Editorial

It is my pleasure to present the readers of the journal the selection of papers devoted to the theme of the Artificial Intelligence, which is one of the breathtaking topics of the Computer Science.

This issue contains twelve papers written by the authors from eight countries that cover various topic of the Artificial Intelligence: image analysis, reasoning, and natural language processing.

The paper "*Automatic Bubble Detection in Cardiac Video Imaging*" by Ismail Burak Parlak, Ahmet Ademoglu, Salih Murat Egi, Costantino Balestra, Peter Germonpre, and Alessandro Marroni (Turkey, Belgium) presents a recognition method for bubble detection that is applied to cardiac videos and allows for emboli candidates revealing.

The work "*Automated Classification of Bitmap Images using Decision Trees*" by Pavel Surynek and Ivana Lukšová (Czech Republic) addresses the design of a method of automated classification of bitmap images on the basis of their natural language descriptions.

A problem of automatic recognition of emotions present in voice data is treated in the paper "*Automatic Emotional Speech Recognition with Alpha-Beta SVM Associative Memories*" by José Francisco Solís Villarreal, Cornelio Yáñez Márquez and Sergio Suárez Guerra (Mexico). The authors present the model of application of a special type of associative memories to the voice data coded with speech energies.

Rafael A.M. Gonçalves, Diego R. Cueva, Marcos R. Pereira-Barretto, and Fabio G. Cozman (Brasil) present a paper "*A Dynamic Model for Identification of Emotional Expressions*" that presents a model for basic emotion recognition on the basis of face analysis at video frames using Kalman filtering.

The paper "*Optical Parameter Extraction using Differential Evolution Rendering in the Loop*" by Mauricio Olguin Carbajal, Ricardo Barrón Fernández, and José Luis Oropeza Rodríguez (Mexico) presents a method for evaluation of optical parameters like illumination, using differential evolution algorithm.

A model of decision making based on persuasion that is applied to massively multiplayer online role-playing games is presented in the paper "*A Model of Decision-Making Based on the Theory of Persuasion used in MMORPGs*" by Helio C. Silva Neto, Leonardo F. B. S. de Carvalho, Fábio Paraguaçu, and Roberta V. V. Lopes (Brasil). The formalism of Petri nets is used.

Andrey Ronzhin, Jesus Savage, and Sergey Glazkov (Russia, Mexico) present in their paper "*User Preference Model for Conscious Services in Smart Environments*" important issues related to the smart environments, namely, the user preferences in given contexts.

The paper "*FPGA Implementation of Fuzzy Mamdani System with Parametric Conjunctions Generated by Monotone Sum of Basic t-Norms*" by Prometeo Cortés Antonio, Ildar Batyrshin, Herón Molina Lozano, Marco Antonio Ramírez Salinas, and Luis Villa Vargas (Mexico) presents an implementation of certain complex logic functions in FPGA.

The problem of automatic music generation and its evaluation by humans is the theme of the paper "*Automatic Music Composition with Simple Probabilistic Generative Grammars*" by Horacio Alberto García Salas, Alexander Gelbukh, Hiram Calvo, and Fernando Galindo Soria (Mexico).

The paper "*An Approach to Cross-Lingual Textual Entailment using Online Machine Translation Systems*" by Julio Castillo and Marina Cardenas (Argentina) deals with the problem of textual entailment that is an edge in natural language processing nowadays. The authors analyze cross-lingual textual entailment using machine translation systems available on the web.

Maria Chondrogianni (UK) in her paper "*Identifying the User's Intentions: Basic Illocutions in Modern Greek*" deals with various manners of expressing user intentions analyzing basic illocutions in Modern Greek and the ways of their expressions.

The paper "*Inference of Fine-grained Attributes of Bengali Corpus for Stylometry Detection*" by Tanmoy Chakraborty and Sivaji Bandyopadhyay (India) presents stylometric features based on lexical markers that are obtained from a corpus. The experiments are made for Bengali language corpus, though the system is language independent.

The papers selected for publication in this thematic issue give the reader a wide panorama of the methods used in Artificial Intelligence.

Grigori Sidorov
Research Professor,
Center for Computing Research,
National Polytechnic Institute,
Mexico City, Mexico

Automatic Bubble Detection in Cardiac Video Imaging

Ismail Burak Parlak, Ahmet Ademoglu, Salih Murat Egi, Costantino Balestra, Peter Germonpre, and Alessandro Marroni

***Abstract*—Bubble recognition is a challenging problem in a broad range from mechanics to medicine. These gas-filled structures whose pattern and morphology alter in their surrounding environment would be counted either manually or with computational recognition procedures. In cardiology, user dependent bubble detection and temporal counting in videos require special trainings and experience due to ultra fast movement, inherent noise and video quality. In this study, we propose an efficient recognition routine to increase the objectivity of emboli detection. Firstly, we started to compare five different methods on two synthetic data sets emulating cardiac chamber environment with increasing speckle noise levels. Secondly, real echocardiographic video records were segmented by variational active contours and Left Atria (LA) were extracted. Finally, three successful methods in simulation were applied to LAs in order to reveal candidate bubbles on video frames. Our detection rate of proposed method was 95.7% and the others were 86.2% and 88.3%. We conclude that our approach would be useful in long lasting video processing and would be applied in other disciplines.**

***Index Terms*—Image thresholding, active contours, venous emboli, echocardiography.**

I. Introduction

In different disciplines, several approaches are developed to detect bubbles and cavitations. These gas-filled structures are generally formed within objects, surfaces, liquids, thin films and inner layers. In solid state mechanics, thermodynamics and metallurgy, bubbles and cavitations are generally locked and have a non-moving nature. On the other hand, their dynamics in fluids are characterized through the viscosity of surrounding environments and physical properties of inner media. Transition between different layers, the effect of non-newtonian fluids would cause considerable variations in their behaviors. Even if bubble models offer a generalization for quantification and motion estimation in these fields, bubble monitoring and recognition in medicine still conserve unresolved problems.

In medicine, bubbles so-called emboli are created in endothelial tissues and are transported with veins to the heart. Though the bubble visualization in time is a challenge for an untrained clinician, human wise bubble recognition is a spatial problem due to turbulence, endocardial tissues, blood transportation and especially high level inherent noise. For healthy subjects embolus would be filtered in circulation, lung shunts or at most they would trigger migraine. However in risky groups, they would cause severe diseases in a broad range from stroke to blindness. For this purpose clinicians are mostly focused on Left Atrium (LA) and Pulmonary Artery (PA). Bubble examination is performed through video streams. Echocardiologists refer to manually selected region of interests (ROI) and try to discriminate moving objects which are labeled as bubbles. These moving objects might have nonlinear nature as Postema et al. [1] schematized such as translation, fragmentation, clustering, jetting and cracking.

Initial attempts to detect bubbles in circulation were based to Doppler ultrasonography. Embolus which are travelling through superior vena cava are classified manually as candidate bubbles when a sound peak from baseline or mean in frequency domain is observed.

Computational procedures were developed to automatize these methods which would cause variations between clinicians or mislead grading methods of bubbles for diagnosis purposes [2], [3].

Moreover, different computational approaches in other disciplines proposed recognition solutions for spatially non-moving bubbles or in low noise levels which would not mislead iterative algorithms or cost functions. Snakes[4], [5], [6], contour based models [7], principal component analysis[8], gradient based thresholding methods[9], [10] were utilized by different groups. Even if these methods would give accurate detection results in single frames, they would cause big time delays in videos. Furthermore, we note that speckle noise, low resolution and partial view of cardiac chambers are other bottlenecks in echocardiography.

Threshold based methods would be preferred in video frames if computational processing time does not cause delay and false alarms are low in recognition results. Researchers in medicine developed frame based semi-automatic approaches.

Manuscript received May 23, 2011. Manuscript accepted for publication August 25, 2011.

I.B. Parlak is with the Institute of Biomedical Engineering, Bogazici University and the Department of Computer Engineering, Galatasaray University, Ciragan Cad. No:36 34257 Istanbul, TURKEY e-mail:bparlak@gsu.edu.tr.

A. Ademoglu is with the Institute of Biomedical Engineering, Bogazici University, Istanbul, TURKEY.

S.M. Egi is with the Department of Computer Engineering, Galatasaray University, Istanbul, TURKEY.

C. Balestra is with the Environmental & Occupational Physiology Lab., Haute Ecole Paul Henri Spaak, Brussels, BELGIUM.

P. Germonpre is with the Centre for Hyperbaric Oxygen Therapy, Military Hospital, Brussels, BELGIUM.

A. Marroni is with Divers Alert Network (DAN) Europe Research Committee, Brussels, BELGIUM.

Brubakk et al. [11] proposed intensity and spatial thresholding algorithm in manually selected ROIs. Norton et al. [12] developed a quantification based on temporal change in opacification through cardiac chambers. These methods brought an expansion in this area as first automated analyzes.

In this study, we developed a spatio-temporal method for bubble quantification. Both simulation and real echogenic records were examined through different procedures with additive noise. In our algorithm, pixel series in acquired frames were thresholded dynamically by separating systolic and diastolic time intervals respectively. Moreover, all detected bubble wise structures were gathered into one frame. On the parallel side four different thresholding procedures were applied onto same data. Performance analysis reveals that our approach is satisfactory through false alarms. We hypothesize that our method provides better detection rates and increases the clinician subjective ease-of-use in terms of decision making in recognition.

II. Methodology

A. Simulation Videos

We started to create two different congenital atrial video records. Simulated frames were set as 160x120 pixels, the average size of segmented LA in real records. Each video stream was 1 sec long and 25 frames/sec (fps). Bubbles on simulation data were placed randomly as it is in real environment. Their contrasts were close to real embolus. In echocardiographic records, bubbles are travelling dynamically and same bubbles generally might be seen in two previous or posterior frames if there is not a massive opacification. This visual procedure is applied by clinicians to lower false alarms. In order to set same echocardiographic environment and bubble behavior, we either placed bubbles in previous and next frames by translating, rotating or removed. This simulation procedure is checked double blinded by two different clinicians. In order to evaluate the performance of recognition algorithms, we generated speckle noise using uniform distribution. This noise with mean μ; 0 and different variances σ was added to simulation frames. In simulation, we also adopted Germonpre et al. [13] criterion for bubble classification in congenital diseases. When bubble numbers in examined area is more or less than 20, subjects are grouped as Type 1 and Type 2, respectively.

B. Cardiac Videos

We acquired two contrast Transoesophageal Echocardiogram (cTEE) video records from two male professional divers. The study protocol was approved in advance by Centre for Hyperbaric Oxygen Therapy, Military Hospital, Ethics Committee. Each subject provided written informed consent to join the study.

Embolus detection and visualization protocol described by Germonpre, et al. [13] is utilized for each subject. Both divers underwent cTEE with agitated saline for contrast. All cTEE video frames were recorded from Ultrasound device (MicroMaxx, SonoSite Inc, WA) in high definition 640x480 pixels, avi format. For all subjects, acquisition was performed three times to ascertain human based grading by two echocardiologists as a double blind study.

C. Segmentation

Initially we started to perform our segmentation using active contours implemented via level set introduced by Caselles et al. [14]. In this approach contours are found using a Lagrangian formulation based on the evolution of parametrized curve. We remarked that the partial differential equation so called evolution is relatively slow in terms of computational time on video sequences. Therefore we adopted a modified level set formulation and combined the methods of Chan et al. and Vemuri et al. [15], [16].

An initial level set by fronts Γ is denoted by a distance function $\phi(x) = \pm d_\Gamma(x)$ A zero level set function is; $\Gamma(t) = (x, y) : \phi(x) = 0$ Given the frant Γ let $F(x)$ be the speed function in the direction of the normal of Γ and $x(t)$ be a point on Γ which evolves progressively then $\phi(x(t), t) \equiv 0$ for all t. When this expression is differentiated through t;

$$\frac{\partial \phi}{\partial t} + \nabla \phi \frac{dx}{dt} = 0 \tag{1}$$

Level set function has both positive and negative terms including zeroes and is called signed distance function;

$$\frac{\partial \phi}{\partial t} = sign(\phi)(1 - |\nabla \phi|) \tag{2}$$

We resolved this equation by interpreting without reinitialization. This approach is based on modified formulation that consists of two energy terms; internal and external. Internal term prevents the deviation of level set from signed distance function whereas external term conducts a motion on zero level set up to the final pattern features especially contours. A consequent evolution of this level set is a gradient flow and it minimizes the energy function as it is expressed in Equation 1 and 2. All digital records were segmented and analyzed in MATLAB 2010a (The MathWorks Inc, Massachusetts).

D. Detection Algorithms

In the review paper of Sezgin et al.[17], distinctive thresholding algorithms were classified through image analysis within different categories.

For our study, we have used four different thresholding methods; Otsu [18], Yen et al. [19], Ramesh et al. [20] and Beghdadi et al. [21] with distinctive nature on image analysis and recognition. Thresholding algorithms create binary level images using either RGB color or gray level images. After thresholding process, blobs would be recognized on frames.

Cardiac patterns and especially LA are composed of gray level patterns. In our simulation and real echogenic videos, bubbles would be easily identified through relevant

thresholds. However, it should be remarked that ultrasonic image processing is vulnerable to inherent speckle noise. This type of interference would cause misleading in detection, affect binary level image or introduce false alarms within target blobs. In our study, these blobs correspond to candidate bubbles.

All recognition results from each method were compared with human wise detection in order to perform the statistical rate of recognition. In simulation phase, we remarked that two methods are reliable in bubble detection for different noise levels. Therefore we applied them to real segmented cardiac videos.

E. Proposed Method

2D Cardiac images form a three dimensional data when frames are acquired sequentially with a device dependent fps. Therefore, all pixels in LA have a time series. It would be foreseen that when a bubble wise structure will be present on this pixel, its gray level will change suddenly. For this reason, a dynamic threshold which is applied to each pixel series would reveal candidate bubbles as it is shown in Fig.3.

This dynamic threshold is set using mean μ and σ of pixel time series. When a pixel value is above $\mu + 2\sigma$, this pixel is recognized as a bubble candidate in time. After this recognition procedure, all bubble candidates are summed up on corresponding frame. After detection, we added all bubbles into one single frame as a novelty. This single frame gathers all candidates and facilitates the visual recognition phase for a clinician. After automatic recognition statistical analysis was performed to compare real bubbles marked by two clinicians with computational recognition.

III. Results

In this paper, results are interpreted as simulation and real echogenic bubbles. In both steps, our proposed method offered better accuracy and low false alarms than existent methods.

In the simulation phase five methods including our method were tested on two different congenital forms through increasing noise levels. Simulation results were compared in Table I and Fig. 1. Only three methods were satisfactory in high level noises. Therefore, we selected them to test their performance in real data.

In real echogenic forms, recognition algorithms offered reliable methods as it is shown in Table II. However, existent methods were vulnerable to inherent noise and endocardial structures. It is noted in Fig. 2 that boundary structures could not be thresholded efficiently with Ramesh et al.[20] or Yen et al.[19]. Our proposed method detects bubbles with low false alarms. It is also evident that bubble detection map which bring all recognized candidates into one frame is a novelty in this paper. In Fig. 4, the visualization of all detected microembolus onto one single frame brings an ease-of-use for bubble movements.

IV. Discussion & Conclusion

Bubbles would be recognized accurately with different approaches in steady state environments without noise. On the other hand, medical imaging introduces always artifacts and inherent noise. Microemboli in cardiology are affected with speckle noise and their patterns are close to boundary structures. Therefore, their recognition which is a diagnostic tool for specialists in cardiology becomes a challenging problem in video processing.

In real echogenic frames, blood circulation in LA translates and rotates bubbles. They would be recognized as easily as in simulated frames in Fig. 1. However, during the circulation bubbles would be clustered, cracked or fragmented due their physical properties and turbulence. In these cases, a severe blurring causes false alarms. Moreover, it is noted that some fragments of endocardial wall would be detected as bubbles. In Fig. 2, recognition with Yen et al.[19] and Ramesh et al.[20] could not filter out endocardial fragments. Suboptimal image quality and acoustic shadowing which are the main artifacts in echogenic records are also another challenge for detection. They lower detection rates by inserting dashed or circular spots whose contrast is identical with real bubbles.

Our method benefits a reliable detection as it is noted in Table I and II with low false alarms. The only pitfall of proposed method is its computational time. It is evident that all pixel series should be interpreted to create frames containing only blobs; candidate bubbles.

As a feature work, we note that bubble behavior would be studied in other cardiac chambers or pulmonary artery using other imaging modalities to build up a computational framework in cardiac analysis.

Acknowledgment

The authors would like to thank Galatasaray and Bogazici Universities for their academical supports and also Divers Alert Network (DAN) for echocardiographic setup and imaging modalities.

References

[1] M. Postema and O. H. Gilja, "Contrast-enhanced and targeted ultrasound," *World J Gastroenterol*, vol. 17, 2011, pp. 28-41.

[2] K. Tufan, A. Ademoglu, E. Kurtaran, G. Yildiz, S. Aydin and S. M. Egi, "Automatic detection of bubbles in the subclavian vein using doppler ultrasound signals," *Aviat Space Environ Med.*, vol. 77, 2006, pp. 957-962.

[3] H. Nakamura, Y. Inoue, T. Kudo, N. Kurihara, N. Sugano and T. Iwai, "Detection of venous emboli using doppler ultrasound," *European Journal of Vascular & Endovascular Surgery*, vol. 35, 2008, pp. 96-101.

[4] J. Wang, X. Huang and Y. Zou, "Bubble detection of railway castings based on snake model," *Jisuanji Gongcheng / Computer Engineering*, vol. 36, 2010, pp. 205-207.

[5] K. H. Chung, M. J Simmons and M. Barigou, "Local gas and liquid phase velocity measurement in a miniature stirred vessel using PIV combined with a new image processing algorithm," *Experimental Thermal and Fluid Science*, vol. 33, 2009, pp. 743-753.

[6] D. C. Cheng and H. Burkhardt, "Bubble recognition from image sequences," in *Inverse Problems and Experimental Design in Thermal and Mechanical Engineering, Eurotherm Seminar N. 68*, 2001.

[7] D. C. Cheng and H. Burkhardt, "Bubble tracking in image sequences," *International Journal of Thermal Sciences*, vol. 42, 2003, pp. 647-655.

TABLE I
EVALUATION OF DETECTION RESULTS IN SIMULATED DATA

	Simulation1			Simulation2		
	$Original$	$\sigma = 0.07$	$\sigma = 0.5$	$Original$	$\sigma = 0.07$	$\sigma = 0.5$
Otsu[18]	100%	96.7%	6.4%	100%	97.3%	3.2%
Yen et al.[19]	100%	99.1%	97.5%	100%	98.4%	97.2%
Ramesh et al.[20]	98.7%	98.3%	96.7%	100%	99.1%	97.4%
Beghdadi et al.[21]	99.6%	5.2%	0%	100%	9.5%	0%
Proposed method	100%	98.8%	97.2%	100%	99.3%	98.1%

TABLE II
EVALUATION OF DETECTION RESULTS IN REAL CARDIAC VIDEOS

	Video 1 Detection Ration	Video 2 Detection Ration	Mean Ratio
Ramesh et al.[20]	85.8%	87.1%	86.2%
Yen et al.[19]	87.6%	88.9%	88.3%
Proposed method	93.8%	96.5%	95.7%

[8] D. C. Cheng and H. Burkhardt, "Template-based bubble identification and tracking in image sequences," *International Journal of Thermal Sciences*, vol. 45, 2006, pp. 321-330.

[9] M. Honkanen, H. Eloranta and P. Saarenrinne, "Digital imaging measurement of dense multiphase flows in industrial processes," *Flow Measurement and Instrumentation*, vol. 21, 2010, pp. 25-32.

[10] M. Honkanen, P. Saarenrinne, T. Stoor and J. Niinimaki, "Recognition of highly overlapping ellipse-like bubble images," *Measurement Science and Technology*, vol. 16, 2005, pp. 1760-1770.

[11] O. Eftedal and A. O. Brubakk, "Detecting intravascular gas bubbles in ultrasonic images," *Med Biol Eng Comput.* vol. 31, 1993, pp. 627-633.

[12] M. S. Norton, A. J. Sims, D. Morris, T. Zaglavara, M. A. Kenny and A. Murray, "Quantification of echo contrast passage across a patent foramen ovale," in:*Computers in Cardiology*, IEEE Press, 1998, pp. 89-92.

[13] P. Germonpre, P. Dendale, P. Unger and C. Balestra, "Patent foramen ovale and decompression sickness in sports divers," *Journal of Applied Physiology*, vol. 84, 1998, pp. 1622-1626.

[14] V. Caselles, F. Catte, T. Coll and F. Dibos, "A geometric model for active contours and image processing," *Numer. Math.* vol. 66, 1993, pp. 1-31.

[15] T. Chan and L. Vese, "Active contours without edges," *IEEE Trans Image Process*, vol. 10, 2001, pp. 266-277.

[16] B. Vemuri and Y. Chen, "Joint image registration and segmentation," in:*Geometric Level Set Methods in Imaging, Vision and Graphics*, Springer, 2003, pp. 251-259.

[17] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," in:*Journal of Electronic Imaging*, vol. 13, 2004, pp. 146-165.

[18] N. Otsu, "A thresholding selection method from gray-level histogram," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, 1979, pp. 62-66.

[19] J. C. Yen, F. J. Chang and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Trans. Image Process.* vol. 4, 1995, pp. 370-378.

[20] N. Ramesh, J. H. Yoo and I. K. Sethi, "Thresholding based on histogram approximation," *IEEE Proceedings Vision, Image and Signal Process.* vol. 142, 1995, pp. 271-279.

[21] A. Beghdadi, A. L. Negrate and P. V. De Lesegno, "Entropic thresholding using a block source model," *Graphical Models in Image Processing* vol. 57, 1995, pp. 197-205.

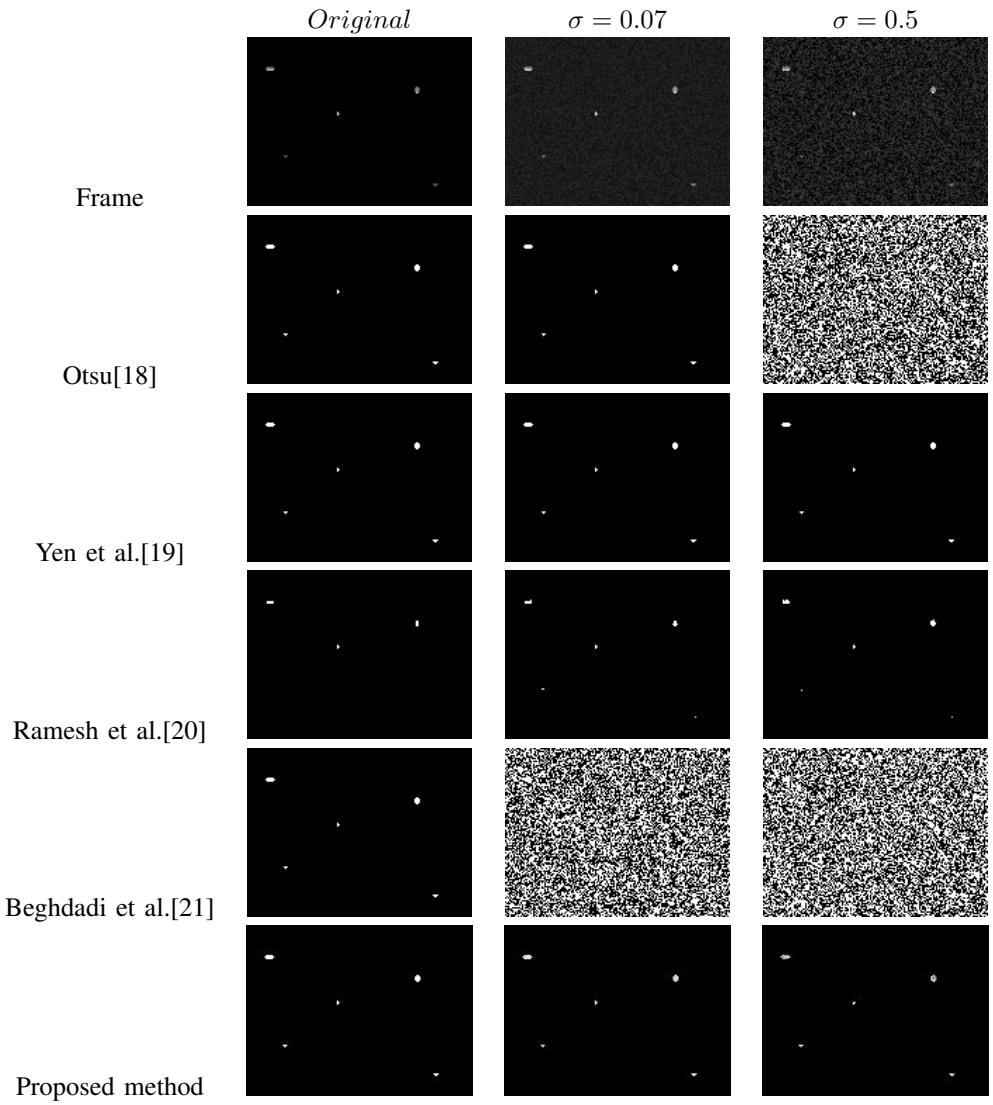


Fig. 1. Comparison of five different methods in simulated LA.

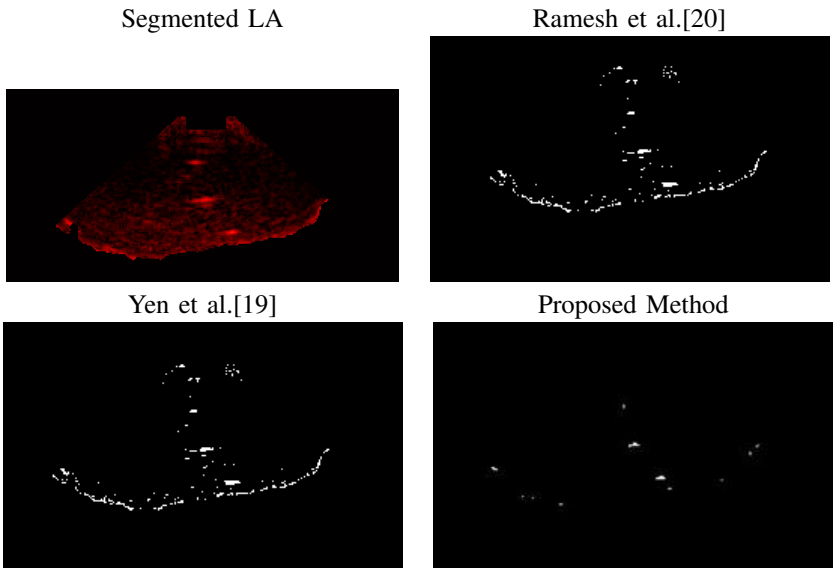


Fig. 2. Comparison of three different methods in TEE.

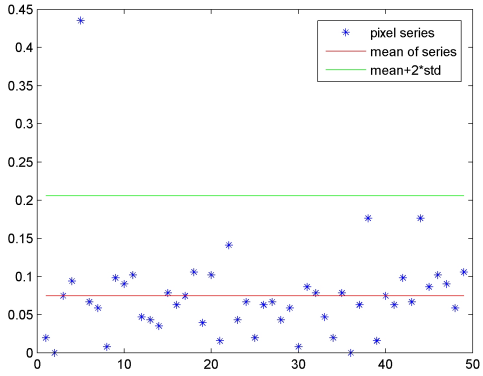


Fig. 3. Dynamic thresholding of pixel through intensity in proposed method.

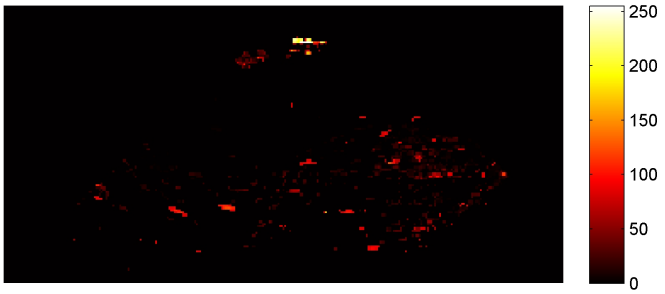


Fig. 4. Bubble map of video sequence.

Automated Classification of Bitmap Images using Decision Trees

Pavel Surynek and Ivana Lukšová

Abstract—**The paper addresses the design of a method for automated classification of bitmap images into classes described by the user in natural language. Examples of such naturally defined classes are images depicting buildings, landscape, artistic images, etc. The proposed classification method is based on the extraction of suitable attributes from a bitmap image such as contrast, histogram, the occurrence of straight lines, etc. Extracted attributes are subsequently processed by a decision tree which has been trained in advance. A performed experimental evaluation with 5 classification classes showed that the proposed method has the accuracy of 75%-85%. The design of the method is general enough to allow the extension of the set of classification classes as well as the number of extracted attributes to increase the accuracy of classification.**

Index Terms—**Image classification, attribute extraction, decision trees, learning.**

I. Introduction

As digital cameras (industrial or personal) are still more common, the demand for tools for the automated processing of recorded data increases. The large amount of data precludes its manual processing and represents the main reason for the requirement on automation. The typical situation many users of personal digital cameras often face is assorting several hundreds or even thousands of pictures from holiday which typically becomes a tedious and time consuming task. Similar tasks arise in industrial applications where it is often necessary to categorize automatically recorded images – for instance images from a surveillance camera that contain an interesting activity need to be distinguished from uninteresting ones.

This work is devoted to an analysis of static images – that is, *bitmap images*. The basic assumption is that the source of bitmap images is not restricted in any way. Thus they can be represented by photographs recorded by the digital camera or by artificially created images such as rendered images. The particular task we are solving in this work is the *classification of bitmap images* into a set of classes defined by an expression in natural language. These classes are represented for example by images depicting landscapes or buildings recorded by the digital camera, or by the artistic images created by an artificial process.

The primary goal was to design a completely automated method that decides what class an input bitmap image belongs to. The set of classes for classification is known in advance (that is, they are not a part of the input). One bitmap image can belong into multiple classification classes (not just one). The secondary goal was to develop such a method which is general enough to allow extending the set of classes for classification as well increasing the accuracy of the process of classification. It is expectable that the prototype implementation will have low accuracy, thus it is necessary to allow further improvements of the method directly by the initial design.

The concept of *decision tree* [6], [8], [11] has been chosen as the basic building block of the method. This choice was guided by the initial analysis of the problem. The decision tree is a structure that describes assignment of a certain value to an input vector of attributes. The assigned value is represented by a classification class in our case and the input vector of attributes consists of properties extracted from the bitmap image such as *brightness*, *histogram* [1] or the *occurrence of straight lines*. It is supposed that all the attributes used for classification are expressed using the integer value. The sufficiently accurate classification using the decision tree is implied by determining a suitable set of attributes extracted from the image in this context. It is necessary to find attributed that characterize the image well with respect to the set of classification classes. It is evident that the choice of attributes has the great impact on the method. On the other side, it provides a great potential to increase the accuracy of the method.

The secondary goal of our design is satisfied by the concept of decision tree as well. The decision tree works completely automatically in the deciding mode (that is, no user interaction is required). The extension of the decision tree with more classification classes and attributes is a routine augmentation in fact.

This article is based on results elaborated within [4]. The text is organized in the following way. The classification task is formally introduced first. Then initial classification classes we need to be classified by the method are described. Techniques from artificial intelligence and computer graphics that are exploited by our classification method and the classification method itself are described in the next section. The last section is devoted to an experimental evaluation of the prototype implementation of the method written in the

Manuscript received June 20, 2011. Manuscript accepted for publication August 20, 2011.

We would like to gratefully thank the Czech Science Foundation and The Ministry of Education, Youth and Sports, Czech Republic for the financial support of this work (contracts 201/09/P318 and MSM 0021620838 respectively).

The authors are with Charles University in Prague, Faculty of Mathematics and Physics, Department of Theoretical Computer Science and Mathematical Logic, Malostranské náměstí 25, Praha, 118 00, Czech Republic, (pavel.surynek@mff.cuni.cz, ivana.luksova@gmail.com).

Java language on a large set of testing images from different sources.

II. The Task of Bitmap Classification

Let K be a finite set of classification classes and let $\mathcal{I}$ be a set of bitmap images. Each classification class $t \in \mathcal{K}$ has assigned a description in the natural language $d(t)$. Next, let a function $c: \mathcal{I} \longrightarrow 2^{\mathcal{K}}$ defines an assignment of classification classes to bitmap images so that $\forall I \in \mathcal{I}\ \ \forall t \in c(I)\ \ d(t)$ describes the image I well from the user's perspective. Notice that the function c is not known in explicit form. It is known implicitly by the user when she or he can give $c(I)$ for an individual input image. There may be also multiple users who together agree on c by some negotiation strategy.

The following definitions provide us with the precise formulation of the classification task and the demands placed on the classification method.

Definition 1 (*classification task, classification method*). A *classification method* is an algorithm that computes a function $\boldsymbol{c'}: \boldsymbol{\mathcal{I}} \longrightarrow \boldsymbol{2^{\mathcal{K}}}$. A *classification task* with bitmap images corresponds to calculation of $\boldsymbol{c'(I)}$ for a given input bitmap image $\boldsymbol{I} \in \boldsymbol{\mathcal{I}}$. □

The classification method as defined above can be completely mistaken with respect to the user's opinion. Therefore we have the following definition allow us to distinguish the accuracy of classification.

Definition 2 (*accuracy of the classification method*). The classification method corresponding to the function $c': \mathcal{I} \longrightarrow 2^{\mathcal{K}}$ is *accurate* for a classification class $t \in \mathcal{K}$ and an input bitmap image $I \in \mathcal{I}$ if and only if $t \in c'(I) \Leftrightarrow t \in c(I)$. Let $\mathcal{S} \subset \mathcal{I}$ be a finite set of bitmap images and let $\mathcal{S}' \subseteq \mathcal{S}$ be a maximum subset of images, where the method is accurate for the classification class $t \in \mathcal{K}$ (that is, there is no such a set $\mathcal{S}'' \supset \mathcal{S}'$ for which the method is accurate with respect to $t \in \mathcal{K}$), then $|\mathcal{S}'|/|\mathcal{S}|$ is called an *accuracy* of the classification method for the classification class t with respect to the set of images $\mathcal{S}$. □

The above definition of the classification task captures large flexibility in terms of extensibility and scalability. Observe that for instance description of classification classes can be represented by a keyword or a set of keywords. Moreover, the mentioned user's point of view can be determined by a preference of users within a social network (such as *flickr* [15] or similar) who can annotate images with some predefined descriptions in the natural language.

The goal of our work is of course to design a classification method with accuracy as high as possible with respect to the given set of classification classes and for the greatest possible set of input bitmap images (ideally, for all the images that the user can submit in the input).

A. Initially Selected Classification Classes

As the initial goal we have required the method to manage the classification task with respect to the following set of five classification classes $\mathcal{K} = \{P, A, B, L, M\}$. These classes were inspired by requirements of many owners of digital cameras who want to automatically classify their databases of images. In the following text, each class $t \in \mathcal{K}$ is described using a brief natural language expression $d(t)$. Images that belongs to the class and that do not are mentioned. Examples of positively and negatively classified images with respect to proposed classes are also shown.

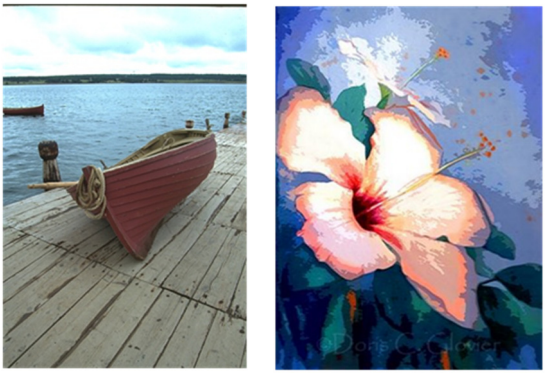


Fig. 1. A bitmap image classified into the class *photography* (P) [left part] and an example of an image not classified into this class [right part].

Fig. 2. An image classified into the class *artistic image* (A).

P: photography

$d(P)$ ="a photography created by a digital or analogue camera"

Positive and negative examples of bitmap images with respect to the class P are shown in Figure 1. Artistic images that are artificially created do not belong to this classification class for instance.

A: artistic image

$d(I)$ ="an image created using a drawing or a painting technique"

An example of artistic bitmap image is shown in Figure 2. For instance, photography does not belong to this class. Pictures with diagrams/schemes and rendered images do not belong to this class as well.

B: buildings

$d(B)$ ="an image depicting building or some architecture"

An example of an image classified into the class B is shown in Figure 3. Landscapes do not belong to this class.

However, a building located in a landscape represents a positive example with respect to this class.

Fig. 3. An example of image classified by the user into the class *buildings* (B).

L: landscapes

$d(L)$ ="an image depicting landscape"

An example of an image depicting a landscape is shown in Figure 4. Images of artificial object are not classified into this class. On the other hand, an image of an artificial object placed in the landscape can represent a landscape as the whole.

Fig. 4. Bitmap image classified into the class *landscapes* (L).

M: macro objects

$d(B)$ ="a photography of an object from the small distance with blurred background"

Fig. 5. An example of bitmap image classified as a *macro object* (M).

An example of an image with a macro object is shown in Figure 5. Artistic images (even though they depot flowers) are not classified as macro objects.

Notice that the textual description $d(t)$ of the classification class $t \in \mathcal{K}$ in the natural language deals with the positive specification of the given class (that is, how does the positive example look like). Negative specification of the class t is given by the situation, that the natural language description does not describe a given image well. Observe, that the proposed set of classification classes allow the existence of a bitmap image $I_1 \in \mathcal{I}$ such that $c(I_1) = \{A, B, L\}$. On the other hand there cannon exist an image $I_2 \in \mathcal{I}$ such that $c(I_2) = \{P, A\}$.

III. The Initial Analysis of the Problem

During the development of the classification method, several key questions needed to be answered. First, it was not clear how to formally handle the condition that $\forall I \in \mathcal{I}$ $\forall t \in c(I)$ $d(t)$ describes the image I well from the user's perspective. It is necessary to take into account that $c: \mathcal{I} \rightarrow 2^{\mathcal{K}}$ is defined by the particular user or the set of users and hence can be biased somehow. Thus we need to design the method so that is should be configurable with respect to different users. Next, it is necessary to take into account that function c is not known explicitly.

Finally, it turned out to be easiest to record function $c: \mathcal{I} \rightarrow 2^{\mathcal{K}}$ partially using an annotated training set of selected images. This solution takes into account the individual users as well as the fact that knowledge of c is implicit only (that is, we can put queries on c but are unable to enumerate it as the whole). The classification method itself implementing assignment $c': \mathcal{I} \rightarrow 2^{\mathcal{K}}$ should be identical with c on all (or on the almost all) input images from the training set. Additionally it should be as much accurate as possible on other images with respect to the function c. To satisfy these requirements the method should have a good **generalization** ability [12].

We decided to use the concept of *decision tree* [8], [11], [12] as the core of our classification method. There exist many successful learning algorithms for decision trees that support its generalization abilities. As it has been mentioned, the decision tree can be easily extended for a larger set of classification classes as well as attributes.

The next important question is the identification and acquiring attributes that characterizes the bitmap images with respect to classification classes well. These attributes are to be used by the decision tree. Intuitively, the following categories of attributes were suggested: *basic color characteristics* (number of colors, relative occurrence of colors, etc.), characteristics based on the *histogram* (contrast, local contrast), and characteristics based on *edge detection* [5], [13], [18] (occurrence of straight lines, occurrence of right angles).

The basic assumption was that basic color characteristics can help to distinguish photographs from artistic images. Next it was expected that characteristics based on histogram can help to distinguish macro object from landscapes that differ in contrast sharply. Finally, characteristics based on edge detection can distinguish images of artificial object from natural ones.

In the complete classification all the attributes play some role. Interestingly, the importance of individual attributes is determined by the learning process of the decision tree.

IV. Techniques for the Process of Classification

One of the major contributions of this work is the suggestion of a set of attributes suitable for image classification and the design of methods for their extraction. The following text is devoted to the existent concept of decision tree – a brief description is given. The process of extraction of attributes is described in more details since it is the original contribution of this work.

A. Binary Decision Tree

Suppose that we have a set of attributes $\mathcal{P}$ extracted from the bitmap image. Each attribute can be assigned a value from the given domain (integers or floating point numbers). Next suppose that we have a classification class $t \in \mathcal{K}$. A *binary decision tree* for the class t is an oriented tree where each node except leaves have two successors. The root and internal nodes are assigned an attribute $p \in \mathcal{P}$. Leaves are assigned the Boolean value $False$ or $True$. Edges from a node which is assigned an attribute p are assigned disjoint subsets of the set of values for the attribute p.

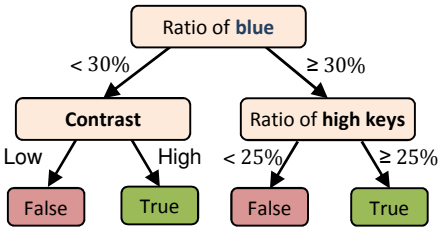


Fig. 6. An example of a simple decision tree for image classification with respect to the classification class *landscapes* (L).

The decision tree assigns the indication of the membership into the given classification class t to a given vector of attributes. Having a vector of attributes, the indication of the membership into the given class is determined by a *controlled traversal* of the decision tree from the root to the leaf. The controlled traversal starts in the root and is defined inductively. Suppose that the current node is assigned an attribute p. The traversal then continues to the endpoint of the edge that starts in the current node and is assigned a subset of values which the value of the attribute p from the input vector belongs to.

The decision tree is said to be *complete*, if the controlled traversal is defined for every input vector of attributes. To represent classification function $c: \mathcal{I} \longrightarrow 2^{\mathcal{K}}$ we will need $|\mathcal{K}|$ decision trees. Observe that the accuracy of the decision tree is now a well-defined concept.

The process of leasing of the decision tree exploited in our method is based on the algorithm ID3 [8]. Although ID3 is outperformed by C4.5 [9] we used ID3 for scholarly purposes. An example of decision tree is shown in Figure 6.

B. Extraction of Attributes from Colors

In the following text we will be working with bitmap images of 8-bit depth for each color component [[1], [14]]. The eventual adaptation for the different color depth is straightforward. The basic characteristic which will be extracted from the image is based on the color information. The following definition formalizes the concept of color information.

Definition 3 (*bitmap image – color, monochrome*). A *color bitmap image* is a 5-tuple $I_{rgb} = (r, g, b, x_{max}, y_{max})$ where $r: \{0,1,\dots,x_{max}\} \times \{0,1,\dots,y_{max}\} \longrightarrow \{0,1,\dots,255\}$, $g: \{0,1,\dots,x_{max}\} \times \{0,1,\dots,y_{max}\} \longrightarrow \{0,1,\dots,255\}$, and $b: \{0,1,\dots,x_{max}\} \times \{0,1,\dots,y_{max}\} \longrightarrow \{0,1,\dots,255\}$ are functions that assign individual pixels of the image the value of *red*, *green*, and *blue* component respectively; x_{max} represents the horizontal size of the image, y_{max} represents the vertical size of the image (the size of the image is thus $(x_{max}+1) \times (y_{max}+1)$ pixels). *A monochrome bitmap image* is a triple $I_{mono} = (i, x_{max}, y_{max})$, where $i: \{0,1,\dots,x_{max}\} \times \{0,1,\dots,y_{max}\} \longrightarrow \{0,1,\dots,255\}$ is a function that assigns the individual pixels of the image their intensity. For a color bitmap image the corresponding monochrome image can be obtained using the following expression: $(x,y) = 0.299\, r(x,y) + 0.587\, g(x,y) + 0.114\, b(x,y)\ \forall x \in \{0,1,\dots,x_{max}\};\ \forall y \in \{0,1,\dots,y_{max}\}$ (this expression has been determined empirically for humans [1], [14]). The size of the image is defined as the number of its pixels, that is $s_{xy} = (x_{max}+1) * (y_{max}+1)$.

A useful characteristic is the total *number of colors* of the image. Having an artistic image which is not affected by any noise it is expectable that the total number of colors will be low. The total number of colors is given by the expression: $c_{rgb} = |\{(r(x,y), g(x,y), b(x,y)) | x \in \{0,1,\dots,x_{max}\} \wedge\ y \in \{0,1,\dots,y_{max}\}\}|$ The next interesting characteristic is represented by the *occurrence of the individual colors*. As the total number of possible colors is too high it is necessary to restrict on a certain color palette [1], [14] and calculate the occurrence of colors that are close to the colors from the palette.

The easiest way how to create a color palette is to represent colors (r, g, b), where

$$\frac{\alpha}{\eta}256 \le r < \frac{\alpha+1}{\eta}256 \ \wedge\ \frac{\beta}{\eta}256 \le g < \frac{\beta+1}{\eta}256 \ \wedge\ \frac{\gamma}{\eta}256 \le b < \frac{\gamma+1}{\eta}256$$

for $\eta \in \mathbb{N}$ and $\alpha, \beta, \gamma \in \{0,1,\dots,\eta-1\}$ using a single color. Since we require the total number of colors in the palette to be low, the value of the parameter η should be low as well. In the implementation of the method we used $\eta = 4$. Occurrence of individual colors is the defined by a function:

$$rgb(\alpha,\beta,\gamma) = |\{(x,y) | x \in \{0,1,\dots,x_{max}\} \wedge y \in \{0,1,\dots,y_{max}\} \wedge \frac{\alpha}{\eta}256 \le r(x,y) < \frac{\alpha+1}{\eta}256 \wedge \frac{\beta}{\eta}256 \le g(x,y) < \frac{\beta+1}{\eta}256 \ \wedge\ \frac{\gamma}{\eta}256 \le b(x,y) < \frac{\gamma+1}{\eta}256\}|/s_{xy},$$

where $\eta \in \mathbb{N}$ and $\alpha, \beta, \gamma \in \{0,1,\dots,\eta-1\}$.

A *typical palette* for each classification class has been determined using the function rgb - it consists of the most frequent colors. The attribute used for classification is represented by the difference of the occurrence of colors in the

input image from the typical palette of the individual classification classes.

C. Attributes Exploiting Contrast and Histogram

Another important characteristic for bitmap images is represented by the *histogram* [1], [14]. In photographic techniques there are frequently used terms as *low-key*, *mid-key*, and *high-key* photograph. These terms express the high relative occurrence of dark colors, middle-tone colors, and light colors respectively. We will quantify these terms exactly using the histogram.

Definition 4 (*histogram*). A *histogram* for the given monochrome image $I_{mono} = (i, x_{max}, y_{max})$ is a function $h_i: \{0,1,\dots,255\} \rightarrow \mathbb{N}_0$ (natural numbers including zero), where $h_i(j) = |\{(x,y) | x \in \{0,1,\dots,x_{max}\} \wedge y \in \{0,1,\dots, y_{max}\} \wedge i(x,y) = j\}|$. For the color image $I_{rgb} = (r, g, b, x_{max}, y_{max})$ we define the histogram in an analogical way using a triple of functions $h_r: \{0,1,\dots,255\} \rightarrow \mathbb{N}_0$, $h_g: \{0,1,\dots,255\} \rightarrow \mathbb{N}_0$, and $h_b: \{0,1,\dots,255\} \rightarrow \mathbb{N}_0$ for individual color components. □

TABLE I
SEVERAL CHARACTERISTICS BASED ON THE HISTOGRAM

Characteristic	Calculation
Ratio of low colors $r_{low}^{h_i}$	$r_{low}^{h_i} = \frac{1}{s_{xy}} \sum_{j=0}^{85} h_i(j)$
Ratio of mid colors $r_{mid}^{h_i}$	$r_{mid}^{h_i} = \frac{1}{s_{xy}} \sum_{j=86}^{170} h_i(j)$
Ratio of high colors $r_{high}^{h_i}$	$r_{high}^{h_i} = \frac{1}{s_{xy}} \sum_{j=170}^{255} h_i(j)$
Contrast by **Michelson** $c_M^{h_i}$	$c_M^{h_i} = \frac{i_{max} - i_{min}}{i_{max} + i_{min}}$
Contrast by **Webera** $c_{rms}^{h_i}$	$c_{rms}^{h_i} = \sqrt{\frac{1}{s_{xy}} \sum_{x=0}^{x_{max}} \sum_{y=0}^{y_{max}} (i(x,y) - i_\mu)^2}$

An important characteristic connected with the histogram is represented by the *contrast*. We consider an image with the great difference between dark and light colors as the contrast one. The histogram of such an image is not concentrated in the narrow interval of the intensity.

The contrast can be formally defined in various ways. In the design of the classification method we used definition from [7], that is definitions by Michelson and Weber. To introduce these concepts we need to know the value of the minimum, maximum, and the average intensity which are defined by the following expressions respectively:
$i_{min} = \min\{j | j \in \{0,1,\dots 255\} \wedge h_i(j) > 0\}$, $i_{max} = \max\{j | j \in \{0,1,\dots 255\} \wedge h_i(j) > 0\}$, and $i_\mu = \frac{1}{s_{xy}} \sum_{j=0}^{255} (j * h_i(j))$.

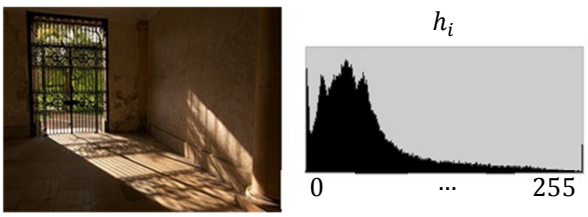


Fig. 7. An example of *low-key* image and its histogram.

Several characteristics based on the histogram for the intensity are shown in Table 1. Notice that all the mentioned characteristics can be defined for individual color components as well (however, the experimental implementation does not use them).

Sometimes it is necessary to use a so called *local contrast*. The local contrast can be found on photographs of objects photographed from the small distance (*macro objects*). The high contrast is concentrated in the central part of the image typically (the object itself) while other parts of the image have the low contrast (background). In our classification method, the input image has been divided by the grid with squares of the size of 25×25 pixels while the contrast (by Michelson $c_M^{h_i}$ as well as by Weber $c_{rms}^{h_i}$) has been calculated and used as an attribute for every square of the grid.

D. Attributes Based on Edge Detection

The core technique which allows us to distinguish the class *buildings* (**B**) from other classes exploits *edge detection* [5], [18]. As the extraction and application of information about edges represents the most complex technique developed within our work, we will describe it more formally using algorithms in pseudo-code.

The goal is to obtain the explicit information about edges occurring in the input bitmap image. This explicit knowledge then allows calculating characteristics such as the *occurrence of right angles*.

The first step consists in using the standard method for **edge detection** based on convolution which the result is the set of edges in the implicit form (edges are merely highlighted in the image). Particularly, the convolution with the *Laplace operator* [1], [14] has been used where the internal convolution matrix is as follows: $h = ((1,1,1); (1,-8,1); (1,1,1))$.

The process of edge detection is formally described using pseudo-code as Algorithm 1. The next phase is the transformation of the implicit information about edges into the **explicit** one. This step is done by a so called *Hough transformation* [3], [10]. The transformation algorithm calculates all the lines that goes through every highlighted point (that can be recognized according to its intensity which is higher than the given threshold θ_H) and are expressed by the equation $\rho = x * \cos\vartheta + y * \sin\vartheta$. More precisely, the parameters ϑ and ρ are calculated for every highlighted point (x, y). As there is infinite number of solutions we use the discrete sampling of the parametric space.

The occurrence of the line is then indicated by the local maximum of the function depending on the parameters ϑ and ρ which expresses the number of highlighted points on the line corresponding to the parameters. The detected lines are subsequently segmented into the line segments according to the intensity of pixels on the line. If the intensity is not high enough the line is segmented (again, the intensity lower than the given threshold θ_S indicates line segmentation). The result is the set of line segments $L_S = \{((a_x^1, a_y^1), (b_x^1, b_y^1), (\varphi^1, \rho^1)), \dots, ((a_x^n, a_y^n), (b_x^n, b_y^n), (b_x^n, b_y^n), (\varphi^n, \rho^n))\}$, where (a_x^k, a_y^k) is the start point, (b_x^k, b_y^k) is an endpoint of the k-th

line segment, and (φ^k, ρ^k) are parameters of the line on that the k-th line segment lies.

Algorithm 1. *Algorithm of* ***edge detection***. The input is a monochromatic image I_{mono} and a convolution matrix h. The output is a monochromatic image with highlighted detected edges.

function *Detect-Edges* $(I_{mono} = (i, x_{max}, y_{max}), h, k_{max})$: **grayscale image**
for $x = 0,1, \dots, x_{max}$ **do**
  for $y = 0,1, \dots, y_{max}$ **do**
    $s \leftarrow 0$
    for $k_x = -k_{max}, \dots, -1,0,1, \dots, k_{max}$ **do**
      for $k_y = -k_{max}, \dots, -1,0,1, \dots, k_{max}$ **do**
        $s \leftarrow s + h(k_{max} + k_x + 1, k_{max} + k_y + 1) * i(x + k_x, y + k_y)$
  $i_E(x, y) \leftarrow s$
return (i_E, x_{max}, y_{max})

Algorithm 2. *Modified* ***Hough transformation***. The input is a monochromatic image I^e_{mono}, obtained by edge detection. The output is a set of line segments in the explicit form.

function *Hough-Transformation* $(I^e_{mono} = (i^e, x^e_{max}, y^e_{max}), \varphi_{max}, \rho_{max}, \theta_H, \theta_S)$: **set**
$a_H \leftarrow \vec{0}$
for $x = 0,1, \dots, x^e_{max}$ **do**
  for $y = 0,1, \dots, y^e_{max}$ **do**
    if $i^e(x, y) \geq \theta_H$ **then**
      for $\varphi = 0,1, \dots, \varphi_{max}$ **do**
        $\rho \leftarrow x * \cos((\varphi/\varphi_{max})\pi) + y * \sin((\varphi/\varphi_{max})\pi)$
        $a_H(\varphi, \rho) \leftarrow a(\varphi, \rho) + 1$
$L \leftarrow \emptyset$
for $\varphi = 0,1, \dots, \varphi_{max}$ **do**
  for $\rho = 0,1, \dots, \rho_{max}$ **do**
    if a_H gains local mixmum in (φ, ρ) **then**
      $L \leftarrow L \cup \{(\varphi, \rho)\}$
$L_S \leftarrow \emptyset$
for each $(\varphi, \rho) \in L_S$ **do**
  let $[(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)]$ is a sequence of (integer) points of line (φ, ρ) in the image I^e_{mono} sorted along the line
  $(a_x, a_y) \leftarrow \perp$
  for $k = 1,2, \dots, m$ **do**
    if $(a_x, a_y) = \perp$ **then**
      if $i_E(x_k, y_k) \geq \theta_S$ **then**
        $(a_x, a_y) \leftarrow (x_k, y_k)$
    else
      if $i_E(x_k, y_k) < \theta_S$ **then**
        $(a_x, a_y) \leftarrow \perp$
    $L_S \leftarrow L_S \cup \{((a_x, a_y), (x_k, y_k), (\varphi, \rho))\}$
return L_S

TABLE II
SEVERAL CHARACTERISTICS DERIVED FROM STRAIGHT LINES

Characteristic	Calculation
Length of longest segment $d^{L_S}_{max}$	$d^{L_S}_{max} = \max\{d_k \mid k = 1,2, \dots, n\}$, where $d_k = \sqrt{(a^k_x - b^k_x)^2 + (a^k_y - b^k_y)^2}$
Average segment length $d^{L_S}_{\mu}$	$d^{L_S}_{\mu} = \frac{1}{n}\sum_{k=1}^{n} d_k$
Length variance $d^{L_S}_{\sigma}$	$d^{L_S}_{\sigma} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(d_k - d^{L_S}_{\mu})^2}$
Ratio of short segments $r^{L_S}_{short}$	$r^{L_S}_{short} = \frac{1}{n}\left\lvert\{d_k \mid 0 < d_k \leq \frac{1}{3}d^{L_S}_{max}\}\right\rvert$
Ratio of medium segments $r^{L_S}_{mid}$	$r^{L_S}_{mid} = \frac{1}{n}\left\lvert\{d_k \mid \frac{1}{3}d^{L_S}_{max} < d_k \leq \frac{2}{3}d^{L_S}_{max}\}\right\rvert$
Ratio of long segments $r^{L_S}_{long}$	$r^{L_S}_{long} = \frac{1}{n}\left\lvert\{d_k \mid \frac{2}{3}d^{L_S}_{max} < d_k \leq d^{L_S}_{max}\}\right\rvert$

Using the explicit knowledge about line segments it is possible to define various characteristics. Some of them are shown in Table 2. We can then calculate the occurrence of right angles which is important for classification of buildings (B). The process of right angle detection is shown as Algorithm 3. The total number of detected right angles is used as the decision attribute, that is the value $c^{A_R} = |A_R|$.

Algorithm 3. ***Detection of right angles***. The input is a set of line segments L_S in the explicit form, an interval of angles φ^H_{min} to φ^H_{max}, that are considered horizontal, and an interval of angles φ^V_{min} to φ^V_{max} that are considered as vertical. The output is a set of pair of line segments that are orthogonal to each other.

function *Detect-Right-Angles* $(L_S, \varphi^H_{min}, \varphi^H_{max}, \varphi^V_{min}, \varphi^V_{max})$: **set**
$A_R \leftarrow \emptyset$
let $L_S = \{((a^1_x, a^1_y), (b^1_x, b^1_y), (\varphi^1, \rho^1)), \dots, ((a^n_x, a^n_y), (b^n_x, b^n_y), (\varphi^n, \rho^n))\}$
for $k_1 = 1,2, \dots, n$ **do**
  for $k_2 = 1,2, \dots, n$ **do**
    if $k_1 \neq k_2$ **then**
      if $\varphi^H_{min} \leq \varphi^{k_1} \leq \varphi^H_{max}$ **and** $\varphi^V_{min} \leq \varphi^{k_2} \leq \varphi^V_{max}$ **then**
        if the line segment given by points $(a^{k_1}_x, a^{k_1}_y), (b^{k_1}_x, b^{k_1}_y)$ intersects with the line segment $(a^{k_1}_x, a^{k_1}_y), (b^{k_1}_x, b^{k_1}_y)$ **then**
          $A_R \leftarrow A_R \cup \{(k_1, k_2)\}$
return A_R

V. Experimental Evaluation

The classification method based on described attributes and the decision tree has been implemented in the Java language. The training sets used for creating decision trees contained 350 manually annotated images while there were at least 40 images for each classification class (the annotation was made by the second author). All the experimental data are available at the web to allow further comparative research: http://ktiml.mff.cuni.cz/~surynek/research/micai2011.

TABLE III
ACCURACY ON THE CLASS *PHOTOGRAPHY* (P)

	Image count	Correctly classified	Accuracy
Training	155	154	99.35%
Set A	297	243	81.82%
Set B	405	300	74.07%

Two sets of testing images called *Set A* and *Set B* can be found at the same web. These sets of images have been used for the experimental evaluation presented in this section. The Set A consists of 297 images and the Set B consists of 405 images. Each set of testing images was collected by a different user (annotation with respect to classification classes was made by the second author again).

In the experimental evaluation, we concentrated on the accuracy of the method with respect to the classification classes P, A, B, L, M (as it is given in Section II). At the same time we investigated what attributes are the most important for individual classes. Fortunately, the importance of attributes can be easily measured by their location in the decision tree – except generalization ability this was another main reason for using decision trees. The most important attributes are located closer to the root of the tree.

Results for the class **photography (P)** are shown in Table 3. The most important attributes were: ratio of colors from the color palette, the total number of colors, the number of local maxima in the histogram, and the variance around local maxima in the histogram. It is also characteristic that the number of colors is not that high for photographs (typically, rendered images contain substantially more colors).

TABLE IV
ACCURACY ON THE CLASS *ARTISTIC IMAGES* (A)

	Image count	Correctly classified	Accuracy
Training	104	104	100.00%
Set A	297	251	84.51%
Set B	405	331	81.73%

TABLE V
ACCURACY ON THE CLASS *BUILDINGS* (B)

	Image count	Correctly classified	Accuracy
Training	104	104	100.00%
Set A	297	232	78.11%
Set B	405	350	86.42%

TABLE VI
ACCURACY ON THE CLASS *LANDSCAPE* (L)

	Image count	Correctly classified	Accuracy
Training	90	89	98,89%
Set A	297	232	84.51%
Set B	405	350	81.73%

TABLE VII
ACCURACY ON THE CLASS *MACRO OBJECTS* (M)

	Image count	Correctly classified	Accuracy
Training	118	118	100,00%
Set A	297	259	87.20%
Set B	405	295	72.84%

Results for the class of **artistic images (A)** are shown in the Table 4. The most important characteristics are: the number of local maxima in the histogram, the variance around local maxima in the histogram, and the total number of colors. The limited number of colors that cover large areas of the image is typical for this class.

Classification results for the class **buildings (B)** are shown in Table 5. As it was expected, the most important attributes are: the number of right angles, the length of the longest line segment, and the ratio of long line segments.

Results regarding the classification class **landscapes (L)** are shown in Table 6. Here, the most important characteristics are: the ratio of blue colors, the ratio of high-keys, the contrast of the image, and the ratio of low-keys.

We can conclude that contrast plays the important role according to the expectation with respect to landscapes – natural objects exhibit high complexity which generates high contrast when photographed. The importance of blue colors can be explained by the fact, that landscape images often depict the blue sky as well.

Finally, results for the classification class of **macro objects (M)** are shown in T 7. The most important attributes are: the area covered by non-contrast parts of the image, the ratio of high, mid, and low keys. In accordance with the expectation, the local contrast was the most important attribute.

VI. COMPARATIVE EVALUATION AND RELATED WORKS

A work most related to our study is *MUFIN* – Multi-feature Indexing Network [16], [17]. It can be regarded as an image classification method that can be used for searching an image according to a keyword or according to another image. The former type of search is based on simple annotation of images using a set of keywords; the latter search is based on the similarity of images. The similarity search uses attributes such as colors and shapes depicted on the image to search for similar images.

The major difference from our work is that we are trying to generalize from a training set of images. Simply said, if we ask our method to find images that are classified as building we are likely to obtain images of various buildings which appearance was generalized from the training set. If we do the same with the *MUFIN* similarity search we will obtain images of buildings of the same shape and color as that in the reference image.

Another related project is *CoPhIR* - Content-based Photo Image Retrieval Test-Collection [1] which provides a test suite for image classification. For capacity reasons we did not used this collection for experimental evaluation and used our own test suite instead.

Notice that image search provided by commercial web search engines is based on annotation of images by keywords while the search is not done over images but merely over the keywords. This represents the most primitive approach of image classification/search as there is no automation of the process.

VII. CONCLUSIONS AND FUTURE WORKS

We proposed a relatively successful method for classification of bitmap images into selected classification classes. The classification classes are determined by the description in the natural language. The classification method itself is based on attribute extraction from the image which is subsequently processed by the decision tree.

The accuracy of the current implementation of the method ranges between 75% and 85%. We have also provided a careful analysis of the classification task for bitmap images and we described key techniques for extraction of complex attributes such as occurrence of right angles. Such complex attributes turned out to be important for distinguishing classification classes such as buildings.

We would like to emphasize that the current set of attributes and classification classes serves as the justification of the concept of using decision tree and attribute extraction for image classification. The current state does not claim to be final in any way. This attitude was kept in mind during the initial design of the concept which we therefore tried to make as extensible as possible with respect to increasing classification accuracy and extending the set of classification classes. It is planned for future work to propose more complex attributes that can be used for classification of difficult characteristics of images such as images depicting people or industrial products such as cars, planes etc. The interesting

problem for future work is also how to allow the user to define her/his own classification class.

REFERENCES

[1] P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti, "CoPhIR: a Test Collection for Content-Based Image Retrieval," *CoRR abs/0905.4627*, http://cophir.isti.cnr.it/, [accessed June, 2011], ISTI CNR, 2009.

[2] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice in C,* Addison-Wesley Professional, 1995.

[3] R. O. Duda and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Communications of the ACM*, ACM Press, 1972.

[4] I. Lukšová, *Klasifikace bitmapových obrázků (Classification of Bitmap Images),* Bachelor thesis, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic, 2010.

[5] D. Marr and E. C. Hildreth, "Theory of edge detection", *Proceedings of the Royal Society of London*, Series B, Volume 207 (1167), pp. 187–217, The Royal Society, 1980.

[6] T. M. Mitchell, *Machine Learning,* McGraw Hill, 1997.

[7] E. Peli, "Contrast in Complex Images", *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, Volume 7 (10), pp. 2032-2040, OSA, 1990.

[8] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Volume 1, pp. 81–106, Springer, 1986.

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning,* Morgan Kaufmann, 1993.

[10] T. Rabbani and F. van den Heuvel, "Efficient Hough transform for automatic detection of cylinders in point clouds", in *ISPRS Workshop Laser scanning 2005*, Institute of Photogrammetry and Remote Sensing, 2005, pp. 60–65.

[11] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers: a survey," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 35, IEEE Press, 2005, pp. 476–487.

[12] S. Russel and P. Norvig, *Artificial Intelligence – A modern approach,* Prentice Hall, 2003.

[13] L. G. Shapiro and G. C. Stockman, *Computer Vision,* Prentice Hall, 2001.

[14] P. Shirley, M. Ashikhmin, and S. Marschner, *Fundamentals of Computer Graphics,* A.K. Peters, 2009.

[15] Yahoo! Inc., "*flickr from Yahoo! - almost certainly the best online photo management and sharing application in the world",* Commercial web site, http://www.flickr.com/, [accessed June, 2011], 2011.

[16] P. Zezula, G. Amato, V. Dohnal, and M. Batko, "Similarity Search - The Metric Space Approach", in *Advances in Database Systems*, Springer, 2006.

[17] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *MUFIN – Multi-feature Indexing Network / Image Search,* Project web site, http://mufin.fi.muni.cz/imgsearch/, [accessed June, 2011], Masaryk University, Czech Republic, 2008.

[18] L. Zhai, S. Dong, and H. Ma, "Recent Methods and Applications on Image Edge Detection", in *International Workshop on Education Technology and Training and International Workshop on Geo-Science and Remote Sensing,* IEEE Press, 2008, pp. 332–335.

Reconocimiento automático de voz emotiva con memorias asociativas Alfa-Beta SVM

José Francisco Solís Villarreal, Cornelio Yáñez Márquez y Sergio Suárez Guerra

***Resumen*—Una de las de investigación de mayor interés y con más crecimiento en la actualidad, dentro del área de procesamiento de voz, es el reconocimiento automático de emociones, el cual consta de 2 etapas; la primera es la extracción de parámetros a partir de la señal de voz y la segunda es la elección del modelo para hacer la tarea de clasificación. La problemática que actualmente existe es que no se han identificado aún los parámetros más representativos del problema ni tampoco se ha encontrado al mejor clasificador para hacer la tarea. En este artículo se introduce un nuevo modelo asociativo de reconocimiento automático de voz emotiva basado en las máquinas asociativas Alfa-Beta SVM, cuyas entradas se han codificado como representaciones bidimensionales de la energía de las señales de voz. Los resultados experimentales muestran que este modelo es competitivo en la tarea de clasificación automática de emociones a partir de señales de voz [1].**

***Palabras Clave*—Reconocimiento de voz emotiva, memorias asociativas Alfa-Beta SVM, procesamiento de voz.**

Automatic Emotional Speech Recognition with Alpha-Beta SVM Associative Memories

***Abstract*—One of the research lines of interest and more growth at present, within the area of voice processing is automatic emotion recognition. It is vitally important the study of speech signal not only to extract information about what is being said, but how is being said, this in order to be closer to the human-machine interaction. In literature the procedure of automatic emotion recognition consists of two stager, the first is the extraction of parameters from the voice signal and the second is the choice of model for the classification task, the problem that currently exists is not yet identified the most representative parameters of the problem nor has found the best classifier for the task, but have not yet been tested several models, this paper presents a two-dimensional representation of energy as data entry for Alpha-Beta associative machines SVM (Support Vector Machine) for the classification of emotions.**

***Index Terms*—Emotional speech recognition, Alpha-Beta SVM associative memories, voice processing.**

Manuscrito recibido el 03 de marzo de 2011. Manuscrito aceptado para su publicación el 22 de junio de 2011.

Los autores trabajan en el Centro de Investigación en Computación del Instituto Politécnico Nacional, México, D.F. (tlilectic.mixtzin@gmail.com, cyanez@cic.ipn.mx, ssuarez@cic.ipn.mx).

I. Introducción

En la última década, el interés en el reconocimiento automático de emociones ha ido creciendo de manera notable. El objetivo de este tema es mejorar la interacción hombre-máquina, haciendo posible que los equipos de cómputo puedan rescatar información afectiva más que el contenido de lo hablado. Esto es necesario para el propósito de tener una comunicación más natural [2].

No obstante que muchos trabajos han contribuido con diferentes enfoques, el rendimiento alcanzado en esta tarea todavía no es suficientemente bueno. En reconocimiento de voz emocional, uno de los principales problemas es localizar los rasgos que se ajusten para llevar a cabo la clasificación. Los investigadores han reportado una gran variedad de parámetros, pero hasta la fecha no se han identificado cuáles son los parámetros más representativos de la información afectiva de la señal de voz que sean útiles para realizar la clasificación de estados emocionales. [3]. A continuación se describen brevemente las aportaciones reciente más importantes en esta área de investigación.

En [4], se usaron 2 bases de datos; la de Berlín [1], que contiene emociones actuadas y la de SmartKom, la cual incluye emociones espontáneas. El enfoque en este trabajo fue el uso de la detección del género previo al reconocimiento de emociones, y se alcanzó un 90% de reconocimiento para la detección del género. Con respecto a la base de datos de Berlín, los autores reportan un desempeño aproximado del 80%.

En otro trabajo publicado en [3], se usaron 102 parámetros de dos bases de datos, la de Berlín y una Polaca. La selección de rasgos fue realizada mediante árboles binarios de decisión, buscando tripletas óptimas, y midiendo la utilidad del conjunto de parámetros mediante su correlación; si la correlación es alta, se desecha el conjunto de parámetros y se selecciona otro al azar. Cada conjunto de parámetros contiene un parámetro de frecuencia, otro de energía y otro de duración. Para la base de datos de Berlín, se alcanzó un rendimiento del 72.04% de reconocimiento usando 6 emociones de este corpus.

En otro enfoque, reportado en [5], se seleccionaron 68 parámetros de la base de datos de Berlín, y se usaron redes neuronales artificiales (backpropagation) como modelo de clasificación; también se realizó una clasificación de género previa a la clasificación de emociones, obteniendo un desempeño del 79.47% de reconocimiento usando 6

emociones de la base de datos (enojado, feliz, miedo, neutral, tristeza y aburrido).

En [14] se reporta el uso de un modelo basado en los k-vecinos más cercanos con una estimación de costo del error. Los investigadores usan las 7 emociones de la base de datos de Berlín [1] y reportan un resultado del 82.44% de reconocimiento de emociones.

El resto del artículo está organizado como sigue: en la sección II se describe el modelo de las memorias asociativas Alfa-Beta SVM, cuyo algoritmo es la base del modelo asociativo de reconocimiento automático de voz emotiva, el cual es introducido en la sección III a través del diseño experimental. Mientras que en la sección IV se detallan los resultados experimentales y la discusión correspondiente, en la sección V se plantean las conclusiones del este trabajo de investigación. Finalmente, se incluyen las referencias.

II. Memorias asociativas Alfa-Beta SVM

El modelo de clasificación usado en este trabajo es la Memoria Asociativa Alfa-Beta SVM [8-9]. Los conceptos básicos concernientes a las memorias asociativas han sido reportados en varios trabajos [10-12]; sin embargo, en este artículo se usan la notación y conceptos introducidos en [10]. Una memoria asociativa *M* relaciona patrones a la manera de un sistema de entrada-salida: $x \to M \to y$ donde x es un patrón de entrada siendo y es un patrón de salida. Por cada patrón de entrada se forma una asociación con el correspondiente patrón de salida. La *k*-ésima asociación está dada por (x^k, y^k), donde *k* es un entero positivo. La Memoria Asociativa *M* está representada por una matriz cuya *ij*-ésima componente es m_{ij}.

La matriz *M* es generada a partir del conjunto fundamental, el cual es representado como: $\{(x^\mu, y^\mu) \mid \mu = 1,2,\ldots,p\}$, donde *p* es la cardinalidad del conjunto.

Si $x^\mu = y^\mu \forall \mu \in \{1,2,\ldots,p\}$, la memoria *M* es autoasociativa; de otro modo, es heteroasociativa. La versión distorsionada del patrón x^k a ser recuperado, está denotada como $\tilde{x}^k$. Se dice que la recuperación es correcta cuando se recupera y^k a partir de $\tilde{x}^k$.

Las memorias asociativas Alfa-Beta se basan en dos operadores binarios; el operador α es usado en la fase de aprendizaje y el operador β es útil para la fase de recuperación. Sean los conjuntos $A = \{0,1\}$ y $B = \{0,1,2\}$; entonces los operadores α y β están definidos de manera tabular en la TABLA I y en la TABLA II, respectivamente.

Los conjuntos *A* y *B*, los operadores α y β, $\wedge$ (mínimo) y $\vee$ (máximo) forma el sistema algebraico que es la base matemática para las Memorias Asosiativas Alfa-Beta.

Todos los conceptos básicos descritos anteriormente [10], son necesarios para describir el algoritmo principal de Alfa-Beta SVM [8,9]. Se tiene un problema de reconocimiento de patrones, donde el conjunto fundamental se describe como $\{(x^\mu, y^\mu) \mid \mu = 1,2,\ldots,p\}$, $x^\mu \in A^n \forall \mu \in \{1,2,\ldots,p\}$, $n, p \in Z^+$ y $A = \{0,1\}$. El algoritmo de Alfa-Beta SVM tiene dos fases, la fase de aprendizaje y la de recuperación.

TABLA I

$\alpha: A \times A \to B$

x	y	$\alpha(x,y)$
0	0	1
0	1	0
1	0	2
1	1	1

TABLA II

$\beta: B \times A \to A$

x	y	$\beta(x,y)$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

Fase de aprendizaje:

1. Del conjunto fundamental, se obtiene el vector soporte S, cuya *i*-ésima componente se calcula así:

$$S_i = \begin{cases} \bigwedge_{k=1}^{p/2} \beta\left(x_i^{2k-1}, x_i^{2k}\right) & si\ p\ es\ par \\ \beta\left[\bigwedge_{k=1}^{(p-1)/2} \beta\left(x_i^{2k-1}, x_i^{2k}\right), x_i^p\right] & si\ p\ es\ non \end{cases}$$

2. Para cada $\mu \in \{1,2,\ldots,p\}$ se forma el vector $x^\mu|_S$, a fin de obtener, a partir de estos resultados, el conjunto fundamental restringido $\{(x^\mu|_S, x^\mu|_S) \mid \mu = 1,2,\ldots,p\}$.

3. Para cada $\mu \in \{1,2,\ldots,p\}$ se forma el vector $\overline{x^\mu}$ (vector negado de x^μ). Con los *p* vectores negados, se obtiene el conjunto fundamental negado $\{(\overline{x^\mu}, \overline{x^\mu}) \mid \mu = 1,2,\ldots,p\}$.

4. Del conjunto fundamental negado, se calcula el vector soporte $\hat{S}$ (como en el paso 1, usando el conjunto fundamental negado).

5. De manera similar a como se hizo en el paso 2, para cada $\mu \in \{1,2,\ldots,p\}$ se forma el vector $\overline{x^\mu}|_{\hat{S}}$, a fin de obtener, a partir de estos resultados, el conjunto fundamental negado restringido $\{(\overline{x^\mu}|_{\hat{S}}, \overline{x^\mu}|_{\hat{S}}) \mid \mu = 1,2,\ldots,p\}$.

Fase de recuperación:

Siendo $\tilde{x} \in A^n$ un patrón de entrada cuyo patrón asociado x^μ es desconocido:

1. Se obtiene la restricción $\tilde{x}|_S$.

2. Para cada $\mu \in \{1,2,\ldots,p\}$ se obtiene $\tau(\tilde{x}|_S, x^\mu|_S)$, donde la transformada τ del vector $x \in A^n$ con respecto al vector $y \in A^n$ da como resultado un vector $\tau(x,y)$ de dimensión *n*, cuya *i*-ésima componente se calcula de la siguiente manera: $[\tau(x,y)]_i = \beta[x_i, \alpha(0, y_i)]$.

3. Para cada $\mu \in \{1,2,\ldots,p\}$ se obtiene $\tau(x^\mu|_S, \tilde{x}|_S)$.

4. Para cada $\mu \in \{1,2,\ldots,p\}$ se obtiene $\theta(\tilde{x}|_S, x^\mu|_S)$, donde la transformada θ del vector $x \in A^n$ con respecto al vector $y \in A^n$ da como resultado un escalar $\theta(x,y)$ definido así: $\theta(x,y) = \sigma_n[\tau(x,y)] + \sigma_n[\tau(y,x)]$, donde el escalar $\sigma_i(x)$ representa el número de componentes con valor 1 que contiene el vector $x \in A^n$ en las primeras i componentes, con $1 \le i \le n$.

5. Encontrar el valor $\psi \in \{1,2,\ldots,p\}$ para el cual se cumple esta expresión: $\theta(\tilde{x}|_S, x^\psi|_S) = \bigwedge_{\mu=1}^{p} \theta(\tilde{x}|_S, x^\mu|_S)$.

6. Se obtiene $\overline{\tilde{x}}$ (el vector negado de $\tilde{x} \in A^n$).

7. Se obtiene la restricción $\overline{\tilde{x}}|_{\hat{S}}$.

8. Para cada $\mu \in \{1,2,\ldots,p\}$ se obtiene $\tau(\overline{\tilde{x}}|_{\hat{S}}, \overline{x^\mu}|_{\hat{S}})$.

9. Para cada $\mu \in \{1,2,\ldots,p\}$ se obtiene $\tau(\overline{x^\mu}|_{\hat{S}}, \overline{\tilde{x}}|_{\hat{S}})$.

10. Para cada $\mu \in \{1,2,\ldots,p\}$ se obtiene $\theta(\overline{\tilde{x}}|_{\hat{S}}, \overline{x^\mu}|_{\hat{S}})$.

11. Encontrar el valor $\varphi \in \{1,2,\ldots,p\}$ para el cual se cumple esta expresión: $\theta(\overline{\tilde{x}}|_{\hat{S}}, \overline{x^\varphi}|_{\hat{S}}) = \bigwedge_{\mu=1}^{p} \theta(\overline{\tilde{x}}|_{\hat{S}}, \overline{x^\mu}|_{\hat{S}})$.

12. Si $\theta(\tilde{x}|_S, x^\psi|_S) \le \theta(\overline{\tilde{x}}|_{\hat{S}}, \overline{x^\varphi}|_{\hat{S}})$, entonces $\omega = \psi$; de otro modo $\omega = \varphi$.

13. Se obtiene $(x^\omega|_S)|^S$, vector que es precisamente x^ω.

No obstante ser de reciente creación, el algoritmo de las Memorias Asociativas Alfa-Beta SVM ha sido aplicado con éxito en algunas tareas de reconocimiento de patrones [8-9], arrojando resultados comparables a los de otros algoritmos de actualidad, y en algiunos casos superando su desempeño.

III. Modelo propuesto y diseño experimental

El modelo de reconocimiento automático de voz emotiva propuesto en este trabajo de investigación, resulta de la fusión del modelo de las Memorias Asociativas Alfa-Beta SVM y la selección, cálculo y representación de los parámetros de señales de voz. Estas señales de voz fueron tomadas de la base de datos de Berlín, la cual se describe en la siguiente sección.

A. Experimentos Tipo 1 (Parámetros Clásicos)

En el primer tipo de experimentos se usan los parámetros clásicos usados en el análisis de voz; se extraen 95 parámetros basados en la energía, la frecuencia fundamental, la duración de los silencios, las primeras 4 formantes y 13 MFCC's. Posteriormente a la selección de estos parámetros, el nuevo modelo incluye el uso de un método de selección de rasgos basados en el encadenamiento hacia adelante, con el fin de mejorar el desempeño del clasificador; como resultado de estos procesos, 14 parámetros fueron seleccionados usando el software WEKA [6]. Finalmente estos datos fueron usados como patrones de entrada para las Memorias Asociativas Alfa-Beta SVM.

Los 95 parámetros que se calcularon para la experimentación de tipo 1, se pueden agrupar en 8 cúmulos diferentes:

- Energía: máximo, promedio, mediana, moda, y desviación estándar.
- Amplitudes de Energía: máximo, mínimo, promedio, mediana, moda, y desviación estándar.
- Duraciones de silencios: máximo, mínimo, promedio, mediana, moda, y desviación estándar.
- Frecuencia fundamental: máximo, mínimo, promedio, mediana, moda, y desviación estándar.
- Duración de frecuencia fundamental: máximo, mínimo, promedio, mediana, moda, y desviación estándar.
- Sonoridad: máximo, mínimo, promedio, mediana, moda, y desviación estándar.
- MFCC's: Del archivo de audio, se extrae una matriz con los coeficientes MFCC's; los renglones representan 13 coeficientes MFCC's y las columnas representan el número de ventanas que tiene el archivo. Seis vectores son calculados de esta matriz; el primero contiene los valores máximos de la matriz, mientras que el segundo contien los mínimos: En el tercero están los promedios y en el cuarto las medianas; el quinto contiene las modas y el sexto las desviaciones estándar. Posteriormente, a fin de reducir la cantidad de estos vectores, se calcula el máximo, mínimo, promedio, mediana, moda y desviación estándar para representar cada uno de los grupos de MFCC´s.
- De los primeros 4 formantes: máximo, mínimo, promedio, mediana, moda, y desviación estándar.

Para el método de selección de rasgos incluido en el nuevo modelo, fue seleccionado el de Encadenamiento hacia adelante, que empieza con un conjunto vacío, se busca luego el atributo que más aporte a la clasificación y se va añadiendo hasta que al añadir el siguiente parámetro haga que caiga el desempeño de la clasificación.

Usando este método en WEKA [6], 14 parámetros fueron extraídos y se enuncian en la TABLA III.

TABLA III
PARÁMETROS SELECCIONADOS POR ENCADENAMIENTO HACIA ADELANTE

ENERGÍA PROMEDIO
AMPLITUD MÁXIMA DE LA ENERGÍA
MÍNIMO DE LA AMPLITUD DE LA ENERGÍA
AMPLITUD PROMEDIO DE LA ENERGÍA
DESVIACIÓN ESTÁNDAR DE LA AMPLITUD DE LA ENERGÍA
MODA DE LAS DURACIONES DE LA FRECUENCIA FUNDAMENTAL
PROMEDIO DE LA SONORIDAD
MFCC MÁXIMO DE LOS PROMEDIOS
MFCC MÍNIMO DE LOS MÍNIMOS
MFCC MÍNIMO DE LOS PROMEDIOS
MFCC MÍNIMO DE LAS DESVIACIONES ESTÁNDAR
MFCC PROMEDIO DE LAS DESVIACIONES ESTÁNDAR
MFCC MEDIANA DE LOS PROMEDIOS
DESVIACIÓN ESTÁNDAR DE LA SEGUNDA FORMANTE

B. Experimentos Tipo 2 (Representaciones Bidimensionales)

El segundo tipo de experimentos incluye el desarrollo de otro enfoque que es original de este trabajo. Como todos los parámetros extraídos en el primer experimento son medidas estadísticas de la energía, frecuencia fundamental, duración de silencios, formantes, 13 MFCC's, tales como el promedio, moda, máximo, mínimo, desviación estándar y mediana, prácticamente son medidas de dispersión que representan en este caso uno o varios vectores de datos; sin embargo, en esta experimentación se observó que el parámetro que más información afectiva extrae de la voz, es la energía, por lo que se decidió trabajar con ese único parámetro: la energía.

Aquí surge la aportación principal del nuevo modelo asociativo de reconocimiento automático de voz emotiva: la representación bidimensional del parámetro energía, cuyo proceso se inicia con la extracción de los valores de energía que genera la envolvente (Figura 1).

Posteriormente, se rellena toda el área bajo la envolvente y se realiza una normalización en el eje de la amplitud para homogeneizar todas las instancias (Figuras 2 y 3).

Alfa-Beta SVM es una herramienta de clasificación efectiva para tareas de clasificación de datos binarios, por lo que se usó la representación bidimensional en forma de matriz como datos de entrada para el proceso de clasificación.

Fig. 1. Representación bidimensional del parámetro energía.

Fig. 2. Relleno de toda el área baljo la envolvente.

Fig. 3. Normalización en el eje de la amplitud.

IV. RESULTADOS EXPERIMENTALES Y DISCUSIÓN

Trabajando en el campo de reconocimiento de voz emotiva, se debe elegir un corpus. Hay algunas bases de datos diseñadas para éste propósito [7], y la mayoría de las emociones comúnmente utilizadas en diferentes corpus de voz son: enojado, tristeza, felicidad, miedo, disgusto, alegría, sorpresa y aburrimiento.

La base de datos de Berlín [1] fue seleccionada por su disponibilidad, ya que la mayoría de las bases de datos reportadas en la literatura son de uso privado o requieren de una onerosa licencia de uso. La base de datos de Berlín contiene 535 instancias, etiquetadas como 127 en estado enojado, 81 de aburrido, 46 de disgustado, 69 de miedo, 71 de felicidad, 62 de tristeza y 79 de neutral. Este corpus fue grabado por 10 actores profesionales, 5 varones y 5 mujeres; contiene registros de 10 diferentes sentencias, con una frecuencia de muestreo de 16,000 datos por segundo en formato wav.

Dado que es la única base de datos disponible sin cargos, muchos investigadores la usan por lo que una de las consecuancias de este hecho es que los resultados de clasificación pueden ser fácilmente comparables y sus emociones son usualmente incluidas en otras bases de datos. La base de datos de Berlín está orientada al reconocimiento de voz emotiva actuada en idioma Alemán.

Al realizar los experimentos tipo 1, se sometieron los 14 parámetros que resultaron de la selección de rasgos de los 95 parámetros extraídos de la base de datos de Berlín, con los modelos que presenta WEKA [6]. Se mustran únicamente los modelos que produjeron los resultados más altos.

El primero de ellos es el Naive-Bayes (ver TABLA IV), que al clasificar toda la base de datos, recuperó correctamente 327 instáncias (61.12%). El segundo modelo que se probó es el de Simple-Logistic, que alcanzó un precisión del 79.81%; es decir, recuperó satisfactoriamente 427 instancias.

El modelo que alcanzó el desempeño más alto de los incluidos en [6], es el Perceptrón Multicapa, que alcanzó un desempeño del (86.54%); es decir, 463 instancias correctamente clasificadas.

TABLA IV
DESEMPEÑO ALCANZADO POR LOS MODELOS EN WEKA [6] Y POR LAS MEMORIAS ASOCIATIVAS ALFA-BETA SVM

MODELO	INSTANCIAS CORRECTAS	INSTANCIAS CORRECTAS	DESEMPEÑO (%)
NAIVE BAYES	327	208	61.12
SIMPLE LOGISTIC	427	108	79.81
PERCEPTRON MULTICAPA	463	72	86.54
ALFA-BETA SVM	**508**	**27**	**94.95**

Es importante hacer notar que el nuevo modelo asociativo de reconocimiento automático de voz emotiva basado en las memorias asociativas Alfa-Beta SVM, el cual ha sido introducido en este artículo, recuperó 508 instancias

correctamente; es decir, alcanzó un desempeño del **94.95%**, con lo cual superó a todos los demás modelos.

Respecto de los experimentos de Tipo 2, al usar las imágenes de la energía normalizadas en el eje de la amplitud, con un relleno por debajo de la envolvente de la señal de energía junto con las máquinas asociativas Alfa-Beta SVM como modelo de clasificación, se clasificaron correctamente 506 instancias; es decir, se logró un desempeño del **94.5%**.

Este resultado está muy cercano al alcanzado por los experimentos de Tipo 1, con la aclaración relevante de que en este caso se usa **sólo un parámetro**: la energía.

V. Conclusiones

La selección de rasgos que se obtuvo de 14 parámetros ha demostrado ser buena. Esto se verifica cuando estos parámetros se usan como entrada en el nuevo modelo asociativo de reconocimiento automático de voz emotiva basado en las memorias asociativas Alfa-Beta SVM.

Al hacer lo anterior, el desempeño resulta casi del 95%, superando a los clasificadores de emociones conocidos, dado que el desempeño más alto reportado en la literatura, con esta misma base de datos, se encuentra alrededor del 80% [13, 14].

Es notable el alto desempeño alcanzado por el nuevo modelo asociativo de reconocimiento automático de voz emotiva basado en las memorias asociativas Alfa-Beta al usar como entrada un único parámetro, consistente en la representación bidimensional de la energía extraída de la señal de voz. El resultado es competitivo con el que se obtiene con 14 parámetros clásicos.

Agradecimientos

Los autores agradecen el apoyo de las siguientes instituciones para la realización de esta obra: Secretaría de Investigación y Posgrado, Secretaría Académica, COFAA y CIC del Instituto Politécnico Nacional, CONACyT y Sistema Nacional de Investigadores (SNI); específicamente, los proyectos SIP-20090807, SIP-20101709 y SIP-20110661.

Referencias

[1] *Berlin emotional speech database*, http//www.expresive-speech.net/

[1] T. Vogt, E. André, and J. Wagner, *Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization, Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, Springer-Verlag, Berlin, Heidelberg, 2008.

[2] J. Cichosz and K. Slot, "Emotion recognition in speech signal using emotion-extracting binary decision trees," in *Proceedings of Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 2007.

[3] T. Vogt, and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proceedings of Language Resources and Evaluation Conference*, 2006.

[4] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Hierarchical classification of emotional speech," *IEEE Transactions on Multimedia*, 2007.

[5] H. Ian, and F. Eibe, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, available online at http://www.cs.waikato.ac.nz/ml/weka/, 2005.

[6] D. Ververidis and C Kotropoulos, "A state of the art review on emotional speech databases," in *Proceedings of 1st Richmedia Conference*, 2003, pp. 109–119.

[7] L. López-Leyva, C. Yáñez-Márquez, R. Flores-Carapia, and O. Camacho-Nieto, "Handwritten Digit Classification Based on Alpha-Beta Associative Model," in *Progress in Pattern Recognition, Image Analysis and Applications. LNCS 5197, Proc. 13th Iberoamerican Congress on Pattern Recognition CIARP 2008*, Havana, Cuba, 2008.

[8] L. López-Leyva, C. Yáñez-Márquez, and I. López-Yáñez, "A new efficient model of support vector machines: ALFA–BETA SVM," in *23rd ISPE International Conference on CAD/CAM, Robotics and Factories of the Future*, Bogotá, Colombia, 2007.

[9] C. Yáñez-Márquez, *Associative Memories Based on Order Relations and Binary Operators* (In Spanish). PhD Thesis. Center for Computing Research, Mexico, 2002.

[10] T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Berlin Heidelberg New York, 1989.

[11] M. H. Hassoun, *Associative Neural Memories*, Oxford University Press, New York, 1993.

[12] T. Kohonen, "Correlation Matrix Memories," *IEEE Transactions on Computers*, 21(4), 353–359, 1972.

[13] M. El Ayadi, M. Kamel, and F. Darray, "Survey on speech emotion recognition: Features, classification schemes and databases," *Pattern Recognition*, vol. 44, March, (2011) 572-587.

[14] S. Zhang, L. Li, and Z. Zhao, "Spoken emotion recognition using kernel discriminant locally linear embedding," *Electronics Letters*, vol 46, 1344–1346, 2010.

A Dynamic Model for Identification of Emotional Expressions

Rafael A.M. Gonçalves, Diego R. Cueva, Marcos R. Pereira-Barretto, and Fabio G. Cozman

***Abstract*—This paper discusses the dynamics of emotion recognition on faces, layering basic capabilities of an emotion sensor. It also introduces a model for the recognition of the overall conveyed emotion during a human-machine interaction, based on the emotional trajectory over an emotional surface.**

***Index Terms*—Emotion dynamics, emotion recognition, emotional surface, Kalman filtering.**

I. Introduction

PERSON-to-person communication is highly non-verbal: face, body and prosody demonstrate much of what is not being said but loudly spoken. Therefore, it is expected that human-machine communication may benefit from non-verbal expressions. This may have already been started as the so-called "user centric experience", by having applications and games with voice and gesture recognition, for instance. But recognizing emotions is not easy not even for humans: it has been shown humans correctly recognize the conveyed emotion in the voice 60% of the cases and 70% to 98% on the face [1], [2]. This paper focuses on emotion recognition on faces.

In the 70´s, Ekman and co-workers established FACS (Facial Action Coding Sys-tem), a seminal work for emotion recognition on faces [3], by decomposing the face into AUs (Action Units) and assembling them together to characterize an emotion. The universality of AUs was strongly debated for the last two decades but inter-cultural studies and experiences with pre-literate populations lead to its acceptance [2]. A state-of-the-art review of emotion detection on faces can be found in [4]. Among the most recent works, we cite eMotion, developed at Universiteit van Amsterdam [5] and FaceDetect, by Fraunhofer Institute [6].

Both eMotion and FaceDetect detect an emotion from each frame on a video (or a small sequence of frames). Therefore, they show excelent results in posed, semi-static situations. But during a conversation, the face is distorted to speak in many ways, leading these softwares to incorrectly detecting the conveyed emotion. Even more, a movement of the mouth during a conversation, similar to a smile, does not mean the speaker is happy; it may be an instantaneous emotion: the speaker saw something not related to the conversation which made him smile. Consider, as an example, the frames from a video, shown in Figure 1 and the outputs from eMotion, on Figure 2.

From the frames on Figure 1, a human would conclude nothing. From eMotion outputs on Figure 2, a human would conclude nothing, also. Or perhaps for Sadness, which seems to display a higher mean value. But by seeing the video, even without sound, a human would easily conclude for Anger.

Therefore, during a conversation, there is a "slow dynamic" related to the overall emotion conveyed, lasting longer than a single video frame. During a conversation, many "emotional modes" (as vibrational modes in Mechanics) may be displayed, invoked by events (internal or external) to the speaker but probably out of the reach for the listener. These modes are interleaved within the conversation, somewhat as it happens with appositive phrases [7]. This work discusses a general model for the detection of emotional modes and presents a model to detect slow dynamic emotions. Some reference material is presented on Section 2, while Section 3 presents the general model and Sections 4 to the end present the proposed model for the detection of emotional modes.

II. Reference Material

Behaviorist theories dominated the Psychology scene from the 30´s to 60´s. According to them, emotions are only a dimension of human behavior, corresponding to a certain degree of energy or activity. The determinist characteristic and one-dimensional associations of event-emotion are on the basis of these theories: for each event, an associated emotion. Appraisal Theories took their place during the 80´s, although started during the 60´s. Simply put, they postulate emotions are elicited from appraisals [8]. Appraisals differ from person to person but the appraisal processes are the same for all persons. Therefore, they offer a model which justifies a common behavior but, at the same time, allows for individual differences.

This work is based on the concept of emotions which, on Scherer (2001) words, are "... an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism".

The appraisal process starts with an event. We argue the perceived emotion should be considered as an event, as much as a strong noise such as an explosion. Therefore, it will be evaluated for its relevance, according to (i) novelty, (ii) intrinsic pleasantness; (iii) Goal/need relevance.

Manuscript received June 10, 2011. Manuscript accepted for publication September 14, 2011.

This work is supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), FAPESP, under research project 2008/03995-5 and the University of São Paulo.

Rafael A.M. Gonçalves, Diego R. Cueva, Prof. Dr. Marcos R. Pereira-Barretto and Prof. Dr. Fabio G. Cozman are with the Decision Making Lab of the Mechatronics Department, University of São Paulo, São Paulo, Brasil. Electronic correspondence regarding this article should be sent to Prof. Dr. Marcos R. Pereira-Barretto to mpbarre@usp.br.

Fig 1. From left to right, eMotion classified the frames as happiness (100%), sadness (70%), fear (83%) and anger (76%).

A specialized sensor such as the one proposed here is need to detect this kind of event. Going further, the perception of an emotion is altered by Attention and conditioned by Memory and Motivation; an emotion sensor should be adaptable as the eyes and ears.

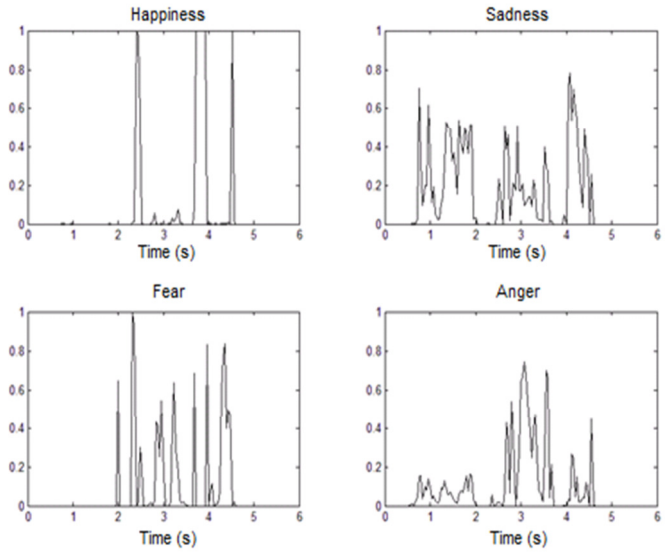


Fig 2. eMotion output for the video of Fig 1.

III. Reference Model

Emotion detection from video frames has been subject to research by many authors, as described on [4], [5] and [6]. Despite the advances, it is still an open subject; re-search on compensation of lightning, speech and body movements are some examples. These works are "raw sensors", on Figure 3. On the top of these sensors, we argue for the need of "emotional mode" detectors, for fast and slow dynamics. Con-sider, for instance, a conversation with a friend: the overall conveyed emotion could be Happiness, the slow dynamics. But suddenly the speaker reminds of someone he hates: Anger may be displayed. The event could be external: the speaker may see someone doing something wrong and also to display Anger. In both cases, Anger is displayed as the fast dynamics, enduring for more than just a few video frames. For the listener, the appraisal process could lead to just continue the conversation, ignoring Anger. Or change the subject to investigate what caused this change in speaker´s face.

IV. Proposed Model for Detection of Emotional Modes

The proposed model to determine the perceived emotion from instantaneous facial expressions is based on the displacement of a particle over a surface, subject to ve-locity changes proportional to the current probability of each emotion, at every mo-ment. This surface will be called here Dynamic Emotional Surface (DES). Over the surface, attractors corresponding to each detectable emotion are placed.

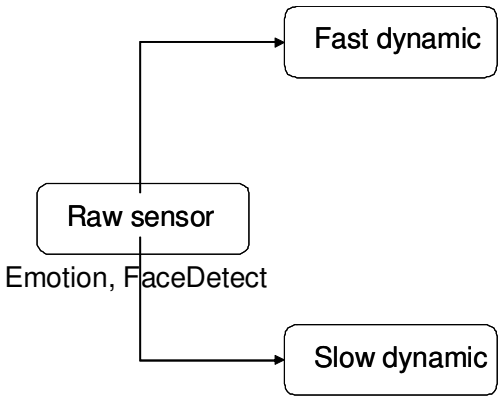


Fig 3. Reference model.

The particle, which represents the instantaneous emotion, moves freely over the DES, subject to attraction to each detectable emotion. It also moves toward the neutral state, placed at the origin of the coordinate system, the point of minimum energy. Therefore, the particle velocity can be determined by Eq. 1.

$$\vec{V}_p = \vec{V}_e + \sum_{a=1}^{N} \vec{V}_a \quad (1)$$

where:

$\vec{V}_p$: particle velocity

$\vec{V}_e$: initial velocity

$\vec{V}_a$: velocity in direction of each attractor or detectable emotion.

The idea, an emotional surface, comes from works such as [10], [11], [12], [13], [14], shown in Figure 4. This surface represents the appraised emotion while DES represents the

perceived emotion; they keep some relationship because the speaker should display a "reasonable" behavior.

DES keeps also some relationship with the Arousal-Valence plane [15], but differs for the same reasons as from Zeeman´s surface.

As an example, suppose the following paraboloid is a DES with the attractors listed in Table I:

$$z = f(x,y) = ax^2 + by^2 \quad (2)$$

$$\gamma(x,y) = (x, y, ax^2 + by^2) \quad (3)$$

$$a = b = 0{,}6 \quad (4)$$

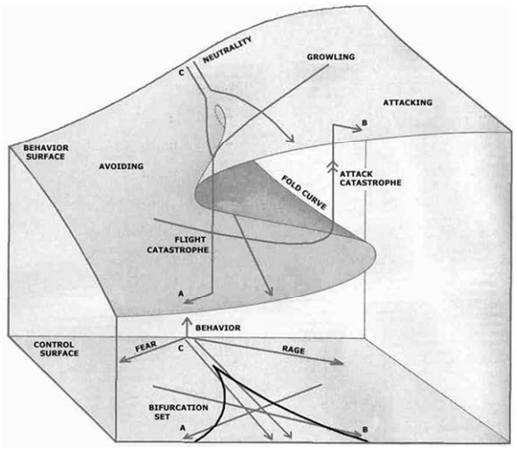


Fig 4. Zeeman´s Emotional Surface [10]

TABLE I
ATTRACTOR PROJECTIONS

Emotion	Attractor Projection
Happiness	$[60, 60, 0]$
Anger	$[-60, 60, 0]$
Sadness	$[-60, -60, 0]$
Fear	$[60, -60, 0]$

This example DES helps in highlighting the differences to the A-V Plane: Fear has been placed on 4th quadrant, while on the A-V Plane is positioned on the 3rd, close to the origin. But this positioning follows Zeeman´s surface, where Fear and Anger are orthogonal.

The velocity in direction of each attractor, $\vec{V}_a$, is proportional to the probability of each emotion as detected by existing software such as eMotion. Considering as $\vec{P}$ the current particle position and $\vec{A}$ the position of the attractor (emotion) being calculated, V_a can be calculated as:

$$\overrightarrow{AP} = \vec{A} - \vec{P} = [a_{px}, a_{py}, a_{pz}] \quad (5)$$

$$S(x) = \gamma(x, rx) \quad (6)$$

$$r = \left|\frac{a_{py}}{a_{px}}\right|, a_{px} \neq 0 \quad (7)$$

$$scale = \left|\frac{dS(x)}{dx}\right| = \sqrt{1 + r^2 + [2(a + br^2) * P_x]^2} \quad (8)$$

$$V_{x,a} = \frac{dS}{dt} * \frac{signal(a_{px})}{scale} * \frac{dS}{dx} \cdot \vec{\imath} \quad (9)$$

$$V_{y,a} = \frac{dS}{dt} * \frac{signal(a_{py})}{scale} * \frac{dS}{dx} \cdot \vec{\jmath} \quad (10)$$

Sensor input is always noisy; that is the case for eMotion also: in this case, noise comes from the frame-by-frame emotion detection. A pre-filtering can be applied to its output prior to submit to the model. Both Kalman filtering and moving-average filtering were tested, as shown in what follows.

V. EXPERIMENTS

Experiments were conducted to test the proposed model for the detection of slow dynamic.

Videos from eNTERFACE'05 Audio-Visual Emotion Database corpus were selected from those available displaying the emotions under study: 7 showing Fear, 9 showing Anger, 5 showing Happiness, 9 showing Sadness and 4 for Neutral. For each video, the facial mesh was adjusted on eMotion and its output collected.

A Kalman filter with process function in the form of Eq. 9 was adjusted, using 4 eMotion outputs for each emotion, given the parameters shown in Table II.

$$\frac{Y(s)}{R(s)} = \frac{K_k}{\tau s + 1} \quad (11)$$

TABLE II
PARAMETERS OF KALMAN FILTER

	Q	R	K_k	τ
Happiness	0.1	0.080	5	1.5
Anger	0.1	0.100	5	1.5
Sadness	0.1	0.035	5	1.5
Fear	0.1	0.010	5	1.5

TABLE III
COMPARISON BETWEEN UNFILTERED SIGNALS, MOVING AVERAGE AND PROPOSED KALMAN FILTERING WITH DES

	Original		Moving Average		Kalman	
Emotion	μ	σ	μ	σ	μ	σ
Happiness	0.175	0.634	0.175	0.237	0.114	0.127
Sadness	0.377	0.532	0.377	0.254	0.207	0.108
Fear	0.211	0.544	0.211	0.206	0.234	0.203
Anger	0.236	0.434	0.236	0.257	0.445	0.434

Applying this filter and moving-average filtering to the video whose frames are displayed on Figure 1 gave the results shown on Figure 5.

For both implementations, mean value and standard deviation were calculated as shown in Table III.

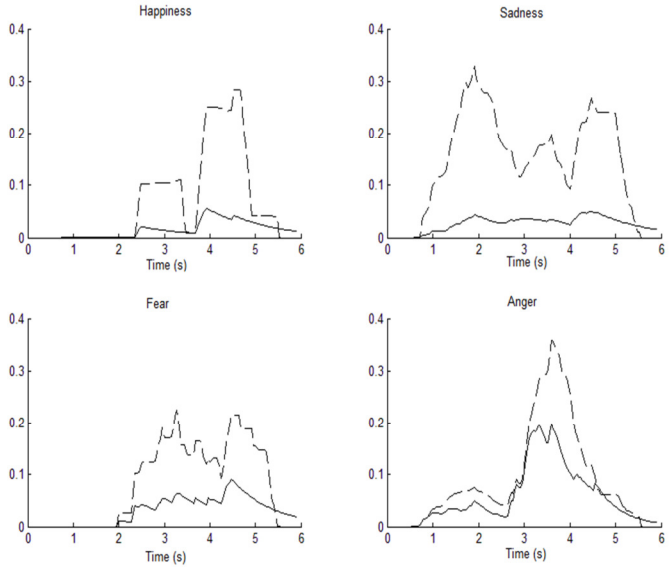


Fig 5. Moving Average (dashed) and Proposed Kalman Filtering with DES algorithm (solid) outputs for example video (see Fig 1).

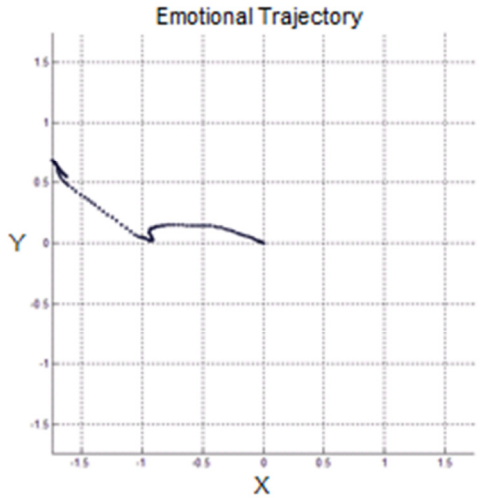


Fig 6. Projection of the Emotional Trajectory of the Sample Video.

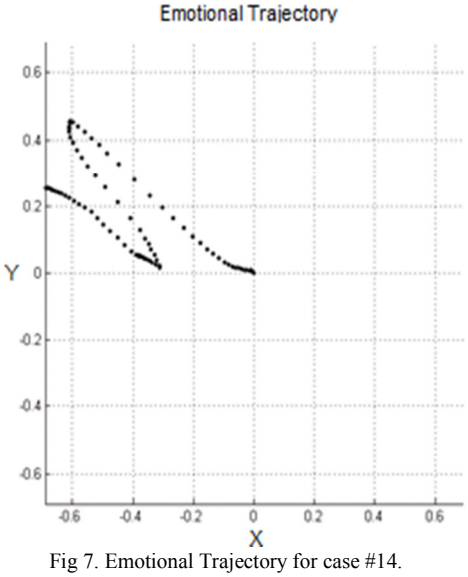


Fig 7. Emotional Trajectory for case #14.

The 14 remaining videos, i.e., those not used for adjusting Kalman filter, were then submitted to the system, yielding to the results shown in Table IV.

As it can be seen, the overall emotion conveyed by the video, Anger, has been correctly detected with Kalman filtering, although with a large std deviation. The projection on X-Y plane of the trajectory over the DES is shown in Figure 6.

As it can be seen, Anger is present almost all the time, starting mild but going stronger as the video continues. This corresponds to the human observation of the video.

TABLE IV
COMPARISON BETWEEN HUMAN EVALUATION AND THE PROPOSED KALMAN FILTERING WITH DES ALGORITHM

#	File	Classifications	
		Human	System
1	S1sa1	Sadness	Sadness
2	S38an1	Anger	Anger
3	S38fe3	Fear	Fear
4	S42sa1	Sadness	Sadness
5	S43ha1	Happiness	Happiness
6	S43an2	Anger	Anger
7	S43an3	Anger	Anger
8	S43an4	Anger	Anger
9	S43fe2	Fear	Fear
10	S42fe1	Fear	Fear
11	S43sa1	Sadness	Sadness
12	S43sa3	Sadness	Sadness
13	S43sa4	Sadness	Sadness
14	S43sa5	Sadness	Anger

As shown before for the sample video (see fig. 6), we may plot the emotional trajectory estimated for S43sa05 (#14):

Note the special case of video #14, showing Anger/Sadness swings, shown in Figure 7, which corresponds to author´s analysis.

VI. CONCLUSION

A reference model for recognition of emotions on faces has been introduced, besides a computational model for computing fast and slow conveyed emotions. The model has been tested for the detection of slow emotions, demonstrating good results.

As future works, the authors plan to test the model for fast emotions. The main obstacle foreseen is the lack of a corpus for this kind of test. The authors also plan to investigate the interaction between the proposed model and CPM processes.

REFERENCES

[1] R.W. Piccard, *Affective Computing*, MIT Press, 1997.
[2] P. Ekman and W.V. Friesen, *Unmasking the face*, Malor Books, 2003.

[3] P. Ekman and W.V. Friesen, *Facial Action Coding System: a technique for the measurement of facial movement*, Consulting Psychologists Press, 1978.

[4] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: state of art," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22 no. 12, 2000.

[5] A. Azcarate, F. Hageloh, K. Sande, and R. Valenti, *Automatic facial emotion recognition*, Universiteit van Amsterdam, 2005.

[6] Fraunhofer Facedetect.
http://www.iis.fraunhofer.de/en/bf/bv/ks/gpe/demo
Accessed 07/04/2001.

[7] R. Cooper, *Quantification and Syntactic Theory*, D. Reidel Publishing Company; 1983.

[8] I.J. Roseman and C.A. Smith, "Appraisal Theory - Overview, Assumptions, Varieties, Controversies," *Appraisal Processes in Emotion - Theory, Methods, Research,* edited by K. Scherer, A. Schorr, and T. Johnstone, Oxford University Press, 2001.

[9] R.R. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal Processes in Emotion - Theory, Methods, Research,* edited by K. Scherer, A. Schorr, and T. Johnstone, Oxford University Press, 2001.

[10] E.C. Zeeman, "Catastrophe theory," *Scientific American,* vol.4 no.254 pages 65−83, 1976.

[11] I.N. Stewart and P.L. Peregoy, "Catastrophe theory modeling in Psychology," *Psychological Bulletin,* vol. 94, no. 2, pp. 336-362, 1983.

[12] K. Scherer, "Emotions as episodes of subsystem synchronization driven by nonlinear appraisal processes," *Dynamic Systems Approaches to Emotional Development,* ed. by M.D. Lewis and I. Granic, Cambridge Press, 2000.

[13] H.L.J. van der Maas and P.C.M. Molenaar, "Stagewise cognitive development: an application of Catastrophe Theory," *Psychological Review,* vol.99, no.2, pp. 395−417, 1992.

[14] D. Sander, D. Grandjean, and K.R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks,* vol.18, pp. 317−352, 2005.

[15] H. Gunes and M. Piccardi, "Observer Annotation of Affective Display and Evaluation of Expressivity: Face vs. Face-and-Body," in *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia, 2006.

Optical Parameter Extraction using Differential Evolution Rendering in the Loop

Mauricio Olguin Carbajal, Ricardo Barrón Fernández, and José Luis Oropeza Rodríguez

Abstract— **Image synthesis is highly dependent on rendering algorithm and optical properties of scenario objects. The goal of this work is to develop a methodology to obtain some illumination parameters of a real scenario represented by an acquired image, and use these parameters for a virtual scenario rendering with the same objects as the original. The proposed methodology consists, first, in acquiring an image of the working scenario, and by using a DE (Differential Evolution) algorithm to render images that gradually approximate to the real acquired image, by some virtual scenario parameter modification based on the DE optimization. We call it "ED Rendering in the loop". Finally we use the obtained parameters to render an image to compare it with similar methods.**

Index terms—**Differential evolution, rendering, images, loop, illumination, model.**

I. Introduction

THE The goal of this study is the development and test for a methodology that can make the parameter extraction needed to render photorealistic images by using DE - rendering algorithm. The illumination model used greatly depends on virtual scenario optical parameters for final image rendering. Some of these parameters are: object color, reflection index, refraction index, texture, etc. Also we must define light position, intensity, color, attenuation index, etc. The illumination model uses these and many other parameters for final image rendering. This present study proposes a scenario parameters extraction methodology from a real scenario by using a more objective evaluation and procedure.

II. Related Work

Realistic image synthesis efforts have been focused in accurate and efficient rendering algorithm development, mainly based on Montecarlo [1] and analytic approximations. This algorithm has parameters which directly or indirectly represents optical properties of scenario objects. To achieve the rendered image having a realistic appearance, it is necessary for these parameters to be exact. There are few methods to measure or estimate optical parameters that are of relevance for computer graphics.

Manuscript received May 26, 2011. Manuscript accepted for publication August 24, 2011.

Mauricio Olguin Carbajal is with Postgraduate department, Centro de Innovacion y Desarrollo Tecnológico en Computo (CIDETEC), IPN, Mexico City, Mexico (e-mail: molguinc@ipn.mx).

Ricardo Barrón Fernández and José Luis Oropeza Rodríguez are with with Center for Computing Research (CIC), National Polytechnic Institute (IPN), UP "Adolfo Lopez Mateos", Mexico City, Mexico (e-mail: rbarron@cic.ipn.mx, joropeza@cic.ipn.mx).

A. Parameter Extraction

Existent methods for optical physics parameter measure require the use of very expensive equipment, for example the one used for scattering properties measure for colloidal chemistry [4] which give very little usable data for computer graphics. More recently there has been made efforts to measure or estimate objects optical physical properties for using these parameters in computer graphics with modern rendering algorithms. In accordance with Srinivasa G. Narasimhan et al. [2] there exist two object optical properties estimation methods: direct measurement and indirect estimation. For direct measure some optical methods like Goniphotometry have been used which measure phase function for translucent media scattering [7]. Indirect estimation uses an analytical approach [5] or numerical solutions for light transport [6]. Fukawa et al. [8] uses an acquisition device, based on laser range scanner to obtain a 3D model of objects texture, later this texture 3D model it will be used in its virtual objects. Also Gero Muller et al. [9] use an acquisition parameters system formed by a hemisphere with fixed cameras and lights which use massive parallel processing with no moving parts, this approach obtains an object 3D texture as well as the object color and reflection index. However the system requires expensive equipment with a lot of processing power.

Wojciech Matusik et al. [10] use a 6 cameras and a light array as an acquisition device, putting objects on a spinning table and uses a multi-background technique for alpha channel acquisition and mate environment for multiple viewpoints which achieves tridimensional object appearance reconstruction with color, reflection and refraction index. Later, these parameters could be used for virtual objects; however equipment and processing are still expensive.

G. Müller et al. [11] have designed a device with a single camera and a single light source which is still, while the camera moves in a semicircle to obtain reflection and color parameters of realistic 2d texture materials varying conditions namely Bidirectional Texture Function (BTF). This technique gives amazing results, however it is still an expensive and time consuming process. The great data amount generated by a system like this also requires a compression for acquired data. Non compressed data can require a lot of space (over 1 gigabyte store space for 81 views and 81 lights for 256x256 texture patch in accordance to Wai Kit Addy Ngan [3]).

Simpler approximations to measure optical parameters have been developed, for example Srinivasa G. Narasimhan et al. [2] made a simple device and technique to estimate scattering

properties for a broad class of different media. An acquisition device is a small media containing tank with a light source inside, and an isolated chamber under the media where a camera measures scattering parameters effects, as well as a phase and absorption light index. Later the obtained parameters will be used on a Montecarlo rendering algorithm for photorealistic image synthesis, with impressive results. In this paper we present a simple acquisition device and use of a heuristic population technique (DE), for optical parameter estimation. Unlike previous approaches that require complicated setup or devices, our method and setup can be used to estimate the illumination parameters that the user requires using very few resources.

B. *Rendering Parameters of Optimization*

There are some previous works related to rendering parameters optimization, and here we show the most relevant. The first is a proposal to design the environment lighting by using optimization techniques applied to a rendering system that uses Radiosity based on an image synthesis system, developed by Kawai, Painter and Cohen [12]. This proposal is based on the illumination parameter optimization and works based on targets and constraints that the user defined to modify the environment lighting. This radioptimization system finds the "best" possible sets for: light source emissivity, elements reflexivity and light source directionality. The Kawai et al. proposal already has a pre-designed scenario where the parameters of objects such as color, hardness, specular reflectivity and diffraction are proposed by the user in advance, unlike ours proposed during the extraction of parameters. As an optimization method Kawai worked with a constrained optimization system that uses the BFGS (Broyden, Fletcher, Goldfarb and Shanno) algorithm, which evaluates the objective function and the gradient in the current step for the design of space to calculate a search direction.

Yu, Debeverec, Malik and Hawkins [13] as well as Yu [14], analyzes the reflection models recover problem for realistic scenes from photographs. This method recovers real scene reflectance properties for all the surfaces by using photograph sets, and then rendering a virtual version of the real scene in which textures are mapped to the scene from the real one, but which responds to virtual lighting conditions. The goal is to find the parameters of the BRDF (Bidirectional Reflectance Distribution Function), for use in rendering. This method allows adding arbitrary changes to the structure and lighting, such as extra items. The system input is a geometric model of the scene and a set of high dynamic range photographs taken with known direct illumination.

Once the parameters are obtained and the image is rendered, we proceed to compare the generated images with real scenario images, both in original and in new conditions; the result is that the methodology adequately predicts the resulting image under new lighting. The parameters obtained are the diffuse and specular reflectivity for red, green and blue color components. The problem of inverse lighting has also been approached by Schonenman et al. in their article "Painting with light". Their solution is proposed so the stage designer uses a "light brush" with which the user can specify areas of an image rendered from the scene which they wish to illuminate with a certain level of intensity. The system looks for the best configuration of lights to illuminate the scene by minimizing the difference between the rendered scene and the desired lighting. As a rendering algorithm they use ray trace.

Finally, Elorza and Rudomin [15], develop a proposal which makes lighting design by image rendering in closed environments, based on a solution to the problem of inverse lighting, using a genetic algorithm as optimization technique and an algorithm for radiosity as a rendering technique. The parameters optimized are the number, position and intensity of light sources used in the scene. The goal is to find the parameters that render an image by minimizing energy use and maximizing lighting.

Our proposal, unlike the previous ones, gets a broader set of parameters such as: rates of diffuse reflection, diffraction, roughness, and brightness, and the parameters of the light source such as position and intensity.

III. RENDERING IN THE LOOP

Our proposal is a methodology for optical parameter estimation, based on a simple acquisition image device, a two image comparison function (one acquired and one rendered), and DE algorithm that gradually finds good optical input parameters for our rendered image and compares with our acquired image (Aimage). Thus, we can obtain our particle fitness value depending on its similarity with reference image. As DE generations pass, rendered images reduce its difference values respect to Aimage. At run end we have a similar set of images and a file with the scenario parameters needed to render these images. These parameters could be used to render new images at greater resolutions and then used to render more complicated environments or animations. As we can see, figure 1, receive two inputs, first, a virtual scenario describing the objects, positions and orientations of the real one; second, a reference image (Aimage), which is our target image.

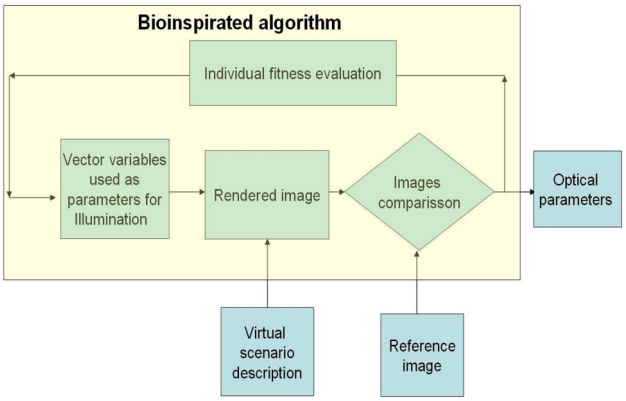


Figure 1. Metodology block diagram.

The normal form of parameter estimation can be conceptualized as an open-loop system, where you have a system that performs the extraction/estimation of parameters,

receiving input data to be used, but making the comparison out of the system. However, our proposed methodology can be conceptualized as a closed-loop in which the comparison step is not carried out, but included in the parameters estimation.

A. *Image Acquiring System*

Image acquiring system, fig.1a, consist of a 15x15x20 cm. box, which has 2 high luminosity led's as a light source, as well as a hole to a camera, for image acquisition. At 6 cm. from the bottom of the box we put a false floor section to sustain objects for parameter extraction.

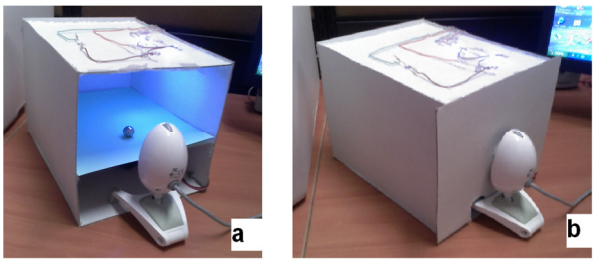


Figure 2. Acquisition image system.

To minimize external reflections we put a front side cover, fig. 1b, in a way that only the lens of the camera is visible, and not all the camera. For comparing both images we propose eq. 1. Comparison function gives one value for each RGB component. With a high component value, the difference between both images is big; with a little component value the difference is small.

$$Dif = \sum_{i=0}^{i=n} | RGBObj_i - RGBTraza_i |$$

where:
n = Xmax * Y max,
$RGBObj_i$ = Each pixel *i* component objective image value,
$RGBTraza_i$ = Each pixel *i* component rendered image value.

B. *Proposed Algorithm*

We present the pseudo code for the rendering in the loop algorithm as follows:

```
ImageO = Acquisition image function
Initialize individuals
  For each individual
    For each variable
    Select one random value between 0.0 and 1.0
      while (Generation_number, or stop criterion is not
      reached)
         for individual=1 to NumberOfIndividuals
           Calculate Fitness
              Use individual variables as raytracer virtual
               world parameters
              Run raytracer and render imageT
              Compare ImageO with ImageT
           Calculate Fitness
           If (U_Test_vector_fitness > XiG_vector_Fitness)
              Assign U_Test_vector_fitness to
              XiG_vector_Fitness
           Calculate Global Best
         End individual loop
      Next Generation until stop criterion is reached
```

The proposed algorithm receives a reference image that has been obtained by the acquisition system and stored into a 128x96 matrix called "MatrizObj". Then we initialize the DE individuals in order to generate a population to evaluate the aptitude of each individual. The evaluation consists of two steps:

- Using the X vector of one chosen individual, we assign to each variable vector a parameter from a virtual world. The virtual world is used as input for a raytrace algorithm to generate an image of 128 x 96, stored in a Matrix named "MatrizTraza".
- We call comparison function for two images, taking as input both matrixes "MatrizObj" and "MatrizTraza" to generate three numeric values, which is the difference between two images RGB values. These values will be stored as individual fitness. At the same time we store in a BMP file the best individual values from each generation. Only one image will be saved per each 10 generations.

The DE evaluates every individual, then selects the best and compares the aptitude from random and test individuals and selects the best from both. The best individual information will be used by DE to calculate the global best. At the same time we store in a TXT file the global best individual values from each 10 generations.

As a rendering algorithm, we use a raytracing with Phong-Blinn shading. This basic illumination model is used due to the low amount of mathematical calculations needed, compared with more complex models. We know that final image realism could be limited, but it is enough for our methodology test purposes. We must consider the great amount of images generated because every individual will generate one raytracing image. Also we have many individuals (60 to 80), test vectors and five hundred generations per test.

C. *Setting Virtual Scenario*

The virtual scenario must be constructed by hand; this means that we must put the plane, the sphere and camera so they are in the same place as the original image. Then we render a sample image to compare with the original. If they have distance and size differences, we adjust the virtual scenario until both images are virtually equal.

IV. Tests and Results

Our virtual world is composed of a simple flat floor, a yellow plastic sphere and a light source. We capture the objective image as follow, fig 2.

Figure 3. System acquired objective image.

The variables vector of each DE individual has a correspondence with the virtual world light source, so the first three variables will have the X, Y and Z values of the light source position, and the fourth variable will be the intensity of the light source. Others variables are the sphere and plane HSV color component values.

The DE we use has a random offspring selection, with one pair of selected solutions and a binary recombination type. This means, that we use a DE/rand/1/bin. Also DE is a multi-objective algorithm looking to minimize RGB difference values, each corresponding to a function minimizing the generated image from the raytracing. We have a light dominance, with two or three elements. We conducted ten program runs, with a stop criterion of 500 generations, with 80 individuals. It uses a CR of 0.7, and F of 0.6 values. Now the results of one run are presented as follows, Fig. 3.

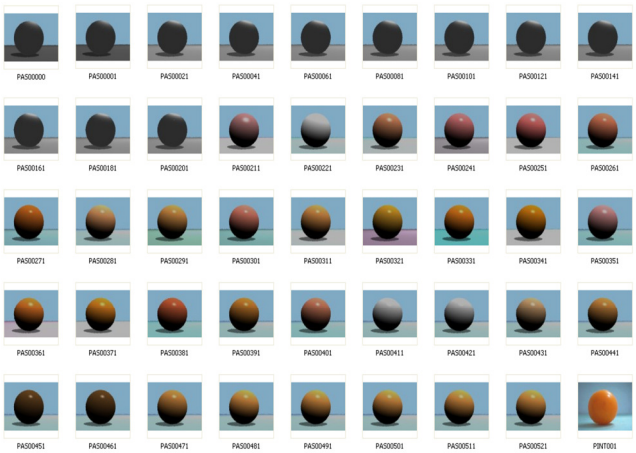


Figure 4. Evolution of generated images, we present the best image of 10 generations, the number of each interaction is shown.

Fig, 3 shows the evolution for parameter estimation; in first place the light source parameters search, running from PAS00000 to PAS00201, two hundred generations to find good position and intensity for the light source. In second place, from PAS00211 to PAS00521 we search for the object optical parameters like color, reflection index, refraction index, roughness, and shininess, for both the sphere and the plane.

The algorithm searches the more fitting positions for the light source as well as intensity; this can be seen in the relative position of the sphere's shadow in the last generation. Finally the sphere's and floor color fitting takes more time but gives good results considering the shading model used, as seen in Fig.4.

The set of experiments includes different sphere colors, in fig. 5. We use red, green and blue spheres to test the primary colors; afterward we use another four spheres to test other common colors such as white, yellow and orange.

Every experiment gives us a final image that we use to compare with the acquired reference image, Table I.

These results show a best average of 5.57 for the red sphere and worst (10.66) for the orange. These are good results considering the shading model. If we use a better model we can obtain better results. Also we made a set of experiments including a rendered scene with the same shading model and used the methodology to estimate the optical parameters user defined. Here are the target objects, Fig. 7.

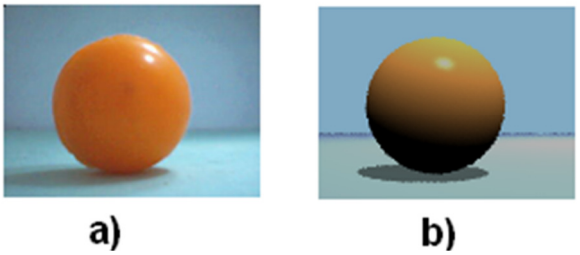


Figure 5. a) Real yellow sphere, b) Virtual rendered sphere.

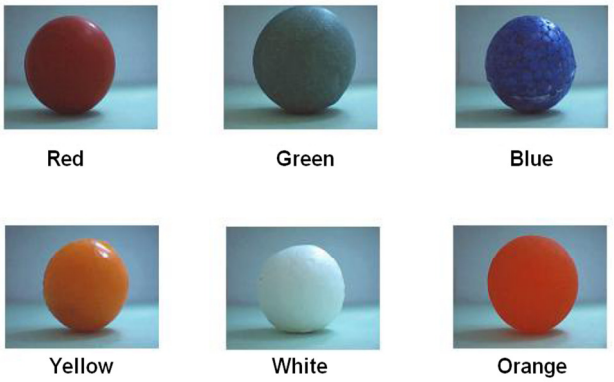


Figure 6. Target real scenarios used to test the metodology.

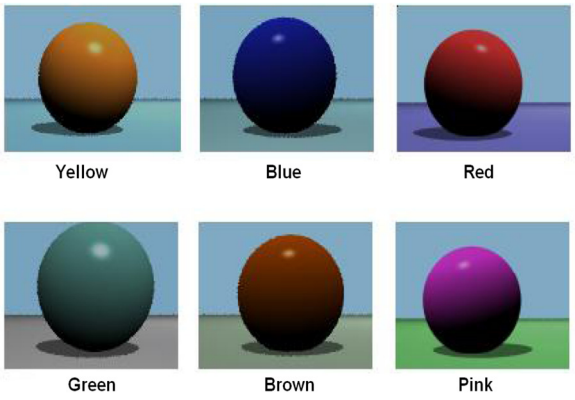


Figure 7. Target virtual scenarios used to test the metodology.

TABLE I
COMPONENT AND AVERAGE DEVIATION FROM ORIGINAL IMAGE COMPARED TO THE BEST GENERATED CORRESPONDING IMAGE

	Deviation			
Sphere	%R	%G	%B	Average
Red	6.25	5.25	5.20	5.57
Green	6.93	6.18	4.39	5.83
Blue	10.84	9.52	7.43	9.26
Yellow	7.29	6.95	10.17	8.14
White	12.61	10.55	6.29	9.82
Orange	8.79	7.23	15.96	10.66

The results for this set of experiments give us lower deviations with respect to the previous set, Table II. The methodology obtains a best average of 2.63 for the blue sphere and a worst 6.79 for the pink.

When we compare the original with the final image the differences are minimal, Figure 8.

TABLE II
COMPONENT AND AVERAGE DEVIATION FROM RENDERED ORIGINAL, COMPARED TO THE BEST GENERATED CORRESPONDING IMAGE

	Deviation			
Sphere	%R	%G	%B	Average
Red	6.25	5.25	5.20	6.21
Green	6.93	6.18	4.39	5.30
Blue	10.84	9.52	7.43	2.63
Yellow	7.29	6.95	10.17	5.77
Brown	12.61	10.55	6.29	3.54
Pink	8.79	7.23	15.96	6.79

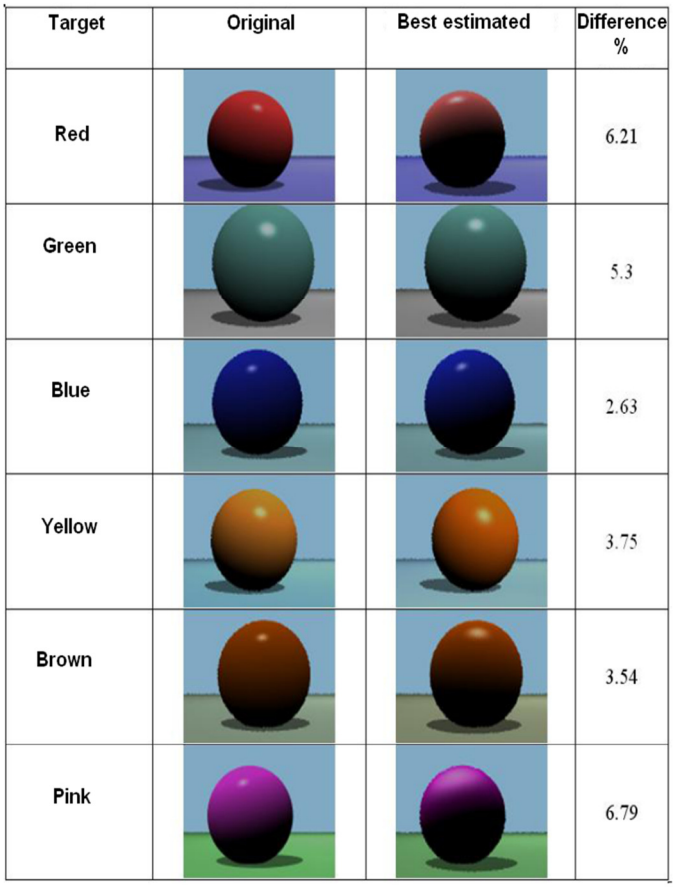


Figure 8. Original image and best individual generated image comparisson.

TABLE III
COMPARISSON, OUR BEST AND WORST VERSUS OTHER EXTRACTION METHODS

	Deviation			
Method	%R	%G	%B	Average
Wai Kit Addy 1	8.92	9.55	8.54	8.97
Wai Kit Addy 2	12.41	9.67	8.58	10.22
Ward	3.30	3.48	3.14	3.30
Ward-Durn	2.20	2.32	2.25	2.25
Our worst	**7.43**	**5.41**	**7.54**	**6.79**
Our best	**1.83**	**2.10**	**3.96**	**2.63**

We made a comparison with other extracting parameter methods. Comparing the original image and the generated one we obtained good results, Table III.

V. CONCLUSIONS

Our methodology searches for two relevant optical parameters, the light source parameters, and the object optical parameters, and obtains good results. The floor's and Sphere's colors is near to the original, nevertheless the position of the light source and its intensity.

The deviation showed by our methodology is close to the best results obtained for other methods (our best 2.63, global best 2.25), but it must be considered that other methods do not search for light source parameters. Our method can obtain more global parameters than any other without expensive equipment, obtaining competitive results. This encourages us to continue with more experiments making adjustments to the algorithm. We think it's possible to include other object parameters like a variable for bump mapping. Also is possible to use a meta-heuristic, and could be good to use an image preprocessing searching for edges, curves and image orientation.

REFERENCES

[1] Y. Willems P. Dutre, F. Suykens, *Optimized Monte Carlo path generation using genetic algorithms*, Report CW267, May 1998.

[2] Srinivasa G. Narasimhan, Mohit Gupta, Craig Donner, Ravi Ramamoorthi, Shree K. Nayar, HenrikWann Jensen, "Acquiring Scattering Properties of Participating Media by Dilution," *ACM Transactions on Graphics* (SIGGRAPH'2006), 2006.

[3] Wai Kit Addy Ngan, *Acquisition and Modeling of Material Appearance*, S.M., Massachusetts Institute of Technology, Phd. Thesis. 2004.

[4] E.G. Finsy, and J. Joosten, "Maximum entropy inversion of static light scattering data for the particle size distribution by number and volume," *Advances in measurements and control of colloidal processes*, Butterworth-Heineman, Ch. 30, 1991.

[5] H. Jensen, S. Marschner, M. Levoy, and P. Hanrahan, "A practical model for subsurface light transport," in *Proc. SIGGRAPH 01*, 2001, pp. 511–518.

[6] B. Sun, R. Ramamorthi, S.G. Narasimhan, and S.K. Nayar, "A practical analytic single scattering model for real time rendering," *ACM Trans. on Graphics (SIGGRAPH)* 24, 3, 1040–1049, 2005.

[7] Fuch E. and Jaffe J. S. "Thin laser light sheet microscope for microbial oceanography," *OPTICS EXPRESS* 10 (2), 145–154, 2002.

[8] R. Furukawa, H. Kawasaki, K. Ikeuchi, and M. Sakauchi, "Appearance based object modeling using texture database: acquisition, compression and rendering," in *Proceedings of the 13th Eurographics workshop on Rendering*, Eurographics Association, 2002, pp. 257–266.

[9] Gero Müller, Gerhard H. Bendels, Reinhard Klein, "Rapid Synchronous Acquisition of Geometry and Appearance of Cultural Heritage Artefacts," in *6th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST*, 2005.

[10] Wojciech Matusik, Hanspeter Pfister, Remo Ziegler, Addy Ngan, and Leonard McMillan, "Acquisition and Rendering of Transparent and Refractive Objects," in *Thirteenth Eurographics Workshop on Rendering,* 2002.

[11] G. Müller, J. Meseth, M. Sattler, R. Sarlette and R. Klein, *Acquisition, Synthesis and Rendering of Bidirectional Texture Functions*, EUROGRAPHICS 2004 State of The Art Report, 2004.

[12] James S. Painter, John K. Kawai, and Michael F. Cohen, *Radioptimization – goal based rendering,* 1993.

[13] Jitendra Malik, Yizhou Yu, Paul Debevec and Tim Hawkins, *Inverse global illumination, Recovering reflectance models of real scenes from photographs*, 1999.

[14] Yizhou Yu. *Modeling and editing real scenes with image-based techniques,* 2000.
[15] Isaac Rudomin Goldberg, Joaquin Elorza Tena, *An interactive system for solving inverse illumination problems using genetic algorithms*, 1997.
[16] Brian Smits, James Arvo, Chris Schoeneman, Julie Dorsey, and Donald Greenberg, "Painting with light," in *SIGGRAPH '93 Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, July 1993.

A Model of Decision-Making Based on the Theory of Persuasion used in MMORPGs

Helio C. Silva Neto, Leonardo F. B. S. de Carvalho, Fábio Paraguaçu, and Roberta V. V. Lopes

***Abstract*—From a videogame perspective, decision-making is a crucial activity that takes place at all times and at different levels of perception. Moreover, this process influences the gamers' performances, which is an interesting feature for RPGs as they are games that are able to work as tools for increasing the improvement of the proximal development zones of players due to their inherent trait of cooperation, which alone, stimulates their skills of socialization, interaction and, consequently, communication. A feat that is achieved by involving players in a kind of plot that requires them to interact and take decisions, hence, favoring decision-making process. For these reasons, the RPG genre was considered as an appropriate test bed to apply the decision-making model proposed by this paper, which was built by using a Petri Net and that combines concepts taken from The Game Theory and from the reciprocity principle from the Theory of Persuasion.**

***Index Terms*—Psychology of persuasion, systems decision making, MMORPG, RPG, Petri net and game theory.**

I. Introduction

At any given time, a person might have to decide over different situations and problems. At those moments, people are likely to use past experiences, values, beliefs, knowledge, or even technical skills to take such decisions. While some people are more conservative, others might have an innovative character and be more willing to accept potential risks [8].

In videogames, making decisions is a crucial matter that must be performed at all times and at different levels of understanding, thus, having a direct influence on a player's performance. In fact, the decision making-process so important that is impossible to think on videogames without considering its occurrence [8].

To ease the decision-making process, this paper employs concepts taken from the Game Theory. The Game Theory of models focused on the analysis of conflicts and of situations that depend of strategic behaviors that, partially, restrict the player's actions. Here, this theory is used to trigger the change of state of a Petri Net, which, is employed in this paper as a tool to model the decision-making process of human players and the different outcomes their decisions might have as they try to maximize their gameplay in a MMORPG (Massively Multiplayer Online Role-Playing Game) [5] [6].

In addition to those theories, this paper also employs the Reciprocity concept found in the Principle of Persuasion as a tool to aid human players to persuade each other in order to achieve their personal goals. In that, players use their communication skills as artifice to make other players voluntarily change their attitudes, beliefs or behaviors, thus, avoiding coercion. In other words, the person that is using of persuasion convinces other players to accept a particular idea. Is at this moment that the Reciprocity principle might standout as it has an inherent meaning of a passed down obligation, which might be applicable to different circumstances [3].

Therefore, this paper models all these different theories and adapts them to a Role Playing Game (RPG) as a way to stimulate them in the creation of their strategies and in the decisions they make to achieve their goals, i.e. their knowledge. In that respect, the objective of this paper may be established as to elaborate a decision-making model (that is applicable to a MMORPG and that is grounded on the Reciprocity concept from the Theory of Persuasion) as a way to promote the interaction of various human players in a environment that favors the acquisition of knowledge, at the same time that permits the application of concepts taken from the Theory of Persuasion and of the Game Theory to aid players in their respective decision-making and in building their own knowledge.

To present this, this paper is divided into five sections. First, as to easy the readers understanding, it is necessary to detail some of the concepts of the RPG genre, including its digital versions, especially, the MMORPG. These subjects are the focus of the second section. Next, the third section presents a discussion regarding concepts of the Theory of Persuasion and of the Principle of Reciprocity.

The fourth section present the authors' model, particularly, how the Reciprocity Principle acts on the game environment, therefore, analyzing the procedures the Petri net implements for this principle. At last, the fifth section presents this paper's final conclusions in respect to the proposed model, noting that such discussion does not have the intention to exhaust the subject presented here, but rather, to emphasize the importance of using the Game Theory and of the Principles of Persuasion in a RPG environment as devices to attend the need human players might have in taking their decisions and searching for knowledge during their gameplay experience.

Manuscript received June 21, 2011. Manuscript accepted for publication August 25, 2011.

The authors are with MSc in Computational Modeling of Knowledge, Federal University of Alagoas (UFAL), Postcode 57072-970, Maceió, AL, Brazil ({helio.hx, lfilipebsc, fabioparagua2000}@gmail.com, rv2l@hotmail.com).

II. Massively Multiplayer Online Role-Playing Game (MMORPG)

The RPG (Role Playing Game) acronym was first appointed in the USA by Dave Arneson at the year of 1973. The situations taking place in a RPG are mainly described by speech representation and by using the players' imagination during the game sessions, which are usually the continuation of an adventure interrupted at a previous session [10] [11] [2].

Like many activities, the RPG has its own language. An example of this is the storyteller being referred as master, the listeners/participants being called of players and the story itself being called adventure. The basic concepts of the RPG are listed below, according to the work found in Debbie [7]:

- Player: the players are the ones in charge of one or more characters (known as PC, player character) of the plot and has freedom of action on the game scenario through his or her character within what is allowed by the game's system of rules;
- Game Master: has control over all factors related to the settings and plot that do not involve the characters' actions (which are exclusive to players). Has control of plot characters that interact with those controlled by the players (NPC Non-Player Character). It also has control over the settings, being able to adjust the game plot according to its needs; and is the sole responsible by the secret objectives of the plot and its progress. Like the others involved, the game master must follow and uphold the system of rules, yet, for the benefit of the plot, is able to change things within a reasonable logic;
- System of Rules: the actions taken by the players' characters are addressed to the game master, who verifies at the system of rules which is the result for the action performed by this character in the circumstances applied to it. Thus, tere are different specific rules for different situations, and specific indications that must be taken into account to deal with unexpected situations. Characters: characters can be built by their own players or provided for the game. However, they must be elaborate in respect to the system of rules and the game scenario. All players are assigned with abilities that define their interaction with the environment. These skills are in accordance with the game's system of rules and achieved by the players as a reflection to their interest in building a particular kind of character;
- NPCs (Non Player character): The term is borrowed from other RPGs to indicate a character that is not controlled by any player, thus controlled by the game master. They usually act as supporting characters for the adventure;

With the release of computer RPG games, which allowed for a multiplayer mode of the game, i.e. allows multiple users to play on a LAN, modem or the Internet, the following characteristics made themselves common in digital RPGs played on Internet:

- Multiplayer interaction;
- Exploration of wide worlds provided with large locations;
- Existence of several sub-plots, allowing the players to create their own history and adventures;
- A great RPG's similarity to the table, because the permission of creation and evolution of the characters;
- For the large majority of games, the user can customize its main character, for example, creating adventures, items, weapons and worlds.

The MMORPGs have as main characteristic the constant intervention of a team of Game Masters, which may be NPCs or human developers in the real world, who work on plots and create challenges for the players' characters. The plots are nonlinear and thus they do not have a beginning, middle or even an end. The concept is that exist a virtual world to be explored, an open story. Another feature of MMORPGs is related to its own name and its idea Massively Multiplayer Online Role-Playing Game that allows for thousands of players together interacting in the "virtual" world.

III. Theory of Persuasion

According to Robert Cialdini [3], persuasion is the use of communication to change the attitudes, beliefs or behaviors of others. However, this change must be voluntary and not through use of force or coercion. Thus, the person using of persuasion convinces the ones it communicates to in accepting a particular idea.

In that sense, the persuasive speech aims to embody "the whole truth" by using of linguistic resources and selecting expressions of "truth" able to introduce a particular assumption. Moreover, the ultimate goal of persuasive speech is to use rhetorical devices in order to convince other people to change their already established attitudes and behaviors [4]. There are six distinct principles of persuasion, each of them comprising a specific characteristic of human interaction. This paper however, focuses exclusively on the Reciprocity Principle, which will be discussed next.

A. Reciprocity

The relevant aspect of the reciprocity principle is the sense of obligation passed down from one person to another and that is ubiquitous in human culture. In this respect, a number of sociologists, such as Alvin Gouldner [9], state that there is no human society that fails to this rule. Furthermore, the archaeologist Richard Leakey [13] gives the essence of what makes the human species prone to reciprocate "We are human because our ancestors learned to share their food and their skills in a community network".

Competitive human societies respect the principle of reciprocity and thus, expect its members to respect and believe in this principle as well. In this sense, as every person was taught to live by this rule the social sanctions and scorn that may be applied on those who violate it are commonly known by everyone. In this sense, people who do not comply by this rule are assigned with derogative labels such as "ungrateful" or

"deadbeat", which is a consequence of the unpleasant felling generated in members of a society by those individuals that are seen as taking a "favor" and refusing to make any efforts to repay it.

Therefore, one of the reasons that make the reciprocity principle so effective is that the rule it stands for is imbued by a force, which often produces a "yes" response to a request that, in lack of a debt, would certainly have been denied. The strategies for applied by reciprocity are plain in appearance, yet extremely efficient and almost undetectable being that, the best solution in fighting them is to think before accepting any favor or concession from someone whose true intentions are unknown. Even though this might sound as a standard practice for exploiting (as a way to manipulate and influence) the reciprocity principle is a fundamental pillar of human society, and one of the reasons for the development of the earliest human communities.

IV. A Model of Decision-Making Based on the Theory of Persuasion

Now that the basic formalisms of the MMORPG and of the Reciprocity Principle from the Theory of Persuasion have been presented (respectively by sections 2 and 3) these concepts can be properly modeled to fit the context that a MMORPG presents.

A. Binary Decision Tree

A MMORPG game environment is prone to present different circumstances where the Reciprocity Principle might be applied, a fact that is due to the MMORPG approach that mixes the real world with a surreal one. Some examples of circumstances found in such games that support the use of strategies based on the Reciprocity Principle are:

- Aid in hunting monsters, locating an item or even completing a given Quest[1], all of which might help a player to increase its level. In those circumstances, the person receiving help will be in debt to the helper, which allows an exchange of knowledge;
- Tips and advice received via forum. Like the above circumstance, the player being aided will be in debt to the ones who made the post as well as to any other helper;
- Providing discounts on sales of some item. In these cases, it is possible that the given discounts are not available anywhere else. Thus, announcing the discounts is likely to attract buyers. These situations may require a little "push" from the seller, or be bounded to acquisition of certain X value resulting in an N discount. The same rules apply to promotions and contests;
- Regular buyers get discounts, thus, creating a loyalty bond.

[1] Mission or purpose connected with RPG

1) Reciprocity principle and the flow module modeled in Petri net

The Reciprocity works as an exchange of favors, i.e. it consists of sharing what was received. In a MMORPG environment this principle might be triggered in several ways, some of which were described in Sec. 4.1. In order to provide a better understanding of the principles that this model applies to the environment, it has its operational flow depicted in Figure 1.

The flow of Figure 1 shows that the Reciprocity Principle is triggered in any circumstance in which, a player asks another one for a different set of simple aids, such as finding an item, completing a *Quest* or help in hunting a monster. For these requests, players have the additional aid of tools for creating a *Party*[2] and for *Trade*[3] an item, all of which are identified on the above figure.

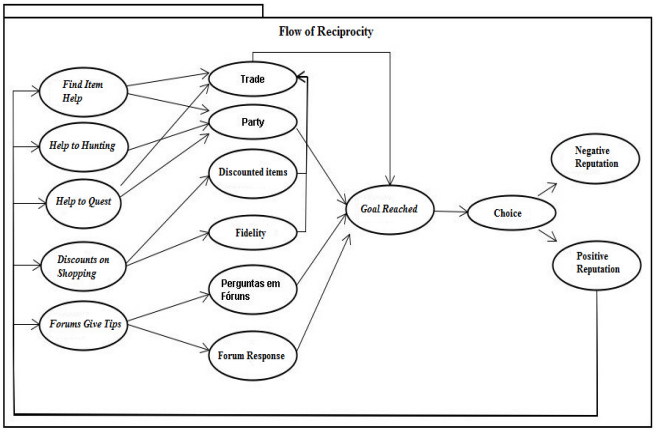


Fig. 1 Flow of Reciprocity.

This flow can be started by any player seeking reciprocity for any of the depicted tasks. Examples of this include different requirements such as *Find Item Help, Help, Help and Hunting Quest*. As a more focused example, should help mean the assistance in finding an item that another player already has, such item can be transferred through the use of the *Trade* tool. In other circumstances a party could be formed to look for that item, as well as in a set of other different circumstances. Additionally, a party grants its members with the ability to share the money and experience points that the game provides as reward for their activities.

As for the *Discounts on Shopping* circumstance, it relates to those cases when a player employs the reciprocity principle to attain another player loyalty, or to purchase a large number of objects being negotiated with other player. Once the involved players reach an agreement the *Trade* tool is used to transfer the goods.

The *Forums Give Tips* situation occurs when players accesses the virtual platform environment, and perform their questions or searches. After achieving their goals, the researchers approve the responses of other members, thus,

[2] A technical term of RPG applied for hunting in group or to creating a task force.

facilitating future searches and motivating other players to improve their reputations.

By using all tools embodied in this principle, the player arrives at the *Goal Reached* circumstance. As the name suggests, this situation corresponds to the player achieving the intend goal and thus being granted with a choice. To take the passive instance of the principle (in that, repaying the received aid and creating a cyclical movement of the Reciprocity), or to ignore the act of returning the received aid, and thus, breaking the principle cycle. In opting for the last choice, the player is aware of the penalties that may be applied to him/her, which were presented in section 3.1.

As mentioned during the introduction, this paper employs a Petri Net to model the decision-making process that the human player must perform during their gameplay and the different outcomes that may arise from these decisions. The Petri Net built from the Reciprocity principles discussed in this paper is show in Figure 2, in which depicts the model proposed here, including its states, transitions and the set of guided arcs.

The formal definition of this Petri Net assumes an initial marking m_1 and supports a cyclic movement of its network that is broken only in those circumstances where, the active Reciprocity player opts to not return the aid that was shared with him/her. In addition, the Petri Net modeled contains several final states that are used to identify the amount of people that completed its cycle, as well as to identify the amount of people that did not completed the network's cycle. The states and variables of this network are:

- Variable *m1 – Start of Reciprocity*: this state addresses when a player uses the reciprocity to get something that desires, by requesting the *Requesting Support* the transition. A player can only proceed when that request is available;
- Variable *m2 – Requested Available*: the state where the player that starts the reciprocity (the initial player) waits for the availability of another player at the same time that it checks for the availability of another player. In case there is no available player, the transition *Supporter Unavailable* is triggered. Otherwise, the *Requesting Support* transition is triggered again;
- Variable *m3 – Await for Support Request*: this state occurs when the active player of the Reciprocity is choosing what desires to ask to the player of the passive persuasion. In this state there are redundant paths within the Petri Net, as it was developed to model parallel, concurrent, asynchronous and non-deterministic systems. This fact is justified by the Petri Net having being modeled with human activities. Thus, choosing which route to follow depends solely of the active player, who may choose any of the following options: *Buy Item*, *Request Help* or *Ask in Forum;*
- Variable *m5 – Wait for Discount*: here, the active player takes advantage of discount in items due to various factors, such as being an active customer, buying a large amount of an item and so on. The passive player will ensure the sale of the product by the exchange of favors. Triggers the *Discount in Item* transition;
- Variable *m4 – Awaits the Order*: the passive player can only grant the discount if the item is effectively requested. In the event that it does not, the *Item Purchase Unrequested* transition is triggered. Otherwise, the item continues to the *Discount in Item* transition, in which the amount for the item is negotiated;
- Variable *m8 - True to Purchase*: This occurs when the order value is set, thus creating a reliable purchase. With this, the process proceeds either to *Trade* or to *Purchase Rejected*;
- Variable *m11 - Requesting for Negative Reputation*: this state can be said to be one of the final states within the network. This condition occurs when the active player of Reciprocity does not want to proceed with the purchase, thereby creating a bond with unreliable connotations. In this case, the player breaks the cycle of reciprocity and is penalized with the loss of points from its reputation. A situation that is not desirable but that, due the circumstances, is possible;
- Variable *m9 - Await for Available Item*: as the name implies, it is related to the availability of an item by a passive player who may negotiate it even if he/she does not currently has it. In this scenario, *the Unavailable Item* transition is triggered and the state persists until the seller comes in possession of such an item as so to sell it. At this moment, the *Request Trade* transition is triggered;
- Variable *m10 - Trade*: is the state in which the actual sale of an item occurs. Once the said item is available and all its negotiation is set, the *Approve Trade* transition is triggered, in which the item and amount of value involved are checked;
- Variable *m12 - Purchase Accomplished*: is the state in which the purchase is concluded and is responsible for triggering the *Reputation Score* transition;
- Variable *m13 - Reputation Score by Sale*: this is the final state for analyzing the amount of sale accomplished by the use of this principle, which also allows for enlarging a seller's reputation in the occurrence of future sales;
- Variable *m14 - Reputation Score by Purchase*: this state is similar to the previous one; however, its focus rests on the buyer;
- Variable *m25 – Reciprocity Goal Achieved*: as the name says, reciprocity does not only occur due to a cycle (albeit this would be the ideal circumstance). In fact it can occur simply by someone, be it the seller or anyone providing the needed assistance, answering to a request at the forums, thus, employing of a Reciprocity that may or not be returned. A characteristic that is due to the work being presented here applying the Principles of Persuasion on its model. That said, next comes the

[3] MMORPG tool that players use to sell or exchange goods.

Activating Reciprocity transitions, which approaches the choice of the active player in repay the aid or not;

- Variable *m26 – Awaits the Choice of a Requester*: this state awaits a player's action within either the *Chooses to Not Repay* or *Chooses to Repay* transitions;
- Variable *m27 - Negative Reputation for the Requester*: it is a final state of the network that occurs when the active player chooses not to return the Reciprocity and thus, acquires a score with negative impact on its reputation;
- Variable *m28 - Reputation for the Requester*: at the *Chooses to Repay* state two parallel routs that may be taken, one of which, takes to the *Reputation for the Requester*. At this state the active player is assigned with a positive score on its reputation, which favors him/her at future activations of the Persuasion. The other route takes to the *Awaits Completion of Reciprocity's Cycle* state that is shown next;
- Variable *m29 - Awaits Completion of Reciprocity's Cycle*: this is a stated that merely waits for its own activation, the one responsible for this being either the passive or active player. Once the state is activated it triggers the *Completion of Reciprocity's Cycle* transition;
- Variable *m30 – Awaits beginning of Reciprocity*: it is a stated that indicates when the Reciprocity process will start (*Reciprocity Start*). It is at this moment that the network effectively starts, triggering both the *m1* and *m2* states;
- Variable *m6 – Awaiting for Help*: this state occurs when the active player chooses the *Request Help* transition and may trigger the *Request Party* transition if there exists a player that intents to help him/her;
- Variable *m15 – Await for Available Party*: this state occurs just as the *m4* state. However, it differs from that due to its need to assert whether or not a *Party* aid is available. If it is not, the *Party Unavailable* transition is set. Otherwise, it triggers the *Request Party* transition;
- Variable *m16 - Party*: this state occurs when members that intent to aid group together and use of the homonymous *Party* tool. Depending on their conducts, the *Aiding in Progress* or the *Party Rejected* transitions might be triggered, the last, indicating the disbanding of the group, due to a lack of commitment in providing aid that might be on either their or the even on the player's part;
- Variable *m17 - Negative Reputation for the Requested*: in case the group or the passive player asks for the disbanding of the *Party*, the reputation of one or of several of the involved may receive a negative score as a penalty;
- Variable *m18 - Aid Provided*: this is the end of the aid process and occurs when the group's goal is achieved, triggering the *Reputation Score* transition;
- Variable *m19 - Reputation Score by Aid*: similar to the *m13* state but the score in this case is set in reason of a previously provided assistance;
- Variable *m20 - Reputation Score for the Assisted*: a state similar to the *m14* one, though the score here is set due to a received aid. Next to this, the previously presented *m25* state is triggered, this continuing the network;
- Variable *m7 – Waiting for a Reply*: it is the other path within the network that is triggered when the active player requests the *Ask in Forum* transition, after which, it awaits at this state until a passive player answers it. After a given answer, the active player may or not trigger *Assert Answer*. In case there is no answer, the *No Given Answers* transition is triggered;
- Variable *m21 - Waiting for a Question*: occurs when passive players are waiting for a question to answer. In case of no question, the *No Given Questions* state is triggered. However, if a question is made and answered the active player is imbued with the task of triggering the *Assert Answer* transition (as stated above);
- Variable *m22 – Tip Provided*: this is a stated that follows the assertion of an answer at the *Assert Answer* state. Next, it triggers the *Reputation Score* transition;
- Variable m23 - *Reputation Scored by Tip*: similar to *m13*. Nevertheless, the largest reputations occur here due to an active participation at the forum in providing someone with requested information;
- Variable *m24 - Reputation Scored by Question*: a state similar to *m14* that carries on the act of attribute a value to someone's reputation due to asking a question. In parallel to this, the *m25* and thus, all the network keeps moving.

Aside from the list of states and variables above, there is another variable that is important to a Petri Net model, which is the weight function responsible for the launching (or not) of the network.

The weight function is given by the strategy of each player, being that each of them, probably, has a distinct number of strategies. Some of the equations provided by The Game Theory were used to develop this strategy. A better understanding of these equations requires a brief overview of some of the Game Theory concepts.

2) Game Theory

Game Theory is a mathematical theory designed to model the phenomena observed when two or more "decision agents" interact. The theory provides a language for the discretion of conscious decision-making processes and the objectives involving more than one individual [5] [6].

Therefore, the application of the principles of this theory to the model proposed in this paper aims to study the choices of optimal decisions under conditions of conflict, precisely, when one person wants to activate the Principles of Reciprocity. For this purpose, the basic game element being considered here is the set of involved players, each of whom has his/her own set of strategies. Additionally, when a player chooses one of his/her strategies, a circumstance or profile is created in the space containing all possible situations (profiles). It must also

be noted that each player has interests or at least preferences focused on different game situations. Mathematically, this ensures that each player has a utility function responsible to assign a real number (the gain of the player) to every game situation [15].

Particularly, the game has the following basic elements: a finite set of players that is given by $G = \{g_1, g_2, ..., g_n\}$, where each $g_i \in G$ player has a finite set of options written as $S_i = \{s_{i1}, s_{i2}, ..., s_{im}\}$, which are known as the pure strategy of the player denoted by $g_i(m_i \geq 2)$. Additionally, a $s = (s_{1j_1}, s_{2_{j_2}}, ..., s_{n_{j_n}})$ vector that has s_{ij_i} as a pure strategy for the $g_i \in G$ player is named a profile of pure strategies. The set of all pure strategies' profiles are a Cartesian product [15] that corresponds to the equation shown in (1) and is known the game's pure strategies' space. For each player corresponding to a $g_i \in G$ value there is a utility function (which is shown in (2)) and that links the $u_i(s)$ gain of the $g_i \in G$ player to each $s \in S$ profile of pure strategy [15].

$$S = \prod_{i=1}^{n} S_i = S_1 \times S_2 \times ... \times S_N,$$

$$u_i : \begin{matrix} S \to R \\ s \mapsto u_i(s) \end{matrix}$$

The two functions above enable a player to choose best strategy to apply the Reciprocity Principle and trigger it at the appropriate moment of the game.

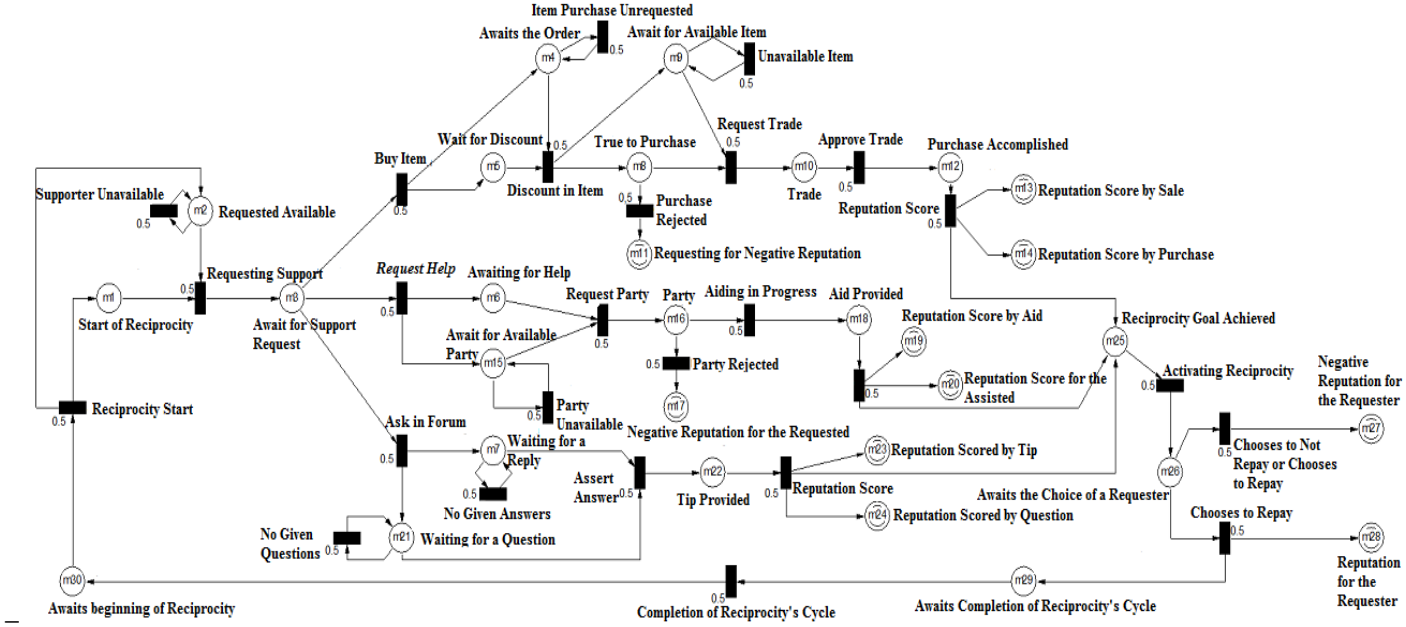


Fig. 2 Reciprocity in Petri Net.

V. Conclusions

In a videogame perspective, the use of the Reciprocity Principle is an underexplored subject and for this reason, few people are aware of its importance to a game's decision-making process. This fact is reflected by the apparent inexistence of models for decision-that use the reciprocity principle, or any of the other principles of the Theory of Persuasion, and became even clearer along the development of this paper as it confirmed that the use of this theory in videogames is, until this point, very incipient.

Additionally, this paper's research also confirms that the implementation of the Theory of Persuasion as tool for decision-making in MMORPGs' environments can change the way that players deal with information (knowledge) due the fact that, they will employ this theory to create their own best strategies and to improve their relationship with each other..

It must also be noted that, much can be done to improve the work performed here (as there is still much work to do). Moreover, the inexistence of a current commercial interest in developing such environments to aid the educational process or any other area of knowledge makes this initiative an academic project that may never come to fruition.

However such a reality might be changed at the moment that the individuals least aware to the benefits brought by the Reciprocity Principle and the other Principles of Persuasion became clear of the players' needs and are encouraged to develop new strategies for obtaining knowledge, as well as share their knowledge with other players, thus making the MMORPG environment an educational partner.

In this sense, this paper pays attention to the real benefits that the Reciprocity Principle may bring for the development of a decision-making environment, and demonstrates its advantages by using of an actual MMORPG environment to create a model that meet the real needs of players.

Thus, it is believed that by applying to MMORPG environments decision-making systems that combine the principle and theory presented along this paper, such systems will act accordingly with the authors' proposed Reciprocity model, supporting this type of games due to their ability to allow players to better build their knowledge. In that, the reciprocity principle will enhance the decision-making systems due to its capacity to transform the decisions taken upon conditions of uncertainty to the ensuring scenario of the certainty conditions.

REFERENCES

[1] E. de Oliveira Batista, *Information System: the conscious use of technology for management,* São Paulo: Saraiva, 2004.

[2] C. Cale, The real truth about dungeons & dragons. http://www.cale.com/paper.htm, 2002.

[3] B.R. Cialdini, *The Psychology of Influence Persuasuion,* Collins, 1998.

[4] A. Citelli, *Language and persuasion*, 2nd ed. New York: Attica, Series Principles, 1986.

[5] J. Conway, "All Games bright and Beautiful," *The American Mathematical Monthly*, 1977.

[6] J. Conway, "The Gamut of Game and Theories," *Mathematics Magazine*, 1978.

[7] M. Debbie and D. Arkanun. 2. ed. Location: Demon, 1998.

[8] E. Garcia and O.P. Garcia, "The importance of management information system for business management," *Journal Social Science Perspective*, Center for Applied Social Sciences of Cascavel, Cascavel, v.2, n.1, 1 wk., 2003.

[9] A.W. Gouldner, "The Norm of Reciprocity: A Preliminary Statement", *American Sociological Review* 25, 1960.

[10] J. Hughes, *Therapy is Fantasy: Roleplaying, healing and the construction of symbolic order*, http://www.rpgstudies.net/hughes/therapy_is_fantasy.html, 1988.

[11] S. Jackson, *Basic module RPG*, GURPS 2 ed. São Paulo: Devir, 1994.

[12] Kenneth C.Laudon and Jane Price Laudon, *Information systems*, 4 ed. LTC: Rio de Janeiro, 1999.

[13] R. Leakey, *People of the Lake*. New York: Anchor Press / Doubleday, 1978.

[14] Maria José Pereira, Lara Breton Fonseca, and João Gabriel Marquez, *Faces of Decision: the paradigm shifts and the power of decision*. São Paulo: Makron Books, 1997.

[15] B. A. Sartini, et al. *Uma Introduction to Game Theory*. Second Biennial of SBM. Federal University of Bahia, 2004.

[16] Ralph M. Stair, *Principles of systems informação*. Rio de Janeiro: LTC, 1998.

[17] P. R. Stephen and Mary Coulter, *Administration*, 5.ed. Prentice Hall Interamericana, 1996.

User Preference Model for Conscious Services in Smart Environments

Andrey Ronzhin, Jesus Savage, and Sergey Glazkov

***Abstract*—Awareness of user preferences and analysis of the current situation makes capable to provide user with invasive services in various applications of smart environments. In smart meeting rooms context-aware systems analyze user behavior based on multimodal sensor data and provide proactive services for meeting support, including active control PTZ (pan, tilt and zoom) cameras, microphone arrays, context dependent automatic archiving and web-transmission of meeting data at the interaction. History of interaction sessions between a user and a service is used for knowledge accumulation in order to forecast user behavior during the next visit. The user preference model based on audiovisual data recorded during interaction and statistics of his/her speech activity, requests, movement trajectories and other parameters was implemented for the developed mobile information robot and smart meeting room.**

***Index Terms*—User preferences, context awareness, action recognition, mobile robot, smart meeting room.**

I. Introduction

THE notions of user model and context are fundamental for artificial intelligence and human-machine interaction in particular. Creation of user model or profile involves gathering user information during his/her interaction with a system. The primary aim of the system personalization is to improve user experience and get relevant service in the current situation [1]. The context change could be caused both a user and environments, in which interaction takes place.

The difference in abilities, interests, roles, location of a user as well as history of previous interaction sessions are main factors considered by context-aware systems concerned with acquisition, understanding of context and action based on the recognized context [2]. The problems of context representation, sensor uncertainty and unreliability are considered in numerous works. However, there is no any accepted opinion on types and number of context spaces and their attributes, as well as there is a lack of universal approaches to the problem of context prediction, especially for acting on predicted context [3].

Manuscript received June 29, 2011. Manuscript accepted for publication August 25, 2011.

This work is supported by Saint-Petersburg State University (project # 31.37.103.2011), the Russian Federal Targeted Program (contracts #P876 and #14.740.11.0357) of the Ministry of Science and Education of Russia.

Andrey Ronzhin is with St. Petersburg State University, 11, Universitetskaya nab., St. Petersburg, 199034, Russia. Jesus Savage is with the Universidad Nacional Autonoma de Mexico, Mexico City, Mexico (e-mail: savage@servidor.unam.mx). Sergey Glazkov is with the Russian Academy of Sciences St. Petersburg Institute for Informatics and Automation RAS, St. Petersburg, 39, 14 Line, 199178, Russia (e-mail: glazkov@iias.spb.su).

Location and time have been the commonly used components of the context. Computing context, user context and physical context were selected by Schilit et al. [4]. User's location, environment, identity and time were analyzed at the context definition by Ryan et al. [5]. Three different categories of contexts were proposed in [6]: (1) real-time (location, orientation, temperature, noise level, phone profile, battery level, proximity, etc.); (2) historical (for instance, previous location, and previous humidity and device settings); (3) reasoned (movement, destination, weather, time, user activity, content format, relationship, etc.).

In [7], the context information used for service personalization and designing of multimedia applications for heterogeneous mobile devices were divided into the five categories: spatio-temporal (place, time), environment, personal, task, social. A personalization service based on user profile retrieves user context and context history information from context management services. It helps the user to get relevant content and services in the current situation.

The human beings, the physical and informational environments were considered by Dai et al. in the framework of two types of contexts [8]: interaction context representing interactive situations among people and environment context describing meeting room settings. They use propositions that the interaction context of a meeting has a hierarchical structure and expresses the context as a tree. User's standing-sitting states, changing user's location, face orientation, head gestures, hand actions, speaker turns and other events are analyzed for the context prediction. A Finite State Machine framework was introduced in order to classify these meaningful participants' actions. However, before the classification an event should be detected, so particular issues of signal capturing and feature extraction are appeared.

The rest of the paper is organized as follows. Section 2 describes the appropriate audio and video processing techniques used for evaluation of user behavior as well as context acquisition and analysis in smart environments including smart meeting rooms and social robots. The issue of evaluation of user behavior and his preferences during interaction with intelligent services equipped by different types of user interface is considered in Section 3. The results of cognitive evaluation of three types of user interfaces for the developed information mobile robot are discussed in Section 4. The architecture of the meeting web-transmission system, which performs selection and transmission of the most actual multimedia content captured from video cameras, whiteboard, presentation slides, based on context analysis

during the meeting in the smart room, is presented in Section 5. Conclusions and plans for future work are outlined in Section 6.

II. Context Acquisition and Analysis in Smart Environments

In a smart meeting environment, to provide conscious services context-aware systems should analyze user behavior based on multimodal sensor data and provide proactive services for meeting support, including active control PTZ (pan, tilt and zoom) cameras, microphone arrays, context dependent automatic archiving and web-transmission of meeting data at the interaction. Automatic analysis of audio and video data recorded during a meeting is not a trivial task, since it is necessary to track a lot of participants, who randomly change positions of their bodies, heads and gazes. Audio-visual tracking has been thoroughly investigated in the framework of CHIL and AMI/AMIDA projects [9, 10].

Use of panoramic and personal cameras is suitable for recording a small-sized meeting, where all the participants are located at one table. In a medium size meeting room (~50 people), a larger space should be processed that affects on the cost of recording technical equipment too [11]. Distributed systems of microphone arrays, intelligent cameras and other sensors were employed for detecting participant's location and selection of a current speaker in the medium meeting room.

Let us consider several recent works devoted to analysis of meeting participant behavior. Zhang et al. proposed a speaker detector for the Microsoft RoundTable distributed meeting device [12]. It has a six-element circular microphone array at the base, and five video cameras at the top. The proposed algorithm fuses audio and visual information at feature level by boosting to select features from a combined pool of both audio and visual features simultaneously. Audio related features are extracted from the output of the maximum likelihood based sound source localization (SSL) algorithm instead of the original audio signal. They achieved a speaker detection rate of 93%, a person detection rate of 96%, and multimodal speaker detection of 98%.

A ceiling 4-camera tracking system, a 360° camera, a single microphone for speaker identification, and a circular 16-microphone array were used in the University of Southern California smart room [13]. A mixture particle filter was used for tracking an unknown number of acoustic sources. The angular estimates of source locations were obtained using a variant of time difference of arrival (TDOA) method for each microphone pair. Speaker detection rate was around 90% during four sessions with approximate length of 15 minutes.

Raykar et al. [14] compared the performance of GCC-PHAT, GCC-ML, Brandstein's pitch-based, and the method based on characteristics of the excitation source during the production of speech using an 8 element microphone array in an office room of dimension 5.67x4.53x2.68 m with an average reverberation time of about 0.2 s and noise level of about 40–50 dB. Signal from each channel is sampled at 8 kHz frequency. Some cases were considered during the experiments: the source was placed at a distance of 2.0 m from the center of microphone pair which are 1 m apart; the speaker moved in such a way that he was always facing the microphones. The error is generally lower for frames where signal energy is high, and also a lower error is obtained when larger frame sizes are used. Using a frame size of 500 ms with frame shift of 50 ms the localization error for the proposed method was lower 30 cm.

Multiband joint position-pitch algorithm for 24 channel circular microphone array was proposed to track a single speaker and multiple concurrent speakers in the meeting room measuring 6.02x5.32x3 m [15]. The array was placed in the center of the room; the loudspeakers were positioned at a constant distance of approximately 2 m from the array. Experiments using real-world recordings in a typically reverberant meeting room showed a frame-wise localization estimation score of about 95% for tracking a single speaker.

The approaches based on the signal (also interaural) level difference between different microphones, and TDOA were tested in a train compartment for aggressive behavior detection [16]. The experiments are concentrated in an area having a length of about 7.5 m with eight predefined candidate locations and four microphones. The mean square error of location estimation for sources near microphones was lower 50 cm, but the performance significantly decreased at the detection of far-field sources.

In the DICIT project the harmonic linear array of 13 microphones was used for detection of up to four persons in a room of dimension 3.4x5.0m, which control an interactive television. 4 person positions were investigated at 2.1 m distance from the microphone array [17]. Adaptive Eigenvalue Decomposition was implemented as an alternative to GCC-PHAT in TDOA estimation. A localization error was labeled either as gross, when it is larger than 0.5m, or as fine otherwise. Localization rate (LR) was defined as the percentage of fine errors over all the localization estimates. Localization accuracy is measured in terms of Root Mean Square Error (RMSE) of all the localization errors (fine and gross). In the 30dB SNR case the localization rate was about 97% and RMSE was lower 25 cm.

Summing up the review, it should be noted that distance between center of microphone array and sound source was lower 3 m in all the considered papers, where the SSL methods were evaluated. Positions of speakers were predefined in most of the applications and position number was up to six. The aim of our study is to select current active speaker in the medium meeting room with the number of sitting participants up to 42. Besides of smart environment, the social mobile robots, which are capable to natural interaction with a user, are actively investigated now.

Robot Neel, developed by an Indian group HitechRoboticSystemz Ltd, is an autonomous reference robot, which provides information services to visitors in shopping mall [18]. The robot navigation system is based on laser

sensors and route planning for a given map. The robot is equipped with a touch screen with graphical menus, menu items can be synthesized by Microsoft Windows TTS. The system of interaction with a user applies speech synthesis and a graphical menu. A user selects goods or services on the touch screen, the robot pronounces his/her choice and the response to the user's query. Neel robot is connected to the information network of the shopping center and notified of all changes, availability of goods and services. Also, when interacting with people the robot creates a database of visitors and their preferences based on analysis of user queries. Currently, the robot is able to independently navigate a given route and to identify obstacles. The user interface is based on JavaFX, which allows quick change of graphical part of the interface.

System with a multimodal user interface, including at least speech recognition and synthesis, in addition to the graphical menu, will benefit for a lot more groups. For example, visually impaired people can interact with the system in a natural way using speech. An example of such systems is a robot FriDA, which was developed by Korean company DASA TECH Co. Ltd. FriDA. This robot is equipped with a touchscreen monitor, speakers, and a microphone array. The monitor has standard graphic menus, as well as speakers and microphones to ensure system of synthesis and speech recognition. The robot is designed to provide reference information at the airport in a verbal dialogue mode and can display and pronounce data required by user.

Systems with three-dimensional avatar of the human head are able to communicate with hearing disabled people. Lip movements of avatars are synchronized with the speech signal, which makes possibility for lip reading. For example, a robot secretary HALA, developed at the University of Carnegie Mellon, is equipped with a touch screen, which displays animated avatars, speaker, microphone and an infrared sensor to determine the presence of a user [19]. HALA can lead voice dialogue with a user in Arabic and English, the avatar is used for verbal expressions (movements of the lips are applied in the process of speech synthesis) and nonverbal means (shaking his head, facial movements).

Recently there was a tendency to create humanoid robots with the approximate shape of the hull, with varying degrees, to the human body shape. Such robots are able to interact with a person, not only through speech but also with gestures. Typically, these robots are not equipped with monitors, therefore they have not any graphical interface. For example, the robot Robotinho, developed at the University of Bonn in Germany, has a humanoid form, and can interact with humans through speech, gestures and facial expressions [20]. The robot uses mixed system of dialogue, and is able to determine position of a user and his face, as well as to recognize and synthesize speech. Robotinho can express its emotional state and communicate with many people simultaneously. Since the robot has a humanoid body shape, it can nonverbally communicate with users through gestures during the dialogue, as well as attract users' attention to itself or to the objects of the environment by gestures or gaze direction. The robot detects a user with two laser range finders, and then he finds a human face with two video cameras. When interacting with users it creates a database of users containing user's face images and his/her preferences, based on the query history. In future the robot will be able to identify user.

Thus, the appointment of the robot and the possibilities of potential users are necessary to consider at the development of multimodal interfaces for a social robot. Ways of interaction must be easy-to-use and do not require special training of users. So, speech and multimodal interfaces with speech processing, are being actively researched and applied in robotic systems. Despite the fact that user interaction with social robots in most cases takes place in an environment with high noises, speech interfaces, and multimodal, including speech and gesture processing, are being actively studied and applied research in robotic systems [21, 22, 23].

III. Influence of User Interface on User Behavior

Fundamental principles of the field of human-computer interaction lays the basis for the design of dialogue models, also the capabilities of modern hardware and software that implement the input, output and processing of information channels available to the user are taken into account. With the development of socially oriented services, it became clear that the interfaces for interaction of the system with a user should be simpler, more intuitive and do not require additional knowledge and training.

The standard interface is a graphical user menu, which includes information inputting by a user in manual mode (keyboard, mouse, touchscreen monitor). The most widespread of such interface has received in a self-service machine, such as payment terminals or ATM services. This kind of interaction is not always convenient for a user, and often even impossible, for example, people with disabilities are not able to interact in this way (blind, armless, etc.). To increase the opportunities of graphical user interface voice prompts to the menu should be used in self-service machines and robots are used.

The standard graphical user interface remained the most common before the appearance of complex interactive systems for mass services. Much greater attention is now given to the development of queuing systems with multimodal user interfaces based on analysis of speech, gestures, and graphical user interface, three-dimensional model of a human head with a strong articulation of speech, facial expressions and other natural means of communication for interpersonal communication.

Besides of interface type, various factors influence on user behavior, for example, the general context and peculiar features of the task; experience of human-computer interaction. The point is that the user usually keeps in his mind all the experience of the same kind, so time after time he/she tends to use one and the same algorithm of interaction,

ignoring new modalities and options of a system. The main purpose of the present investigation is to assess user behavior and his preferences during interaction with intelligent services. Let us consider several types of interface, which are used in our experiments during testing an inquiry system.

Visual interface gives complete information; an inquiry is outputted to the screen, variants of answers to choose by pushing the menu items on the touchscreen. In this instance minimum of speech actions is expected from the user, especially speech communication with the robot. The potential client sees the interface assuming tactile-visual interaction, and is not ready to think of possibility of speech modality, even if this function is available. However predisposition to a choice of a touch modality instead of the speech one depends on the visual components of the interface.

In the case of a visual-speech interface, questions are synthesized by voice without any text duplication on the screen, and variants of answers are outputted to the display. Presence of output speech modality stimulated the user to give speech responses.

In speech interface (even combined with a visual component of the dialogue-system), both questions and variants of answers are voiced, a speech modality can be preferred, even if a touch modality is available. The choice of the speech interface can be made as the most natural.

In the case of both speech and visual interface with a full duplication of speech by the text on the touchscreen, it is expect that user behavior will similar to the variant with a completely visual interface owing to more informativity of the visual modality.

All the described interface cases are suitable for those tasks of dialogue interaction when there is no obvious requirement for a combination of interfaces (for example, speech and touch ones). Depending on type of a problem and type of information used during the interaction, as well as user experience, the necessary modality combination will be chosen by the user.

IV. Use Case: User Interface for Information Mobile Robot

The developed mobile robot consists of a mobile information platform and information desk. Multimodal user interface, developed earlier for the stationary information kiosk, was used in the design of the mobile version [24]. First of all the combination of audio source localization, voice activity detection and face tracking technologies was realized in the developed multimodal infokiosk equipped by the standard means for information input/output (touch-screen and loudspeaker) and the devices for contactless HCI (microphone array and web cameras). This test-bed model is able to determine the client's mouth coordinates and to detect boundaries of speech signal appeared in the kiosk speech dialogue area. The model was used for cognitive evaluation of three types of user interfaces: a) a speech interface; b) a speech-and-text interface; c) text interface.

Experiments were performed by questioning users with help of different types of the interface. There were questions of two kinds: with some variants of answer and without them. Testing of the three variants of the interface was carried out by means of questions of the first category only. For a reception of a spontaneous answer from the user and assessing his/her behavior in the limits of a spontaneous interaction the second kind of the questions was used, by means of the text-speech interface. To define influence of experience on the subsequent interactions for different groups of users' sets of questions were alternated. All the informants were students. Each student had 20 questions to answer; the first 10 questions had variants of answer, and the last 10 implied spontaneous and long answers. The test bench asked the students in three modes: 1) question in a synthesized voice; 2) question in a synthesized voice, duplicated upon the screen with a text; 3) text only.

TABLE I
User behavior during the experiments

Symbol	Number of phenomena	Number of students
Question to the associates	24	15
Attempt to control the dialogue	6	4
Silence	37	15
Voiced pause	41	15
Thoughts aloud	82	22
Self-correction	17	8
Multiple pressing the button	3	3
Repeated answer	5	3

The students were distributed into three groups, 10 students in the group, and each group was questioned in one mode. The progress bar and announcement about speech recording were outputted to the display. The informants were not instructed about behavior, all the decisions were to be made in the course of the test. The informants were tested one-by-one and did not see previous sessions. During the experiments a constant record of answers and monitoring of button-pressing was made. Table 1 presents the types and number of phenomena (i.e. reaction of informants) registered during the test. As it is well shown in the table, a half of the students asked their associates for help — perhaps, they did not trust the computer completely or just could not find ways to ask the computer itself. It was very typical of situations of hesitations about modality choice ("Should I press *the button* here?"), type of required answer ("Should I *just name the number*, eh?") or when the informant just did not know what to do ("*What must I tell* if I know no answer?").

The majority of the students kept silence if they knew no answer. Sometimes the pause was vocalized by sustaining some sound (a vowel or a sonant), cough, laugh and so on. But if the informant knew the answer, it was given in no time. Sometimes the students expressed their thoughts aloud.

A few students acted fussily, they pushed buttons several times and repeated answers. It is to be noted especially, that the dialogue was "one-sided", i.e. the computer just received some information and confirmed it. The informant did not want anything from the machine, so he had no fear, that

interaction would not be very successful. Some students told after the test, that they were confused by a long time given for answering.

During the experiments an answer to question means, the answer of any type and by any means was received: by speech, by pressing buttons; for questions with variants repetition of a variant or naming only its number was allowed; or just a "dunno-answer". Unlike usual examinations or test, the students knowing the answer, recognized it without trying to think up something, or kept silence, expecting a following question. In questions with offered variants of answers uncertainty was expressed by words like, "appears", "it's something like" "maybe", etc. More tangled and dim answers were recorded. The majority of answers was not similar to short orders and commands, they were supplemented with other words, characterizing the degree of confidence to the machine, reflexions etc. Also for answers of users (especially at the speech interface) were peculiar if the question or variants of answers was badly remembered. Thus respondents did not look forward to hearing to these questions, and used them only as the discourses markers often used in dialogue between people.

V. Use Case: Smart Meeting Room

The developed smart room is intended for holding small and medium events with up to forty-two participants. Two groups of devices are used for tracking participants and recording speakers: (1) personal web-cameras serve for observation of participants located at the conference table; (2) four microphone arrays with different configurations and five video cameras of three types are used for audio source localization and video capturing of participants, who sit in rows of chairs in another part of the room.

In our research, three major types of conscious services are studied: (1) an active controlling PTZ camera to point on active speakers; (2) an automatic archiving of meeting data, including photos of participants' faces, video records of speakers, presentation slides and whiteboard sketches based on online context analysis; (3) selection and web-transmission of the most actual multimedia content during the meeting in the smart room. The meeting web-transmission system, which deals with the latter service and uses some results of other services, is considered here.

The developed meeting web-transmission system (MWTS) consists of five main software complexes and one control server. Figure 1 presents all six modules, which are marked by digits. The first complex is Multimedia Device Control System (MDCS), which joins modules that control all multimedia hard-software. This multimedia hard-software records behavior of participants and displays some presentation data. Second complex is Multichannel Personal Web Camera Processing System (PWPS), which captures and processes both audio and video streams from the personal web-cameras. The third complex stores the recorded audio and video data of the meeting in the smart room. The fourth complex is a database, which includes information about the meeting. Meeting Control Server (MCS) (№ 6 in Figure 1) receives and analyses data from all other modules and gives information about received data to displaying web-system (DWS) (№ 5 in Figure 1). DWS joins modules, which transmit multimedia content to remote participants. Content Management System (CMS) consists of third, fifth and sixth complexes.

The first complex MDCS is responsible for multimedia devises work. Sketch Board System (SBS) allows subjects to use the plasma panel with the touch screen for drawing and writing notes. Presentation Control is responsible for loading, displaying and switching presentation slides. Multichannel Sound Localization System (MSLS) gives information about audio activity in the smart room. Multichannel Video Processing System (MVPS) is responsible for processing and recording of video streams incoming from the cameras, which are focused on the auditorium, presenter and sitting participants in the zone of chairs.

MPVPS consists of client modules, such as PWC, which supports work of personal cameras located on the conference table, as well as PWPS, which processes data from the PWCS modules. Audio files in the *wav* format and video files in the *avi* format, which were received from the personal cameras and processed by the MCS (change of the format, resolution and file name) images from MVPS, PCS, SB and PWC are added to the file storage. The meeting database is realized by MySQL server and includes two tables: (1) basic information on all scheduled meetings and; (2) information about the current meeting, which includes some data for the meeting display system. DWS works as a web-page with several forms [25]. The data about form content are processed based on the AJAX technology. The transmitting of audio data to the client-computer is based on the RTMP stream server and the Abobe Flash technology. MCS receives and analyses data from all the modules, as well as chooses of audio and video content for DWS. This analysis is based on the logical-time model. Software modules of MWTS were installed on several personal computers joined in one local network, connection between them is based on transmitting messages in a string format by UDP packets.

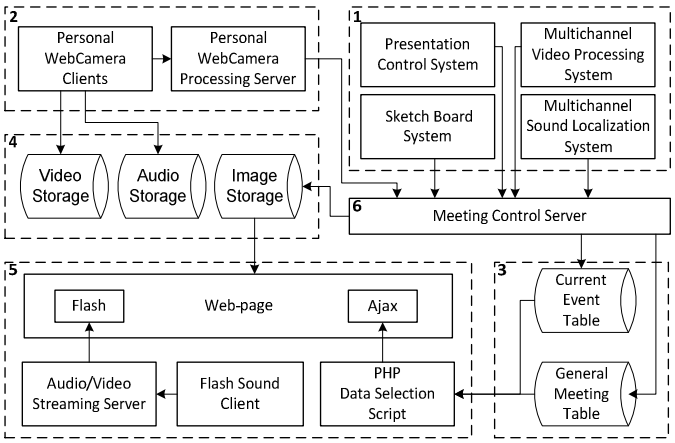


Fig. 1. Architecture of the meeting web-transmission system.

The work of the meeting web-transmission system and its components depends on the situation in the room. The component status and synchronization of audio and video content depend on the incoming events from the modules for audio localization, video monitoring, multimedia devices control. CMS manages by the multimedia content output, which is accessible for remote meeting participant. The events, which are generated by MCS and influenced on the meeting web-transmission system work, can be divided into four types by the following criteria: (1) by time; (2) by activity of the main speaker; (3) by activity of sitting participants; (4) by use of the presentation devices.

Experimental results were obtained with a natural scenario, where several people discussed a problem in the meeting room of 8.85x7.15x4.80m. One of the participants stayed in the presentation area and used the smart desk and the multimedia projector. Other participants were located at the conference table. The main speaker started his talk, when all the participants came together in the meeting room. Every participant could ask any questions after finish of the presentation. During the experiments the most of errors were made by the algorithm for detection of the active speaker, such errors occur when a participant at the conference table asks a question, but an image of other participant, which sits nearly, was displayed on the web-page. The accuracy of switching between the active participant and the presenter is higher. In total, about 97% of whole meeting time the graphical content were correctly selected at the analysis of the current situation in the meeting room.

VI. Conclusion

User profile and context modeling are the most important challenges of the ambient intelligence design. Development of the context-aware meeting processing systems gives appreciable benefits for automation of recording, archiving and translation of the meeting stream. The analysis of user behavior and multimedia equipment statuses is used for the context prediction and selection of audio and video sources, which transmit the most actual multimedia content for perception of the meeting and user provision with the relevant service. The developed meeting web-transmission system allows remote participants to perceive whole events in the meeting room via personal computers or smartphones. Further work will be focused on enhancement of abilities of remote participation during events in the intelligent meeting room and interaction with mobile information robot.

References

[1] T. Laakko, "Context-Aware Web Content Adaptation for Mobile User Agents," in *Studies in Computational Intelligence*, R. Nayak et al. (Eds.): SCI 130, Evolution of the Web in Artificial Intelligence Environments, 2008, pp. 69–99.

[2] C. Bolchini, C.A. Curino, E. Quintarelli, F.A. Schreiber, and L. Tanca, "A data-oriented survey of context models," *SIGMOD*, 36(4), 2007, pp. 19–26.

[3] A. Boytsov and A. Zaslavsky, "Extending context spaces theory by proactive adaptation," S. Balandin et al. (Eds.): *NEW2AN/ruSMART 2010*, LNCS 6294, Springer, 2010, pp. 1–12.

[4] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Proc. of the Workshop on Mobile Computing Systems and Applications*, Santa Cruz, CA, USA, 1994, pp. 85–90.

[5] D.R. Morse, N.S. Ryan, and J. Pascoe, "Enhanced reality fieldwork using hand-held computers in the field," *Life Sciences Educational Computing*, 9 (1), 1998, pp. 18–20.

[6] B. Moltchanov, C. Mannweiler, and J. Simoes, "Context-Awareness Enabling New Business Models in Smart Spaces," S. Balandin et al. (Eds.): *NEW2AN/ruSMART 2010*, LNCS 6294, Springer, 2010, pp. 13–25.

[7] K.H. Goh, J.Y. Tham, T. Zhang, and T. Laakko, "Context-Aware Scalable Multimedia Content Delivery Platform for Heterogeneous Mobile Devices," in *Proc. of MMEDIA 2011*, Budapest, Hungary, 2011, pp. 1–6.

[8] P. Dai, L. Tao and G. Xu, "Audio-Visual Fused Online Context Analysis Toward Smart Meeting Room," J. Indulska et al. (Eds.): *UIC 2007*, LNCS 4611, Springer, 2007, pp. 868–877.

[9] *Computers in the human interaction loop*. Ed. A. Waibel and R. Stiefelhagen, Berlin: Springer, 2009.

[10] G. Garau and H. Bourlard, "Using Audio and Visual Cues for Speaker Diarisation Initialisation," in *Proc. of ICASSP'2010*, 2010, pp. 4942–4945.

[11] Y. Rui, A. Gupta, J. Grudin, and L. He, "Automating lecture capture and broadcast: Technology and videography," *Multimedia Systems*, 10, 2004, pp. 3–15.

[12] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-Based Multimodal Speaker Detection for Distributed Meeting Videos," *IEEE Transactions on Multimedia*, Vol.10, No.8, 2008, pp.1541–1552.

[13] V. Rozgic, C. Busso, P.G. Georgiou, and S.S. Narayanan, "Multimodal meeting monitoring: Improvements on speaker tracking and segmentation through a modified mixture particle filter," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2007, pp. 60–65.

[14] V.C. Raykar, B. Yegnanarayana, S.R. Prasanna, and R. Duraiswami, "Speaker Localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, Volume 13, Issue 5, Part 2, 2005, pp. 751–761.

[15] T. Habib and H. Romsdorfer, "Concurrent Speaker Localization Using Multi-Band Position-Pitch (M-PoPi) Algorithm with Spectro-Temporal Pre-Processing," in *Proc. of Interspeech 2010*, Makuhari, Japan, 2010, pp. 2774–2777.

[16] J. Voordouw, C. Yang, L. Rothkrantz, and M. Capg, "A Comparison of the ILD and TDOA Sound Source Localization Algorithms in a Train Environment," in *Proc. of EuroMedia 2007*, Delft, 2007.

[17] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Proc. of Hands-Free Speech Communication and Microphone Arrays (HSCMA'2008)*, Trento, Italy, 2008.

[18] C. Datta, A. Kapuria, and R. Vijay, "A pilot study to understand requirements of a shopping mall robot," in *Proc. of HRI'2011*, 2011, pp. 127–128.

[19] M. Makatchev, I. Fanaswala, A. Abdulsalam, B. Browning, W. Ghazzawi, M. Sakr, and R. Simmons, "Dialogue Patterns of an Arabic Robot Receptionist," in *Proc. of HRI'2010*, 2010, pp. 167–168.

[20] M. Nieuwenhuisen, J. Stuckler, and S. Behnke, "Intuitive Multimodal Interaction for Service Robots," in *Proc. of HRI'2010*, 2010, pp. 177–178.

[21] A.C. Tenorio-Gonzalez, E.F. Morales, and L. Villaseñor-Pineda, "Teaching a robot to perform tasks with voice commands," in Grigori Sidorov, Arturo Hernandez Aguirre, Carlos Alberto Reyes Garcia (Eds.): *Proc. of the 9th Mexican international conference on Advances in artificial intelligence: Part I (MICAI'10)*, Springer-Verlag, 2010, pp. 105–116.

[22] G. Carrera J. Savage, and W. Mayol-Cuevas, "Robust feature descriptors for efficient vision-based tracking," in Luis Rueda, Domingo Mery, and Josef Kittler (Eds.): *Proc. of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications (CIARP'07)*, Springer-Verlag, 2007, pp. 251–260.

[23] A.C. Ramirez-Hernandez, J.A. Rivera-Bautista, A. Marin-Hernandez, and V.A. Garcia-Vega, "Detection and Interpretation of Human Walking Gestures for Human-Robot Interaction," in *Proc. of the 2009 Eighth Mexican International Conference on Artificial Intelligence (MICAI '09)*, IEEE Computer Society, Washington, DC, USA, 2009, pp. 41–46.

[24] V. Budkov, M. Prischepa, and A. Ronzhin, "Dialog Model Development of a Mobile Information and Reference Robot," *Pattern Recognition and Image Analysis*, Pleiades Publishing, Vol. 21, No. 3, 2011, pp. 442–445.

[25] V.Yu. Budkov, Al.L. Ronzhin, S.V. Glazkov, and An.L. Ronzhin, "Event-Driven Content Management System for Smart Meeting Room," S. Balandin et al. (Eds.): *NEW2AN/ruSMART 2011*, LNCS 6869, Springer-Verlag, 2011, pp. 550–560.

FPGA Implementation of Fuzzy Mamdani System with Parametric Conjunctions Generated by Monotone Sum of Basic t-Norms

Prometeo Cortés Antonio, Ildar Batyrshin, Herón Molina Lozano,
Marco Antonio Ramírez Salinas, and Luis Villa Vargas

***Abstract*—The paper presents the results of FPGA implementation of fuzzy Mamdani system with parametric conjunctions generated by monotone sum of basic t-norms. The system is implemented on the DE2 Altera development board using VHDL language. The system contains reconfigurable fuzzy Mamdani model with parametric membership functions and parametric operations that gives possibility to adjust the system to specific application.**

***Index terms*— Fuzzy Mamdani model, parametric conjunction, t-norm, FPGA, Altera.**

I. Introduction

THE fuzzy models and soft computing systems based on them have wide applications in the solution of real world problems [1, 2]. For this reason it is increasing the need in hardware implementation of highly reconfigurable fuzzy systems that can be easy adapted to various applications or to change of environment where fuzzy system is operated. Such reconfigurable fuzzy systems can be developed on two levels: on the level of fuzzy model and on the level of hardware implementation. This paper presents a method of hardware implementation of Mamdani fuzzy systems that reconfigurable on both levels. On the level of fuzzy model we consider fuzzy systems with parametric membership functions and parametric operations [3-5]. Such parameterization of fuzzy systems gives possibility to construct highly reconfigurable fuzzy systems with high adaptive possibilities. On the level of hardware we consider FPGA (Field Programmable Gate Array) implementation of fuzzy systems, i.e. an easily reprogrammable integrated circuit [6] that can be adapted to the change of parameters of fuzzy system and moreover to the change of the structure of fuzzy system.

We consider Mamdani fuzzy models with two input variables with fuzzy values given by triangular membership functions and one output variable with fuzzy values given by singletons. Such models are very popular in applications of fuzzy systems [2, 3]. Fuzzy models with parametric operations have been considered in various papers [4, 5, 7-10]. The methods of FPGA implementation of fuzzy systems have been studied in [11-16]. FPGA implementation of parametric conjunctions based on generator functions have been proposed in [17, 18]. In [19] it was proposed the method of generation of parametric conjunctions based on (p)-monotone sum of basic t-norms that has simpler FPGA implementation [20] than the method based on generator functions. In our paper we extend results obtained in [20] and we propose the method of FPGA implementation of Mamdani fuzzy models with parametric conjunctions based on (p)-monotone sum of basic t-norms.

The paper has the following structure. In the section II we discuss Mamdani fuzzy models that we consider in this work. The (p)-monotone sum of basic t-norms is discussed in Section III. Section IV presents logic diagrams of modules used in FPGA implementation of fuzzy system with parameterized membership functions and operations. Section V contains conclusions and directions of future work.

II. Mandani Fuzzy Models

Mandani fuzzy models with two inputs consist of the following rules [3]:

$$R_i\text{: If } x \text{ is } A_i \text{ AND } y \text{ is } B_i \text{ then } z \text{ is } C_i$$

where x,y are input variables, A_i and B_i are fuzzy sets defined on domains X and Y of x and y respectively, z is the output variable and C_i is its fuzzy value or singleton. In this paper we suppose that C_i is a singleton. Fuzzy sets are defined by their membership functions $A_i{:}X\rightarrow L$, $B_i{:}Y\rightarrow L$ where L is a set of membership values. In traditional fuzzy systems it is used the set of membership values L= [0,1]. In digital representation of membership values with m bits as in [19] we use the set of membership values L= {0,1,2,…, 2m-1} with maximal value 2m-1 denoted as I. This value will represent the full membership corresponding to the value 1 in traditional set of membership values [0,1]. For example, I= 15 if m= 4 and I= 255 if m= 8. Many concepts of fuzzy systems have straightforward extension on digital case when we replace the set of membership values [0, 1] by L= {0, 1, 2,…, 2m-1} and maximal membership value 1 by I = 2m-1.

Manuscript received May 30, 2011. Manuscript accepted for publication August 26, 2011.

Prometeo Cortés Antonio, Herón Molina Lozano, Marco Antonio Ramírez Salinas, and Luis Villa Vargas are with Center for Computing Research (CIC), National Polytechnic Institute (IPN), UP "Adolfo Lopez Mateos", Mexico City, Mexico (e-mail: acorteo@hotmail.com, lvilla@cic.ipn.mx, hmolina@cic.ipn.mx, mars@cic.ipn.mx).

Ildar Batyrshin is with Mexican Petroleum Institute, Mexico City, Mexico (e-mail: batyr1@gmail.com).

In our implementation we use m= 8 bits for representation of membership values and hence I= 255. Input variables x, y have three fuzzy values $\{XS, XM, XL\}$, $\{YS, YM, YL\}$ respectively with parameter values P_x, P_y defined as shown in Fig. 1 for variable x. These fuzzy values can be considered as formalizations of linguistic values *SMALL, MIDDLE, LARGE* of variables x, y respectively. For simplicity of fuzzy system implementation we use also m= 8 bits for representation of domains X, Y such that $X = Y = L$. Generally it can be used more bits for such representation, but in any case we can consider such x,y as normalized inputs of fuzzy system. The fuzzy sets A_i, B_i in the rules R_i take values from the sets $\{XS, XM, XL\}$, $\{YS, YM, YL\}$ respectively and the system generally contains 9 rules corresponding to all possible combinations of fuzzy values A_i and B_i.

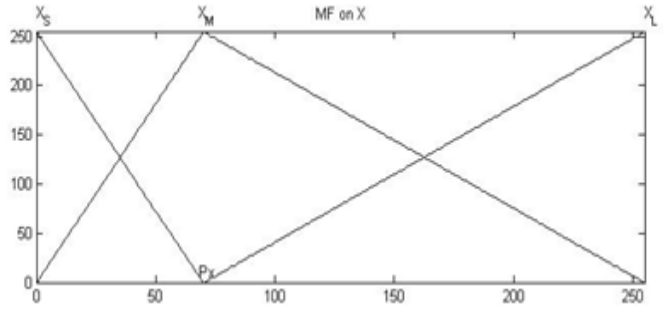


Fig. 1. The membership functions of fuzzy values X_S, X_M, X_L defined by parameter value P_x.

The details of inference, aggregation and defuzzification procedures for Mamdani fuzzy systems can be found in [3]. As a conjunction operation AND in Mamdani fuzzy model we use parametric conjunction defined by (p)-monote sum of basic t-norms discussed in the following section.

III. (P)-MONOTONE SUM OF BASIC T-NORMS

Fuzzy conjunction operation is a function $T : L\times L \rightarrow L$ satisfying on L conditions:

$T(x,I) = x, \quad T(I,y) = y,$ (boundary conditions)
$T(x,y) \le T(u,v)$ *if* $x \le u, y \le v.$ (monotonicity)

Commutative and associative conjunctions are called t-norms [21]. Usually in fuzzy systems as conjunction operation they are used the simplest conjunctions such as min o product operations. As parametric conjunction operations in fuzzy models it is possible to use parametric t-norms [21] or parametric fuzzy conjunctions introduced in [4,5] and having more simple form. But both types of parametric conjunctions are complicated for hardware implementation. In [17,18] it has been introduced a wide class of parametric fuzzy conjunctions based on simple generator functions and suitable for hardware implementation. In [19] it was proposed more simple class of parametric conjunctions called (p)-monotone sum of basic t-norms. We consider here the following basic t-norms:

$$T_M(x,y) = min\{x,y\}$$
(minimum)

$$T_L(x,y) = max\{x+y-I, 0\},$$

$$T_D(x,y) = \begin{cases} x, & if\ y = I \\ y, & if\ x = I \\ 0, & if\ x,y < I \end{cases}$$
(drastic t-norm)

Note that any fuzzy conjunction T satisfies the following inequalities:

$T_D(x,y) \le T(x,y) \le T_M(x,y)$.

We say that $T_1 \le T_2$ if $T_1(x,y) \le T_2(x,y)$ for all x,y from L, e.g. we have: $T_D \le T_L \le T_M$.

Suppose $a,b \in L$ and $a \le b$. The set of all elements $c \in L$ such that $a \le c \le b$ is denoted as $[a,b]$. Suppose $p \in \{0,1,\ldots, 2^m\text{-}2\}$ is a parameter. (p)-monotone sum of basic t-norms is defined as follows [19]. Define a partition of L on two sets: $X_1 = [0,p]$ and $X_2 = [p+1,I]$. These sets define a partition of $L\times L$ on four sections: $D_{ij}=X_i\times X_j$, $i,j \in \{1,2\}$ as shown in Fig. 2. Select a sequence of fuzzy conjunctions (T_{11}, T_{21}, T_{12}, T_{22}) ordered as follows: $T_{11} \le T_{12} \le T_{22}$, $T_{11} \le T_{21} \le T_{22}$. Define a function T on $L\times L$ by $T(x,y) = T_{ij}(x,y)$ if $(x,y) \in D_{ij}$, $i,j \in \{1,2\}$. This function T is called a (p)-monotone sum of fuzzy conjunctions T_{ij}, $i,j \in \{1,2\}$ and it will be a fuzzy conjunction [19]. It is clear that this fuzzy conjunction T can be defined as follows:

$$T(x,y) = \begin{cases} T_{11}(x,y), & if \quad x \le p, \quad y \le p, \\ T_{21}(x,y), & if \quad x > p, \quad y \le p, \\ T_{12}(x,y), & if \quad x \le p, \quad y > p, \\ T_{22}(x,y), & if \quad x > p, \quad y > p, \end{cases}$$

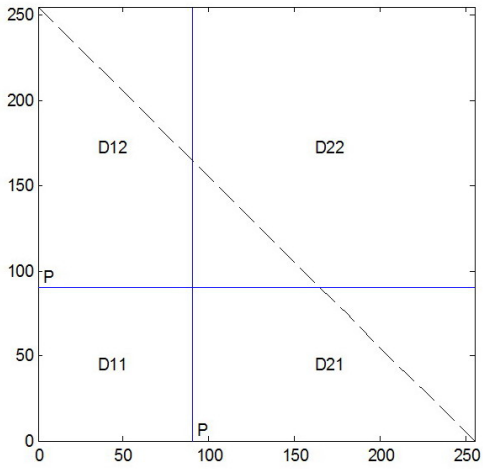


Fig. 2. Partition of $L\times L$ on sections $D_{ij}=X_i\times X_j$ defined by parameter value p.

As conjunctions (T_{11}, T_{21}, T_{12}, T_{22}) in the definition of (p)-monotone sum we will use basic t-norms T_D, T_L, T_M. For example, (p)-monotone sum defined by basic t-norms (T_D, T_L, T_L, T_M) will be denoted as conjunction T_{DLLM} and defined as follows:

$$T(x,y) = \begin{cases} T_D(x,y), & if \quad x \le p, \quad y \le p, \\ T_L(x,y), & if \quad x > p, \quad y \le p, \\ T_L(x,y), & if \quad x \le p, \quad y > p, \\ T_M(x,y), & if \quad x > p, \quad y > p, \end{cases}$$

Fig. 3 depicts on the left the locations of basic t-norms used in the construction of T_{DLLM} in sections D_{11}, D_{21}, D_{12}, D_{22}. On the right it is shown the shape of this digital fuzzy conjunction when the membership scale L is defined by m= 4 bits with parameter value p = 9.

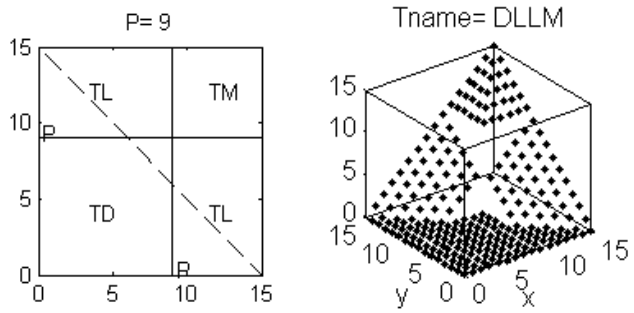


Fig. 3. Parametric conjunction TDLLM obtained by (p)-monotone sum of basic t-norms: TD- drastic, TL- Lukasiewicz and TM– minimum t-norms.

(p)-monotone sum will be commutative if T21 = T12. By means of basic t-norms it can be constructed 7 different non-trivial (p)-monotone sums defined by the following sequences of basic t-norms: DDDL, DDDM, DLLL, DLLM, DMMM, LLLM, LMMM. All of them have been implemented in [20] in FPGA. In the following section we will propose the method of FPGA implementation of these operations in Mamdani model.

IV. FPGA Implementation of Mamdani Model

The paper presents the results of 8-bit FPGA implementation of nine rule Mamdani fuzzy system developed by means of VHDL language [22] in Quartus II with ModelSim software of Altera [23,24]. FPGA implementation of Mamdani fuzzy system consists of various modules. First, for each input values of x and y we need to calculate their membership values in corresponding three fuzzy sets $\{X_S, X_M, X_L\}$ and $\{Y_S, Y_M, Y_L\}$ respectively. Fig. 4 shows a diagram of module calculating membership values of input variable x for all fuzzy sets $\{X_S, X_M, X_L\}$ by means of one divisor.

Fig. 5 presents the logic diagrams for implementing the basic t-norms, which are used to generate parametric conjunction defined by (p)-monotone sum of basic t-norms developed in [20].

As mentioned above, there are 7 possible combination operators generated by (p)-monotone sum, which are shown in Table I.

The methodology used to implement the parametric operators is based on the principle of partial reconfiguration, so that the implementation of the operators will be done according to the block diagram shown in Fig. 6, where a parametric operator consists of five basic modules: the modules of drastic, Lukasiewicz and minimum t-norms described above that connected to a multiplexer, which controls the output through the comparator module, which is responsible for comparing the values of the weights that generate the membership functions, with parameter p, and determine the combination used to obtain the result.

TABLE I
(P) PARAMETRIC CONJUNCTION OPERATOR

Code	Operator
1	DDDL
2	DDDM
3	DLLL
4	DLLM
5	DMMM
6	LLLM
7	LMMM

As an example the logic diagram that implements the Comparator module to DLLM operator implementation is shown in Fig. 7, consists of two comparators and a multiplexer. The implementation of the other operators generated by (p)-monotonic sum method are made in much the same way, varying only the allocation of the output and the code of comparators shown in Table 1.

The output crip value calculation in Mamdani fuzzy system with output membership functions given by singletons z_i is shown in equation

$$z_C = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i}, \tag{1}$$

where w_i are true values of antecedents of the rules. The implementation of this module can be divided in two parts. First, the module shown in Fig. 8a multiplies the nine outputs of the parametric conjunction operators with the corresponding singletons of consequents of the rules z_i, $F_i = w_i * z_i$. Second, the module shown in Fig. 8, calculates the rest of the formula (1).

Mamdani system implementation, with nine rules is shown in Fig. 9. The modules Fuzzy_X, Fuzzy_Y implement the membership functions of fuzzy sets. The basic operations are presented by modules Drastic, Lukasiewicz and Minimum. Comparator determines the sector of monotone sum where the operands of parametric conjunction are located and defines what outputs of basic t-norms should be used in multiplexor for all nine rules in parallel. The last module calculates the output of Mamdani model using singletons that stores in registers A_i.

V. Conclusion

The paper presents FPGA implementation of 8-bit parameterized fuzzy Mamdani system with two input variables and one output variable with singleton values. The fuzzy variables have three linguistic values used in antecedents of nine rules and presented by parameterized triangular membership functions. The system is implemented on the DE2 Altera development board using VHDL language. The system contains reconfigurable fuzzy Mamdani model with parametric membership functions and parametric operations that gives possibility to adjust the system to specific

application. In the future we plan to extend the class of generated parametric operations by including associative operations (t-norms), and to join together Mamdani y Sugeno models in one architecture.

REFERENCES

[1] J. Yen, R. Langari, and L.A. Zadeh, *Industrial Applications of Fuzzy Logic and Intelligent Systems*. NJ: IEEE Press, 1995.

[2] R.A. Aliev and R.R. Aliev, *Soft Computing and its Applications*. World Scientific, New Jersey, 2001.

[3] J.-S.R. Jang, C.T. Sun, E. Mizutani, *Neuro-Fuzzy and Soft Computing. A Computational Approach to Learning and Machine Intelligence*. Prentice-Hall International, 1997.

[4] I. Batyrshin, and O. Kaynak, "Parametric classes of generalized conjunction and disjunction operations for fuzzy modeling," *IEEE Transactions on Fuzzy Systems*, vol. 7, pp. 586–596, 1999.

[5] I. Batyrshin, O. Kaynak, and I. Rudas, "Fuzzy modeling based on generalized conjunction operations," *IEEE Transactions on Fuzzy Systems*, vol. 10, pp. 678–683, 2002.

[6] S. Kilts, *Advanced FPGA Design. Architecture, Implementation, and Optimization*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2007.

[7] A. Bikbulatov, and I. Batyrshin, "Tuning of operations in fuzzy models by neural nets," in *7th Zittau Fuzzy Colloquium*, Zittau, Germany, 1999, pp. 142–147.

[8] P.D. Koprinkova-Hristova, "Fuzzy operations' parameters versus membership functions' parameters influence on fuzzy control systems properties," in *2nd IEEE Int. Conf. on Intelligent Systems*, 2004, pp. 219–224.

[9] J. Alcalá-Fdez, F. Herrera, F. Márquez, and A. Peregrín, "Increasing fuzzy rules cooperation based on evolutionary adaptive inference systems," *Intern. Journal of Intelligent Systems*, vol. 22, pp. 1035–1064, 2007.

[10] A.C. Aras, O. Kaynak, and I.Z. Batyrshin, "A comparison of fuzzy methods for modeling," in *IECON 2008, 34th Annual Conf. of the IEEE Industrial Electronics Society*, Orlando, USA, 2008, pp. 43–48.

[11] M. McKenna, and B.M. Wilamowski, "Implementing a fuzzy system on a field programmable gate array," in *IJCNN'01, International Joint Conf. Neural Networks*, Washington, DC, 2001, vol.1, pp. 189–194.

[12] G. Mermoud, A. Upegui, C. Peña, and E. Sanchez, "A dynamically-reconfigurable FPGA platform for evolving fuzzy systems," in *Computational Intelligence and Bioinspired Systems*, LNCS, vol. 3512, Berlin Heidelberg: Springer, 2005, pp. 572–581.

[13] A. Di Stefano, and C. Giaconia, "An FPGA-based adaptive fuzzy coprocessor," in *Computational Intelligence and Bioinspired Systems*, LNCS, vol. 3512, Berlin Heidelberg: Springer, 2005, pp. 590–597.

[14] K.M Deliparaschos, F.I Nenedakis, and S.G Tzafestas, "Design and implementation of a fast digital fuzzy logic controller using FPGA technology," *J. Intelligent Robotic Systems*, vol. 45, pp 77-96, 2006.

[15] S. Sanchez-Solano, A.J. Cabrera, I. Baturone, F.J. Moreno-Velo, and M. Brox, "FPGA implementation of embedded fuzzy controllers for robotic applications," *IEEE Transactions on Industrial Electronics*, vol. 54, pp. 1937–1945, 2007.

[16] G. Lizarraga, R. Sepulveda, O. Montiel, and O. Castillo, "Modeling and simulation of the defuzzification stage using Xilinx system generator and Simulink," in *Soft Computing for Hybrid Intelligent Systems*, Studies in Computational Intelligence, vol. 154, Berlin Heidelberg: Springer, 2008, pp. 333–343.

[17] H. Zavala, I.Z. Batyrshin, I.J. Rudas, L. Villa Vargas, and O. Camacho Nieto, "Parametric operations for digital hardware implementation of fuzzy systems," in *MICAI 2009*, LNAI, vol. 5845, Berlin Heidelberg: Springer, 2009, pp. 432–443.

[18] I.J. Rudas, I.Z. Batyrshin, A. Hernández Zavala, O. Camacho Nieto, and L. Villa Vargas, "Digital fuzzy parametric conjunctions for hardware implementation of fuzzy systems," in *ICCC2009, IEEE 7th Int. Conf. Computational Cybernetics*, Palma de Mallorca, Spain, 2009, pp. 157–166.

[19] I.Z. Batyrshin, I.J. Rudas, and A. Panova, "On generation of digital fuzzy parametric conjunctions," in *Towards Intelligent Engineering and Information Technology, Studies in Computational Intelligence*, vol. 243, Berlin Heidelberg: Springer, 2009, pp. 79–89.

[20] P. Cortés Antonio, I. Batyrshin, I. Rudas, A. Panova, L.A. Villa Vargas, "FPGA Implementation of (p)-Monotone Sum of Basic t-norms," in *WCCI 2010, FUZZ-IEEE*, Barcelona, 2010.

[21] E.P. Klement, R.Mesiar, and E. Pap, *Triangular Norms*. Dordrecht: Kluwer, 2000.

[22] S. Brown, and Z. Vranesic, *Fundamentals of Digital Logic with VHDL Design*. Second Edition. Mc Graw Hill, 2005.

[23] Cyclone II Device Handbook, Vol. 1, Altera, 2008. Available: http://www.altera.com/literature/hb/cyc2/cyc2_cii5v1.pdf

[24] DE2 Development and Education Board User Manual. User Manual. Versión 1.4. Altera, 2006. Available: ftp://ftp.altera.com/up/pub/Webdocs/DE2_UserManual.pdf

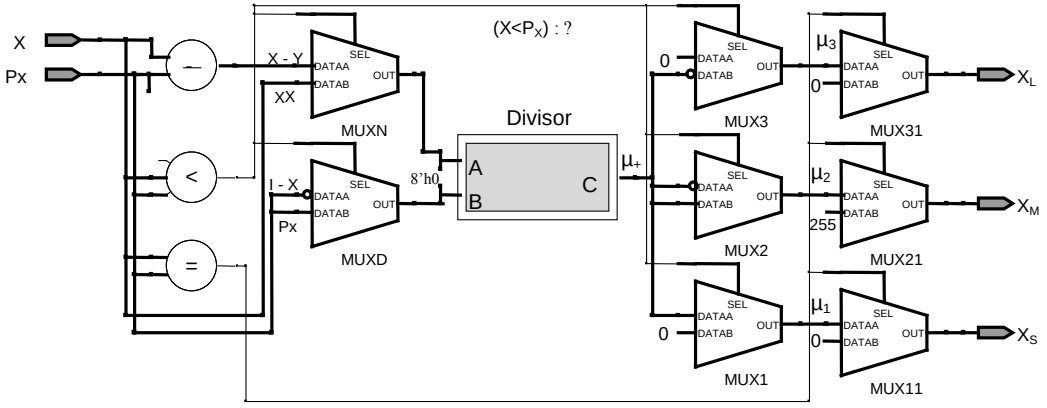
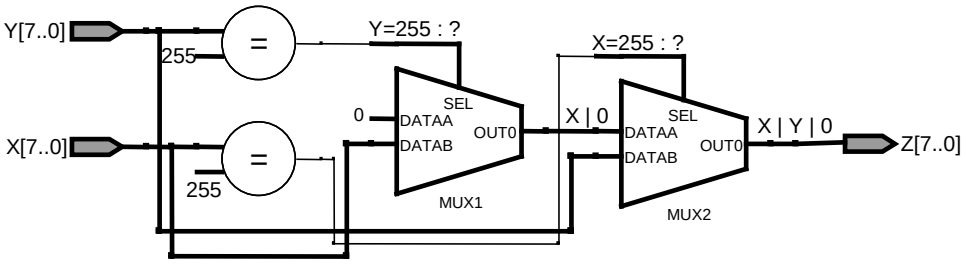
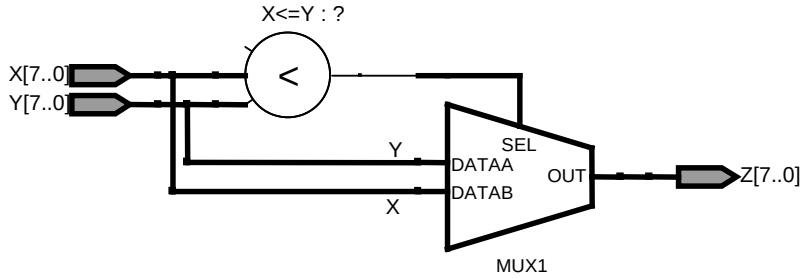


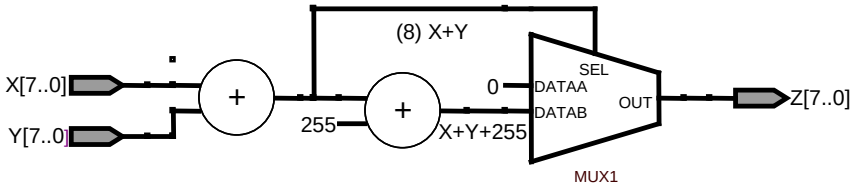
Fig. 4. Diagram of calculation of membership values of input variable *x*.



a) Drastic operator.



b) Min operator.



c) Lukasiewizt operator.

Fig. 5 Diagrams of basic t-norm calculation: a) Drástic, b) Lukasewicz y c) min.

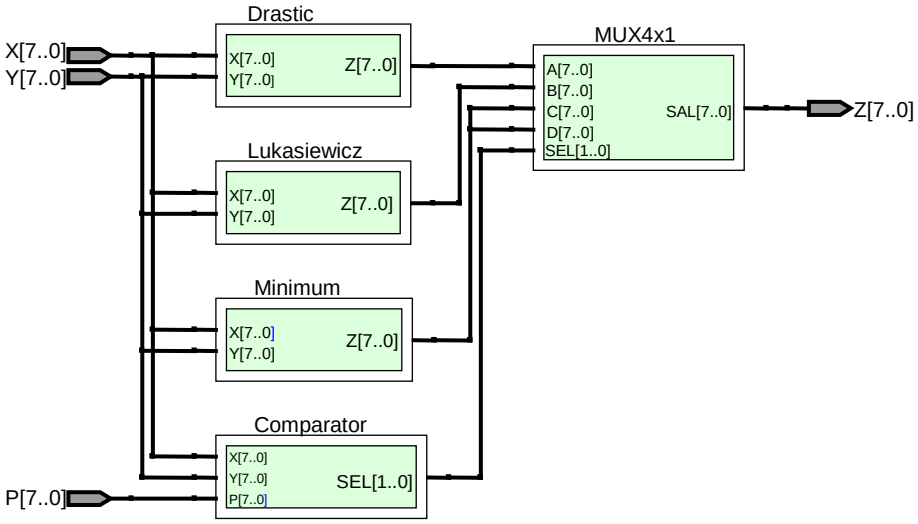


Fig. 6 Diagram of calculation of parametric conjunction used by (p) monotonic sum.

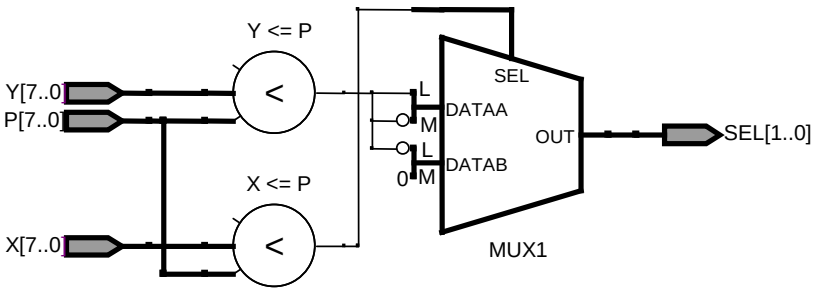
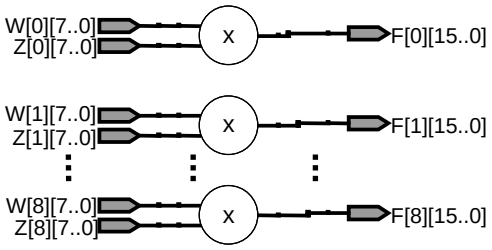
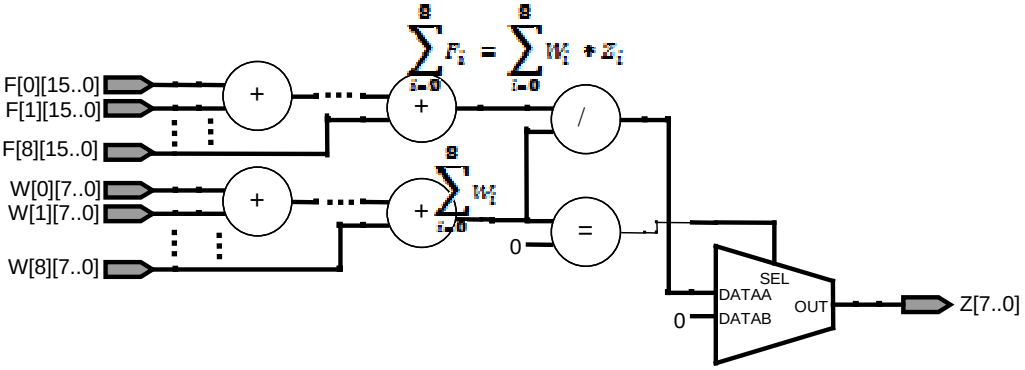


Fig. 7. Diagram of comparator module.



a) $w_i * z_i$



b) implementation of the rest of the formula (1).

Fig. 8 Diagrams of calculation of the output of the system.

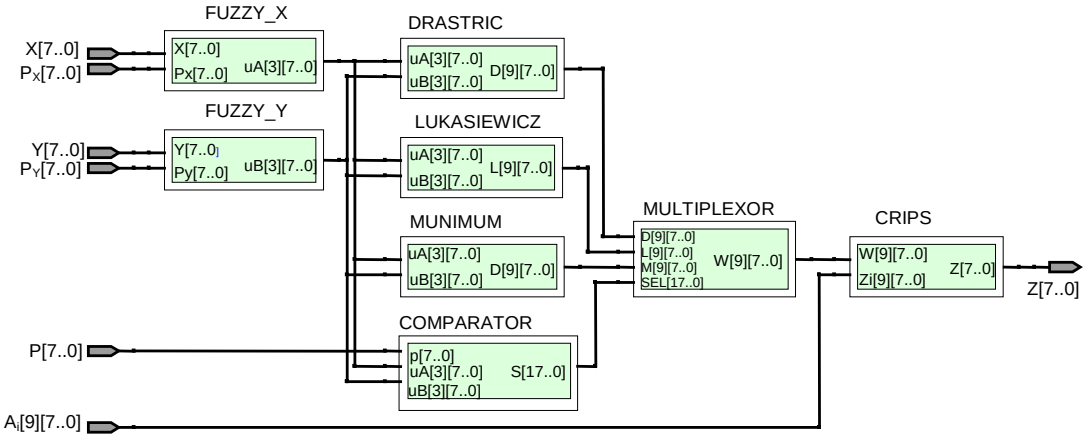


Fig. 9. General diagram of Mandani system.

Automatic Music Composition with Simple Probabilistic Generative Grammars

Horacio Alberto García Salas, Alexander Gelbukh, Hiram Calvo, and Fernando Galindo Soria

***Abstract*—We propose a model to generate music following a linguistic approach. Musical melodies form the training corpus where each of them is considered a phrase of a language. Implementing an unsupervised technique we infer a grammar of this language. We do not use predefined rules. Music generation is based on music knowledge represented by probabilistic matrices, which we call evolutionary matrices because they are changing constantly, even while they are generating new compositions. We show that the information coded by these matrices can be represented at any time by a probabilistic grammar; however we keep the representation of matrices because they are easier to update, while it is possible to keep separated matrices for generation of different elements of expressivity such as velocity, changes of rhythm, or timbre, adding several elements of expressiveness to the automatically generated compositions. We present the melodies generated by our model to a group of subjects and they ranked our compositions among and sometimes above human composed melodies.**

***Index Terms*—Evolutionary systems, evolutionary matrix, generative grammars, linguistic approach, generative music, affective computing.**

I. Introduction

Music generation does not have a definite solution. We regard this task as the challenge to develop a system to generate a pleasant sequence of notes to human beings and also this system should be capable of generating several kinds of music while resembling human expressivity. In literature, several problems for developing models for fine arts, especially music have been noted. Some of them are: How to evaluate the results of a music generator? How to determine if what such a system produces is music or not? How to say if a music generator system is better than other? Can a machine model expressivity?

Different models have been applied for developing automatic music composers; for example, those based on neural networks [15], genetic algorithms [2, 25] and swarms [4] among other methods.

Manuscript received February 10, 2011. Manuscript accepted for publication July 30, 2011.

H. A. García Salas is with the Natural Language Laboratory, Center for Computing Research, National Polytechnic Institute, CIC-IPN, 07738, DF, México (e-mail: itztzin@gmail.com).

A. Gelbukh was, at the time of submitting this paper, with the Waseda University, Tokyo, Japan, on Sabbatical leave from the Natural Language Laboratory, Center for Computing Research, National Polytechnic Institute, CIC-IPN, 07738, DF, México (e-mail: gelbukh@gelbukh.com).

H. Calvo is with the Natural Language Laboratory, Center for Computing Research, National Polytechnic Institute, CIC-IPN, 07738, DF, México (e-mail: hcalvo@cic.ipn.mx).

F. Galindo Soria is with Informatics Development Network, REDI (e-mail: fgalindo@ipn.mx).

In order to generate music automatically we developed a model that describes music by means of a linguistic approach; each musical composition is considered a phrase that is used to learn the musical language by inferring its grammar. We use a learning algorithm that extracts musical features and forms probabilistic rules that afterwards are used by a note generator algorithm to compose music. We propose a method to generate linguistic rules [24] finding musical patterns on human music compositions. These patterns consist of sequences of notes that characterize a melody, an author, a style or a music genre. The likelihood of these patterns of being part of a musical work is used by our algorithm to generate a new musical composition.

To model the process of musical composition we rely on the concept of *evolutionary systems* [8], in the sense that systems evolve as a result of constant change caused by flow of matter, energy and information [10]. Genetic algorithms, evolutionary neural networks, evolutionary grammars, evolutionary cellular automata, evolutionary matrices, and others are examples of evolutionary systems. In this work we follow the approach of evolutionary matrices [11].

This paper is organized as follows. In Section II we present works related to automatic music composition. In Section III, we describe our model. In Section IV, we describe an algorithm to transform a matrix into a grammar. In Section V we show how we handle expressivity in our model. In Section VI, we present results of a test to evaluate generated music. Finally, in Section VII, we present some conclusions of our model and future work to improve our model.

II. Related Work

A. Review Stage

An outcome of development of computational models applied to humanistic branches as fine arts like music is generative music or music generated from algorithms.

Different methods have been used to develop music composition systems, for example: noise [5], cellular automata [20], grammars [13, 22], evolutionary methods [13], fractals [14, 16], genetic algorithms [1], case based reasoning [19], agents [21] and neural networks [7, 15]. Some systems are called hybrid since they combine some of these techniques. For a comprehensive study please refer to [23] and [17].

Harmonet [15] is a system based on connectionist networks, which has been trained to produce chorale style of J. S. Bach. It focuses on the essence of musical information, rather than restrictions on music structure. Eck and Shmidhuber [7] believe that music composed by recurrent neural networks lacks structure, and do not maintain memory of distant events.

They developed a model based on LSTM (Long Short Term Memory) to represent the overall and local music structure, generating blues compositions.

Kosina [18] describes a system for automatic music genre recognition based on audio content signal, focusing on musical compositions of three music genres: classical, metal and dance. Blackburm and DeRoure [3] present a system to recognize through the contents of a music database, with the idea to make search based on music contours, i.e. in a relative changes representation in a musical composition frequencies, regardless of tone or time.

There is a number of works based on evolutionary ideas for music composition. For example, Ortega *et al.* [22] used generative context-free grammars for modeling the musical composition. Implementing genetic algorithms they made grammar evolve to improve the musical generation. GenJam [1] is a system based on a genetic algorithm that models a novice jazz musician learning to improvise. It depends on user feedback to improve new compositions through several generations.

Todd and Werner [25] developed a genetic algorithm based on co-evolution, learning and rules. In their music composer system there are male individuals that produce music and female critics that evaluate it to mate them. After several generations they create new musical compositions.

In our approach we focus on the following points:

- The evolutionary aspect—to keep learning while generating;
- Stressing the linguistic metaphor of musical phrases and textual phrases, words and sets of notes;
- Adding expressiveness to achieve a more human aspect;
- Studying the equivalence between a subset of grammar rules and matrices [11].

III. Music Generation

A musical composition is a structure of note sequences made of other structures built over time. How many times a musical note is used after another reflects patterns of sequences of notes that characterizes a genre, style or an author of a musical composition. We focus on finding patterns on monophonic music.

A. Linguistic approach

Our model is based on a linguistic approach [9]. We describe musical compositions as phrases made up of sequences of notes as lexical items that represent sounds and silences throughout time. The set of all musical compositions forms the musical language.

In the following paragraphs we define some basic concepts that we will use in the rest of this paper.

Definition 1: A *note* is a representation of tone and duration of musical sound.

Definition 2: The *alphabet* is the set of all notes:
alphabet = {notes}.

Definition 3: A *musical composition m* is an arrangement of musical notes: Musical composition = $a_1\ a_2\ a_3 \ldots a_n$ where $a_i \in$ {notes}.

In our research we work with musical compositions *m* of monophonic melodies, modeling two variables of notes: musical frequencies and musical tempos. We split these variables to form a sequence of symbols with each of them.

Definition 4: The *Musical Language* is the set of all musical compositions: Musical Language = {musical compositions}.

For example, having the sequence of notes (frequencies) of musical composition "*El cóndor pasa*" (the condor passes by):
b e $d_\#$ e $f_\#$ g $f_\#$ g a b_2 d_2 b_2 e_2 d_2 b_2 a g e g e b e $d_\#$ e $f_\#$ g $f_\#$ g a b_2 d_2 b_2 e_2 d_2 b_2 a g e g e b_2 e_2 d_2 e_2 d_2 e_2 g_2 e_2 d_2 e_2 d_2 b_2 g e_2 d_2 e_2 d_2 e_2 g_2 e_2 d_2 e_2 d_2 b_2 a g e g e
We assume this sequence is a phrase of musical language.

B. Musical Evolutionary System

Evolutionary systems interact with their environment finding rules to describe phenomena and use functions that allow them to learn and adapt to changes. A scheme of our evolutionary model is shown in Fig. 1.

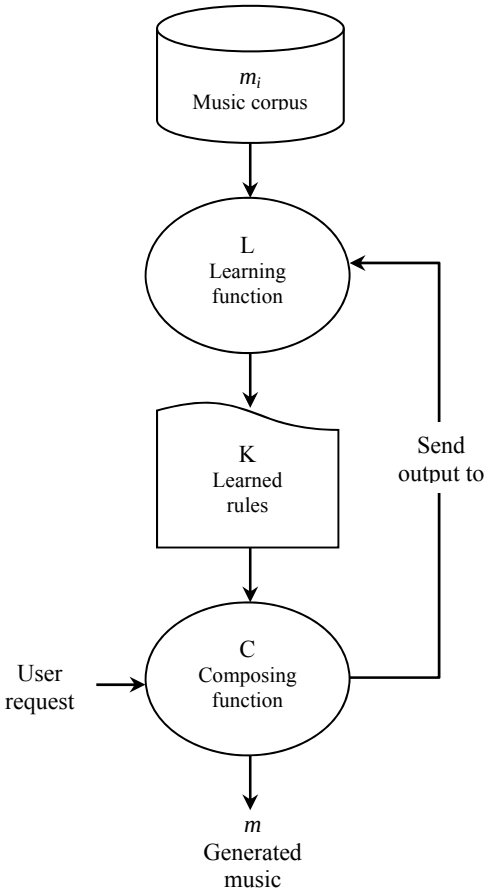


Fig. 1. Model.

The *workspace* of musical language rules is represented by K and there exist many ways to make this representation, e.g. grammars, matrices, neural nets, swarms and others. Each musical genre, style and author has its own rules of

composition. Not all of these rules are described in music theory. To make automatic music composition we use an evolutionary system to find rules *K* in an unsupervised way.

The function L is a learning process that generates rules from each musical composition m_i creating a representation of musical knowledge. The evolutionary system originally does not have any rule. We call K_0 when K is empty. While new musical examples m_0, m_1, …, m_i are learned K is modified from K_0 to K_{i+1}.

$$L(m_i, K_i) = K_{i+1}$$

Function L extracts musical features of m_i and integrates them to K_i generating a new representation K_{i+1}. This makes knowledge representation K evolves according to the learned examples.

These learned rules K are used to generate musical composition *m* automatically. It is possible to construct a function C(K) where C is called musical composer. Function C uses K to produce a novel musical composition *m*.

$$C(K) = m$$

For listening of the new music composition there is a function I called musical interpreter or performer that generates the sound.

$$I(m) = \text{sound}$$

Function I takes music *m* generated by function C to stream it to the sound device. We will not discuss this function in this paper.

C. Learning Module based on Evolutionary Matrices

To describe our music learning module we need to define several concepts. Let L be a learning process as the function that extracts musical features and adds this information into K. There are different ways to represent K. In our work we use a matrix representation. We will show in Section IV that this is equivalent to a probabilistic grammar.

Definition 5: Musical Frequency = {musical frequencies} where musical frequencies (*mf*) are the number of vibrations per second (Hz) of notes.

Definition 6: Musical Time = {*musical times*} where *musical times* (mt) are durations of notes.

Function L receives musical compositions *m*. Musical Composition $m = a_1\ a_2\ a_3 \ldots\ a_n$ where $a_i=\{f_i,t_i\}$, $i \in [1,n]$, $f_i \in$ Musical Frequency, $t_i \in$ Musical Time, $[1,n] \subset N$

To represent rules K we use matrices for musical frequencies and for musical times. We refer to them as rules M. Originally these matrices are empty; they are modified with every musical example.

Rules M are divided by function L into MF and MT where MF is the component of musical frequencies (*mf*) rules extracted from musical compositions and MT is the component of musical time (*mt*) rules.

We are going to explain how L works with musical frequency matrix MF. Time matrix MT works the same way.

Definition 7: MF is a workspace formed by two matrices. One of them is a *frequency distribution matrix* (FDM) and the other one is a *cumulative frequency distribution matrix* (CFM).

Each time a musical composition m_i arrives, L upgrades FDM. Then it recalculates CFM, as follows:

Definition 8: Let *FrequencyNotes* be an array in which are stored the numbers corresponding to a musical composition notes.

Definition 9: Let *n* be the number of notes recognized by the system, $n \in N$.

Definition 10: Frequency Distribution Matrix (FDM) is a matrix with *n* rows and *n* columns.

Given the musical composition $m = f_1\ f_2\ f_3 \ldots\ f_r$ where $f_i \in$ *FrequencyNotes*. The learning algorithm of L to generate the frequency distribution matrix FDM is:

$$\forall i \in [1,r],\ [1,r] \subset N,\ FDM_{fi,fi+1} = FDM_{fi,fi+1}+1,$$

where $FDM_{fi,fi+1} \in FDM$.

Definition 11: Cumulative Frequency Distribution Matrix CFM is a matrix with *n* rows and *n* columns.

The algorithm of L to generate cumulative frequency distribution matrix CFM is:

$$\forall i \in [1,n],\ \forall j \in [1,n],\ [1,n] \subset N,\ \forall\ FDM_{i,j} \neq 0$$

$$CFM_{i,j} = \sum_{k=1}^{j} FDM_{i,k}$$

These algorithms to generate MF, the workspace formed by FDM and CFM, are used by function L with every musical composition m_i. This makes the system evolve recursively according to musical compositions m_0, m_1, m_2,…, m_i.

$$L(m_i,\ldots L(m_2, L(m_1, L(m_0, MF_0))))=MF_{i+1}$$

D. Composer Function C: Music Generator Module

Monophonic music composition is the art of creating a single melodic line with no accompaniment. To compose a melody a human composer uses his/her creativity and musical knowledge. In our model composer function C generates a melodic line based on knowledge represented by cumulative frequency distribution matrix CFM.

For music generation is necessary to choose next note. In our model each *i* row of CFM represents a probability function for each *i* note on which is based the decision of the next note. Each *j* column different of zero represents possible notes to follow the *i* note. The most probable notes form characteristic musical patterns.

Definition 12: T_i and T.
Let T_i to be an element where it is store the total of cumulative frequency sum of each *i* row of FDM.

$$\forall i \in [1,n], [1,n] \subset \mathrm{N}, \quad \mathrm{T}_i = \sum_{k=1}^{n} FDM_{i,k}$$

Let T be a column with *n* elements where it is store the total of cumulative frequency sum of FDM.

Note generation algorithm:

```
while(not end)
{
  p=random(T_i)
  while(CFM_i,j < p)
    j=j+1
  next note=j
  i=j
}
```

E. Example

Let us take the sequence of frequencies of musical composition "*El cóndor pasa*":
b e $d_{\#}$ e $f_{\#}$ g $f_{\#}$ g a b_2 d_2 b_2 e_2 d_2 b_2 a g e g e b e $d_{\#}$ e $f_{\#}$ g $f_{\#}$ g a b_2 d_2 b_2 e_2 d_2 b_2 a g e g e b_2 e_2 d_2 e_2 d_2 e_2 g_2 e_2 d_2 e_2 d_2 b_2 g e_2 d_2 e_2 d_2 e_2 g_2 e_2 d_2 e_2 d_2 b_2 a g e g e

FrequencyNotes = {**b**, $\mathbf{d}_{\#}$, **e**, $\mathbf{f}_{\#}$, **g**, **a**, $\mathbf{b}_2$, $\mathbf{d}_2$, $\mathbf{e}_2$, $\mathbf{g}_2$} are the terminal symbols or alphabet of this musical composition. They are used to tag each row and column of frequency distribution matrix FDM. Each number stored in FDM of Fig. 2, represents how many times a row note was followed by a column note in *condor pasa* melody. To store the first note of each musical composition S row is added, it represents the axiom or initial symbol. Applying the learning algorithm of L we generate frequency distribution matrix FDM of Fig. 2.

	b	$d_{\#}$	e	$f_{\#}$	g	a	b_2	d_2	e_2	g_2
S	1									
b			2							
$d_{\#}$			2							
e	1	2		2	3		1			
$f_{\#}$					4					
g			6	2		2			1	
a					3		2			
b_2					1	3		2	3	
d_2							6		6	
e_2								10		2
g_2									2	

Fig. 2. Frequency distribution matrix FDM.

We apply the algorithm of L to calculate cumulative frequency distribution matrix CFM of Fig. 3 from frequency distribution matrix FDM of Fig. 2. Then we calculate each T_i of T column.

For generation of a musical composition we use note generator algorithm. Music generation begins by choosing the first composition note. S row of matrix of Fig. 3 contains all possible beginning notes. In our example only the **b** note can be chosen. Then **b** is the first note and the *i* row of $CFM_{i,j}$ which we use to determine second note. Only the **e** note can be chosen after the first note **b**.

So the first two notes of this new musical melody are m_{i+1}={**b**, **e**}. Applying note generator algorithm to determine third note: We take the value of column T_e=9. A *p* random number between zero and 9 is generated, suppose *p*=6. To find next note we compare *p* random number with each non-zero value of **e** row until one greater than or equal to this number is found. Then column **g** is the next note since $M_{e,g}$=8 is greater than *p* = 6. The column *j* = **g** is where it is stored this number that indicates the following composition note and the following *i* row to be processed. The third note of new musical composition m_{i+1} is **g**. So m_{i+1} = {**b**, **e**, **g**,…}. Then to determine the fourth note we must apply the note generator algorithm to *i* = **g** row.

Since each non-zero value of *i* row represents notes that used to follow *i* note, then we will generate patterns according to probabilities learned from musical compositions examples.

	b	$d_{\#}$	e	$f_{\#}$	g	a	b_2	d_2	e_2	g_2	T
S	1										1
b			2								2
$d_{\#}$			2								2
e	1	3		5	8		9				9
$f_{\#}$					4						4
g			6	8		10			11		11
a					3		5				5
b_2					1	4		6	9		9
d_2							6		12		12
e_2								10		12	12
g_2									2		2

Fig. 3. Cumulative frequency distribution matrix CFM.

IV. Matrices and Grammar

Our work is based on a linguistic approach and we have used a workspace represented by matrices to manipulate music information. Now we show that this information representation is equivalent to a probabilistic generative grammar.

There are different ways to obtain a generative grammar G,. From frequency distribution matrix FDM and total column T, it is possible to construct a probabilistic generative grammar.

Definition 13: MG is a workspace formed by FDM and a probabilistic grammar G.

To generate a grammar first we generate a probability matrix PM determined from frequency distribution matrix FDM.

Definition 14: Probability Matrix (PM) is a matrix with *n* rows and *n* columns.

The algorithm to generate probability matrix PM is:

$$\forall i \in [1,n], \forall j \in [1,n], \forall \mathrm{FDM}_{i,j} \neq 0 \qquad \mathrm{PM}_{i,j} = \mathrm{FDM}_{i,j}/\mathrm{T}_i$$

There is a probabilistic generative grammar G{V_n, V_t, S, P, Pr} such that G can be generated from PM. V_n is the set of nonterminals symbols, V_t is the set of all terminal symbols or alphabet which represents musical composition notes. S is the axiom or initial symbol, P is the set of rules generated and Pr is the set of rules probabilities represented by values of matrix PM.

For transforming the PM matrix in a grammar we use the following algorithm:

1. Build the auxiliary matrix AM from PM:
    a. substitute each row *i* tag of PM with a nonterminal symbol X_i except S row which is copied as it is
    b. substitute each column *j* tag by its note f_j and a nonterminal symbol X_j
    c. copy all values of cells of matrix PM into corresponding cells of matrix AM
2. For each row *i* and each column *j* such that $AM_{i,j} \neq 0$
    a. *i* row corresponds to grammar rule X_i
    b. *j* column corresponds to a terminal symbol f_j and a nonterminal symbol X_j with probability $p_{i,j}$

Then rules of grammar G are of the form $X_i \rightarrow f_j\ X_j\ (p_{i,j})$. This is a grammatical representation of our model. For each music composition m_i a MG, the workspace formed by FDM and grammar G, can be recursively generated.

$$L(m_i,\ldots L(m_2, L(m_1, L(m_0, MG_0))))=MG_{i+1}$$

A. *Example*

From frequency distribution matrix FDM of Fig. 2 it is generated probability matrix PM of Fig. 4.

	b	$d_\#$	e	$f_\#$	g	a	b_2	d_2	e_2	g_2
S	1									
b			1							
$d_\#$			1							
e	1/9	2/9		2/9	3/9		1/9			
$f_\#$					1					
g			6/11	2/11		2/11			1/11	
a					3/5		2/5			
b_2					1/9	3/9		2/9	3/9	
d_2							6/12		6/12	
e_2								10/12		2/12
g_2									1	

Fig. 4. Probability matrix PM.

	bX_1	$d_\#X_2$	eX_3	$f_\#X_4$	gX_5	aX_6	b_2X_7	d_2X_8	e_2X_9	g_2X_{10}
S	1									
X_1			1							
X_2			1							
X_3	1/9	2/9		2/9	3/9		1/9			
X_4					1					
X_5			6/11	2/11		2/11			1/11	
X_6					3/5		2/5			
X_7					1/9	3/9		2/9	3/9	
X_8							6/12		6/12	
X_9								10/12		2/12
X_{10}									1	

Fig. 5. Auxiliary matrix AM.

From matrix PM of Fig. 4 the auxiliary matrix AM of Fig. 5 is generated. From given AM matrix of Fig. 5 We can generate grammar $G\{V_n, V_t, S, P, Pr\}$. Where $V_n=\{S, X_1, X_2 X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}\}$ is the set of non-terminals symbols. $V_t=\{b, d_\#, e, f_\#, g, a, b_2, d_2, e_2, g_2\}$ is the set of all terminal symbols or alphabet. S is the axiom or initial symbol. Pr is the set of rules probabilities represented by values of matrix AM. Rules P are listed in Fig. 6.

$S \rightarrow b\ X_1(1)$
$X_1 \rightarrow e\ X_3(1)$
$X_2 \rightarrow e\ X_3(1)$
$X_3 \rightarrow b\ X_1(1/9) \mid d_\#\ X_2(2/9) \mid f_\#\ X_4(2/9) \mid g\ X_5(3/9) \mid b_2\ X_7(1/9)$
$X_4 \rightarrow g\ X_5(1)$
$X_5 \rightarrow e\ X_3(6/11) \mid f_\#\ X_4(2/11) \mid a\ X_6(2/11) \mid e_2\ X_9(1/11)$
$X_6 \rightarrow g\ X_5(3/5) \mid b_2\ X_7(2/5)$
$X_7 \rightarrow g\ X_5(1/9) \mid a\ X_6(3/9) \mid d_2\ X_8(2/9) \mid e_2\ X_9(3/9)$
$X_8 \rightarrow b_2\ X_7(6/12) \mid g_2\ X_{10}(6/12)$
$X_9 \rightarrow d_2\ X_8(10/12) \mid g_2\ X_{10}(2/12)$
$X_{10} \rightarrow e_2\ X_9(1)$

Fig. 6. Probabilistic generative grammar.

V. Expressivity

Expressivity can be regarded as a mechanism that displays transmission and interpretation vividness of feelings and emotions. For example fear in front of a threat. Physical factors interfere like cardiac rhythm, changes in respiratory system, in endocrine system, in muscular system, in circulatory system, secretion of neurotransmitters, etc. Another important factor is empathy which is the capacity of feelings and emotions recognition in others [6]. It is out of our research to explain how these physical changes are made or how empathy takes place among living beings. We just simulate expressivity in music generation.

A. *Expressivity within our Model*

Music can be broken down into different functions that characterize it like frequency, time and intensity. So each note of a melody is a symbol with several features or semantic descriptors that give the meaning of a long or short sound, low, high, intense, soft, of a guitar or of a piano.

With our model is possible to represent each of these variables using matrices or grammars that reflect their probabilistic behavior. In this paper we have presented how to model frequency and time. We can build an intensity matrix the same way. With more variables more expressivity the generated music will reflect.

Using our model we can characterize different kinds of music based on its expressivity, for example in happy music or sad music. Besides we have the possibility of mixing features of distinct kinds of music, for example frequency functions of happy music with time functions of sad music. Also we can combine different genres like classic times with rock frequencies. So in addition of generating music we can invent new genres and music styles.

VI. Results

In order to evaluate whether our algorithm is generating music or not, we decided to conduct a Turing-like test. Participants of this test had to tell us if they like music generated by our model, without them knowing that it was automatically music generated. This way we sought the answer to two questions: whether or not we are doing music and whether or not our music is pleasant.

We compiled 10 melodies, 5 of them generated by our model and another 5 by human composers and we asked human subjects to rank melodies according to whether they liked them or not, with numbers between 1 and 10 being number 1 the most they liked. None of subjects knew about the order of music compositions. These 10 melodies were presented as in Table I.

TABLE I
ORDER OF MELODIES AS THEY WERE PRESENTED TO SUBJECTS

ID	Melody	Author
A	*Zanya*	(generated)
B	*Fell*	Nathan Fake
C	*Alucin*	(generated)
D	*Idiot*	James Holden
E	*Ciclos*	(generated)
F	*Dali*	Astrix
G	*Ritual Cibernetico*	(generated)
H	*Feelin' Electro*	Rob Mooney
I	*Infinito*	(generated)
J	*Lost Town*	Kraftwerk

We presented this test to more than 30 participants in different places and events. We sought that the characteristics of these participants were as varied as possible (age, gender and education), however most of them come from a related IT background. Test results were encouraging, since automatically generated melodies were ranked at 3rd and 4th place above human compositions. Table II shows the ranking of melodies as a result of the Turing-like test we developed.

TABLE II
ORDER OF MELODIES OBTAINED AFTER THE TURING-LIKE TEST

ID	Ranking	Melody	Author
B	1	*Fell*	Nathan Fake
D	2	*Idiot*	James Holden
C	3	*Alucín*	(generated)
A	4	*Zanya*	(generated)
F	5	*Dali*	Astrix
H	6	*Feelin' Electro*	Rob Mooney
J	7	*Lost Town*	Kraftwerk
E	8	*Ciclos*	(generated)
G	9	*Ritual Cibernético*	(generated)
I	10	*Infinito*	(generated)

VII. CONCLUSIONS AND FUTURE WORK

We proposed an evolutionary model based on evolutionary matrices for musical composition. Our model is learning constantly, increasing its knowledge for generating music while more data is presented. It does not need any predefined rules. It generates them from phrases of the seen language (musical compositions) in an unsupervised way.

As we shown, our matrices can be expressed as probabilistic grammar rules, so that we can say that our systems extracts grammar rules dynamically from musical compositions. These rules generate a musical language based on the compositions presented to the system. These rules can be used to generate different musical phrases, meaning new musical compositions. Because the probabilistic grammars learned can generalize a language beyond the seen examples of it, our model has what can be called innovation, which is what we are looking for music creation, while keeping the patterns learned from human music.

As a short-term future work we plan to characterize different kinds of music, from sad to happy, or from classic to electronic in order to find functions for generating this kind of music. We are also developing the use of other matrices to consider more variables involved in a musical work, such as velocity, fine-graded tempo changes, etc., thus adding more expressivity to the music created by our model.

ACKNOWLEDGEMENTS

The work was done under partial support of Mexican Government (CONACYT 50206-H, SIP-IPN 20113295, COFAA-IPN, PIFI-IPN, SNI).

REFERENCES

[1] J. A. Biles, "GenJam: Evolution of a Jazz Improviser," *Creative Evolutionary Systems, Section: Evolutionary Music*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 2001, pp. 165–187.

[2] D. Birchfield, "Generative Model for the Creation of Musical Emotion, Meaning and Form," in *Proceedings of the 2003 International Multimedia Conference ACM SIGMM*, Berkeley, California: Workshop on Experiential Telepresence, Session: Playing experience, 2003, pp. 99–104.

[3] S. Blackburm, and D. DeRoure, "A tool for content based navigation of music," in *Source International Multimedia Conference. Proceedings of the sixth ACM international conference on Multimedia*. Bristol, United Kingdom, 1998, pp. 361–368.

[4] T. Blackwell, "Swarming and Music," *Evolutionary Computer Music*. Springer London, 2007, pp. 194–217.

[5] M. Bulmer, "Music From Fractal Noise," in *Proceedings of the Mathematics 2000 Festival*, University of Queensland, Melbourne, 2000.

[6] T. Cochrane, "A Simulation Theory of Musical Expressivity," *The Australasian Journal of Philosophy*, Volume 88, Issue 2, 191–207, 2010.

[7] D. Eck, and J. Schmidhuber, *A First Look at Music Composition using LSTM Recurrent Neural Networks*, Source Technical Report: IDSIA-07-02. Publisher Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 2002.

[8] F. Galindo Soria, "Sistemas Evolutivos: Nuevo Paradigma de la Informática," en *Memorias XVII Conferencia Latinoamericana de Informática*, Caracas Venezuela, 1991.

[9] F. Galindo Soria, "Enfoque Lingüístico," en *Memorias del Simposio Internacional de Computación de 1995*, Cd. de México: Instituto Politécnico Nacional CENAC, 1995.

[10] F. Galindo Soria, *Teoría y Práctica de los Sistemas Evolutivos*, Cd. de México, 1997.

[11] F. Galindo Soria, "Matrices Evolutivas," en *Memorias de la Cuarta Conferencia de Ingeniería Eléctrica CIE/98,* Cd. de México: Instituto Politécnico Nacional, CINVESTAV, 1998, pp. 17–22.

[12] A. García Salas, *Aplicación de los Sistemas Evolutivos a la Composición Musical*, México D.F: Tesis de maestría, Instituto Politécnico Nacional UPIICSA, 1998.

[13] A. García Salas, A. Gelbukh, and H. Calvo, "*Music Composition Based on Linguistic Approach*," in *Proceedings of the 9th Mexican International Conference on Artificial Intelligence*, Pachuca, México, 2010, pp. 117–128.

[14] M. Gardner, "Mathematical Games: White and Brown Music, Fractal Curves and One-Over-f Fluctuations," *Scientific American*, 4, 16–32, 1978.

[15] H. Hild, J. Feulner, and W. Menzel, "Harmonet: A Neural Net for Harmonizing Chorales in the Style of J. S. Bach," *Neural Information Processing 4*. Germany: Morgan Kaufmann Publishers Inc., 1992, pp. 267–274.

[16] R. Hinojosa, *Realtime Algorithmic Music Systems From Fractals and Chaotic Functions: Toward an Active Musical Instrument*, Barcelona: PhD Thesis, Universitat Pompeu Fabra, 2003.

[17] H. Järveläinen, "Algorithmic Musical Composition," in *Seminar on content creation Art@Science*, Helsinki: University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2000.

[18] K. Kosina, "Music Genre Recognition." *Diplomarbeit. Eingereicht am Fachhochschul-Studiengang. Mediente Chnik Und Design in Hagenberg*, 2002.

[19] G. Maarten, J.L. Arcos, and R. López de Mántaras, "A Case Based Approach to Expressivity-Aware Tempo Transformation," *Machine Learning*, 65(2-3): 411–437, 2006.

[20] K. McAlpine, E. Miranda, and S. Hoggar, "Making Music with Algorithms: A Case-Study System," *Computer Music Journal*, 23(2): 19–30, 1999.

[21] M. Minsky, "Music, Mind, and Meaning." *Computer Music Journal*, 5(3), 1981.

[22] A P. Ortega, A.R. Sánchez, and M. M. Alfonseca, "Automatic composition of music by means of Grammatical Evolution," *ACM SIGAPL APL*, 32(4): 148–155, 2002.

[23] G. Papadopoulos, and G. Wiggins, "AI Methods for Algorithmic Composition: A Survey, a Critical View and Future Prospects," in *Symposium on Musical Creativity 1999,* University of Edinburgh, School of Artificial Intelligence Division of Informatics, 1999, pp. 110–117.

[24] Y. Ledeneva and G. Sidorov, "Recent Advances in Computational Linguistics," *Informatica. International Journal of Computing and Informatics*, 34, 3–18, 2010.

[25] P.M. Todd and G.M. Werner, "Frankensteinian Methods for Evolutionary Music Composition," *Musical networks: Parallel distributed perception and performance*, MA, USA: Cambridge, MIT, Press Bradford Books, 1999.

An Approach to Cross-Lingual Textual Entailment using Online Machine Translation Systems

Julio Castillo and Marina Cardenas

Abstract—**In this paper, we show an approach to cross-lingual textual entailment (CLTE) by using machine translation systems such as Bing Translator and Google Translate. We experiment with a wide variety of data sets to the task of textual Entailment (TE) and evaluate the contribution of an algorithm that expands a monolingual TE corpus that seems promising for the task of CLTE. We built a CLTE corpus and we report a procedure that can be used to create a CLTE corpus in any pair of languages. We also report the results obtained in our experiments with the three-way classification task for CLTE and we show that this result outperform the average score of RTE (Recognizing Textual Entailment) systems. Finally, we find that using WordNet as the only source of lexical-semantic knowledge it is possibly to build a system for CLTE, which achieves comparable results with the average score of RTE systems for both two-way and three-way tasks.**

Index Terms—**Cross-lingual textual entailment, textual entailment, WordNet, bilingual textual entailment corpus.**

I. Introduction

THE objective of the Recognizing Textual Entailment (RTE) task [1] is determining whether the meaning of a hypothesis *H* can be inferred from a text *T*. Thus, we say that *T entails H*, if a person reading *T* would infer that *H* is most likely true.

Therefore, this definition assumes common human understanding of language and common background knowledge. Below, we provide an example of a *T-H* pair:

T= "*Dawson is currently a Professorial Fellow at the University of Melbourne, and an Adjunct Professor at Monash University*".

H= "*Dawson teaches at Monash University*".

In that context, Cross-Lingual Textual Entailment has been recently proposed in [2] as a generalization of Textual Entailment task (also Monolingual Textual Entailment) that consists in determining if the meaning of *H* can be inferred from the meaning of *T* when *T* and *H* are in different languages.

This new task has to face more additional issues than monolingual TE. Among them, we emphasize the ambiguity, polysemy, and coverage of the resources. Another additional problem is the necessity for semantic inference across languages, and the limited availability of multilingual knowledge resources. In RTE the most common resources used are WordNet, VerbOcean, Wikipedia, FrameNet, and DIRT. From them, only WordNet and Wikipedia are available in other languages different than English, but again, naturally with problems of coverage.

However, it is interesting to remark that, from the ablation test reported on TAC2010[1][3], some RTE systems had a positive impact using such resources, but other had a negative impact, thus the important thing is the way in which the systems utilize the available knowledge resources.

In this paper, we conduct experiments for CLTE, taking English as source language and Spanish as target language in the task of deciding the entailment among multiple languages. We chose this pair of languages due to the well-known accuracy of the translation models between Spanish and English and also due to our availability of translators whose first language is Spanish. In our work, the CLTE problem is addressed by using a machine learning approach, in which all features are WordNet-based, with the aim of measuring the benefit of WordNet as a knowledge resource for the CLTE task.

We know that the coverage of WordNet is not very good for narrow domains [4], and that also provides limited coverage of proper names. However, we are interested in evaluating the effectiveness of WordNet for CLTE, because this is the most widely used in TE. Despite these limitations, our system achieves a performance above the average score, and provides a promising direction for this line of research.

Thus, we tested a MLP and SVM classifier over two and three way decision tasks. Our focus to CLTE is based on free online (web) machine translation systems, so we chose Microsoft Bing Translator[2], because it has a good efficiency when translating English to Spanish or vice-versa, and also because provides a wide range of language pairs for translation. In addition, we use Google Translate[3], because his high efficiency has been tested in other NLP tasks [5] and [6].

This decoupled approach between Textual Entailment and Machine Translation has several advantages, such as taking

Manuscript received July 1, 2011. Manuscript accepted for publication October 2, 2011.

J. Castillo is with National University of Cordoba - FaMAF, Cordoba, Argentina and also with the National Technological University-Regional Faculty of Cordoba, Argentina (email: jotacastillo@gmail.com).

M. Cardenas is with the National Technological University-Regional Faculty of Cordoba, Argentina (email: ing.marinacardenas@gmail.com).

[1] http://www.nist.gov/tac/2010/RTE/index.html
[2] http://www.microsofttranslator.com/
[3] http://translate.google.com/

benefits of the most recent advances in machine translation, the ability to test the efficiency of different MT systems, as well as the ability to scale the system easily to any language pair.

Our approach is similar to that described in [2], because it uses a machine translation approach to CLTE. But, while they use an English-French CLTE engine with the TE engine based on edit distance algorithms, in contrast, our approach is English-Spanish CLTE, and it is completely based on semantics, because our TE engine only uses WordNet-based semantic similarity measures.

We also present the first results on assessing CLTE for the three-way decision task proposed by [7] for monolingual TE, with the idea of building a CLTE system whose outputs provide more precise informational distinctions of the judgments, making a three-way decision among *YES*, *NO*, and *UKNOWN*.

Additionally, to our knowledge, we present the first available bilingual entailment corpus aimed for the task of CLTE, which is released to the community.

This paper continues on Section 2 showing the creation of the CLTE datasets. Section 3 describes the system architecture. In section 4 we provide an experimental evaluation and discussion of the results achieved for CLTE in the two and three way tasks. Finally, Section 5 summarizes some conclusions and future work.

II. Creating the Dataset for CLTE

In order to perform experiments in CLTE, we first needed to create a corpus. Thus, we started creating a bilingual English-Spanish textual entailment corpus which was based on the original monolingual corpus from previous RTE Campaigns. We built a training set and a test set, both based on the technique of human-aided machine translation.

A. Training Set

In our experiments, we built three training sets that were generated according to the following procedure.

First, we started by selecting the original RTE3 development set, and then the hypothesis was translated from English into Spanish, using Microsoft Bing Translator as machine translation system. As a result, we generated the dataset denoted by RTE3_DS_ENtoSP.

Second, all hypotheses H are manually classified in one of three classes: Good, Regular and Bad, according to the following heuristic definition:

Good: One hypothesis *H* is classified as *Good* if its meaning is perfectly understandable for a native Spanish speaker and has the same meaning as the original hypothesis *H* that belongs to the RTE3 dataset.

Regular: One of the hypotheses *H* is classified as *Regular* if its meaning is understandable for a native Spanish speaker with little effort, or if it contains less than three syntactic errors, and it has the same meaning as the original hypothesis *H* that belongs to the RTE3 dataset.

Bad: One hypothesis *H* is classified as *Bad* if its meaning is not comprehensible to a native Spanish speaker, or has three or more syntactic errors, or if its meaning is different from the original hypothesis *H* that belongs to the RTE3 dataset.

The above procedure involved the participation of three translators whose native language is Spanish, and the classification decision was obtained from a consensus of the translators themselves. For convenience, we say that a *T-H* pair belongs to any of the above categories if the hypothesis *H* belongs to one of them. As a result, we obtained a sets of *T-H* pairs, which are denoted as RTE3_DS_ENtoSP to indicate that the dataset is composed by *T-H_Sp* pairs, where the hypothesis *H_Sp* is the translated version to Spanish from the original hypothesis *H*, and here we adopted the notation $\text{RTE3\_DS\_ENtoSP} = \{Bad\} \cup \{Regular\} \cup \{Good\}$. In a similar way, for those *T-H* pairs classified as *Good* or *Regular*, we generated the dataset: $\text{RTE3\_DS\_ENtoSP\_Good} + \text{RegPairs} = \text{RTE3\_DS\_ENtoSP} - \{Bad\}$, and finally, for those *T-H* pairs classified as *Good*, we generated the dataset: $\text{RTE3\_DS\_ENtoSP\_Good} = \text{RTE3\_DS\_ENtoSP} - \{Bad\} - \{Regular\}$.

TABLE I
EXAMPLES OF THE CLASSIFIED PAIRS

Pair ID	CLASS	Hypothesis	Comment
454	BAD	En 1945, se *eliminó* una bomba atómica sobre Hiroshima.	Wrong verb.
537	BAD	El *faro de faros* estaba situado en Alejandría.	Wrong NER.
788	BAD	Los miembros de Democrat tenían expedientes de votación fuertes de la pequeña empresa.	Don't make sense.
766	REG	Molly Walsh planea *parar el comprar* de los productos genéricos.	*parar de comprar*
18	Good	La aspirina previene la hemorragia gastrointestinal.	
756	Good	Las píldoras contaminadas contuvieron fragmentos del metal.	

The use of these training sets is motivated by the need of assessing the impact of automatic translations and manual translations performed by native Spanish speakers in the task of CLTE. Also, we are especially interested in measuring the effect of the pairs classified as *Bad* in the overall accuracy of the system.

As result, the RTE3_TS_ENtoSP_Good dataset is composed by 542 pairs, the $\text{RTE3\_TS\_ENtoSP} + \text{RegPair}$ dataset is composed by 704 pairs, and the RTE3_TS_ENtoSP dataset is composed by 800 pairs.

Table 1 illustrates some examples of the pairs classified as *Good*, *Bad* and *Regular*. When the hypothesis belongs to the class *Bad*, it is provided the justification of the human translators.

B. Test Set

In test set, we conducted a separate classification process for each annotator. The reason for this is that we are interested in assessing the agreement between the annotators on the test set built. Thus, each hypothesis *H* of the dataset was judged as *Good*, *Regular* or *Bad*, following the previous definition. We

note that pairs on which the annotators disagreed were filtered-out of the class *Good*.

We started selecting the original RTE3 test set, and then the hypothesis is translated from English into Spanish. Thus, the test set named RTE3_TS_ENtoSP is created.

First, three annotators judged each pair of the RTE3_TS_ENtoSP testset generated by Google Translate. Then, we applied the Fleiss' kappa statistical measure with the aim of assessing the reliability of agreement among the annotators. As a result, the annotators agreed in 82% of their judgment, and disagreed in 18% which corresponded to Kappa level of 0.68, regarded as substantial agreement according to [8]. The disagreement was generally found when classifying a hypothesis *H* as Regular, due to the fact that some errors in *H* could by easily corrected and thus include *H* into the class *Good*. Whereas other times, the hypothesis *H* presented some errors that justified the inclusion to the class *Bad*, for one annotator, but it was classified as *Regular* according to the criteria of another annotator. We also remark that the classes *Good* and *Bad* has high degree of agreement among annotators.

For that reason, we filtered-out the pairs classified as *Regular*, eliminating about 19% from the original pairs, and then we removed the pairs classified as *Bad*, which is an additional elimination of 10% and it is because we suppose that these pairs are not useful for inference purposes. As result, we built the dataset RTE3_TS_ENtoSP_Good. Furthermore, one annotator performed a final proofreading editing the dataset. Finally, this corpus is composed by 558 pairs, which represent a 69% of the original dataset.

In the experiments, we adopted the notation: $\text{RTE3\_TS\_ENtoSP} = \{Bad\} \cup \{\text{Regular}\} \cup \{Good\}$, and $\text{RTE3\_TS\_ENtoSP\_Good} = \text{RTE3\_TS\_ENtoSP} - \{Bad\} - \{Regular\}$.

III. System Architecture

Our system is based on a machine learning approach for CLTE. The system produces feature vectors for all datasets defined in the previous section. We experimented with SVM and MLP classifiers because of their well known performance in natural language applications. The architecture of the system is shown in Figure 1.

From Figure 1 we can see that two Online Machine Translation systems are used. Also, we note that an adaptation layer has been built in order to convert a bilingual TE task into a monolingual TE task. The datasets created on Section 2.2 are required to be in bilingual English-Spanish as inputs to the CLTE layer. In opposite, the other datasets are in monolingual English-English.

This is because some of them are used at the level of CLTE layer, and other are used at the TE level.

In all experiments it was necessary a bilingual test set in English-Spanish language.

We used the following training sets: RTE3-4C[4], and RTE4-4C[4], as proposed by the authors in [9] in order to extend the RTE data sets by using machine translation engines following a variation of the round trip translation technique. We remark that all corpus used in this paper are available to the community[4].

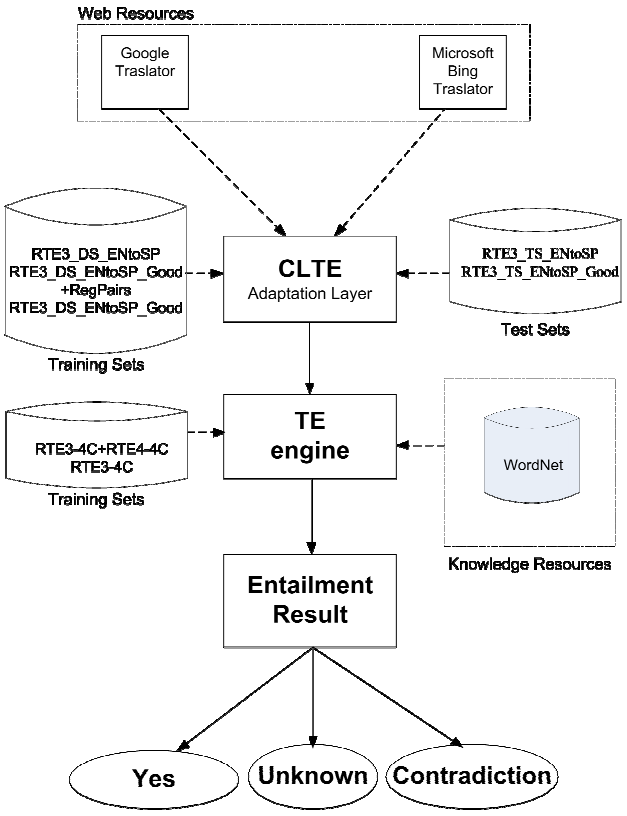


Fig. 1. General architecture of the system.

Round trip translation is defined as the process of starting with an *S* (string in English) and translating it into a foreign language *F(S)* (for example Spanish) and finally back into the English source language $F^{-1}(S)$. The Spanish language was chosen as the intermediate language, and Microsoft Bing Translator as the only MT system in this process. It was built upon the idea of providing a tool to increase the corpus size aiming to acquire more semantic variability.

The expanded corpus is denoted RTE3-4C and the three-way task is composed of: 340 pairs *Contradiction*, 1520 pairs *Yes*, and 1114 pairs *Unknown*. Thus, the two-way task is composed of: 1454 pairs *No*(No Entailment), and 1520 pairs *Yes*(Entailment). On the other hand, the RTE4-4C dataset has the following composition: 546 pairs *Contradiction*, 1812 pairs *Entailment*, and 1272 pairs *Unknown*. Therefore, in the two-way task, there are 1818 pairs *No* and 1812 pairs *Yes* in this dataset.

The sign "+" represents the union operation of sets and "4C" means "four combinations" denoting that the dataset was generated using the algorithm to expand datasets [9] and using only one Translator engine.

In addition, we also converted the three-way corpus into only two classes: *Yes* (Entailment), and *No* (No Entailment). For this purpose, both *Contradiction* and *Unknown* classes

[4]http://www.investigacion.frc.utn.edu.ar/mslabs/~jcastillo/Sagan-test-suite/

were conflated and retagged as the class *No* Entailment.

It is important to note that the dataset RTE3-4C+RTE4-4C+RTE3_DS_ENtoSP is not present in the Figure 1 because is a result of the union of dataset of both CLTE and TE layers.

Finally, our Textual Entailment engine utilizes eight WordNet-based similarity measures, such as proposed by the authors in [10], with the purpose of obtaining the maximum similarity between two concepts. These text-to-text similarity measures are based on the followings word-to-word similarity metrics: Resnik [11], Lin [12], Jiang & Conrath [13], Pirrò & Seco [14], Wu & Palmer [15], Path Metric, Leacock & Chodorow [16], and a semantic similarity to sentence level named SemSim [10].

A. Features

In this section we provide a brief resume of the text-text similarity measures which are the features of our system.

WordNet is used to calculate the semantic similarity between a T (Text) and an H (Hypothesis). The following procedure is applied:

Step 1. Perform WSD based on WordNet glosses.

Step 2. A semantic similarity matrix between *T* and *H* is defined.

Step 3. A function *Fsim* is applied to *T* and *H*.

Where the Function *Fsim* could be one of the followings seven functions over concepts *s*, and *t*:

Function 1. The Resnik similarity metric is calculated as: $RES(s,t) = IC(LCS(s,t))$, where *IC* (information content) is defined as: $IC(w) = -\log P(w)$

The function *P(w)* is the probability of selecting *w* in a large corpus, and the function *LCS(s,t)* is the least common subsume of s and t.

Function 2. The Lin similarity metric is calculated as follows:

$$LIN(s,t) = \frac{2 * IC(LCS(s,t))}{IC(s,t)}$$

Function 3. The Jiang & Conrath metric is computed as follows:

$$JICO(s,t) = \frac{1}{IC(s) + IC(t) - 2 * IC(LCS(s,t))}$$

Function 4. The Pirro & Seco (PISE) similarity metric is computed as follows:

$$PISE(s,t) = \begin{cases} 3 * IC(msca(s,t)) - IC(s) - IC(t), if\ s \neq t \\ 1 \qquad\qquad\qquad\qquad\quad , if\ s = t \end{cases}$$

The function *msca* is the most specific common abstraction value for the two given synsets (Lucene documents).

Function 5. The Wu & Palmer measure is computed as follows:

$$WUPA(C_1(s), C_2(t)) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

Where: C_1 and C_2 are the synsets to which *s* and *t* belong, respectively. C_3 is the least common superconcept of C_1 and C_2. N_1 is the number of nodes of the path from C_1 to C_3. N_2 is the number of nodes of the path from C_2 to C_3. N_3 is the number of nodes on the path from C_3 to root.

Function 6. The metric Path is reciprocal to the length of the shortest path between 2 synsets. Note that we count the 'nodes' (synsets) in the path, not the links. The allowed POS types are nouns and verbs.

$$PA(s,t) = Min_i(PathLength_i(s,t))$$

where: $PathLength_i(s,t)$ gives the length of the i-Path between *s* and *t*.

Function 7. The Leacock & Chodorow metric finds the path length between s and t in the "is-a" hierarchy of WordNet, and is computed as follows:

$$LECH(C_1(s), C_2(t)) = -\log\left(\frac{Min_i(PathLength_i(s,t)}{2 * D}\right)$$

where: *D* = is the maximum depth of the taxonomy (considering only nouns and verbs).

Step 4. Finally, the string similarity between two lists of words is reduced to the problem of bipartite graph matching by using the Hungarian algorithm over this bipartite graph. Then, we find the assignment that maximizes the sum of ratings of each token. Note that each graph node is a token/word of the list.
At the end, the final score is calculated by:

$$finalscore = \frac{\sum_{s \in T, t \in H} opt(Fsim(s,t))}{Max\ (Length(T), Length\ (H))}$$

where: *opt(F)* is the optimal assignment in the graph.
Length (T) is the number of tokens in *T*, *Length (H)* is the number of tokens in *H*, and

$$F_{sim} \in \{RES, LIN, JICO, PISE, WUPA, PA, LECH\}$$

Finally, note that the partial influence of each of the individual similarities will be reflected on the overall similarity.

Function 8. Additionally, the SemSim metric is defined and calculated as follows:
Step 1. Perform WSD based on WordNet definitions.
Step 2. Compute a semantic similarity matrix between words in *T* and *H*, using only synonym and hyperonym relationship. The Breadth First Search algorithm is used over these tokens. Then, the semantic similarity between two words/concepts *s* and *t*, is computed as:

$$Sim\ (s,t) = 2 \times \frac{Depth\ (LCS(s,t))}{Depth\ (s) + Depth\ (t)}$$

where: *Depth(s)* is the shortest distance from the root node to the current node.
Step 3. In this step the Function 8 is computed. Thus, in order to obtain the final score, the matching average between two sentences *T* and *H* is calculated as follows:

$$SemSim(T,H) = 2 \times \frac{Match\ (T,H)}{Length\ (T) + Length\ (H)}$$

Finally, this procedure produces eight WordNet-based semantic similarity measures, which have been tested over monolingual textual entailment [10] achieving results that outperformed the average accuracy of the RTE systems.

IV. Results and Discussion

In this section, we test the system to predict the following test sets: RTE3_TS_ENtoSP and RTE3_TS_ENtoSP_Good. In the experiments performed we used the training sets given below:

- RTE3_DS_ENtoSP,
- RTE3_DS_ENtoSP_Good+RegPairs, and
- RTE3_DS_ENtoSP_Good.

Additionally, we utilize the RTE3-4C, and RTE3-4C+RTE4-4C datasets.

We generated a feature vector for every *T-H* pair with both training and test sets. The feature vector is composed of the following eight components: F_{RES}, F_{LIN}, F_{JICO}, F_{PISE}, F_{WUPA}, F_{PA}, F_{LECH}, and SemSim. The achieved results are shown in Table 2 and Table 3.

Results reported in both tables show that we achieved the best performance, or nearly the best, with the dataset RTE3-4C+RTE4-4C in the majority of the cases.

It is interesting to note that our best result in the two-way task is obtained to predict the RTE3_TS_ENtoSP_Good test set, which is actually the realistic case, because this dataset contains only pairs validated by humans. On the other hand, the test set RTE3_TS_ENtoSP contains *BAD* pairs, and we obtained results comparables to those obtained with the previous case.

On the contrary, in the case of three-way task, the highest results are achieved considering RTE3_TS_ENtoSP as test set.

In both cases, the difference found when predicting RTE3_TS_ENtoSP and RTE3_TS_ENtoSP_Good is not statistical significant.

Surprisingly, the worse results in all the cases were obtained with the RTE3_DS_ENtoSP_Good as training set. This can be caused by the size of this dataset, which is composed by only 542 pairs.

As we previously note, the datasets RTE3-4C and RTE4-4C have been created for monolingual textual entailment, however the system is able to use these datasets because of our decoupled approach for CLTE. Thus, this result suggests that the corpus used on monolingual task improves the result of the CLTE system.

As a term of comparison, in the RTE3 Challenge [17] the average score achieved in the two-way task for the monolingual textual entailment was 62.37% of accuracy reached by the competing systems, which is 0.75% and 1.13% below our accuracy levels of 63.12% and 63.5% obtained with the SVM classifier and using the RTE3-4C+RTE4-4C and RTE3-4C+RTE4-4C+ RTE3_DS_ENtoSP as the training sets, but not resulting in a significant statistical difference.

In the RTE4 Challenge, the average score achieved in the three-way task was 50.65%, and thus our system outperforms on 9.63% when using SVM and RTE3-4C+RTE4-4C as training set, which is a significant statistical difference, although these sets are not actually comparable.

Although the elements belonging to the class *Bad* are included in RTE3_DS_ENtoSP, surprisingly, better performances are achieved in comparison with other data sets with neither *Regular* nor *Bad* pairs. The *T-H* pairs included in the set *Bad* have some syntax errors and, even more, are not understandable by the translators. However, many of the words "w" in the text *T* are also present in the hypothesis the *H* as "w", or are present as synonyms of "w", which increases the semantic correspondence between the *T-H* pair. This could be a reason for the increasing in efficiency when using RTE3_DS_ENtoSP as training set.

Interestingly, if we analyze only the size of data sets, we see that the larger the training set, the greater the efficiency gains. This highlights the need for larger datasets for the purpose of building more accurate models. It is also showed by the best accuracy that is found in our system when using the expanded dataset RTE3-4C+RTE4-4C.

TABLE II
ACCURACY OBTAINED CONSIDERING RTE3_TS_ENTOSP AND RTE3_TS_ENTOSP_GOOD AS TEST SET IN THE TWO-WAY TASK

Datasets	RTE3_TS_ENtoSP		RTE3_TS_ENtoSP_ Good	
Classifiers	2-way MLP Classifier	2-way SVM Classifier	2-way MLP Classifier	2-way SVM Classifier
RTE3_DS_ENtoSP	59.75	62.12	60.46	61.53
RTE3_DS_ENtoSP_Good +RegPairs	58.37	60.62	59.39	61.53
RTE3_DS_ENtoSP_Good	57.62	58.12	57.96	61.53
RTE3-4C+RTE4-4C	60.37	**63.12**	**63.32**	62.96
RTE3-4C	**62.62**	61.75	62.61	62.43
RTE3-4C+RTE4-4C+ RTE3_DS_ENtoSP	62.62	62.25	62.79	**63.50**

TABLE III
ACCURACY OBTAINED CONSIDERING RTE3_TS_ENTOSP AND RTE3_TS_ENTOSP_GOOD AS TEST SET IN THE THREE-WAY TASK

Datasets	RTE3_TS_ENTOSP		RTE3_TS_ENtoSP_ Good	
Classifiers	3-way MLP Classifier	3-way SVM Classifier	3-way MLP Classifier	3-way SVM Classifier
RTE3_DS_ENtoSP	57.96	58.31	57.96	**58.31**
RTE3_DS_ENtoSP_Good +RegPairs	58.49	**58.85**	58.14	56.35
RTE3_DS_ENtoSP_Good	54.75	56.62	55.09	55.28
RTE3-4C+RTE4-4C	**60.28**	58.14	58.32	58.14
RTE3-4C	58.75	57.50	58.32	58.14
RTE3-4C+RTE4-4C+ RTE3_DS_ENtoSP	59.87	57.50	**59.57**	58.14

V. Conclusion

From our experiments, we conclude that a promising algorithm to expand an RTE Corpus yielded significant statistical differences when predicting RTE test sets. We also show that although WordNet is not enough to build a competitive TE system, an average score could be reached or outperformed for the CLTE task.

As a further contribution, our experiments suggest that using the expanded method for the corpus can increase the accuracy of CLTE systems, in both two-way and three-way tasks. All results obtained in these tasks are comparable (or outperformed) with the average score of existing RTE systems. As additional result, we present the first CLTE corpus, and a procedure to create a corpus with the technique of human-aided machine translation, which also could be used to create a bilingual TE corpus in any language pairs. This corpus reaches an inter-annotator agreement corresponding to Kappa level of 0.68, regarded as substantial agreement.

Furthermore, the results obtained for the three-way task in CLTE outperforms the score of an average system by 9.63% accuracy when predicting the RTE3_TS_ENtoSP dataset.

Our future work will address the incorporation of additional knowledge resources and will incorporate additional lexical similarities features and semantic resources and assess the improvements they may yield. Finally, we aim at releasing additional CLTE corpus to the community in the future.

References

[1] L. Bentivogli, I. Dagan, H. Dang, D. Giampiccolo, and B. Magnini, "The Fifth PASCAL RTE Challenge," in *Proceedings of the Text Analysis Conference*, 2009.

[2] Y. Mehdad, M. Negri, and M. Federico, "Towards Cross-Lingual Textual entailment," in *Proceedings of the 11th NAACL HLT*, 2010.

[3] L. Bentivogli, P. Clark, I. Dagan, H. Dang, D. Giampiccolo, "The Sixth Pascal Recognizing Textual Entailment Challenge," in *Proceedings of Textual Analysis Conference*, NIST, Maryland USA, 2010.

[4] R. Richardson and A. Smeaton, "Using WordNet in a Knowledge-Based Approach to Information Retrieval," *Techn. Report Working Paper*: CA-0395, Dublin City University, Dublin, Ireland, 1995.

[5] J. Marlow, P. Clough, J. Recuero, and J. Artiles, "Exploring the Effects of Language Skills on Multilingual Web Search," in *Proceedings of the 30th European Conference on IR Research (ECIR'08)*, Glasgow, UK. LNCS, Volume 4956, Springer, Heidelberg, 2008, pp. 126−137.

[6] J. Lilleng and S. Tomassen, "Cross-lingual information retrieval by feature vectors", *NLDB 2007, LNCS*, pp. 229−239, 2007.

[7] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The Third PASCAL Recognizing Textual Entailment Challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic, 2007.

[8] J. Landis and G. Koch, "The measurements of observer agreement for categorical data," *Biometrics*, 33:159–174, 1997.

[9] J. Castillo, "Using Machine Translation to expand a Corpus in Textual Entailment," in *Proceedings of the 7th ICETAL*, Reykjavik, Iceland. LNCS, vol. 6233, Springer, Heidelberg, 2010, pp. 97–102.

[10] J. Castillo, "A Semantic Oriented Approach to Textual Entailment using WordNet-based Measures," in *Proceedings of the MICAI 2010*, Pachuca, Mexico, LNCS, vol. 6437, Springer, Heidelberg, 2010, pp. 44–55.

[11] P. Resnik, "Information Content to Evaluate Semantic Similarity in a Taxonomy," in *Proceedings of IJCAI 1995*, 1995, pp. 448−453.

[12] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of Conference on Machine Learning*, 1997, pp. 296–304.

[13] J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in *Proceedings of the ROCLING X*, 1997.

[14] G. Pirrò and N. Seco, "Design, Implementation and Evaluation of a New Similarity Metric Combining Feature and Intrinsic Information Content," *ODBASE 2008*, Springer LNCS, 2008.

[15] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proceedings of the 32nd ACL*, 1994.

[16] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, MIT Press, pp. 265−283, 1998.

[17] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The Third PASCAL Recognizing Textual Entailment Challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic, 2007.

[18] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts," in *Proceedings of the AAAI-04*, 2004.

[19] C. Quirk, C. Brockett, and W. Dolan, "Monolingual Machine Translation for Paraphrase Generation," in *Proceedings of the ACL-HLT*, 2004.

Identifying the User's Intentions: Basic Illocutions in Modern Greek

Maria Chondrogianni

Abstract—**This paper presents a comprehensive classification of basic illocutions in Modern Greek, extracted following the linguistic choices speakers make when they formulate an utterance, provided such choices form part of a language's grammar. Our approach lies on the interface between Morphosyntax, Pragmatics and Phonology and allows for basic illocutions to be established depending on the particular verb mood, particle, number, person, aspect and segmental marker, as well as the prosodic contour used when an utterance is realized. Our results show that Indicative uses, for example, are mostly associated with propositional illocutions, consisting of declarative uses, including assertions, miratives, and assertions in disguise; interrogative uses, including polar and content interrogatives; and behavioral illocutions i.e. exhortations (expressed in first person plural only). Secondary sentence types, (involving additional segmental marking) include requests for confirmation, wondering, expression of uncertainty and proffer. In this paper we discuss propositional uses only. Such a theoretical approach can have a direct impact on applications involving Human-Computer Interaction, including intention-based dialogue systems' modeling, natural language interfaces to Data Bases and Intelligent Agents as well as Belief, Desire and Intention systems, which require the computer to be able to interpret what a user's objective (intention) is, so that the users' needs can be best served.**

Index Terms—**Pragmatics, basic illocutions, Modern Greek.**

I. Introduction

THE ability of machines to communicate with humans (or even to provide content in a co-operative way), in a manner that reflects or mimics human communication has been at the core of AI research for some decades. As natural languages are viewed as the input of choice for a series of soon to appear applications (including user interfaces to Data Bases, e-commerce systems, and gaming applications among others) the need to improve the way computers communicate with humans is ever more pertinent. Fundamental to this quest is to come up with techniques which will allow for the user's goals to be identified, based on greater interaction and collaboration between theoretical linguists and natural language engineers.

In the theoretical linguistics-focused research below, we take the position that, whether for dialogue modeling applications or natural language user interfaces, the user's intentions can be identified based on a Pragmatics analysis of the linguistic input provided by the user themselves. Earlier attempts, where illocution was considered, can be seen in Allen [1] or the DDML team's work [11], who married XML with Pragmatics and provided the opportunity for personalized human-computer interaction. Our analysis can form the basis for a computer implementation of users' intentions. The linguistic choices users make to express/phrase their query, for example, and the particular verb forms and particles they use are crucial in identifying their intention.

The focus of our research is on the way illocution is codified in a Speaker's message, through the grammatical/phonological choices a Speaker makes. The natural language of application for our research is Modern Greek (MG), a language with rich morphology. The outcome of our research consists of a comprehensive classification of the basic illocutions of MG, based on markers that have an illocutionary impact, such as the verb mood, the negation, the clitic placement, the intonation patterns and any additional segmental strategies used by MG speakers.

In our approach we share a similar perspective with Steuten [10], who undertook a linguistic analysis of business conversations; we share her fundamental view that a conversation consists of a series of communicative acts [7], expressed through basic illocutions, connected with each other, 'with the purpose of defining a goal and reaching that goal'. We are interested in the basic illocutions, which form part of a grammatical system that a speaker (and their addressee) have at their disposal, which will allow them to reach their goal. We consider phonology as being part of a language's grammatical system, hence the prosodic contour (intonation patterns) described below is crucial in identifying basic illocutions.

II. Criteria for the Identification of Basic Illocutions: Intonation Patterns

Crucial to the identification of MG basic illocutions is the specification of intonation patterns that speakers adopt [2] in specific instances of utterances at *Utterance* level (as per the layered structure of the FDG Phonological component [8]. We distinguish among 5 MG intonation patterns [4], briefly described below.

A. Intonation Pattern 1 (INT1)

The characteristic of this pattern is its broad focus and a high level of the accented syllable. Its Fundamental Frequency (FO) includes a heightening of the pitch starting at the first accented syllable, followed by a small dip and a fall for the last word. The boundary is low. Schematically, the tonal structure of our INT1 pattern is illustrated in Fig. 1 below. The nucleus

Manuscript received June 25, 2011. Manuscript accepted for publication August 20, 2011.

Maria Chondrogianni is with the School of Electronics and Computer Science, University of Westminster, 115 New Cavendish Street, London, W1W 6UW, UK (e-mail: M.N.Chondrogianni@westminster.ac.uk).

might create variations on this pattern; in some cases it can be used interchangeably with INT2, when focality affects the way an utterance is expressed. INT1 characterizes broad focus.

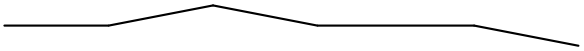


Fig. 1: Intonation Pattern 1 (INT1).

B. Intonation Pattern 2 (INT2)

INT2 starts with a plateau followed by a rise on the nucleus, followed by a fall from the post-nuclear syllable onwards. Schematically, INT2 tonal structure is illustrated in Fig. 2 below. It characterizes narrow focus.

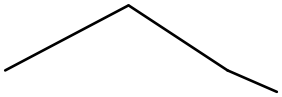


Fig. 2: Intonation Pattern 2 (INT2).

C. Intonation Pattern 3 (INT3)

This is the typical pattern for content interrogatives. It starts high, with the first accented syllable and it starts dropping immediately after it, with a potential slight rise at the end. Although typical questions are expected to finish with rising intonation, the question word here provides the key to the addressee on how the utterance is to be interpreted, hence a variation with a slightly rising, level or slightly falling end syllable is not unexpected. INT3 can schematically be illustrated in Fig. 3 below.

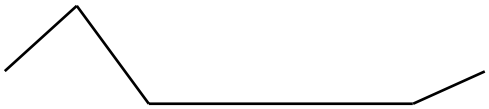


Fig. 3: Intonation Pattern 3 (INT3).

D. Intonation Pattern 4 (INT4)

This is the typical polar question intonation pattern. The pick is on the last stressed syllable of the final word. Following a gradual fall, we have a low plateau followed by a rise (with a possible slight fall at the end). The boundary is Rise-fall. Schematically we present its tonal structure in Fig. 4 below.

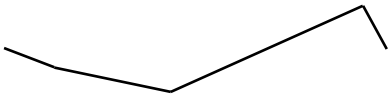


Fig. 4: Intonation Pattern 4 (INT4).

E. Intonation Pattern 5 (INT5)

This pattern starts with a small fall, followed by a rise (and possibly a high plateau), and followed by a fall (and a potential small rise at the end). The boundary is low-high. It is the typical prosodic contour for curses. Schematically we are illustrating INT5 in Figure 5 below.

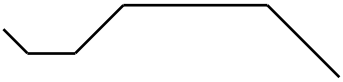


Fig. 5: Intonation Pattern 5 (INT5).

III. Basic Illocutions of Modern Greek

Each illocutionary function included below is described in terms of:

- The grammatical mood used; in propositional uses, we encounter the Indicative, optionally introduced by the future marker *θα* (tha); the Subjunctive, introduced by the subjunctive particle *να* (na); and the Hortative, introduced by the hortative particle *as* (as); in behavioral uses, which are not covered in the present paper, we encounter the Indicative, the Subjunctive, the Imperative, the Hortative and the Prohibitive verb moods.
- The prosodic contour it is expressed with; the five intonation patterns identified in section 2 are used as part of each illocutions' characteristics.
- The associated negation i.e. *δε(ν)* ('de(n)') for Indicative and *μη(ν)* ('mi(n)') for Subjunctive and Hortative.
- Potential segmental markers which provide cues to the addressee on how a certain utterance is to be interpreted such as *ίσως* ('isos') for uncertainty and *άραγε* ('araye') for wondering.
- Grammatical tense restrictions, for example the choice of tense in wishes, which characterizes the fufillability of a wish.
- Aspectual restrictions (where appropriate); for example, the sole possibility of imperfective aspect with past in wishes.

In addition, where appropriate, we refer to number and person restrictions and to frequent lexical additions. All basic illocutions are associated with their relevant intonation patterns, as distinguished in section 2.

A. Propositional uses

Following the basic illocution classification proposal in [9], we present the MG propositional illocutions, consisting of assertive uses, mirative uses, wishes and curses, expressions of wondering, uncertainty and estimating. The verb forms used for propositional uses include the Indicative, the Subjunctive, and the Hortative moods.

B. Assertions

Assertions are signaled by the use of the Indicative [3],[6]. Although we demonstrate that there is no one-to-one relationship between the Indicative mood and the Declarative sentence type, since Indicative presents a rich variety of uses, we maintain that the reverse presents a one-to-one relationship:

the declarative sentence type can only be expressed in Indicative. Intonation Patterns INT1 and INT2 apply (depending on the broadness or narrowness of focus).

Type	Propositional
Function	Assertion
Grammatical Mood	Indicative (optional particle *θα*, optional negation *δε(ν)*)
Tense	Present/Past/Future
Aspect	Perfective and Imperfective
Person	Any
Number	Singular or Plural
Intonation Pattern	INT1/INT2

C. Assertions in disguise-contrastive statements

The unique character of assertions in disguise-contrastive statements is based on the use of the 1st person as well as the fact that a tag question is used as a compulsory element of the utterance's structure; alternatively, this illocution is marked by the compulsory use of the segmental marker *μήπως* ('mipos', perhaps), usually followed by the Indicative negation *δε(ν)*.

Type	Propositional
Function	Assertions in disguise- contrastive statements
Grammatical Mood	Indicative (optional particle *θα*, optional negation *δε(ν)*)
Tense	Present/Past/Future
Aspect	Perfective and Imperfective
Person	1st
Number	Singular or Plural
Segmental Marker	Tag or μήπως (usually followed by negation)
Intonation Pattern	INT2 + INT4 with tag INT4 with μήπως

D. Requests for confirmation

Requests for confirmation also involve the compulsory use of a tag; through such utterances the Speaker seeks to confirm the truth of the State of Affairs described. Requests for confirmation are expressed in indicative, with the optional use of particle *θα* and negation *δε(ν)*, usually in the 2nd person (3rd person uses are also possible), using INT2 for the assertion and INT4 for the tag.

Type	Propositional
Function	Request for Confirmation
Grammatical Mood	Indicative (optional particle *θα*, optional negation *δε(ν)*, use of tag question)
Tense	Present/Past/Future
Aspect	Perfective and Imperfective
Person	Usually 2nd, 3rd possible
Number	Singular or Plural
Intonation Pattern	INT2 + INT4

E. Miratives

Mirative uses are a very interesting category of basic illocution, in that the Speaker expresses a qualitative view on a State of Affairs, and the positivity or negativity of their stance is formally expressed through the use of a particular grammatical element (verb mood). Mirative uses of approval are expressed in Indicative, whilst those of disapproval are expressed in Subjunctive [4].

Type	Propositional
Function	Mirative uses
Grammatical Mood	-Indicative (approval, optional particle *θα*, optional negation *δε(ν)*) -Subjunctive (disapproval, particle *να*, optional negation *μη(ν)*)
Tense	Present (also Past is possible but unusual; Future is common in the Indicative)
Aspect	Perfective/Imperfective
Person	2nd /3rd (1st possible)
Number	Singular or Plural
Intonation Pattern	INT3

F. Wishes

MG Wishes are expressed either in Subjunctive or in Hortative [5]. A Subjunctive use is introduced by the particle *να*, while a Hortative one by the particle *ας*. In Subjunctive wishes are potentially preceded by the segmental marker *μακάρι* ('makari'); the negation *μη(ν)* might optionally apply to either uses. Any person and number might be used, while aspectual and tense (present or past) differences affect a wish's fulfillability or unfulfillability. Intonation pattern INT1 and INT2 apply.

Type	Propositional
Function	Wishes
Grammatical Mood	-Subjunctive (particle *να*, optional negation *μη(ν)*, optional segmental marker *μακάρι*) -Hortative (particle *ας*, optional negation *μη(ν)*)
Tense	Present (fulfillable) Past (unfulfillable)
Aspect	Imperfective Present, Past) Perfective (Present only)
Person	1st, 2nd and 3rd
Number	Singular or Plural
Intonation Pattern	INT1 (INT2 when introduced by *μακάρι*)

G. Curses

Curses are expressed in the Subjunctive. They are introduced by the Subjunctive particle *να*; the optional Subjunctive negation *μη(ν)* might be used, while a speaker might opt to use the segmental marker *που* at the beginning of a curse. Present tense with perfect aspect characterizes their most common uses, which are expressed in the 2nd or 3rd person. In the 1st person, they are similar to an oath. They are expressed using a dedicated intonation pattern, INT5.

Type	Propositional
Function	Curses (Negative Wishes)
Grammatical Mood	Subjunctive(particle *να*,optional negation *μη(ν)*, optional segmental marker *που*).
Tense	Present (fulfillable)
Aspect	Perfective (imperfective not excluded, But uncommon)
Person	2nd /3rd (1st not excluded)
Number	Singular or Plural
Intonation Pattern	INT5

H. Wondering

MG wondering is expressed in the Indicative or in the Subjunctive. In the Indicative the use of the wondering particle *άραγε* (araye) is compulsory. The wondering particle's placement in the clause is not fixed i.e. it might precede or it might follow the verb. Wondering in Subjunctive can be expressed without the use of a specific segmental marker (other than the subjunctive marker *να*); or by the combination of *άραγε* + *να* (which strengthens the wondering illocution). Here again *άραγε* might precede the subjunctive marker, or it might follow the verb.

Type	Propositional
Function	Wondering
Grammatical Mood	-Indicative (segmental marker *άραγε*, optional negation *δε(ν),* optional particle *θα*) -Subjunctive (particle *να*, or combination of *άραγε* and *να*, optional negation *μη(ν), question word with INT3*)
Tense	Present/Past (also Future in Indicative)
Aspect	Perfective/Imperfective
Person	3rd
Number	Singular or Plural
Intonation Pattern	INT4 (also INT3 in Subjunctive)

I. Uncertainty

Uncertainty is a built-in characteristic of MG Subjunctive, similar to other languages. In many ways, wondering in Subjunctive expresses the Speaker's uncertainty about the validity of the described State of Affairs; such an uncertainty forms the impetus behind the Speaker's wondering. In addition to pragmatically relatively ambiguous uses (i.e. implying wondering as well as uncertainty), MG uncertainty is expressed through the use of particle *ίσως* ('isos', maybe), which might be followed by Indicative or by Subjunctive (the latter use expresses reinforced uncertainty). *Ισως* is most likely to be placed ahead of the Indicative verb, although it is not uncommon for it to follow the verb. Its position in a Subjunctive utterance is fixed, always preceding the subjunctive marker.

Type	Propositional
Function	Expression of uncertainty
Grammatical Mood	-Indicative (uncertainty particle *ίσως*, optional particle *θα*, optional negation *δε(ν)*, usually precedes the verb but position after the verb acceptable) -Subjunctive (particle *να*, uncertainty particle *ίσως*, optional negation *μη(ν)*)
Tense	Present/Past (Future in indicative acceptable by some speakers)
Aspect	Perfective/ Imperfective
Person	Any
Number	Singular or Plural
Intonation Pattern	INT1 (Subjunctive) INT2 (Indicative)

J. Polar and Content Interrogatives

MG Questions are expressed in Indicative. Polar interrogatives are differentiated by assertions because of the combination of the Indicative mood with intonation pattern INT4 and the expectation that the addressee will confirm or reject the validity of the proposition through a positive or a negative response. A response denoting consent to a polar interrogative would be inappropriate.

In content interrogatives a question word is involved (such as who, when, where among others) to identify the particular information the speaker is seeking. The question word might be introducing the content interrogative, or might be placed in different positions in the utterance depending on focality, which affects their intonation pattern; more than one element of the utterance can be questioned. INT3 applies to content interrogatives. The speaker's expectation is that the addressee will provide information on the slot denoted by the question word.

Type	Propositional
Function	Interrogatives
Grammatical Mood	Indicative (optional particle *θα*, optional negation *δε(ν)*) Question word(s)
Tense	Present/Past/Future
Aspect	Perfective and Imperfective
Person	Any
Number	Singular or Plural
Intonation Pattern	INT3 (content interrogatives); INT4 (polar interrogatives)

IV. Conclusions

We described above an original classification of the MG propositional basic illocutions, based on the functions' formal characteristics, which form part of the grammatical system and we placed the focus on function, rather than form.

All indicative uses are marked by the optional particle *θα* and the optional negation *δε(ν)*. Assertions are distinguished by the use of the Indicative and the use of intonation patterns INT1/INT2 (based on whether a broad or narrow focus applies). Mirative uses of approval are distinguished by the use of the Indicative, the use of intonation pattern INT3, and the lack of a question word related response from the addressee (when compared with the content interrogatives, also uttered in INT3). Content interrogatives are distinguished by the use of

Indicative mood, a question word (such as who, what, when where, how), the use of intonation pattern INT3 and the expectation that the addressee's response will provide information on the questioned element of the utterance. Polar interrogatives are distinguished by the use of Indicative mood, the intonation pattern INT4, and the expectation that a positive or negative response (or a response expressing a degree of certainty or uncertainty) will be provided by the addressee. Mitigated questions/proffer are expressed in Indicative, introduced by the segmental marker *μήπως,* expressed in INT4, in the 2nd person. Wondering uses are distinguished by the use of Indicative, the segmental marker *άραγε*, and the most common use of 3rd person (also the use of 1st person in deliberative questions). Assertions in disguise-contrastive statements are expressed in Indicative, they include either a compulsory tag (when their intonation involves intonation patterns INT2 for the assertive part and INT4 for the tag) or by *μήπως,* in the 1st person. When in the second or third person (excluding *μήπως* uses), the use expresses a request for confirmation.

Subjunctive propositional uses are marked by the Subjunctive particle *να* and the optional negation *μη(ν)*. Wishes are marked by the use of Subjunctive, the optional use of the segmental marker *μακάρι* and the intonation pattern INT1. Curses are marked by the distinct intonation pattern INT5 and the optional use of the segmental marker *που*. Uncertainty in Subjunctive is marked by the segmental marker *ίσως* and the intonation pattern INT1. Wondering uses in Subjunctive are optionally introduced by the segmental marker *άραγε*, marked by intonation INT4 and the use of 3rd person; 1st person deliberative uses require the compulsory presence of *άραγε*. Mirative uses (of disapproval) are marked by intonation. Hortative wishes are marked by the Hortative particle *άς* and intonation INT1/INT2; they exclude 1st person plural uses, which characterize expressions of exhortation.

REFERENCES

[1] J. Allan, "Recognizing intention from natural language utterances," in Brady, M. Berwick, R. (eds.) *Computational Models of Discourse*, Cambridge MA: MIT Press, 1983, pp.107–186.

[2] A. Arvaniti, & Baltazani, M., "Intonation analysis and prosodic annotation of Greek spoken corpora," in Sun-Ah Jun (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford: Oxford University Press, 2005, pp. 84–117.

[3] M. Chondrogianni, "Basic illocutions of the MG Indicative," in *10th International Conference on Greek Linguistics*, Komotini, Greece, (forthcoming[a]) September 2011.

[4] M. Chondrogianni, "Basic Illocutions of the MG Subjunctive (to appear in the Selected papers volume of ISTAL 20)," in *20th International Symposium on Theoretical and Applied Linguistics*, Thessaloniki, Greece, April 2011.

[5] M. Chondrogianni, "The Pragmatics of Prohibitive and Hortative in MG," in Kitis E., Lavidas N., Topintzi N. & Tsangalidis T. (eds.) *Selected papers from the 19th International Symposium on Theoretical and Applied Linguistics* (19 ISTAL, April 2009), Thessaloniki: Monochromia, 2011, pp. 135–142.

[6] M. Chondrogianni, "The Indicative in Modern Greek," in Tsangalidis A, (ed.) *Selected papers from the 18th International Symposium on Theoretical and Applied Linguistics* (18th ISTAL, May 2007), Thessaloniki: Monochromia, 2009, pp. 123–130.

[7] J. Habermas, *The theory of communicative action, London,* Beacon Press, 1981.

[8] K. Hengeveld and J. L. Mackenzie, *Functional Discourse Grammar: A typologically-based theory of language structure*, Oxford: Oxford University Press, 2008.

[9] K. Hengeveld, E. Nazareth Bechara, R. Gomes Camacho A. Regina Guerra, T. Peres de Oliveira, E. Penhavel, Goreti E. Pezatti, L. Santana, E. R. F. de Souza, & M. L. Teixeira, "Basic illocutions in the native languages of Brazil," in Mattos Dall'Aglio M. Hattnher, & K. Hengeveld, (eds) *Advances in Functional Discourse Grammar.* Special issue of Alfa-Revista de Lingüística 51 (2) 73–90, 2007.

[10] A.A.G. Steuten, R.P. van de Riet & Dietz, J.L.G., "Linguistically based conceptual modeling of business communication," *Data Knowledge Engineering* 35 (2) 121–136, 2000.

[11] W. Zadrozny, M. Budzikowska, J. Chai, N Kambhatla, S. Levesque, & N. Nicolov, "Natural Language Dialogue for Personalized Interaction", *Communications of the ACM* (CACM) 43 (8) 116–120, 2000.

Inference of Fine-grained Attributes of Bengali Corpus for Stylometry Detection

Tanmoy Chakraborty and Sivaji Bandyopadhyay

Abstract—**Stylometry, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and belongs to the core task of Text categorization that involves authorship identification, plagiarism detection, forensic investigation, computer security, copyright and estate disputes etc. In this work, we present a strategy for stylometry detection of documents written in Bengali. We adopt a set of fine-grained attribute features with a set of lexical markers for the analysis of the text and use three semi-supervised measures for making decisions. Finally, a majority voting approach has been taken for final classification. The system is fully automatic and language-independent. Evaluation results of our attempt for Bengali author's stylometry detection show reasonably promising accuracy in comparison to the baseline model.**

Index terms—**Stylometry, stylistic markers, cosine-similarity, chi-square measure, Euclidean distance.**

I. Introduction

STYLOMETRY is an approach that analyses text in text mining e.g. novels, stories, dramas written by authors, trying to measure the author's style, rhythm of his pen, subjection of his desire, prosody of his mind by choosing some attributes that are consistent throughout his writing and play the linguistic fingerprint of that author. In other words, stylometry is the application of the study of linguistic style, usually with reference to written text that concerns the way of writing rather than its contents. Computational Stylometry is focused on subconscious elements of style less easy to imitate or falsify.

Stylistic analysis that has been done by Croft [2] claimed that for a given author, the habits "*of style*" are not affected "*by passage of time, change of subject matter or literary form. They are thus stable within an author's writing, but they have been found to vary from one author to another*" [8]. However, stylometric authorship attribution can be considered as a typical clustering, classification and association rule problem, where a set of documents with known authorships are used for training and the aim is to automatically determine the corresponding author of an anonymous text, but the way of selecting the appropriate features is not focused in that sense and vary from one research to other.

Most of the authorship identification studies are better at dealing with some closed questions like (i) who wrote this, A or B, (ii) if A wrote these, did he also writes this, (iii) how likely is it that A wrote this etc. The main target in this study is to build a decision making system that enables users to predict and to choose the right author from an anonymous author's novel under consideration, by choosing various lexical, syntactic, analytical features known as ***stylistic markers***. The system uses three semi-supervised, reference based measurements (Cosine-similarity, Chi-square measurement and Euclidean distance) which behave as an expert opinion to map the testing documents to the appropriate authors. Without focusing much on the distributional lexical measures like vocabulary richness or frequency of individual word counts, we mainly focus on some low-level measures (sentence count, word count, punctuation count, length of words and sentences etc.), phrase level measures (noun chunk, verb chunk, etc.) and context level measures (number of dialog, length of dialog, sentence structure analysis etc.). Additionally, we propose a baseline system for Bengali stylometry analysis using ***vocabulary richness function***. The present attempt basically deals with the microscopic observation for the stylistic behaviours of the articles written by the famous novel laureate Rabindranath Tagore long years back and tries to disambiguate them from the anonymous articles written by some other authors in that period.

The paper is organized as follows. In Section 2, related researches on stylometry and authorship identification in other language and their approaches are described. In Section 3, our approach is detailed along with the extracted features and classification models used in this experiment. The experimental results compared to the baseline system are described in Section 4. The experimental results are analysed in Section 5 and the conclusions are drawn in Section 6.

II. Related Works

Stylometry, which may be considered as an investigation of "Who was behind the keyboard when the document was produced?" or "Did Mr. X wrote the document or not?" is a long term study mainly in forensic investigation department that started from late Nineties. In the past, where stylometry emphasized the rarest or most striking elements of a text, contemporary techniques can isolate identifying patterns even in common parts of speech. The pioneering study on authorship attributes identification using word-length histograms appeared at the very end of nineteen century [6]. After that, a number of studies based on content analysis [5], computational stylistic approach [4], exponential gradient learn algorithm [7], Winnow regularized algorithm [9] , Support Vector Machine based approach [3] etc. have been

Manuscript received November 7, 2010. Manuscript accepted for publication February 6, 2011.

The authors are with Department of Computer Science and Engineering, Jadavpur University, Kolkata, India (e-mail: its_tanmoy@yahoo.co.in, sivaji_cse_ju@yahoo.com).

proposed for various languages like English and Portuguese. Recently, research has started to focus on authorship attribution on larger sets of authors: 8 [11], 20 [12], 114 [13], or up to thousands of authors [14]. The use of computers regarding the extraction of stylometrics has been limited to auxiliary tools (i.e. simple program for counting user-defined features fast and reliably). Hence, authorship attribution studies so far may be looked like *computer-assisted*, not *compute-based*. As a beginning of Indian language stylometry analysis, our research does not consider any manual intervention for extracting features (like identification of some high frequent start-up words), moreover we have dealt with a number of large-size non-homogeneous texts since they are composed of dialogues, narrative parts etc. and try to build a language and text-length independent system for attribute analysis.

III. Our Approach

The methodology used in this work generally depends on the combination of 76 fine-grained style-markers for feature engineering and three semi-supervised approaches for decision making. As an initial attempt, we have decided to work with the simple approach like statistical measurement, analyze the drawbacks and further go beyond for working with other machine learning or hybrid approaches. Furthermore, the reasons for not attempting with the methods described in the related work section are as follows: the content analysis is one of the earliest types of computations, also for exponential and Winnow algorithms as both are purely mathematical models and the SVM based method has a strong affinity to the language for which the system is designed. Currently, authorship attribute studies are dominated by the use of lexical measures. In a review paper [1], the author asserted that:

" ..... yet, to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than based on lexical items."

For this reason, in order to set a baseline for the evaluation of the proposed method, we have decided to implement a lexical based approach called *vocabulary richness*. Detailed discussion about the baseline system and our approach are mentioned in Section 4.

A. Proposed Methodology Design

As mentioned, the proposed stylistic markers used in this study take full advantage of the analysis of the distributed contextual clues as well as full analysis by natural language processing tools. The system architecture of the proposed stylometry detection system is shown in Figure 1. In this section, we first describe brief properties of different components of the system architecture and then the set of stylistic features is analytically presented. Finally the classification methods are elaborated with brief description of their functionalities.

1) Textual Analysis

Basic pre-processing before actual textual analysis has been done so that stylistic markers are clearly viewed for further analysis by the system. Token-level markers discussed in the next Section, are extracted from the pre-processed corpus. Then parsing using Shallow parser[1] has been done to separate the sentence and the chunk boundaries and parts-of-speech. From this parsed text, chunk-level and context-level markers are identified.

2) Stylistic Features Extraction

Stylistic features have been proposed as more reliable style markers than for example, word-level features since they are not under the conscious control of the author. To allow the selection of the linguistic features rather than n-gram terms, robust and accurate text analysis tools such as lemmatizers, part-of-speech (POS) taggers, chunkers etc. are needed. We have used the Shallow parser, which gives a parsed output of a raw input corpus. It tokenizes the input, performs a part-of-speech analysis, looks for chunks and inflections and a number of other grammatical relations. The stylistic markers which have been selected in this experiment are coarsely classified into three categories and discussed in the Table I.

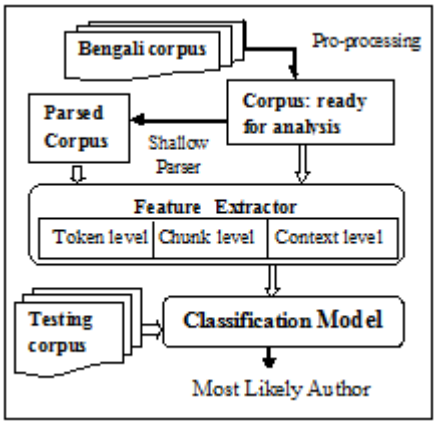


Fig. 1. Architecture of the stylometry detection system.

Sentences are detected using the sentence boundary markers mainly *'dari'* or *'viram'* ('।'), question marks ('?') or exclamation notation ('!') in Bengali. Sentence length and word count are the traditional and well-defined measures in authorship attribute studies and punctuation count is another interesting characteristics of the personal style of a writer. Chunk or phrase level markers are indications of various stylistic aspects, e.g., syntactic complexity, formality etc. Out of all detected chunk sets, mainly nine chunk types have been considered in this experiment. They are noun chunk (NP), verb-finite chunk (VGF), verb-non-finite chunk (VGNF), gerunds (VGNN), adjective chunk (JJP), adverb chunk (RBP), conjunct phrase (CCP), chunk fragment (FRAGP) and others (OTHERS). Shallow parser identifies 25 Part Of Speech (POS) categories. Among them, 24 POSs have been taken into consideration except UNK. Words tagged with UNK are unknown words and are verified by Bengali monolingual dictionary. Since Shallow parser is an automated text-processing tool, the style markers of the above levels are measured approximately. Depending on the complexity of the text, the provided measures may vary from real values which

[1] http://ltrc.iiit.ac.in/analyzer/bengali

can only be measured using manual intervention. Making the system fully automated, the system performance depends on the performance of the parser. As we can see in the Table I that each marker is defined as a percentage measure of the ratio of two relevant measures, this approach was followed in order to work with text-length independent style markers as possible. However, it is worth noting that we do not claim that the proposed set of 76 markers is the final one. It could be possible to split them into more fine-grained measures e.g. F21 can be split into separate measures i.e. individual occurrence of the punctuation symbols (comma per word, colon per word, dari per word etc.). Here, our goal is to make an attempt towards the investigation of Bengali author's writing style and to prove that an appropriately defined set of such style markers performs better than the traditional lexical based approaches.

TABLE I
FINE-GRAINED STYLISTIC FEATURES

Coarse-grained Classification	Stylistic Markers	Description	Total
Token-level	F1 to F10	Word length (1 to 9 and above) in %	10
	F11 to F20	Words per sentence (0-10, 10-20 and so on, up to 80-90 and above) in %	10
	F21	Punctuations per word in %	1
	F22 to F31	Individual punctuations in % (10 punctuations)	10
Chunk-level	F32 to F40	Detected NP, VGF, VGNF, VGNN, JJP, RBP, CCP, FRAGP, OTHER out of total chunks in %	9
	F41 to F49	Average words included in all above mentioned chunks in %	9
	F50 to F73	Individual percentage of detected POS (24) by Shallow parser	24
Context-level	F74	Average words per dialog in %	1
	F75	Words not included in the dictionary including Named-Entities in %	1
	F76	Hapax-legomena count out of all words in %	1

3) Classification Model

A number of discriminative models based on statistical and machine learning measures, such as Bayesian Network, decision trees, neural networks, support vector machines, K-nearest neighbour approach etc. are available for text categorization. In this experiment, three semi-supervised, reference-based classification models have been used: (1) Cosine-similarity measurement, (2) Chi-square measure and (3) Euclidean distance. These are briefly discussed below.

Cosine-similarity measurement: Cosine-similarity is a measure of similarity between two vectors of ***n*** dimensions by finding the cosine of the angle between them, often used to compare documents in text mining. Given two vectors of attributes, ***R*** and ***T***, the cosine similarity, θ is represented using a dot product and magnitude as:

$$Similarity = \cos(\theta) = \frac{R.T}{|R|.|T|} = \frac{\sum_{i=1}^{n} r_i.t_i}{\sqrt{\sum_{i=1}^{n} r_i^2} * \sqrt{\sum_{i=1}^{n} t_i^2}}$$

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity. Here, n is the number of features (i.e., 76) that act as dimensions of the vectors and r_i and t_i are the features of reference and test vectors respectively.

Chi-square measure*:** Chi-square is a statistical test commonly used to compare observed data with the expected data according to a specific hypothesis. That is, chi-square (χ^2) is the sum of the squared differences between observed (O***) and the expected (***E***) data (or the deviation, ***d***), divided by the expected data in all possible categories.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Here, the mean of each cluster is used as the observation data for that cluster and used as reference ***O***. ***n*** is the number of features and O_i is the observed value of the i^{th} feature. The Chi Square test gives a value of χ^2 that can be converted to Chi Square (c^2) using chi-square table which is an n×n matrix with row representing the degree of freedom (i.e., difference between the number of rows and columns of the contingency matrix) and column representing the probability we expect. This can be used to determine whether there is a significant difference from the null hypothesis or whether the results support the null hypothesis. After comparing the chi-squared value in the cell with our calculated χ^2 value, if the χ^2 value is greater than the 0.05, 0.01 or 0.001 column, then the goodness-of-fit null hypothesis can be rejected, otherwise accepted.

Euclidean distance: The Euclidean distance between two points, ***p*** and ***q*** is the length of the line segment. In Cartesian coordinates, if $\boldsymbol{p} = (p_1, p_2 \ldots p_n)$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$ are two points in Euclidean n-space, then the distance from ***p*** to ***q*** is given by:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

where, ***n*** is the number of features or dimension of a point, ***p*** is the reference point (i.e. mean vector) of each cluster and ***q*** is the testing vector. For every test vector, three distances from three reference points have been calculated and smallest distance defines the probable cluster.

IV. EXPERIMENT

A. Corpus

Resource acquisition is one of the most important challenges to work with resource constrained languages like Bengali. The system has used thirty stories in Bengali written by the noted Indian Nobel laureate Rabindranath Tagore[2]. Among them, we have selected twenty stories for training purpose and rest for testing. We choose this domain for two reasons: firstly, in such writings the idiosyncratic style of the author is not likely to be overshadowed by the characteristics of the corresponding text genre; secondly, an earlier work [10] has worked on the corpus of Rabindranath Tagore to explore some of the stylistic behaviours of his writing. To differentiate them from other author's articles, we have selected 30 articles from author A and 30 articles of other authors[3]. In this way, we have three clustered set of documents identified as articles of Author R (Tagore's articles), Author A and others (O). This paper focuses on two topics: (i) the effort of earlier works on feature selection and learning and (ii) the effort of limited data in authorship detection.

B. Baseline System

In order to set up a baseline system, we proposed traditional lexical based methodology called ***vocabulary richness***. Among the various measures like Yule's K measure, Honore's R measure, we have taken most typical one as the type-token ratio ***(V/N)*** where ***V*** is the size of the vocabulary of the sample text and ***N*** is the number of tokens which forms the simple text. We have gathered dimensional features of the articles of each cluster and averaged them to make a mean vector for every cluster.

TABLE II
CONFUSION MATRIX OF THE BASELINE SYSTEM

	R	A	O	e (Error)
R	6	0	4	0.40
A	7	2	1	0.80
O	5	2	3	0.70
Average error				0.63

So these three mean vectors indicate the references of three clusters respectively. Now, for every testing document, similar features have been extracted and a test vector has been developed. Now, using Nearest-neighbour algorithm, we have tried to identify the author of the test documents. The results of the baseline system are shown using confusion matrix in Table II. Each row shows classification of the ten texts of the corresponding authors. The diagonal contains the correct classification. The baseline system achieves 37% average accuracy. Approximately 60% of average accuracy error (for author A and O) is due to the wrong identification of the author as Author R.

[2] http://www.rabindra-rachanabali.nltr.org
[3] http://banglalibrary.evergreenbangla.com

C. Performance of Our System

We have discussed earlier that our classification model is based on three statistical techniques. A voting approach combining the decision of the three models for each test document have also been measured for expecting better results.

The confusion matrix in Table III and IV shows that Chi-square measure has relatively less error (46%) rate compared to other measures. A majority voting technique has an accuracy rate of 63% which is relatively better than others. In the case when the three statistical techniques produce different results, the result of Chi-square measure has been taken as correct result because it has given more accuracy compared to the others when measured individually.

TABLE III
CONFUSION MATRIX OF OUR SYSTEM (PART I)

	Cosine-similarity				Chi-square measure			
	R	**A**	**O**	**e**	**R**	**A**	**O**	**e**
R	5	2	3	0.5	7	3	0	0.3
A	3	6	1	0.4	5	4	1	0.6
O	4	1	5	0.5	4	1	5	0.5
Average error				0.46	Average error			0.46

TABLE IV
CONFUSION MATRIX OF OUR SYSTEM (PART II)

	Euclidean distance				Combined voting			
	R	**A**	**O**	**e**	**R**	**A**	**O**	**e**
R	6	2	2	0.4	8	2	0	0.2
A	4	4	2	0.6	4	5	1	0.5
O	3	2	5	0.5	2	2	6	0.4
Average error				0.5	Average error			0.37

V. DISCUSSION

Form the experimental results, it is clear that statistical approaches show nearly similar performance and accuracies of all of them are around 50%. Also the major sources of the errors are for the inappropriate identification of author as Author R. From the figure 2, we can see that the system looks little bit biased towards the identification of Rabindranath Tagore as author of the test documents. In all cases, the bar graphs for Author R are higher than others. The reason behind this is the acquisition of resources. Developing appropriate corpus for this study is itself a separate research area and takes huge amount of time. Furthermore, the collected articles from other authors are heterogeneous and not domain constrained.

Our studies in future will be planned to focus on the identification of the unpublished articles of Rabindranath Tagore. For this, more microscopic observation in various fields of his writings will be needed. Here we only try our experiments on the stories of the writer. The success of the system lies not on the correct mapping of the articles to their corresponding three authors but to filter all the inventions of

Rabindranath Tagore from a bag of documents and the more the accuracy of the filtering, the more the accuracy of the system. Apart from being the first work of its kind for Bengali language, the contributions of this experiment can be identified as: (i) application of statistical approach in the field of stylometry, (ii) development of classification algorithm in n-dimensional vector space, (iii) developing a baseline system in this field and (iv) more importantly, working with the great writings of Rabindranath Tagore to reveal his swinging of thought and dexterity of pen when writing articles.

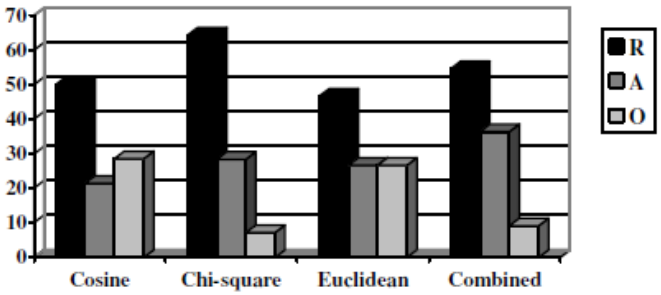


Fig. 2. Error analysis of different approaches.

VI. Conclusions

This paper introduced the use of a large number of fine-grained features for stylometry detection. The presented methodology can also be used in author verification task i.e. the verification of the hypothesis whether or not a given person is the author of the text under study. The methodology can be adopted for other languages since maximum of the features are language independent. The classification is very fast since it is based on the calculation of some simple statistical measurements. Particularly, it appears from our experiments that texts with less word are less likely to be classified correctly. For that, our system is little biased towards the stylometry of Rabindranath Tagore. It is due to the lack of the large number of resources of other authors under study. However from this preliminary study, future works are planned to increase the database with more fine-grained features and to identify more context dependent attributes for further improvement.

References

[1] D. Holmes, "Authorship Attribution," *Computers and the Humanities*, 28, 87–106, 1994.

[2] D.J. Croft, "Book of Mormon word prints reexamined," *Sunstone Publish*., 6, 15–22, 1981.

[3] D. Pavelec, E. Justino, and L.S. Oliveira, "Author Identification using Stylometric features," *Inteligencia Artificial, Revista Ideroamericana de Inteligencia Artifical,* 11, 59–65, 2007.

[4] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic authorship attribution," in *Proc. of the 9th Conference on European Chapter of the ACL*, 1999, pp. 158–165.

[5] K. H. Krippendorf, *Content Analysis-An Introduction to its Methodology*, Sage Publications Inc., 2nd Edition, 440 p., 2003.

[6] M.B. Malyutov, "Authorship attribution of texts: A review," *Lecture Notes in Computer Science*, vol. 4123, 362–380, 2006.

[7] S. Argamon, M. Saric, and S.S. Stien, "Style mining of electronic messages for multiple authorship discrimination: First results," in *Proc. 9th ACM SIGKDD*, 2003, pp. 475–480.

[8] T.K. Mustafa, N. Mustapha, M.A. Azmi, and N.B. Sulaiman, "Dropping down the Maximum Item Set: Improving the Stylometric Authorship Attribution Algorithm in the Text Mining for Authorship Investigation," *Journal of Computer Science,* 6 (3), 235–243, 2010.

[9] T. Zhang, F. Damerau, and D. Johnson, "Text chunking using regularized winnow," in *Proc. 39th Annual Meeting on ACL*, 2002, pp. 539–546.

[10] T. Chakraborty and S. Bandyopadhyay, "Identification of Reduplication in Bengali Corpus and their semantic Analysis: A Rule Based Approach," in *Proc. of the COLING (MWE 2010)*, Beijing, 2010, pp. 72–75.

[11] V. H. Halteren, "Linguistic profiling for author recognition and verification," in *Proceedings of the 2005 Meeting of the Association for Computational Linguistics (ACL)*, 2005.

[12] S. Argamon, M. Saric, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: First results," in *Proceedings of the 2003 Association for Computing Machinery Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, 2003, pp. 475–480.

[13] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye, "Author identification on the large scale," in *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*, 2005.

[14] M. Koppel, J. Schler, and E. Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors," *Journal of Machine Learning Research*, 8, 1261–1276, 2007.

Journal Information and Instructions for Authors

I. Journal Information

"*Polibits*" is a half-yearly research journal published since 1989 by the Center for Technological Design and Development in Computer Science (CIDETEC) of the National Polytechnic (Technical) Institute (IPN) in Mexico City, Mexico. The journal solicits original research papers in all areas of computer science and computer engineering, with emphasis on applied research.

The journal has double-blind review procedure. It publishes papers in English and Spanish.

Publication has no cost for the authors.

A. Main Topics of Interest

The journal publishes research papers in all areas of computer science and computer engineering, with emphasis on applied research.

More specifically, the main topics of interest include, though are not limited to, the following:

- Artificial Intelligence
- Natural Language Processing
- Fuzzy Logic
- Computer Vision
- Multiagent Systems
- Bioinformatics
- Neural Networks
- Evolutionary algorithms
- Knowledge Representation
- Expert Systems
- Intelligent Interfaces: Multimedia, Virtual Reality
- Machine Learning
- Pattern Recognition
- Intelligent Tutoring Systems
- Semantic Web
- Database Systems
- Data Mining
- Software Engineering
- Web Design
- Compilers
- Formal Languages
- Operating Systems
- Distributed Systems
- Parallelism
- Real Time Systems
- Algorithm Theory
- Scientific Computing
- High-Performance Computing
- Geo-processing
- Networks and Connectivity
- Cryptography
- Informatics Security
- Digital Systems Design
- Digital Signal Processing
- Control Systems
- Robotics
- Virtual Instrumentation
- Computer Architecture
- other.

B. Indexing

Index of excellence of CONACYT, LatIndex, Periódica, e-revistas.

II. Instructions for Authors

A. Submission

Papers ready to review are received through the Web submission system www.easychair.org/polibits. See also the updated information at the web page of the journal www.cidetec.ipn.mx/polibits.

The papers can be written in English or Spanish.

Since the review procedure is double-blind, the full text of the papers should be submitted without names and affiliations of the authors and without any other data that reveals the authors' identity.

For review, a file in one of the following formats is to be submitted: PDF (preferred), PS, Word. In case of acceptance, you will need to upload your source file in Word or TeX. We will send you further instructions on uploading your camera-ready source files upon acceptance notification.

You can submit your paper at any moment. It will be considered for the forthcoming issues.

B. Format

Please, use IEEE format[1], see section "Template for all Transactions (except IEEE Transactions on Magnetics)". The editors keep the right to modify the format and style of the final version of the paper if necessary.

We do not have any specific page limit: we welcome both short and long papers, provided the quality and novelty of the paper adequately justifies the length.

In case of being written in Spanish, the paper should also contain the title, abstract, and keywords in English.

[1] www.ieee.org/web/publications/authors/transjnl/index.html