graph until arriving at the back edge. The first node is the destination of the back edge, then follow the normal course of recurrence equations, until the source node of the back edge.
2. You must find the values that unsatisfiable the clause that represents the back edge.
3. Is again make to traverse from the destination node to the source node of the back edge, but taking into account only the value of the ordered pair that contains the destination node which unsatisfied the back edge. The other side will get a zero.
4. When you reach the the source node back edge only subtract the negative part of the ordered pair of the new tour route to normal.
5. This new ordered pair used to traverse the whole graph.
6. Should you find another cycle or back edge repeat the process.

Example 3 *Let* $\Sigma = \{(x_1, \overline{x}_2), (\overline{x}_2, x_3), (x_2, x_4), (x_3, x_4), (\overline{x}_4, \overline{x}_5)\}$ *be a boolean formula in 2-CF.*

Step 1. The graph obtained of Σ is shown in Figure 3a. The values of recurrence equations are below the graph, without regard to cycle or back edge. In this way, there are 11 assignments that satisfy Σ.

Step 2. The back edge that forms a simple cycle is (x_2, x_4). Previous clause is unsatisfied with $x_2 = 0$ and $x_4 = 0$, if x_4 is source node and x_2 is destination node.

Step 3. Make to traverse from the node x_2, but its value will $(0, 2)$, as this node unsatisfiable in the negative part and therefore the positive part is converted in zero. This is shows in Figure 3b.

Step 4. Substraction is made of the negative part obtained in the steps 1 and 3,

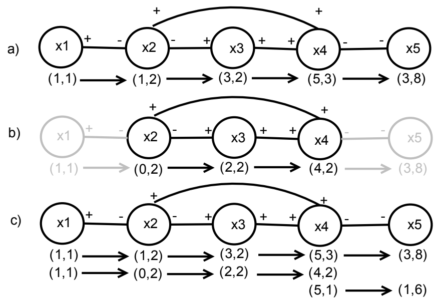


Fig. 3. a) Graph with simple cycle. Normal tour. b) Traversing the cycle. c)Count models of a graph with a simple cycle.

respectively, $(5, 3) - (0, 2) = (5, 1)$.
Step 5. We ends the tour which is shown in figure 3c. The result is 7 assignments that satisfy Σ.

4. Conclusions and Future Work

Most companies today use work teams to reach a particular purpose, which obtains effectively results for certain type of project under a certain time.

The task of project managers is to consolidate staff to create a team that can achieve the desired results, under certain types of restrictions, that might be empathy, capabilities, abilities and knowledge to accomplish the project. These relations or restrictions are often complex for the project manager and he need to take the best decision.

Our solution is based on graph theory and the satisfiability solutions of boolean formulas that represent the graphs. They can solve the problem by finding the number of ways to form work teams and the individuals that are part of those teams, eliminating errors and reducing time in selecting the right staff.

Until now only we solve graphs with simple edges, parallel edges, simple cycles and combinations between them. The solution of graphs with nested loops is in process and will help to have a complete scene and can model any job relation.

References

1. D.J.M. Garey: Computers and Intractability a Guide to the Theory of NP-Completeness (1979)
2. Fred S. Roberts: Graph Theory and Its Applications to Problems of Society, Edit. SIAM Collection, (1978)
3. Germina K.A.: Hameed S.K.: On signed paths, signed cycles and their energies, Applied Mathematical Sciences, Vol. 4:(70),pp. 3455-3466 (2010)

4. Johnsonbaugh Richard: Matemticas Discretas, Editorial Prentice Hall. 4ta edicin (1993)
5. Kota P. S., Subramanya M.S.: Note on path signed graphs, Notes on Number Theory and Discrete Mathematics 15:(4), pp. 1-6 (2009)
6. Roth, D.: "On the hardness of approximate reasoning", Artificial Intelligence 82,pp 273-302 (1996)
7. R. Tarjan: Depth-First Search and linear Graph Algorithms, SIAM Journal on Computing 1:146-160 (1972)
8. Vadhan, S. P.: "The complexity of Counting in Sparse, Regular, and Planar Graphs", *SIAM Journal on Computing*,pp. 398-427, V31, N2 (2001)

Sistema Auxiliar Basado en Android para el Tránsito de Usuarios del Sistema de Transporte Colectivo Metro de la Ciudad de México

Yaxkin Flores Mendoza, Axel Ernesto Moreno Cervantes, Miguel Ángel Moran Flores

Escuela Superior de Cómputo-IPN, Av. Juan de Dios Bátiz s/n esquina Miguel Othón de Mendizabal. Unidad Profesional Adolfo López Mateos. Col. Lindavista C.P. 07738, México, D.F.
floresya@hotmail.com, axelernesto@gmail.com,
mikesaurio@hotmail.com

Abstract. This paper shows the development of a guiding Android-based system for passengers to transit through México City's Subway System, "SmartWay", which uses technologies based on user location, such as: geocells and / or mobile phone's intensity power vectors. Moreover, the system's features includes: the generation of the shortest path between two stations, subsystem of warning proximity to the target station, voice recognition commands, posting messages on Facebook and image processing to manipulate Subway´s map. The goal of SmartWay is to assist passengers to travel around the city, indicating the nearest stations to the user location and generate the shortest path.

Keywords: Geocell, information system, mobile device, power vector.

1 Introducción

En la actualidad el Sistema de Transporte Colectivo Metro (STC Metro), conocido coloquialmente como Metro (por la contracción de tren metropolitano), constituye la infraestructura física, técnica y humana más importante con la que cuenta el Gobierno de la Ciudad de México para enfrentar la demanda de servicios de transporte. De acuerdo con los datos revisados en la referencia [1], el Metro tiene una red de 200 kilómetros de vías dobles, distribuidas en 11 líneas y 175 estaciones, cuenta con 302 trenes, de los cuales 201 son para la operación diaria, y realiza 1 millón 157 mil 490 vueltas, lo que se traduce en una oferta de servicio de 3.4 millones de lugares anuales. Además, de los casi 9 millones de personas que habitan en la Ciudad de México [2], el 80% utiliza el transporte público [3], y casi 1.5 millones corresponde a la afluencia de usuarios del Metro durante el año 2010. Estos datos reflejan el constante movimiento de la población usando este medio de transporte; sin embargo, es precisamente debido a su extensión que para algunas personas, llámense turistas nacionales o extranjeros, que por alguna razón acuden a la Ciudad de México, o incluso residentes de la misma, resulta complicado planear su trayecto de un punto a otro, convirtiéndose esto en un problema. Por otro lado, el avance de la tecnología ha hecho posible que 9

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 151–160
Paper Recived 22-09-2013 and Acepted 21-10-2013

de cada 10 personas posean celulares y 1 de cada 7 cuenten con Smartphones [4]-[6], los cuales tienen la misma funcionalidad de los celulares pero además pueden manejar cualquier tipo de cuenta de correo electrónico, acceder a redes inalámbricas (Wi-Fi), visualizar varios tipos de archivos de oficina y, por supuesto, tienen una alta capacidad de personalización en cuanto a la adición de aplicaciones y gadgets que los hace prácticamente ilimitados en funciones. Debido a esto se están convirtiendo en una herramienta primordial para la gente común y de negocios, es decir, no son de uso exclusivo de los estándares sociales altos. Con lo anterior podemos observar que se necesita el uso de tecnología en las actividades cotidianas, actualmente, el STC Metro cuenta con un programa en su página web que obtiene de la ruta más corta entre una estación origen y una destino, sin embargo lo que nosotros queremos de resolver es que la obtención de la ruta más corta, los sitios interés cercanos a una estación y la ubicación de la estación más cercana entre otras funcionalidades sea en cualquier momento y lugar, todo esto haciendo uso de los Smartphones.

2 Estado del arte

Dadas las características de los Smartphones, actualmente existen aplicaciones, las cuales se han dividido en globales, nacionales e institucionales, basadas en estos dispositivos que ayudan a los usuarios del Metro en su tránsito. En la tabla 1 se muestran algunos ejemplos:

Tabla 1. Resumen de aplicaciones similares a SmartWay.

NIVEL INTERNACIONAL

SOFTWARE	CARACTERÍSTICAS
Metro	Aplicación móvil elaborada en el S.O. *Android* para ver los mapas de los sistemas de tránsito, Metro, autobuses y trenes de 180 ciudades de todo el mundo. Con esta aplicación se puede: Ampliar los mapas, encontrar el camino más corto entre las estaciones seleccionadas, así como encontrar la estación más cercana junto a la ubicación del usuario, además de obtener información sobre las estaciones. La interfaz está desarrollada en varios idiomas particularmente inglés, ruso y francés, la actualización de mapas es de forma automática [7].
Java Metro Madrid 2.0	Con ésta aplicación se puede visualizar el mapa con todas los trasbordes de metros y realizar búsquedas de las estaciones. Cuenta con todas las líneas y estaciones ordenadas y organizadas alfabéticamente, pudiendo localizar cualquiera de estas. Funciona en todos los teléfonos móviles que soporten java [8].
Metro de Madrid en tu iPhone / iPod Touch	Ofrece características de cálculo de rutas según trasbordes y duración del trayecto, diagramas del Metro de Madrid y localización de la estación más cercana [9], [10].

NIVEL NACIONAL

Metro DF	Es una aplicación desarrollada para iPhone, la cual provee información sobre las 11 líneas del Metro de la Ciudad de México y sus 175 estaciones y un buscador en el cual se escribe la dirección y da información de la estación más cercana a la ubicación del usuario usando Google Maps [11].
Aplicación Metro México	Es una aplicación con la que se puede observar los diagramas de Metro del DF, Guadalajara y Monterrey, permite localizar las estaciones en un mapa para la parte de DF [12].

NIVEL POLITÉCNICO

Orientación visual de las rutas del STC Metro (2010)	Se trata de un sistema de orientación tipo kiosko que, además de detallar a usuarios de este transporte aspectos de las rutas, brinda datos de los sitios de interés cultural, más cercanos a cada estación del Metro y de eventos artísticos a celebrarse en el Distrito Federal. Otra innovación de este producto tecnológico es que integra una cama termográfica que capta la temperatura corporal del usuario para conocer su nivel de estrés [13].

Como se puede observar, los sistemas tienen características similares como: la visualización de los mapas del STC Metro, así como encontrar la estación cercana a la ubicación del usuario, además de generar la ruta más corta entre los puntos seleccionados y mostrar los lugares de interés alrededor de la estación destino.

3 Metodología utilizada

"SmartWay" es un sistema auxiliar para el tránsito de usuarios del STC Metro de la Ciudad de México para dispositivos con Sistema Operativo Móvil Android, está conformado por tres módulos mismos que son mostrados en la figura 1 y descritos a continuación:

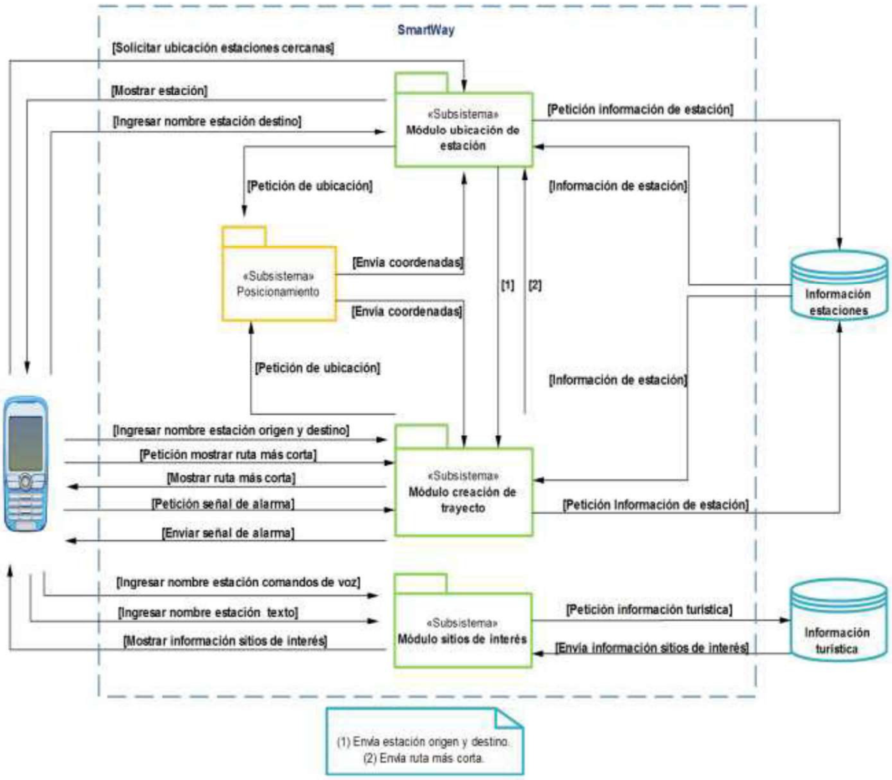


Fig. 1. Arquitectura de la aplicación Smartway

3.1 Ubicación de estación.

En éste módulo se localizan las estaciones más cercanas de acuerdo al punto donde se encuentre el usuario por medio de GPS (Sistema de Posicionamiento Global), por medio de la API de Google Maps. El usuario visualiza su ubicación y las estaciones del metro cercanos a él. También se cuenta con la opción en donde se ingresa una estación destino, de ser así, el segundo módulo de la aplicación obtendrá el cálculo de la ruta más corta, de cada una de las estaciones cercanas. Son consideradas 2 tecnologías para la localización: GPS y Network. Al realizar pruebas se determinó que el GPS aunque tarda un poco más, es mucho más exacto, la variante de exactitud utilizando Network es que si hay suficientes antenas resulta una buena localización, de lo contrario la localización tiene fallas notables. Se considera como cercano los metros que encuentra a 3 km a la redonda de la posición del usuario, de no haber estaciones de Metro en ese radio éste irá creciendo hasta encontrar 3 estaciones cercanas.

3.2 Creación de trayectos

Permite al usuario ingresar una estación origen y destino para mostrarle la ruta que tenga el menor número de transbordos y distancia entre estas. Además se podrá hacer uso de una alarma cuando el trayecto se haya creado, esta se activará una estación antes del destino, si estas últimas pertenecen a las líneas 1, 2, 3 o 5 del STC-Metro. Para la obtención de la ruta más corta se utilizó el algoritmo de A*, que trata de una especialización o concreción del algoritmo general de búsqueda de grafos en su versión BF (Best First), en la que la función de evaluación se define como una estimación del coste del camino solución condicionado a pasar por el nodo a evaluar [14]. La idea principal de este algoritmo es: Minimizar el coste estimado total de un camino en el árbol de búsqueda combinando, el coste para llegar al nodo n (se conoce exactamente: g), y el coste aproximado para llegar a un nodo meta desde el nodo n (estimado por el valor heurístico h*). La función heurística de A* está dada por:

- Coste real del plan (camino) de mínimo coste que pasa por n.

$$f(n) = g(n) + h(n) \quad (1)$$

- Estimación de f.

$$f^*(n) = g(n) + h^*(n) \quad (2)$$

A* mantiene dos estructuras de datos auxiliares, que podemos denominar abiertos, implementado como una cola de prioridad (ordenada por el valor f de cada nodo), y cerrados, donde se guarda la información de los nodos que ya han sido visitados, eligiendo el nodo de valor f* mínimo [13]. En la figura 2 puede apreciarse un ejemplo del uso de este algoritmo.

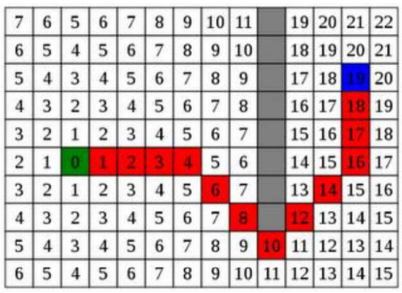


Fig. 2. Ejemplo de aplicación del algoritmo A*

La aplicación toma en cuenta distancias y tiempos entre estaciones y transbordos, ésta información es obtenida de una base de datos dentro del Smartphone, se decidió esto debido a que si se optaba por un servidor la comunicación entre el dicho dispositivo y este dentro de las estaciones subterráneas no se podría realizar.

Por otro lado el funcionamiento de la alarma se debe a dos factores: para la ubicación de Estación externa se tienen las geoceldas y para la ubicación de Estación subterránea lo hace a través del ID de antenas de telefonía celular.

Una geocelda está marcado por una cadena hexadecimal que define una región rectangular dos dimensiones dentro de la [-90,90] x [-180,180] de latitud / longitud del espacio. Resolución de un geocelda se mide por la longitud de su cadena de etiqueta. Se empieza con 16 geoceldas marcados por el número de hexdecimal con un solo dígito y, posteriormente, dividir cada geocelda en más de 16 geoceldas etiquetados por cadenas hexdecimal con más dígitos.

A medida que se aumenta la resolución, la etiqueta de una geocelda se hace más largo y el área cubierta por la geocelda más granular [15] tal como se puede ver en la figura 3.

Fig. 3. Visualización gráfica de Geoceldas

Para que el funcionamiento de la alarma en estaciones externas e internas haya tenido un buen funcionamiento, el servicio creado deberá comparar la geocelda generada o la captura del ID de antena en el momento, con la que se encuentra almacenada en el Smartphone.

3.3 Sitios de interés

Permite al usuario ya sea por medio de comandos de voz o de una lista seleccionar una estación de la cual se muestran los sitios de interés categorizados y así poder recibir información turística de estos, además el usuario puede compartir dicha información a través de redes sociales como: Facebook y Twitter.

La aplicación utiliza el reconocimiento de voz que maneja el sistema operativo Android en versiones 2.2 o superior. La clasificación de categorías y subcategorías es la siguiente: Entretenimiento (cines, teatros, centro comercial, restaurantes & bares y centro de espectáculo), Cultura (Museo, Casa de Cultura e Iglesia), Educación (Escuela y Biblioteca), Salud (Hospital, Clínica y Farmacia),Áreas Verdes (Parques & Jardines, reservas ecológicas y Deportivo), Gobierno (Jefatura Delegacional, Juzgado Cívico, Ministerio Público, Oficina de gobierno y Oficina Postal), Servicios (Banco), Transporte (Central Camionera, Aeropuerto, Metrobús y Suburbano);

4 Resultados

Para la verificación del buen funcionamiento del sistema se realizaron diferentes casos de estudio, mismos que dadas su extensión no fue posible describir detalladamente aquí, pero pueden ser consultados vía web en la siguiente dirección: http://148.204.58.221/axel/smartway/pruebas_SMARTWAY.pdf. En estas pruebas se propusieron diferentes problemas apegados a la vida cotidiana y dando solución con la aplicación SmartWay. Se realizaron un total de 100 pruebas de forma independiente, es decir, cada módulo fue sometido a diferentes situaciones críticas, midiendo en cada una de ellas el tiempo en segundos que se tardaba en dar respuesta a cada petición hecha y posteriormente se compararon contra mediciones manuales que fueron hechas a diferentes personas sin el uso de la aplicación. Dado lo anterior, se observó que en el módulo de reconocimiento de voz se obtuvo un tiempo promedio de respuesta de 6.5 segundos. Además, se presentaron casos en donde no se reconocieron las siguientes estaciones: Miguel Ángel de Quevedo, UAM-I y Constitución de 1917. Estas fallas se produjeron por la extensión del nombre de estación, así como la presencia de caracteres especiales y/o números. Finalmente, se recomienda mantener el teléfono a una distancia no mayor de diez centímetros para obtener mejores resultados en el reconocimiento de las estaciones.

Por otro lado, el módulo de alerta de proximidad a la estación destino presenta una fuerte dependencia al identificador asociado a las antenas de telefonía instalas en cada estación. Por lo tanto, algún cambio que se presente al respecto provocará que las alertas no funcionen apropiadamente.

En el caso de la búsqueda de la estación más próxima se obtuvo un tiempo promedio de respuesta de 0.91 segundos, siendo un tiempo aceptable.

Posteriormente se hicieron pruebas de integración de todos los módulos participantes para verificar que el funcionamiento de toda la aplicación fuera el adecuado.

Las pruebas se realizaron de esta forma para obtener resultados en tiempo real, y detectar las fallas que pudiesen presentarse de forma modular para evitar que afectara al desempeño funcional del sistema en general.

Por otra parte, la mayoría de las personas a quienes se les pedio realizar la dinámica realizaron rutas por estaciones conocidas no importando si éstas tenían más transbordos, debido a esto invertían más tiempo de lo planeado en realizar y seleccionar la ruta de forma empírica, de igual manera la búsqueda de una Línea o Estación desconocida era más tardado.

Usando la aplicación SmartWay se comprobó el correcto funcionamiento de los modulos (Un video tomado durante el uso de la herramienta SmartWay puede ser visto desde la siguiente dirección: http://youtu.be/eazbys15cRE), facilita a los usuarios su tránsito durante algún recorrido en el Metro y ahorra tiempo en el viaje para llegar de un punto a otro. Los resultados obtenidos de las pruebas realizadas fueron exitosos y el cumplimiento de los requerimientos fue alcanzado en su totalidad.

5 Conclusiones y trabajo futuro

Con la culminación del presente trabajo se obtuvo un sistema que funge como una guía de apoyo para los usuarios que transitan por el Sistema de Transporte Colectivo Metro de la Ciudad de México. Habiendo realizado las pruebas a la aplicación móvil se comprobó un buen desempeño, obteniendo en lo general tiempos de respuesta aceptables en cada módulo.

Se puede considerar en un futuro enlazar Smartway con un medio de actualización periódico el cual permita a la aplicación móvil mantenerse vigente y con ello brindar un mejor servicio al usuario. El tipo de información que se puede contemplar son: suspensión de estaciones debido a eventos, manifestaciones, inundaciones, etc. De igual manera incorporar nuevos sitios de interés, así como la suspensión de servicio de alguno de ellos debido a remodelaciones, bloqueos, etc.

Una extensión natural del sistema smartway es rediseñar la aplicación móvil para poder incluir algún otro sistema de transporte público de la ciudad de México. Tales como el Metrobus y tren suburbano. De tal manera se tendría una aplicación amplia la cual sería una herramienta más útil para que los pasajeros puedan viajar más fácilmente por el área metropolitana.

Finalmente, la aplicación móvil en un futuro puede ser adaptada como herramienta oficial del STC Metro, ayudando no sólo a las personas que cotidianamente viajan por este transporte, si no, también incrementando el turismo, debido a los sitios de interés que se puedan agregar posteriormente, y a su vez los usuarios puedan difundir con otras personas en sus redes sociales u otros medios de la existencia de la aplicación y así generar publicidad sin tener que invertir en medios físicos.

Agradecimientos

Se agradece a la Escuela Superior de Cómputo del Instituto Politécnico Nacional por las facilidades otorgadas para el desarrollo del sistema "Smartway".

Referencias

1. STC METRO. (2011). Equipo del Metro de la Ciudad de México. [online] Retrieved from: http://www.metro.df.gob.mx. [Accessed: 3 October 2011].
2. INEGI. (2011). INEGI. [online] Retrieved from: http://www.inegi.org.mx. [Accessed: 3 October 2011].
3. Lobo A. (2011). Transporte público sustentable: Movilidad con calidad de vida. [online] Retrieved from: http://www.circuloverde.com.mx/artman2/uploads/1/8._transporte_sustentable.pdf. [Accessed: 29 October 2013].
4. Solís V. (2011). Bajo la sombra de los Smartphones. [online] Retrieved from: http://www.eluniversal.com.mx/notas/735597.html. [Accessed: 5 Enero 2011].

5. Cristal P. (2008). Nuevas estadísticas telefónicas en México. [online] Retrieved from: http://vivirmexico.com/2008/05/nuevas-estadisticas-telefonicas-en-mexico. [Accessed: 29 October 2011].
6. Uber D. (2010). Estadísticas del uso de los móviles en América Latina. [online] Retrieved from: http://www.celularis.com/mercado/estadisticas-moviles-america-latina.php. [Accessed: 4 October 2011].
7. Roman G. (2010). aMetro. Subway on Android. [online] Retrieved from: http://www.ametro.org. [Accessed: 4 October 2011].
8. MuchoMoviles team. (2009). Metro Madrid 2.0 Aplicación Java Para Móvil. [online] Retrieved from: Disponible en: http://www.muchomoviles.com/metro-madrid-20-aplicacion-java-para-movil. [Accessed: 2 November 2010].
9. Apple Inc. (n.d.). Madrid Subway by Presselite. [online] Retrieved from: http://itunes.apple.com/app/madrid-subway/id303811539?mt=8. [Accessed: 29 October 2013].
10. Apple Inc. (2010). Metro Madrid by PhoneSoftware. [online] Retrieved from: http://itunes.apple.com/app/metro-madrid/id292665921?mt=8. [Accessed: 29 October 2010].
11. Apple Inc. (2010). Metro DF y algo más... Por AMT Desarrollos. [online] Retrieved from: http://itunes.apple.com/mx/app/metro-df-y-algo-mas/id367589411?mt=8. [Accessed: 29 October 2010].
12. Apple Inc. (2010). Metro México by Nahum Jaimes Nava. [online] Retrieved from: http://itunes.apple.com/app/metro-mexico/id314015529?mt=8. [Accessed: 29 October 2013].
13. IPN. (2010). Diseñan un Módulo de Orientación Visual de las Rutas del STC Metro. Gaceta Politécnica, Noviembre (2010).
14. Adler V., Jared F. and Tikhon B. (2007). Comparación Estrategias de Búsqueda. [online] Retrieved from: http://es.scribd.com/doc/53453308/Comparacion-de-Estrategias-de-Busqueda-Admisibles. [Accessed: 29 October 2012].
15. Nurik R. and Shen S. (2009). Geospatial Queries with Google App Engine using GeoModel. [online] Retrieved from: https://developers.google.com/maps/articles/geospatial. [Accessed: 29 October 2012].

Hacia el Diseño de una Plataforma de Desarrollo de Aplicaciones Colaborativas en Ambientes Inteligentes

Luis G. Montané-Jiménez, Carmen Mezura-Godoy, Edgard Benítez-Guerrero, Adrián García-Arcos

Facultad de Estadística e Informática, Universidad Veracruzana, Xalapa, México 91020
lmontane@uv.mx, cmezura@uv.mx, edbenitez@uv.mx,
zs12015363@estudiantes.uv.mx

Resumen. Este documento presenta una revisión de los trabajos que actualmente existen para el diseño y creación de ambientes inteligentes que apoyan el trabajo colaborativo. En esta revisión se han identificado las principales ventajas y desventajas con los mecanismo propuestos, y de igual manera se observa la dificultad en la construcción de este tipo de sistemas. Por ello, en este trabajo se propone una arquitectura conceptual para la construcción de sistemas colaborativos que integren inteligencia ambiental, incorporando el reconocimiento de recursos físicos o lógicos que apoyen el desarrollo de actividades colaborativas.

Palabras clave. Trabajo colaborativo asistido por computadora, inteligencia ambiental, actividades colaborativas, descubrimiento de recursos.

1 Introducción

En los últimos años los seres humanos vivimos en ambientes rodeados de pequeños dispositivos o recursos que ofrecen mayor capacidad computacional, por ejemplo: sensores, celulares, tablets, lectores RFID, etc. La integración de estos elementos es útil para construir aplicaciones o sistemas que puedan programar y configurar comportamientos inteligentes con respecto al contexto de las actividades realizadas, con el fin de proveer espacios físicos automatizados y transparentes para los usuarios. En estos sistemas lo importante no sólo es la parte interactiva, sino también la parte sensible del contexto, o lo que se denomina también consciencia contextual. Sin embargo, es posible observar que la mayoría de estos sistemas están orientados a trabajar en ambientes que no soportan el trabajo colaborativo.

Para ejemplificar una actividad en un ambiente colaborativo imaginemos el siguiente escenario (s_1): Hugo, Francisco y Luis son tres investigadores de la facultad de Ingeniería que forman un grupo de trabajo (t_1), para realizar una actividad colaborativa redacción de artículo (a_1). El grupo se ha fijado una meta para la actividad a_1, la cual consiste en someter el artículo en tiempo y forma (m_1). Para cumplir con la m_1 los investigadores realizan tareas individuales como: discutir resultados (it_1), escribir el estado del arte (it_2), editar imágenes (it_3), escribir conclusiones (it_4) o editar tablas

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 161–171
Paper Recived 22-09-2013 and Acepted 21-10-2013

(it_5). Igualmente para coordinarse y discutir sobre los avances se reúnen en la sala de profesores *(r_1)* mientras se toman un café.

Para alcanzar la meta m_1, los miembros del grupo de trabajo t_1 tienen que comunicares, coordinarse y colaborar, para ello pueden hacer uso de los recursos que se encuentren disponibles en la r_1, tales como: computadoras, impresoras, pizarrones virtuales, televisiones, redes inalámbricas, etc. Puede ser que al llegar al r_1 los investigadores desconozcan los dispositivos que pueden utilizar, por lo que tendrán que averiguar por ellos mismos los recursos disponibles a los que tienen acceso. Esta tarea adicional les haría tener que invertir tiempo adicional en una tarea no prevista inicialmente ya que no es parte de la activad del grupo, entonces, porqué no pensar que este tipo tareas pueden excluirse de la actividad principal del grupo. Tomando en consideración lo anterior, sería conveniente que la sala de reunión r_1 detectara automáticamente que la reunión se llevará a cabo a las 11:00 am con el grupo de trabajo t_1, que tiene como meta m1. La sala r_1 o de una manera más general el ambiente que rodea las tareas del grupo, podría percibir recursos disponibles en el momento y punto de encuentro (sala de reunión), y dependiendo de reglas, gustos, intereses definidos por los usuarios, el ambiente podría disparar acciones o secuencia de acciones (p.ej. encender luces, abrir ventanas, etc.) que ayuden en el trabajo colaborativo. La respuesta del ambiente hacia los usuarios podría ser realizada a través de dispositivos móviles (p.ej. celulares, tablets) que ejecuten aplicaciones cliente para acceder y manipular el ambiente.

La construcción de este tipo de sistemas es difícil, ya que es necesario integrar varios elementos, tales como: i) modelado y detección de actividades, ii) razonamiento social, iii) aprendizaje y adaptación automatizada, etc. Es decir que la transición de los espacios de trabajo tradicionales a los sistemas ubicuos involucraría las áreas del CSCW y el Cómputo Ubicuo.

En este documento se presenta un análisis respecto a la generación de ambientes inteligentes que apoyan el Trabajo Colaborativo Asistido por Computadora[1]. Este documento está estructurado de la siguiente forma: la Sección 2 presenta una discusión acerca del CSCW y el Cómputo Ubicuo, la Sección 3 presenta una discusión respecto a los Ambientes Inteligentes que incorporen mecanismos sociales, la Sección 4 discute ventajas y desventajas de los trabajos actuales en la construcción de ambientes con inteligencia social e individual, en la Sección 5 se propone una arquitectura conceptual que considere estos aspectos. Finalmente, la Sección 6 presenta las conclusiones y trabajos futuros.

2 CSCW y cómputo ubicuo

El área de la computación dedicada al estudio de la colaboración y propuesta de herramientas de apoyo al trabajo en grupo se le denomina CSCW, particularmente se observa un énfasis en el estudio del trabajo en equipo, los espacios compartidos de información y la adaptación de la tecnología en las organizaciones [18]. Para lograr

[1] Computer-Supported Cooperative Work, por sus siglas en inglés.

una colaboración efectiva en sistemas de este tipo, los usuarios deben poder saber en todo momento que es lo que pasa en el grupo, por ejemplo: quién está en línea, qué están haciendo los demás o cuáles son las metas en común, qué roles asumen los participantes, qué objetos manipulan o producen, etc. Considerando estas características, resulta conveniente proponer un avance considerable en los ambientes de trabajo tradicionales. Los ambientes de trabajo inteligentes podrían mejorar la interacción de los usuarios mediante la integración de objetos lógicos y físicos (p.ej. mediante el uso de sensores u otros tipos de hardware) a la actividad del grupo.

En el Cómputo Ubicuo, se observa que las preferencias y otras variables relacionadas a los usuarios (p.ej. ubicación, actividad física) han sido ampliamente estudiadas en los últimos años, debido a que el Cómputo Ubicuo tiene como objetivo incrementar el uso de sistemas computacionales a través de entornos físicos; haciéndolos disponibles y transparentes al usuario [19]. Para lograr esto, los sistemas ubicuos contemplan la inclusión de consciencia contextual (contextawareness) como un concepto para la detección de actividades y eventos generados por los usuarios, logrando de esta forma proveer ambientes sensibles que se adapten a las situaciones desarrolladas. Por lo tanto, dispositivos y tecnología ubicua son utilizados para adquirir datos ambientales y proporcionar información personalizada relacionada al contexto del usuario; por esta razón el contexto es un aspecto clave. La forma en cómo los usuarios interactúan con los sistemas de información ha ido evolucionando, ahora ya es posible la creación de ambientes inteligentes que contemplan interfaces en lenguaje natural que faciliten el uso de los recursos [14].

La tecnología hoy en día busca proveer espacios de interacción sensibles al contexto. La sensibilidad al contexto normalmente incluye diversos elementos, tales como: presencia, actividad y contenido. Estos elementos son esenciales para reconocer situaciones y relaciones entre los colaboradores, con el fin de reducir la brecha entre las necesidades de los usuarios finales y las funcionalidades proporcionadas por los entornos de colaboración. Trabajos como el de [16], presenta una visión respecto a la necesidad de proponer modelos multiagentes dinámicos que sirvan para crear ambientes inteligentes que apoyen a los usuarios en la creación de planes de acción. En un entorno de colaboración de índole ubicuo, implementar mecanismos que contemplen el contexto de los usuarios ayudará a los desarrolladores a crear sistemas colaborativos más personalizados en diversos dominios [4].

3 Inteligencia ambiental en escenarios colaborativos

La creciente demanda de servicios y acceso a la información ocasiona que el paradigma de cómputo ubicuo evolucione con el fin de proveer a los usuarios ambientes inteligentes que integren dispositivos cotidianos (p.ej. celulares, tablets, impresoras). El concepto de Inteligencia Ambiental (AmI[2], por sus siglas en inglés) hace hincapié en ofrecer mayor facilidad y apoyo en las interacciones humano-computadoras, logrando que las personas estén rodeadas de interfaces inteligentes e intuitivas incrusta-

[2] Ambient Intelligence

das con todo tipo de objetos y entornos capaces de reconocer y responder de una forma sencilla, discreta, y a menudo transparente [3]. En [17] la visión de AmI se caracteriza por: i) inteligencia e ii) incrustación. La primera se refiere al hecho de que el entorno digital sea capaz de analizar el contexto, adaptar y aprender comportamientos o emociones. Mientras que la incrustación está orientada a la disponibilidad de dispositivos. En este sentido la inteligencia ambiental propone la creación de ambientes sensibles y receptivos a la presencia de los usuarios [17, 1].

Los ambientes inteligentes que apoyan el trabajo colaborativo consideran diversos factores: usuarios, actividades, metas, recursos y tareas. En un ambiente colaborativo, los usuarios interactúan con el ambiente social a través de objetos que acceden a recursos disponibles. De este modo, los recursos, son aquellas entidades físicas o lógicas que ayudan a un conjunto de usuarios a realizar actividades en común. Los recursos físicos son dispositivos computacionales (p.ej. computadoras, cámaras digitales, proyectores, etc.) que pueden ser accedidos mediante interfaz física, mientras que los recursos lógicos son todas aquellas aplicaciones que pueden ser accedidas mediante una interfaz de comunicación del tipo WDSL[3], IDL[44], etc.

Al igual como se hace en los sistemas lógicos, un sistema ubicuo o pervasivo colaborativo busca proveer componentes que sean utilizados para asignar roles o establecer reglas dependiendo de los usuarios disponibles y la actividad desempeñada. El análisis del estado actual motiva a explorar y proponer nuevas técnicas para medir el nivel de relevancia que los usuarios tienen dentro de un grupo determinado.

4 Arquitecturas para ambientes inteligentes

A continuación se presenta una revisión y una clasificación de los trabajos que abordan el diseño y construcción de ambientes inteligentes. En esta revisión se encontraron trabajos enfocados a: i) descubrimiento de recursos, ii) asignación de permisos, iii) planificación de actividades y iv) recomendación de información.

4.1 Descubrimiento de recursos

Existen trabajos que han explorado la inteligencia ambiental desde varios enfoques. Por ejemplo, trabajos enfocados a la utilización de sistemas basados en agentes [8, 7, 9] para encontrar recursos físicos (p.ej. impresoras, puertas, calentadores, persianas). Particularmente, en [8] se busca que el usuario sea capaz de configurar un ambiente inteligente de acuerdo a sus gustos y necesidades, haciendo uso de artefactos o computadoras que tienen capacidad de razonamiento, planificación y aprendizaje, sin embargo, este trabajo contempla espacios monousuarios y un dominio específico. Por otro lado, en [7] se contemplan elementos sociales y se proponen aplicaciones para publicar recursos que apoyen a los usuarios en la realización de sus actividades, sin embargo, ésta perspectiva únicamente ha sido probada en ambientes locales con

---

3 Web Service De_nition Language

4 Interface Description Language

escenarios simples de colaboración y utilizando elementos sociales tradicionales como roles y grupos. En [9] presentan un sistema que muestra el acceso a electrodomésticos (p.ej. microondas, refrigeradores, tv etc.) a través de medios multimedia controlados con un PDA, para ello utilizan agentes que proporcionan datos y servicios a los usuarios, no obstante, esta arquitectura de software no contempla interacciones multiusuarios.

4.2 Asignación de permisos y seguridad

Existen enfoques que proponen la generación de modelos y arquitecturas [15, 13, 12] centrados en la seguridad, acceso y distribución de los diversos procesos individuales y colaborativos que pueden ser ejecutados en un sistema, en [15] es posible generar instancias para gestionar recursos físicos y lógicos con el _n de replicar y asignar permisos a diferentes usuarios a partir del proceso colaborativo, no obstante, aún carece de una evaluación con usuarios reales. En [13] el enfoque principal es la gestión de recursos lógicos por medio de la web, los usuarios son encargados de asignar permisos a los recursos protegidos que se encuentran en un ambiente distribuido. Por otra parte, [12] detalla una arquitectura para crear ambientes inteligentes que utilicen dispositivos bluetooth para que los usuarios puedan autenticarse e identificarse.

La finalidad de proveer soluciones de este tipo es para controlar diversos dispositivos y ofrecer mecanismos donde los usuarios puedan personalizar las funcionalidades de acuerdo a sus intereses y preferencias. Aunque algunos de estos trabajos están enfocados a la conectividad, muchas de estas arquitecturas carecen de una representación del flujo de información y de elementos sociales, dejando una brecha entre el uso de los diferentes recursos (p.ej. refrigeradores, microondas, lámparas e impresoras) y los usuarios.

4.3 Planificación de actividades

Soluciones utilizadas para facilitar el desarrollo de sistemas context-aware con índole ubicuo son abordadas en [6, 10, 11], donde se utiliza la ubicación de los usuarios para proveer los recursos físicos o lógicos. Particularmente, en [10] se propone la utilización de agentes para el seguimiento y el control de planes de actividades, buscando adaptar el medio ambiente a los usuarios a través de la asignación de tareas dependiendo de la zona geográfica, sin embargo, abordan la problemática desde ambientes mono-usuarios. En [6] se propone compartir recursos mediante middlewares que faciliten la colaboración entre los usuarios y utiliza agentes especializados en una tarea, sin embargo, este enfoque aún carece de un proceso de regulación para establecer permisos de acceso. Trabajos como el de [3] aborda el concepto de inteligencia ambiental con el objetivo de implementar ambientes para personas con discapacidad o personas mayores, en este trabajo se relacionan diferentes áreas, tales como: consciencia contextual e inteligencia ambiental. Por otra parte, [11] está orientado al análisis de sistemas de investigación para ambientes inteligentes, considerando módulos para la adquisición y almacenamiento de conocimiento, razonamiento y apoyo en la toma de decisiones, no obstante, las solución propuesta aún faltan ser validada y probada.

4.4 Recomendación de usuarios

Enfoques que implementen sistemas de recomendación o sistemas de reputación son tratados en [2, 5]. En [2] se presenta una arquitectura basada en puntuación de niveles de confianza que proporciona mecanismos para la creación de identidades o relaciones de confianza, la arquitectura utiliza componentes para obtener datos de los usuarios, tales como el historial y nivel de confianza de los integrantes, sin embargo, en [2] no se detallan las fuentes de datos donde proviene la información correspondiente al clasificador de reputación, y de igual forma la colaboración únicamente se ve reflejada por la comunicación entre dos personas. Por otra parte, en [5] se define un sistema de recomendación consciente del contexto destinado a sugerir puntos de interés (POIs) a los turistas, sin embargo, los puntos de interés no son fáciles de visualizar debido a que aparecen como objetos independientes al usuario, el dominio donde se ha probado esta solución es turismo.

A partir de la discusión anterior, una comparativa de los trabajos relacionados es presentada en la Tabla 1. Debido a que estos trabajos no contemplan a profundidad el uso de variables relacionadas a entornos colaborativos, se ha diseñado una arquitectura conceptual que apoye la generación de ambientes inteligentes que soporten el trabajo colaborativo.

5 Arquitectura propuesta

La construcción de un marco de trabajo que permita detectar recursos físicos y lógicos en un ambiente inteligente colaborativo podría mejorar eventualmente la productividad de los usuarios que faciliten las actividades grupales. Por ello, en este trabajo se propone una arquitectura para crear sistemas colaborativos ubicuos que incorporen el reconocimiento de recursos físicos o lógicos. En la figura 1 se muestra la arquitectura propuesta. A continuación se describen los elementos que la componen.

Recursos. Los recursos son artefactos físicos o lógicos utilizados por los usuarios cuando desempeñan una actividad colaborativa dentro de un espacio físico, por ejemplo, en oficinas o casas se podrían ofrecer luces, pantallas, impresoras, pizarrones digitales u otros equipos de cómputo que los usuarios utilizarían para realizar sus actividades. También se contemplan los recursos lógicos que apoyen el trabajo en grupo, por ejemplo: servicios web o sistemas multiagentes.

Explorador de recursos. En ambientes tradiciones, los usuarios son los que por cuenta propia exploran y notifican los recursos que pueden ser utilizados para desempeñar sus actividades colaborativas en espacios físicos. De modo que los usuarios que estuvieron previamente en el lugar son quienes conocen los objetos y recursos a los que pueden tener acceso, mientras no haya cambios en el ambiente, y son los mismos grupos de usuarios quienes deciden que recursos utilizar en un momento determinado.

Sin embargo, este procedimiento ocasiona un retardo en los procesos de coordinación, cooperación y comunicación correspondientes al trabajo colaborativo. Por lo tanto, a través de un sistema ubico que utilice información contextual este procedimiento podría ser automatizado y ejecutado antes y durante la actividad grupal. Para ello, se propone que la arquitectura utilice protocolos de descubrimiento de servicios en ambientes ubicuos.

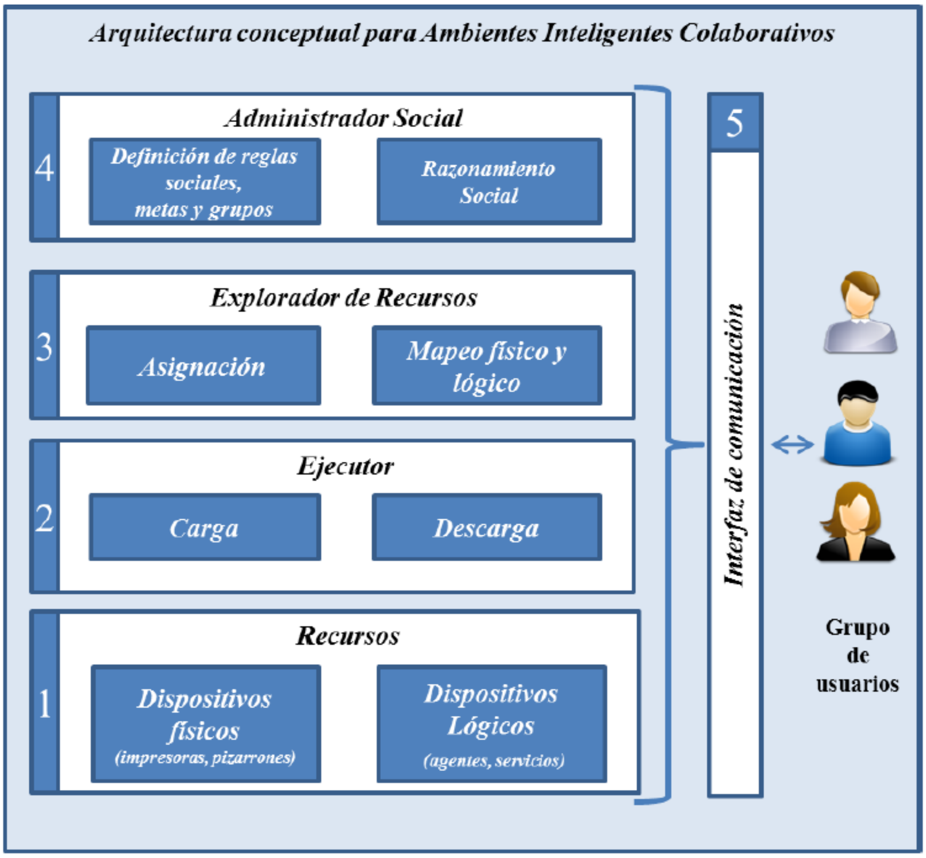


Fig. 1. Marco de trabajo propuesto.

Ejecutor. En este elemento se crean relaciones entre componentes lógicos y componentes físicos. De modo que a través de los componentes lógicos (p.ej. aplicaciones móviles) se manipulan o utilizan dispositivos físicos (p.ej. pizarrones, proyectores, laptops, servicio lógico, etc.). La finalidad del Ejecutor es poder tener un ambiente donde se suscriban servicios o recursos físicos que sean accedidos a través de componentes lógicos. Por ejemplo, en una sala de junta existiría una infraestructura para registrar los componentes que implementen una interfaz establecida previamente. Por lo que cuando un usuario llegue a ese lugar, de forma natural ellos utilizarían interfaces en sus dispositivos (p.ej. celulares, tablets) para saber de forma automática los recursos disponibles que están registrados y que son pertinentes para la actividad colaborativa que está en curso. En la figura 2 se muestra a mayor detalle las partes del ejecutor.

Administrador social. Este elemento permite definir y establecer reglas sociales de los grupos de trabajo. Aquí se definen preferencias, roles, metas, objetos, grupos y comportamientos (ver Tabla 2) con los que se podrían hacer inferencias del grupo y de los usuarios, con la finalidad de proveer servicios y recursos en tiempo y de forma adecuada. En este sentido también se espera tener un almacén de información con datos de los grupos y usuarios. Interfaz de Comunicación. La interfaz de comunicación es utilizada por las aplicaciones clientes que son ejecutadas en dispositivos móviles o computadoras personales de los usuarios. Estas aplicaciones clientes utilizan esta interfaz para comunicarse con los componentes suscritos al ambiente de ejecución.

Tabla 1. Comparativa de trabajos relacionados.

Nombre	Colaborativo	Prototipo	Variables	Desventajas
RAMS [7]	•	•	Ubicación	No soporta multiactividades.
AYPUY [15]	•		Ubicación	Carece de una arquitectura conceptual.
Kimovski [10]	•		Ubicación	Carece de un mecanismo de seguridad en datos.
HoCCAC [6]		•	Ubicación, roles	Arquitectura difícil de implementar.
CCB [11]	•		Ubicación	Carece de un mecanismo de razonamiento.
SRBNC[2]		•	Tiempo, confidencialidad	No soporta mas de dos usuarios.
SRT [5]	•		Ubicación, tiempo	Interacciones con usuarios indefinidas.

Tabla 2. Variables en ambientes colaborativos.

Categoría	Elementos
Interactivos	Objetos, tareas, eventos, usuarios, ubicaciones.
Cohesivos	Grupos, roles, metas, alianzas, actividades, reglas.

Interfaz de comunicación. La interfaz de comunicación es utilizada por las aplicaciones clientes que son ejecutadas en dispositivos móviles o computadoras personales de los usuarios. Estas aplicaciones clientes utilizan esta interfaz para comunicarse con los componentes suscritos al ambiente de ejecución.

6 Conclusiones y trabajos futuros

Con la arquitectura de reconocimiento de recursos físicos o lógicos se busca mejorar los ambientes colaborativos para que los usuarios realicen de forma más eficiente sus tareas comunes. Actualmente diferentes grupos de investigación trabajan en la búsqueda de métodos, plataformas y herramientas que apoyen a mejorar los entornos colaborativos en diferentes contextos (p.ej. cuartos inteligentes). Sin embargo, la mayoría de los trabajos desarrollados dentro del área de ubicuidad abordan la temática desde un enfoque monousuario y no desde un enfoque social, por lo tanto, existe un campo de oportunidad que converge en el área del CSCW y el Computo Ubicuo.

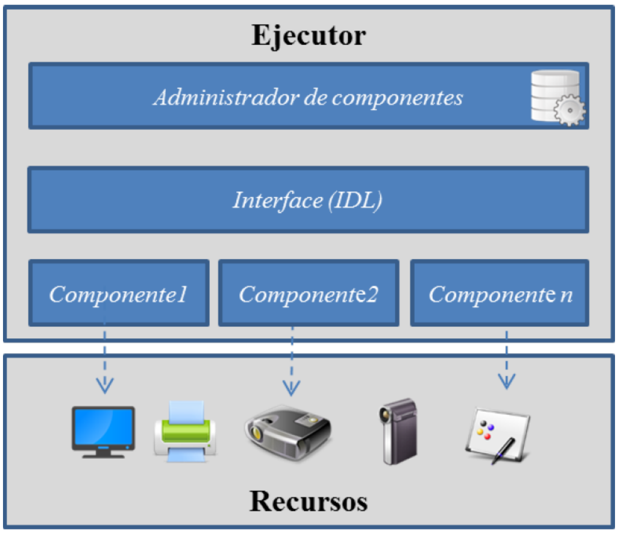


Fig. 2. Ambiente de ejecución.

Dentro del beneficio que se pretende alcanzar con la plataforma propuesta es la de proveer funcionalidades integradas por diferentes componentes computacionales. Para ello, se busca diseñar nuevos mecanismos que adquieran información con base en el perfil y necesidades del usuario, con el fin de proveer herramientas que faciliten y mejoren el proceso colaborativo. Como trabajo futuro se implementará y validará la plataforma en un ambiente inteligente con un escenario educativo; midiendo efectividad y eficiencia.

Referencias

1. Aarts, E.: Ambient intelligence: a multimedia perspective. IEEE MultiMedia 11(1), 12_19 (2004).
2. Carullo, G., Castiglione, A., Cattaneo, G., De Santis, A., Fiore, U., Palmieri, F.: Feeltrust: Providing trustworthy communications in ubiquitous mobile environment. In: Proc. of the IEEE 27th International Conference on Advanced Information Networking and Applications (AINA'13). pp. 1113_1120 (2013).
3. Emiliani, P.L., Stephanidis, C.: Universal access to ambient intelligence environments: opportunities and challenges for people with disabilities. IBM Systems Journal 44(3), 605_619 (Aug 2005).

4. Fadzillah, N.A., Omar, N., Abidin, S.: Application based context awareness in collaborative workspaces: A review. In: Proc. of 2012 IEEE International Power Engineering and Optimization Conference (PEOCO'12). pp. 411_415 (2012).
5. Fenza, G., Fischetti, E., Fumo, D., Loia, V.: A hybrid context aware system for tourist guidance based on collaborative _ltering. In: Proc. of the IEEE International Conference on Fuzzy Systems (FUZZ'11). pp. 131_138 (2011).
6. Fraile Nieto, J.A.: Sistema multiagente sensible al contexto. Universidad de Salamanca, España (2011).
7. Garcia Garcia, E.K.: Disponibilidad de recursos compartidos en un ambiente colaborativo con caracter ubicuo. Centro de Investigacion y de Estudios Avanzados del Instituto Politecnico Nacional, Mexico (Jan 2009).
8. Hagras, H., Callaghan, V., Colley, M., Clarke, G., Pounds-Cornish, A., Duman, H.: Creating an ambient-intelligence environment using embedded agents. IEEE Intelligent Systems 19(6), 12_20 (Nov 2004).
9. Jian, H., Yumin, D., Yong, Z., Zhangqin, H.: Creating an ambient-intelligence environment using multi-agent system. In: Proc. of the International Conference on Embedded Software and Systems Symposia (ICESS'08). pp. 253_258 (2008).
10. Kimovski, G., Trajkovic, V., Davcev, D.: Resource manager for distance education systems. In: Proc. of the IEEE International Conference on Advanced Learning Technologies. pp. 387_390 (2001).
11. Ko, H., Ramos, C.: A survey of context class_cation for intelligent systems research for ambient intelligence. In: Proc. of the International Conference on Complex, Intelligent and Software Intensive Systems (CISIS'10). pp. 746_751 (2010).
12. Lu, L.X., Huang, Z.Q., Hou, Y.B., Li, J., feng Wang, S.: A safe and fast connecting strategy of the bluetooth identi_cation sensors in ambient intelligence environment. In: Proc. of the International Symposiums on Information Processing (ISIP'08). pp. 538_542 (2008).
13. Machulak, M.P., Maler, E.L., Catalano, D., van Moorsel, A.: User-managed access to web resources. In: Proc. of the 6th ACM workshop on Digital identity management (DIM '10). pp. 35_44. ACM, New York, NY, USA (2010).
14. Marcos, M.C.: HCI (human computer interaction): concepto y desarrollo. El Profesional de la Información 10(6) (2001).
15. Quimbaya, A., Torres-Ribero, L., Carrillo-Ramos, A., Arias-Baez, M., Gonzalez, E.: Aypuy: A resource management system for collaborative environments. In: Proc. on 7th Colombian Computing Congress (CCC'12). pp. 1_6 (2012).
16. Rosales, R., Rodriguez, D., Flores, D., Palafox, L., Castanon-Puga, M., Gaxiola-Pacheco, C.: Model of interaction among embedded agents in ubiquitous computing environments. In: Proc. of the International Conference on Information Society (i- Society'12). pp. 501_502 (2012).
17. Sadri, F.: Ambient intelligence: A survey. ACM Computing Surveys (CSUR) 43(4), 36:1_36:66 (Oct 2011).
18. Schmidt, K., Bannon, L.: Taking cscw seriously: Supporting articulation work. Computer Supported Cooperative Work (CSCW). pp. 1_2 (1992).
19. Weiser, M.: Some computer science issues in ubiquitous computing. Commun. ACM 36(7), 75_84 (Jul 1993).

Consultas Móviles Conscientes del Contexto: Retos y Oportunidades

Jorge G. Hernández-Calderón, Edgard Benítez-Guerrero, Carmen Mezura-Godoy

Facultad de Estadística e Informática, Universidad Veracruzana, Xalapa, México 91020
scholar.gc@gmail.com, edbenitez@uv.mx, cmezura@uv.mx

Resumen. Un Sistema de Información Móvil Consciente del Contexto (SIM-CC) aprovecha el contexto del usuario nómada para proporcionarle información y/o servicios adecuados y oportunos. Este artículo ofrece un panorama general de los trabajos actuales sobre la expresión y el procesamiento de consultas a bases de datos en los SIM-CC, e identifica oportunidades de investigación en estos temas.

Palabras clave. Cómputo consciente del contexto, procesamiento de consultas, cómputo móvil.

1 Introducción

Un Sistema de Información Móvil (SIM) provee a sus usuarios acceso a información y/o servicios en cualquier lugar y en cualquier momento mediante una combinación de tecnologías de cómputo estacionarias y móviles [1]. Por ejemplo, hoy en día son comunes aquellos sistemas que permiten al usuario nómada encontrar un restaurante, reservar un vuelo o un hotel, comprar boletos para el cine o acceder a información empresarial desde su teléfono o tablet. En un ambiente móvil, los sistemas deben enfrentar diversas dificultades, como desconexión de la red, recursos (batería, memoria) limitados, e incluso problemas ligados a la seguridad y la privacidad del usuario.

Los SIM actuales han demostrado su utilidad en diversos ámbitos aplicativos; sin embargo, se espera que en la siguiente generación estos sean"conscientes del contexto". La consciencia contextual es una característica que se le puede dotar a un software, de forma que detecte su contexto de ejecución y lo use para ofrecerles a sus usuarios información y/o servicios adecuados y oportunos. Esta característica puede aplicarse a sistemas estacionarios, ubicuos o, como en este caso, móviles. Por ejemplo, un Sistema de Información Móvil Consciente del Contexto (SIM-CC) sobre restaurantes puede proveerle a un usuario información sobre restaurantes cercanos a su ubicación, considerando que este es el dato contextual relevante a los requerimientos del usuario. Los sistemas conscientes del contexto (particularmente los basados en la ubicación) han proliferado, debido a la posibilidad de acceder a ellos en ambientes móviles y ubicuos.

Es importante señalar que contar con un SIM-CC puede ser ventajoso, pero su desarrollo es un trabajo arduo. Además de tener que realizar las tareas propias del desarrollo de software (análisis de requerimientos, diseño, implementación, pruebas), se hace necesario diseñar y poner en operación mecanismos para detectar automáticamente el contexto en el que se está usando el sistema y para utilizarlo. Desde una perspectiva de Bases de Datos, el contexto puede ser usado durante la gestión transaccional (para, por ejemplo, saber qué hacer con una transacción bancaria cuando se está acabando la batería del móvil) o bien durante la consulta de datos (para obtener, por ejemplo, los restaurantes más cercanos a la ubicación del usuario y su móvil).

Una consulta móvil consciente del contexto entonces es una consulta móvil en la que se considera el contexto con el objetivo de obtener resultados que se ajusten mejor a las necesidades del usuario. Las aplicaciones potenciales de este tipo de consultas son numerosas, por ejemplo: buscar los amigos más cercanos, listar las películas que sean del género preferido por el usuario y que sean exhibidas en el centro comercial donde se encuentra o listar las refaccionarias que estén abiertas en este momento. Procesar una consulta de este tipo requiere resolver varios problemas, que van desde recuperar el contexto actual, hasta utilizarlo en la consulta y en su procesamiento. Este documento presenta un panorama general del procesamiento de consultas móviles conscientes del contexto, identificando los retos y las oportunidades de investigación en esta área.

El resto de este documento se encuentra organizado de la siguiente manera. Primero, se presentan una serie de conceptos generales. Después, se presenta un marco de referencia para las consultas conscientes del contexto y su procesamiento, para posteriormente presentar oportunidades de investigación identificadas a partir de ese marco general. Finalmente, se concluye este documento.

2 Conceptos generales

Esta sección presenta brevemente algunos conceptos generales sobre el contexto y sus características, así como sobre los sistemas conscientes del contexto.

2.1 Contexto y sistemas conscientes del contexto

No existe una única definición de contexto. Diversos autores lo definen de manera diferente para diferentes propósitos, por ejemplo, una definición típica es la de “cualquier información usada para caracterizar la situación de una entidad, siendo una entidad una persona, lugar o cosa que se pueda considerar relevante para la interacción entre un usuario y una aplicación” [3]. Una definición alternativa, desde el punto de vista del manejo y la administración de los datos es la propuesta por que indica que el contexto es “la situación bajo la cual un usuario accede a una base de datos” [5].

Independientemente de cómo se defina al contexto, éste se caracteriza como un conjunto de datos relacionados al usuario y a su entorno. Estos datos pueden ser recuperados de fuentes de datos físicas, como los sensores, o lógicas, como la agenda del usuario. La Tabla 1 muestra algunos ejemplos de tipos de datos contextuales, así como sus posibles fuentes. Cabe señalar que, en el caso de los datos contextuales ”lógicos”, como el estado de ánimo del usuario o la actividad que éste realiza, se han hecho propuestas para determinarlo automáticamente a partir de datos obtenidos de sensores físicos, pero el problema aún no ha sido completamente resuelto.

Tabla 1. Clasificación de los datos contextuales.

	Tipo de dato contextual	Fuente
Fuente física	Ubicación	GPS
	Tiempo	Reloj
	Luz ambiental	Fotómetro
	Ruido	Micrófono
	Proximidad	Sensor de proximidad
	Velocidad	Acelerómetro
	Dirección	Brújula
	Posición del dispositivo	Giroscopio
Fuente Lógica	Contactos de usuario	Agenda de contactos
	Estados de ánimo	Perfil del usuario
	Actividades	Correo electrónico/agenda

Por su parte, un Sistema de Información Móvil Consciente del Contexto es un SIM que aprovecha los datos contextuales. En los SIM-CC, el contexto se emplea para diversos propósitos, que van desde adaptar interfaces de usuario, entrelazar datos, mejorar la precisión de la información entregada al usuario hasta descubrir servicios. Así mismo, las características presentes en estos sistemas pueden ser todas o algunas de las siguientes [8]:

1. Autonomía para operar con la intervención mínima del usuario.
2. Proactividad para manejar el contexto actual pero también para actuar en anticipación de metas o problemas.
3. Reconocimiento de otros dispositivos en proximidad.
4. Provisión transparente de servicios de localización para usuarios nómadas.

En la actualidad es posible identificar un número creciente de SIM-CC; en [4] estos sistemas se categorizan en dos clases: (i) de uso personal cotidiano, como teléfonos móviles conscientes del contexto, administración de información personal o aprendizaje móvil, y, (ii) sistemas orientados a servicios ofrecidos por instituciones dentro de una área geográfica específica; por ejemplo, guías de turistas, publicidad móvil, o de apoyo al cuidado de la salud.

2.2 Características de los datos contextuales

En [2] se identifican categorías de datos contextuales comúnmente modeladas: tiempo, espacio, sujeto del contexto, perfil de usuario e historial del contexto. El tiempo y el espacio son significativos para los SIM-CC, ya que el contexto varía no solo en función del tiempo, sino también con respecto a la ubicación de los usuarios. El sujeto de contexto se refiere al punto de vista empleado para describir el contexto, ya sea desde la perspectiva del usuario o de la aplicación. El perfil del usuario se considera como parte del contexto, ya que hay variaciones en la forma de representar un perfil y asignarlo a un usuario, ya sea por medio de roles o de manera individual. Finalmente, el historial de contexto se modela para reconocer comportamientos del usuario al paso del tiempo.

Los datos contextuales se caracterizan por tener las siguientes propiedades [7]:

- **Naturaleza estática/dinámica.** Los datos contextuales pueden ser estáticos como un nombre de usuario, o dinámico como la ubicación.
- **Imperfección.** Pueden ser erróneos, ambiguos, indisponibles o incompletos. Por ejemplo, no es tan sencillo obtener la ubicación exacta de un dispositivo móvil y así se vuelve incierto.
- **Diferentes niveles de precisión.** Los datos contextuales pueden ser descritos a diferentes niveles de precisión; por ejemplo, la ubicación puede ser descrita como una coordenada geográfica o como un código postal.
- **Derivación.** Se pueden derivar a partir de otros; por ejemplo, se puede derivar el nombre de la ciudad en la que se encuentra un usuario a partir de una coordenada geográfica.
- **Cambios discretos/continuos.** Los datos contextuales pueden cambiar de forma discreta, como un cambio de actividad de un usuario o modificaciones a su perfil, o de forma continua, como el nivel de ruido en el ambiente. Cabe mencionar que hay datos que se pueden manejar tanto de forma discreta como continua, y su uso dependerá del tipo de aplicación que los utiliza, por ejemplo, la ubicación puede representarse en ambas modalidades, continua como coordenadas geográficas o discreta como un código postal.

2.3 Arquitectura conceptual de un SIM-CC

Las funciones que un SIM-CC debe implementar para ser considerado como consciente del contexto son la adquisición, la organización y la administración y uso de los datos contextuales. La figura 1 muestra la arquitectura conceptual típica de un SIM-CC [1].

En el extremo izquierdo de la figura 1 se encuentran las fuentes de datos contextuales, que son potencialmente heterogéneas, exponiendo cada una de ellas los datos en su propio formato y proveyendo cada una su propio método de acceso, y autónomas, actuando independientemente una de la otra. En un SIM-CC, el componente responsable de la adquisición del contexto tiene que interactuar con las fuentes de datos para colectar los datos contextuales básicos, transformarlos (de ser necesario) y producir información contextual nueva de más alto nivel.

Al centro de la figura 1 se pueden encontrar aspectos relacionados al almacenamiento, gestión y consulta de datos contextuales. Cabe señalar que la Base de Datos Contextuales es un componente lógico en la arquitectura y que puede o no ser materializada físicamente. En todo caso, se considera que tal base requiere de un modelo para representar los datos y un lenguaje de consulta asociado.

En el extremo derecho de la figura 1 lo relativo al uso del contexto. Desde el punto de vista de la gestión de datos móviles, el contexto puede ser empleado durante la expresión y procesamiento de consultas, así como para la gestión transaccional. Con respecto a la consulta de bases de datos, el contexto puede ayudar a obtener resultados que se ajusten mejor a las necesidades del usuario; por ejemplo, si un usuario hace una consulta sobre restaurantes, es probable que le interesen aquellos que están cercanos a su ubicación; así, una consulta convencional se puede convertir en una donde el contexto sea considerado para darle información potencialmente más útil. Por otra parte, el uso del contexto puede ayudar a mejorar la tasa de transacciones exitosamente validadas (*commited*), al tratar con la variabilidad del ambiente móvil.

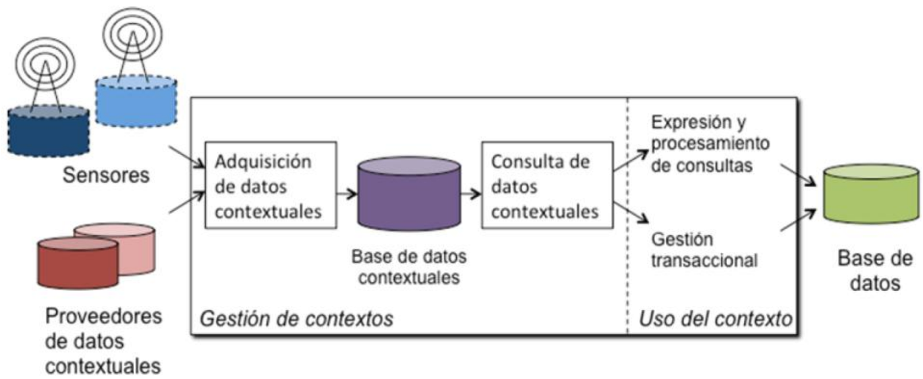


Fig. 1. Arquitectura conceptual de los SIM-CC [1].

3 Consultas móviles conscientes del contexto y su procesamiento

Esta sección presenta los tipos de consulta que se pueden ejecutar en un ambiente móvil, así como el procesamiento tradicional de consultas conscientes del contexto.

3.1 Tipos de consultas y modos de ejecución

Las consultas que se pueden ejecutar en un ambiente móvil pueden dividirse en dos grandes tipos: las no relativas al contexto y las conscientes en el contexto. El primer tipo son consultas tradicionales independientes del contexto (por ejemplo “recuperar los artículos vendidos en el mes”). El segundo tipo se refiere a las consultas cuyo resultado dependerá del contexto del usuario; por ejemplo “recuperar los restaurantes más cercanos”. Es precisamente la ubicación la que ha sido más utilizada como dato contextual, aunque este, para ciertas aplicaciones, puede resultar irrelevante (por ej. ”listar los nombres de los artículos que puedan ser útiles a la discusión que tengo ahora con mis colegas”).

Los modos de ejecución de una consulta (independientemente si es o no consciente del contexto) son dos: discreto (también llamado *snapshot*) y continuo. En el modo discreto una consulta se ejecuta una sola vez, mientras que en el modo continuo una consulta se sigue ejecutando ya sea a intervalos regulares de tiempo o cuando nuevos datos llegan a formar parte del resultado de la consulta. Por ejemplo, la ejecución de "buscar los hoteles más cercanos a mi ubicación" daría, en modo discreto, los datos de los hoteles que cumplan con la condición en el punto geográfico en el que se hizo la consulta, mientras que la consulta "reportar cada 5 minutos los hoteles más cercanos a mi ubicación", en modo continuo, regresaría los hoteles que cumplan la condición en el tiempo especificado.

3.2 Procesamiento de consultas conscientes del contexto

El procesamiento de consultas tradicionales ha sido plenamente estudiado en el área de Bases de Datos, donde se han propuesto técnicas de análisis sintáctico y semántico, así como de optimización heurística y basada en costos, para tratar consultas a bases de datos locales y distribuidas. Más recientemente, dichas técnicas se han extendido para procesar consultas en ambientes móviles, donde las desconexiones no se tratan como errores, sino más bien como una característica normal del ámbito móvil. Por su parte, el procesamiento de consultas conscientes del contexto comienza a ser estudiado.

La figura 2 muestra un marco general para el procesamiento de consultas conscientes del contexto propuesto en [5]. Cabe señalar que esta propuesta fue originalmente realizada para ambientes ubicuos pero es lo sufientemente general como para aplicarla a ambientes estacionarios y móviles. En la figura 2 se puede observar que una consulta *q* (por ejemplo, "listar información sobre restaurantes"), independientemente del lenguaje en el que esté escrita (por ej. SQL), pasa por tres etapas: pre-procesamiento, ejecución y post-procesamiento.

En la etapa de pre-procesamiento, el Coordinador de Consultas Conscientes del Contexto refina q con restricciones generales descritas en términos de atributos contextuales para obtener una consulta *q'* (por ej., "listar información sobre restaurantes en 100 metros a la redonda de mi posición P"). Dicha consulta *q'* es convertida en *q''* asignando valores concretos a los atributos contextuales (por ej., "listar información sobre restaurantes en 100 metros a la redonda de mi posición <20 N, -98 O>"); los valores concretos son proporcionados por el Administrador de Contextos, que es el encargado de recuperar y gestionar datos contextuales. En la etapa de ejecución, la consulta *q''* es enviada a las bases de datos públicas, a través del Comunicador con Servidores de Bases de Datos, para obtener la respuesta correspondiente. La etapa de post-procesamiento comienza cuando la respuesta a la consulta está lista para entregarse al usuario. Primero, el resultado *a* es ordenado de acuerdo al contexto para obtener *a'* (por ej., los restaurantes listados de acuerdo a la distancia a la posición del usuario), para finalmente entregarle este resultado al usuario de acuerdo al contexto (por ej. si está manejando, el resultado podría ser "leído" por el móvil utilizando una función de síntesis de voz).

Cabe señalar que esta propuesta de marco de referencia es importante porque identifica explícitamente los elementos necesarios para procesar una consulta consciente del contexto. Sin embargo, existen un conjunto de problemas que no se consideran, los cuales se discuten en la siguiente sección.

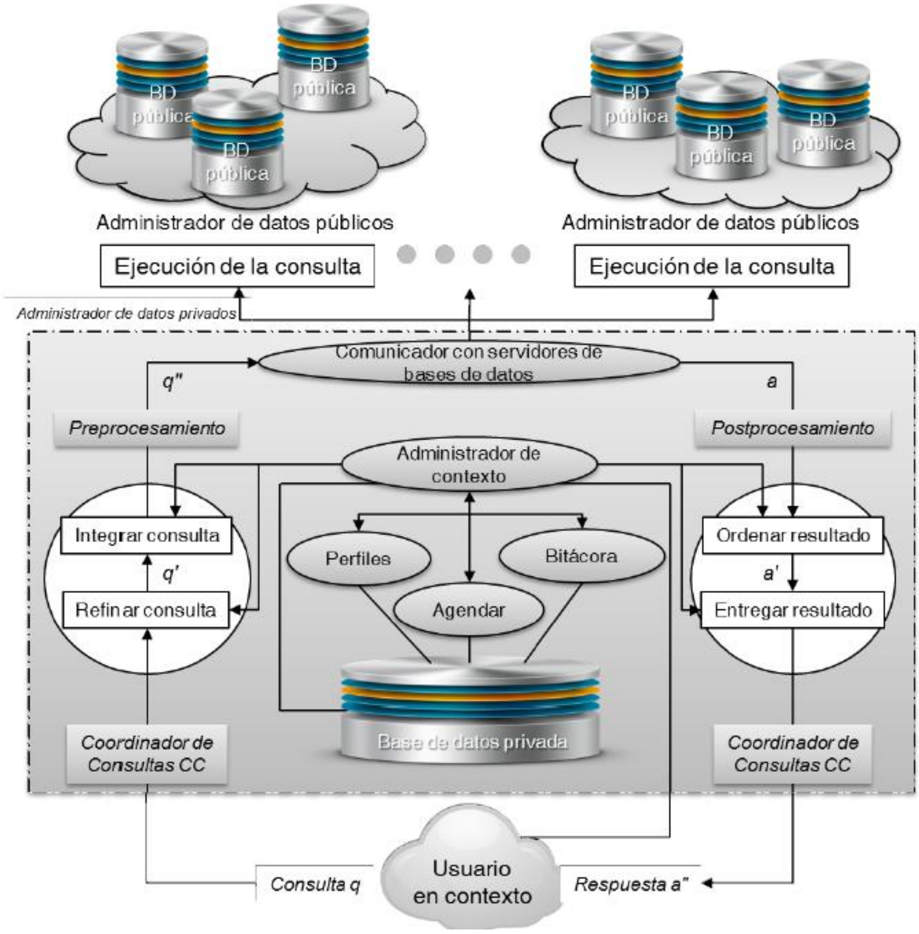


Fig. 2. Procesamiento de Consultas Conscientes del Contexto [5].

4 Retos y oportunidades detectados

Esta sección identifica retos y oportunidades relacionados con las consultas móviles conscientes del contexto y su procesamiento, particularmente en lo que respecta a la adquisición y consulta del contexto así como el procesamiento de consultas ante cambios contextuales a nivel de disponibilidad de recursos de cómputo o a nivel de aspectos del propio usuario y sus necesidades.

4.1 Adquisición y consulta del contexto

Uno de los problemas fundamentales de los sistemas conscientes del contexto es la adquisición del mismo. En algunos casos, los datos contextuales requeridos pueden ser relativamente fáciles de obtener (p. ej., los cambios de posición en el espacio de un dispositivo móvil por medio de un acelerómetro), pero en otros casos, como por ejemplo la actividad realizada por el usuario, se vuelve más difícil. En todo caso, como se mencionó anteriormente, estos provienen de múltiples fuentes cuya heterogeneidad y autonomía hacen que se vuelva difícil tener datos integrados en el formato requerido en el momento oportuno. Una posible solución que ha sido propuesta es la de establecer una infraestructura para la adquisición del contexto. En la actualidad, existen varias propuestas las cuales ofrecen APIs para que las aplicaciones hagan uso de los datos contextuales que captan e infieren. Sin embargo, no existe un método de acceso uniforme que pueda ser aprendido una vez por los desarrolladores de sistemas y utilizado sin importar diferencias entre infraestructuras y/o fuentes[1].

Dada esta situación, el considerar que lógicamente existe una base de datos contextuales que puede ser accedida mediante un lenguaje de consulta estándar (como indicado en la figura 1) parece una solución razonable. Esta situación ha comenzado a ser estudiada y varios modelos y sistemas de bases de datos (lógicos, relacionales, semiestruturados, entre otros) se han propuesto para organizar y almacenar tal base [2]. No obstante, aun hacen falta trabajos que investiguen la expresión y la ejecución de consultas discretas y continuas sobre la base de datos contextuales que traten con la imperfección y los diferentes niveles de derivación de estos.

4.2 Cambios en el contexto de cómputo

En un sistema móvil, tradicionalmente se usa una arquitectura cliente-servidor en el que al realizar una consulta desde el dispositivo móvil, ésta se envía al servidor para su procesamiento y los resultados son enviados al móvil. Este esquema ha sido ampliamente utilizado dadas las limitaciones en cuanto a recursos (batería, memoria y procesador) de los dispositivos móviles de gama baja. Dado que el ambiente en el que se realiza este proceso es móvil, surgen complicaciones tales como estados de conexión inconsistentes, insuficiente ancho de banda, falta de cobertura, elevados costos de transmisión de datos, etc. En este momento, la arquitectura cliente-servidor se vuelve ineficiente, ya que al realizar consultas de esta manera se corre el riesgo de no recibir respuesta alguna.

Una posible solución sería la de permitir que el dispositivo móvil participe en el procesamiento de la consulta, por ejemplo derivando datos localmente o incluso calculando respuestas aunque sea parciales. Esto es posible dado que los dispositivos actuales de gama media y alta cuentan con recursos suficientes para participar; sin embargo, aún se debe considerar que la batería o el tiempo de CPU pueden no ser suficientes y esto debe considerarse en el procesamiento de la consulta [6]. En este momento se hacen necesarias estrategias que utilicen adecuadamente los recursos en el móvil y en el servidor, teniendo que cambiar el lugar de la ejecución de una consulta incluso en tiempo de ejecución. Un primer esfuerzo en este sentido se encuentra en [9].

---

[1] Esta situación se torna similar a la que se tenía en los años 1960s cuando no existían aún los sistemas de gestión de bases de datos y existían diferentes clases de sistemas de archivos.

4.3 Cambios en el contexto del usuario

Este problema se refiere al hecho de que el contexto (mental, físico, social) del usuario también es cambiante. Supóngase que se emplea la ubicación como dato de contexto para una aplicación. Si el usuario realiza una consulta Q cuando está en un punto A mientras se dirige a un punto B, cuando lleguen los resultados de Q a B estos serán válidos para A, pero no así para B, lo que provocaría frustración en el usuario porque los datos no satisfacen sus requerimientos. Luego, es necesario incluir el cambio de ubicación cuando se procesa la consulta [6]. En [10] se propone un algoritmo que permite considerar esta situación incluyendo la proyección de futuras ubicaciones del usuario buscando minimizar el impacto del cambio de ubicación. Investigaciones que generalicen estos resultados para considerar no tan solo a la ubicación sino a cualquier otro tipo de dato contextual, como el estado de ánimo del usuario, sus actividades o las personas con las que se encuentre o interactúe, son necesarias.

5 Conclusión

Este documento muestra el estado del arte del procesamiento de consultas conscientes del contexto para la consulta efectiva de información en sistemas móviles. La idea es básicamente extraer información transparentemente para que, al momento de que el usuario realice una consulta, ésta pueda ser complementada con información de su contexto para devolverle resultados relevantes en un momento dado. Sin embargo, aparte de las inherentes dificultades propias de los ambientes móviles, se hace necesario tratar con problemas propios a la consciencia contextual, como lo son la adquisición y consulta del contexto así como el procesamiento de consultas ante condiciones contextuales cambiantes. En el primer caso, se hacen necesarios mecanismos que permitan el acceso homogéneo y uniforme a múltiples fuentes de datos contextuales potencialmente heterogéneas y autónomas; en este momento, el considerar una base de datos contextuales (virtual o física) con un lenguaje de consulta (y su procesador asociado) capaz de contender con la naturaleza de los datos contextuales parece una solución prometedora. En el segundo caso, se hace necesario adaptar el procesamiento de consultas tradicional para que pueda tratar cambios contextuales repentinos y de esta forma se entreguen a los usuarios resultados que sean válidos y/o oportunos en nuevos contextos. Resolver estos problemas permitirá construir los sistemas móviles conscientes del contexto del futuro.

Reconocimientos

Este trabajo es parte del proyecto "Herramientas para la gestión de datos en sistemas móviles dependientes del contexto"(ref. CONACYTSEP 83619). El primer autor es apoyado por CONACYT (beca no. 350127).

Referencias

1. Benítez-Guerrero, E.: Context-Aware Mobile Information Systems: Data Management Issues and Opportunities. In: Proc. of the 2010 Int. Conference on Information & Knowledge Engineering (IKE 2010). pp. 127–133. CSREA Press, Las Vegas, USA (2010).
2. Bolchini, C., Curino, C.A., Quintarelli, E., Schreiber, F.A., Tanca, L.: A Data-Oriented Survey of Context Models. ACM SIGMOD Record 36(4) (2007).
3. Dey, A., Abowd, G.: Towards A Better Understanding of Context and Context-Awareness. In: CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness. pp. 304–307 (2000).
4. Dongsong, Z., Adipat, B., Mowafi, Y.: User-Centered Context-Aware Mobile Applications: The Next Generation of Personal Mobile Computing. Communications of the Association for Information Systems 24(3), 27–46 (Dec 2009).
5. Feng, L., Apers, P.M., Jonker, W.: Towards Context-Aware Data Management for Ambient Intelligence. In: Proc. of DEXA '04. pp. 422–431 (2004).
6. Grine, H., Delot, T., Lecomte, S.: Adaptive Query Processing in Mobile Environments. In: Proc. of the 3rd Int Workshop on Middleware for Pervasive and Ad-hoc Computing. pp. 1–8. MPAC '05, ACM, New York, NY, USA (2005).
7. Haghighi, P.D., Zaslavsky, A., Krishnaswamy, S.: An Evaluation of Query Languages for Context-Aware Computing. In: Proc. DEXA 2006. pp. 455–462 (2006).
8. Hong, J.Y., Suh, E.H., Kim, S.J.: Context-Aware Aystems: A Literature Review and Classification. Expert Systems and Applications 36(4), 8509–8522 (May 2009).
9. Iglesias, J., Bernardos, A., Alvarez, A., Sacristan, M.: A Light Reasoning Infrastructure to Enable Context-Aware Mobile Applications. In: 2010 IEEE IFIP Int. Conf. on Embedded and Ubiquitous Computing. pp. 386–391 (2010).
10. Mokbel, M.F., Xiong, X., Aref, W.G.: SINA: Scalable Incremental Processing of Continuous Queries in Spatio-Temporal Databases. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data. pp. 623–634. SIGMOD '04, ACM, New York, NY, USA (2004).

Review Committee

Comité revisor

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
noviembre de 2013
Printing 500 / Edición 500 ejemplares

www.ipn.mx
www.cic.ipn.mx

ISSN: 1870-4069

Research in Computing Science