Workshop on Multimodal Semantics for Robotic Systems (MuSRobS)
IEEE/RSJ International Conference on Intelligent Robots and Systems 2015

A Novel Multimodal Emotion Recognition Approach for Affective Human Robot Interaction

Felipe Cid, Luis J. Manso and Pedro Núñez

***Abstract*— Facial expressions and speech are elements that provide emotional information about the user through multiple communication channels. In this paper, a novel multimodal emotion recognition system based on visual and auditory information processing is proposed. The proposed approach is used in real affective human robot communication in order to estimate five different emotional states (i.e., happiness, anger, fear, sadness and neutral), and it consists of two subsystems with similar structure. The first subsystem achieves a robust facial feature extraction based on consecutively applied filters to the edge image and the use of a Dynamic Bayessian Classifier. A similar classifier is used in the second subsystem, where the input is associated to a set of speech descriptors, such as *speech-rate, energy and pitch*. Both subsystems are finally combined in real time. The results of this multimodal approach show the robustness and accuracy of the methodology respect to single emotion recognition systems.**

I. Introduction

In the last decade, Human Robot Interaction (HRI) has become one of the most important issues in social robotics. Within HRI, one of the main objectives is the development of methodologies focused on non-invasive techniques based on natural language. This would allow the robot to interact with users in a similar way to humans, through multimodal systems that combine information from different channels. In order to socially interact with humans, a robotic system should be able not only to understand users behaviour and intentions, but also to estimate their emotional state. Knowing and understanding these human emotions helps social robots adapting the communication in real time, improving and enriching the interaction [1]. This kind of HRI is usually known as affective HRI.

The design of social robots looks for developing natural interfaces for an affective interaction. In this context, most of the current affective HRI techniques use an unique information channel, called *mode*, such as speech, facial expressiveness or body language. However, an emotion is generally expressed through several modalities. In multimodal affective interaction, the user can communicate his/her emotional state to the robot using different several input channels [2]. Contrary to approaches where each channel introduces complementary information in the emotion recognition system, most of the multimodal systems use these channels as redundancy information. This redundancy is useful in real interaction scenarios, where for instance, errors associated to noise or occlusions can be reduced.

This work presents a novel multimodal emotion recognition system for affective HRI. The proposed approach is based on a real-time multimodal system that integrates speech and facial expression analysis. The main contribution of this work is a robust feature extraction approach for facial expression recognition. In this paper, these facial features are a combination of independent and antagonistic distortions of the face. Besides, a novel acoustic analysis of speech is used to extract features relevant to emotion. Both visual and acoustic features constitute the input of two Dynamic Bayessian Network (DBN) [3], which classify them into a set of basic emotions. The output of each DBN estimates the user's emotional state based on the extracted features, the emotional states available in each DBN being: $happiness$, $sadness$, $anger$, $fear$ and $neutral$ (non-emotional state). Finally, the purpose of this system is to integrate the output information of each subsystem, in a third DBN to analyze the information associated with each mode. This multimodal methodology uses redundant information to estimate accurate and robust results on the user's emotional state.

This paper is organized as follows: after discussing known approaches to emotion recognition systems from facial expressions, speech or multimodal systems in Section II, Section III presents an overview of the proposed Multimodal Emotion Recognition System. In Section IV, the experimental results are pointed out, and finally, Section V describes the conclusions and future work of the approach.

II. Related Works

In the field of HRI, different automatic emotion recognition systems have been studied. Most of these approaches are based on single information channel analysis, such as video sequences or audio signals. Independent of the nature of the information source, the raw data is processed and a set of features are extracted. Then, these features are classified into different categories, *i.e.,* emotions. On one hand, facial expressions have been commonly used to detect and recognize human emotions. An interesting and updated review was shown in [4]. Commonly, these frameworks use the Facial Action Coding System (FACS) proposed by Ekman et al's [5], which is based on facial muscle deformations. On the other hand, speech has been also used for emotion recognition (see the review [6]). These systems are usually focused on acoustic variables variations that are related to emotions.

L. J. Manso and P. Núñez are members of Robotics and Artificial Vision Lab. *Robolab* Group, University of Extremadura, Spain (e-mail: lmanso@unex.es; pnuntru@unex.es)

F. Cid is with Institute of Electrical and Electronics, Universidad Austral de Chile, Chile. (e-mail:felipe.cid@uach.cl

Approaches that only use visual or speech information individually usually fail in real scenarios. Light conditions, shadows or occlusions, among others, are typical situations where the accuracy of the results decreases for visual systems. In a similar way, environmental noise or people moving while talking are error sources in audio systems. Therefore, for an efficient affective HRI, several authors have focused their attention on multimodal systems that recognize the emotional state of the user from different modalities or information sources such as: face, gesture, body language, speech or physiological signals, among others. For instance, in [3] the authors develop an architecture for multimodal emotion recognition, where facial expresiveness and speech are used. In [7], Kessous et al. propose a system that fuse body and facial languages, and speech. Also, interesting reviews are presented in [2], [8]. Most of these approaches use a dominant mode in the classifier strategy, that is, when the probability of the detected emotion is low, other channel information is used. In contrast to these works, this paper presents a multimodal approach where all the input modes are analyzed in real-time and the fusion strategy consists on a Dynamic Bayessian Network classifier. This multimodal fusion is produced at the decision level (*i.e.,* the information is integrated from the single modalities after being interpreted by their own classifiers).

III. Multimodal Emotion Recognition System

In this section, the proposed multimodal emotion recognition system is presented. The framework consists on two subsystems running in parallel that estimate human emotions using two independent DBN. The facial expression recognition subsystem (see Fig. 1) is based on a fast and robust feature extraction approach where consecutively morphological and convolutional filters are applied to reduce the noise and the dependence against luminosity changes. After that, a Gabor filter is used for efficient edge detection. The output edge image of this filter bank is used to detect and extract scale-invariant facial features, which will be the input variables of the DBN to estimate the user's emotional state. In the second system, the user's speech is analyzed in order to extract a set of independent descriptors. These descriptors are the input of a second DBN. Finally, the proposed system integrates the information associated to both methods in a third DBN, which estimates the final user emotion. The output of both subsystems and a third DBN has as a result an estimate of the emotional state of the user, within four possible emotional states (happiness, sadness, anger, fear) and a non-emotional state (neutral). Each subsystem is described in detail in the following subsections.

A. Emotion recognition from Facial Expressions

The facial expression recognition system proposed in this work uses a video sequence acquired by the robot in real time. An overview of the proposed methodology is shown in Fig. 1. Each video frame is processed and a set of robust and invariant features of the user's face is detected. The proposed method consists on the following stages:

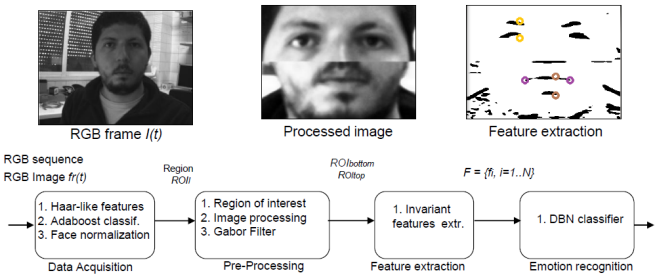


Fig. 1. Overview of the proposed facial expression recognition system. The approach flows from left to right. See the text for more details.

1) Data acquisition: Data acquisition for the recognition of facial expressions is based on the processing of a sequence of RGB images S_{fr} obtained from the robot camera for real-time interaction. First, the system recognizes the user's face within each frame $fr(t)$ in the instant of time t, using the well-known Viola and Jones' method [9] to obtain the initial region of interest of the face ROI_I. The ROI_I is normalized to a fixed size and converted to gray scale to be pre-processed.

2) Pre-processing: In this stage, the ROI_I is processed to remove noise, reducing its light dependence and eliminating unnecessary information. To eliminate errors in the detection and reduce the processing time of the system, the ROI_I is divided in two sub-regions, ROI_{top} and ROI_{bottom}, respectively. The usage of both sub-regions allow the elimination of irrelevant information (*e.g.,* nose) and divides the feature extraction process in two elements (eyes and mouth). Let ROI_I being the face image of size W x H, and let $p_0 = (n, m)$ being the central pixel in the image, which estimates the approximated position of the nose in the image. Then, ROI_{top} and ROI_{bottom} are defined as selective copies of ROI_I as follows: ROI_{top} of size W x $(m - N_{Th})$ and ROI_{bottom} of size W x $(m + N_{Th}))$, where N_{Th} is an user-fixed threshold. In order to detect and extract robust facial features in different environments, both ROI_{top} and ROI_{bottom} images are processed to reduce the effects of the light dependence. The method is based on the approach described in [10]. The processing sequence follows a set of consecutive stages: 1) gamma correction; 2) difference of Gaussian (DoG) filtering; 3) masking; and 4) contrast equalization. Next, a filter bank of Median, Blur and Gaussian filters, is applied to mitigate the noise effect in the images by the beard, wounds or similar facial elements.

3) Gabor Filter: The Gabor filter is a fast and effective linear filter for the detection of edges with different orientations. In the proposed approach, the Gabor filter is used as a previous stage to the detection and extraction of facial features, which are extracted using the contours of the facial elements (i.e., the eyes, the mouth or the eyebrows). Gabor impulse response in the spatial domain consists of a sinusoidal plane wave of some orientation and frequency, modulated by a two-dimensional Gaussian envelope. Let $I(u, v)$ be the input image, then the output of the Gabor

Fig. 2. Action Units (AUs) used in this paper.

filter, $G(u, v)$, is given by:

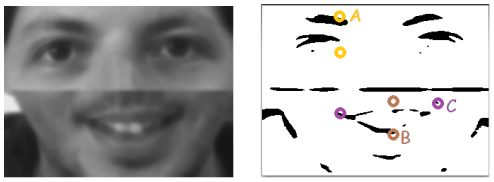
$$G(u,v) = exp(-\frac{1}{2}(\frac{u_\theta^2 + v_\theta^2}{\sigma^2}))cos(2\pi\frac{u_\theta}{\lambda} + \psi) \quad (1)$$

where θ, λ and ψ are associated to the sinusoidal plane wave (orientation, wavelength and phase, respectively), and being u_θ and v_θ described as:

$$\begin{aligned} u_\theta &= ucos\theta + vsin\theta \\ v_\theta &= -usin\theta + vcos\theta \end{aligned} \quad (2)$$

4) Feature extraction: The crucial step in an automatic facial expressions recognition system is the extraction of relevant features from the image, F^I=$\{f_i^I \mid i = 1..m\}$. In the proposed work, a set of edge-based features is extracted, which is invariant to scale or distance from the user to the robot. Each one of the features are directly related to the Action Units (AUs) described by the Facial Action Coding System (FACS) [5]. A set of independent and antagonistic AUs has been used in this paper (see Fig. 2, AU1 and AU4 are related to distortions of the eyebrows, and they are antagonistic and independent). In this approach, only three features are defined in the edge face image, which are associated to the Euclidean distance between the upper contour of the eyebrows and the lower edge of the eyes (dA), lip corners (dB) and upper and lower contour of the mouth (dC), respectively. These features are normalized using the values extracted from the neutral state, which allows the system to be independent of the scale or distance of the user to the sensor. Fig. 3(a) illustrates the ROI of the face in the image. The image is processed according to the method described in this section. Results after applying light normalization, noise removal methods and Gabor filtering are shown in Fig. 3(b). Fig. 3(b) also illustrates the set of extracted features, labelled as dA (yellow), dB (brown) and dC (violet).

5) Dynamic Bayesian Network classifier : In order to classify the Facial Expression (FE) produced by the user's face, a *Dynamic Bayesian network* is proposed, where the overall classification result achieved is the one foreseen by the belief variable FE, in the scope ($FE_{[neutral]}$, $FE_{[happiness]}$, $FE_{[sadness]}$, $FE_{[fear]}$, $FE_{[anger]}$).



(a) (b)

Fig. 3. Edge-based facial feature extraction: a) Region of interest in the face image (ROI_{top} and ROI_{bottom}); and b) Features extracted in the image.

This bayesian approach is based on the detection of 11 AUs with antagonistic and exclusive features as leaves of the DBN, which allows us to reduce the size of the DBN to 7 variables that group these AUs. Thus, these 7 variables are obtained from combinations of the extracted features of the previous process: dA, dB and dC. In this case, it is assumed that these 7 leaf variables are independent given the facial expression (FE). Although some muscular movements from one area of the face may slightly affect other areas, this small influence could not be detected by the robot cameras. Besides, for correct detection of each emotional state it is necessary that each AU achieves a minimum threshold intensity of B (i.e., slight evidence), within the intensity range described in FACS [5].

Fig. 4 illustrates the structure of the two level bayesian network, and the time influence that characterizes this DBN. The first level shows the estimation of the user's emotional state FE, while the second level shows the 7 variables based on the AUs. In addition, one of the main features worth mentioning about the bayesian network is the need to provide it with learning data. The most common method is that each new sample is matched using a threshold. In this work, to avoid the extant gaps, a pre-processing stage is done before the learning stage, fitting a Gaussian distribution to the data. The learning data acquisition was performed by a supervisor, who collected the samples of the 7 random variables manually, correctly classifying them. The leaf random variables of the model, and their respective virtual-scopes are shown in the Table I.

Table II the relationship between the basic emotional states of a user, and the $Action\ Units$ associated with different facial distortions. In this case, these AUs depend on the few deformable or mobile elements of the face, such as: the mouth, eyes and eyebrows. This feature of deformation or movement of facial elements is what allows analyze a muscular change associated with a specific emotion, otherwise there would be no facial expressions. However, elements such as the nose does not present specific changes in facial expressions.

In this approach, the data (D) in the classification process is obtained according to the following setup:

$$D = ((x_1, y_1)...(x_n, y_n)), x_i \in \mathbb{R}^d, y_i \in \mathbb{R} \quad (3)$$

Consider that y_1 to y_5 are the five possible emotional states ($FE_{[neutral]}$, $FE_{[happiness]}$, $FE_{[sadness]}$, $FE_{[fear]}$,

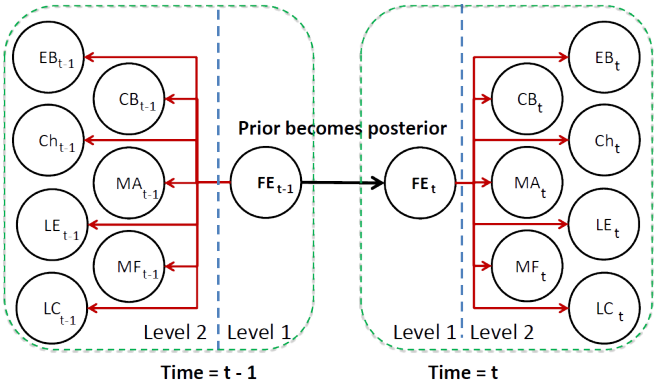


Fig. 4. Facial Expression Dynamic Bayesian Network. Two time intervals are shown.

Variable	Action Units AUs	Element of the face
EB	AU1, AU4	Eye-Brows
Ch	AU6	Cheeks
LE	AU7	Lower Eyelids
LC	AU12, AU15	Lips Corners
CB	AU17	Chin Boss
CB	AU20, AU23	Mouth's Form
MA	AU24, AU25	Mouth's Aperture

TABLE I
LEAF VARIABLES OF THE DBN

Emotion	Action Units AUs
Sadness	AU1 - AU4 - AU15 -AU17
Happiness	AU6 - AU12 - AU25
Fear	AU1 - AU4 - AU20 - AU25
Anger	AU4 + AU7 - AU17 - AU23 - AU24
Neutral	—

TABLE II
RELATIONSHIP BETWEEN ACTION UNITS (FACIAL FEATURES) AND THE DIFFERENT EMOTIONAL STATES OF THIS SYSTEM. (INFORMATION COLLECTED FROM: [3]).

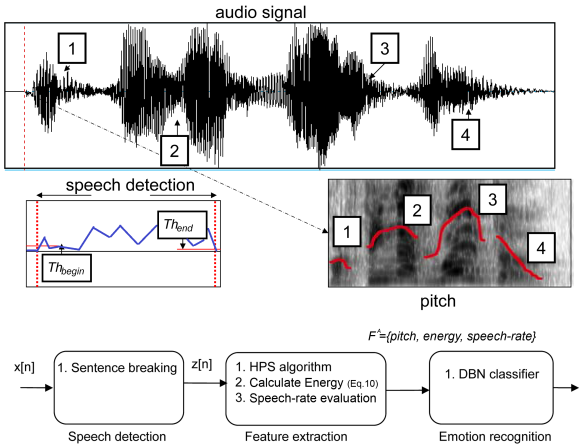


Fig. 5. Overview of the proposed emotion recognition system from speech. (Image acquired from the publication: [12])

$FE_{[anger]}$); and each dimension of x, corresponds to one of the previously described random variables, namely: EB, Ch, LE, LC, CB, MF and MA. Since the learning data may have gaps between its samples, a model is built assuming that $(X_1, ..., X_n)$ are independent given FE, and

$$X_i \sim N(prior^T x_i, \sigma^2) \tag{4}$$

At first, $prior \sim U(1/n)$, however throughout the iterations, the posterior of $t-1$ becomes the prior on t.

Finally, the posterior equation is obtained using Bayes' rule:

$$P(FE|x_m) = \frac{\prod_1^n P(x_i|FE) * P(FE)}{P(x_m)}, \tag{5}$$

where x_m is the most recent visual information acquired. The last dividend can be computed using the Bayesian marginalization rule:

$$P(x_m) = \sum_{FE} \prod_1^q P(x_i|FE) * P(FE), \tag{6}$$

being $q = 7$, the number of random variables of the system.

The dynamic properties of the network cause a convergence over time. The resultant histogram from the previous frame is used as prior knowledge for the current frame. Each classification is considered correct, if it converges in a maximum number of 5 frames exceeding a threshold of 80%. Otherwise, if after 5 frames no value is higher than the threshold, the classifier selects the highest probability value (usually referred to as the *Maximum a posteriori decision* in Bayesian theory) as the result.

B. Emotion recognition from Speech

Emotion recognition using speech as input poses several subproblems, such as: detection, extraction and characterization of a set of significant acoustic features to recognize emotions. Therefore, the proposed approach has a similar structure of the subsystem described in Sec. III-A, thus, like the former, a set of acoustic features is extracted and used as input to a DBN classifier. The proposed methodology is based on the work presented in [11], which studied the influence of pitch, energy and speech-rate over the activation and also the intensity of some emotions. Fig. 5 illustrates an overview of the proposed system. The system is divided in the following stages:

1) Detection of the Speech: The audio signal is pre-processed in order to detect the presence or absence of speech in real time, within a communication between a human and a robot. This stage was achieved by using audio library SoX [13], which allows the detection of voice and silences of agreement with the sensitivity of the microphone. Specifically, the library analyzes and processes the stream original audio from the microphone through a feature VAD (Voice Activity Detection) based on the measurement of the power cepstral. Since this function is responsible for removing the noise, silence or any type of sound not related to the human voice. Thus, a signal in the time, $x(t)$, corresponding to an audio signal acquired by the robot, the power $Cepstrum$

$C(\tau)$ [14] is given by the equation:

$$C(\tau) = \mathcal{F}^{-1}\left(\log(|\mathcal{F}(x(t))|^2)\right) \quad (7)$$

Where $\mathcal{F}$ and $\mathcal{F}^{-}1$ represent the Fourier Transform direct and reverse, respectively. The output of this process is a signal that is composed of frames that contain the information of the user's voice,at a sampling frequency f_s of 44,100 KHz. Finally, these frames are input in the function responsible for extracting acoustic features.

2) Acoustic feature extraction: From the selected spoken sentence, a set of acoustic features, able to characterize the user's emotional state, are extracted. The selection of these characteristics follows a common theme with other methods of literature, which are related to the use of elements of prosody, such as: [3],[15], [16] and [17].

Due the previous delivery process the audio frames, it is possible to extract a set m of features represented by: F^A=$\{f_i^A \mid i = 1..m\}$, as described in [11]. In this system, it is considered 3 characteristic elements of prosody (m= 3), such as the pitch, energy and the tempo. In [11] the relationship between the different features of the speech and the emotional states of an interlocutor is studied, coming to the conclusion that many of the elements of prosody are affected by the intensity and valencia of each of the emotions. For example, emotions with high intensity features have high values (Energy, Pitch and Tempo). While the emotions with low intensity have lower values in these same features.

In this section, the following features to extract and the respective justification of its relevance for emotion recognition are:

- **Pitch**: also called *Fundamental Frequency*, is the rate of vibration of the vocal cords to produce sound. The pitch range is a feature that allows to identify not only the gender of users, but also their emotional states.
- **Energy**: distribution of the signal amplitude values in time. In an audio signal, the energy in the voice is a determining factor in the generation of emotions. Emotions with higher intensity are associated to higher energy values in the voice. In a similar way, emotions with lower intensity show lower energy values.
- **Speech-Rate**: also called *velocity* or *BPM (beats per minute)*, is diction speed or number of words in a time period. In a similar way to the Energy, the Speech-Rate is a feature associated directly to the intensity of emotions. Thus, emotions with high intensity present high Speech-Rate values, and emotions with low intensity are associated with low Speech-Rate values.

The methods to extract each of these three acoustic features are briefly described bellow:

Pitch: The pitch range is calculated using the HPS (Harmonic Product Spectrum) algorithm [18]. This algorithm uses the Hann function to create windows of short duration on the input signal, separating it into frames $X(\omega)$. The HPS algorithm measures the maximum coincidence for harmonics according to eq. 8 for each spectral frame.

$$Y(\omega) = \prod_{r=1}^{i=R} |X(\omega r)|^2, \qquad 0 \leq w < N \quad (8)$$

where R is the number of harmonics to be considered, and frequency ω_i is in the range of possible fundamental frequencies. Pitch value, $Y^p(\omega_i)$, is calculated as the maximum value of the resulting periodic correlation array, $Y(\omega)$.

$$Y^p(\omega_i) = \max(Y(\omega)) \quad (9)$$

Energy: The quantification of the signal's energy is defined as:

$$E = \frac{1}{N} \cdot \sum_{x=0}^{x=i} x[i]^2 \quad (10)$$

Speech-Rate: In order to calculate the Speech-Rate, the beats are evaluated according to the following process: i) the Fast Fourier Transform of the signal is performed; ii) the signal is multiplied with trains of pulses with different speech-rates; and iii) the amount of Energy at each rate is analyzed. Finally, the Speech-Rate is calculated as the signal with the highest Energy value.

These three features are the input of the Bayesian network that estimates the user's emotional state. The table III shows the relation between the acoustic features and the emotional states of the user. In the table, it is seen as low-intensity emotions, such as sadness and neutral state, have similar features. Similarly, the emotions with a high intensity, such as anger or fear, show also common characteristics, but with slight differences perceptible to the user.

Emotion	*Pitch*	*Energy*	*Tempo*
Sadness	Slightly narrower	Lower	Slightly slower
Happiness	Much wider	Higher	Faster or slower
Fear	Much wider	Normal	Much faster
Anger	Much wider	Higher	Slightly faster
Neutral	much narrower	Normal	Slower

TABLE III
RELATIONSHIP BETWEEN ACOUSTIC FEATURES AND THE DIFFERENT EMOTIONAL STATES OF THIS SYSTEM. (INFORMATION COLLECTED FROM: [2] AND [3]).

3) Dynamic Bayesian Network classifier: For the classification of the emotional state of the user from the voice, a second DBN is proposed. This second network, represented in Fig. 6, has a two-tier structure, and the scope of the possible emotional states are identical to the facial expressions DBN (mentioned in section III-A.5). The first level is a single node that represents the variable associated with the result of the classified auditory emotion ($AE[Neutral, Anger, Happines, Sadness, Fear]$). The second level of the network corresponds to three nodes associated with three independent variables between them, such as: PT- Pitch; EN- Energy.and TE- Speech-Rate. These 3 variables quantify the elements of the speech as: speed and the intensity of talk, and are directly related to

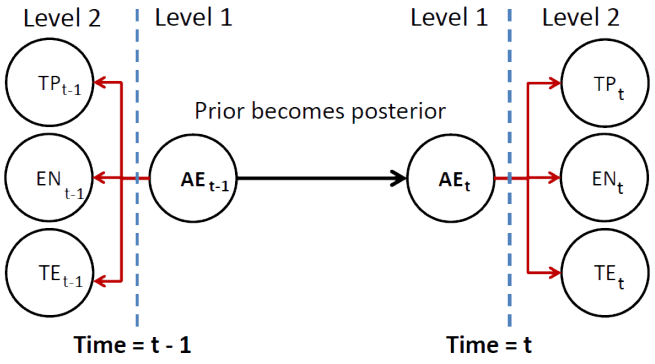


Fig. 6. Dynamic Bayessian Network classifier. Two time interval are shown.

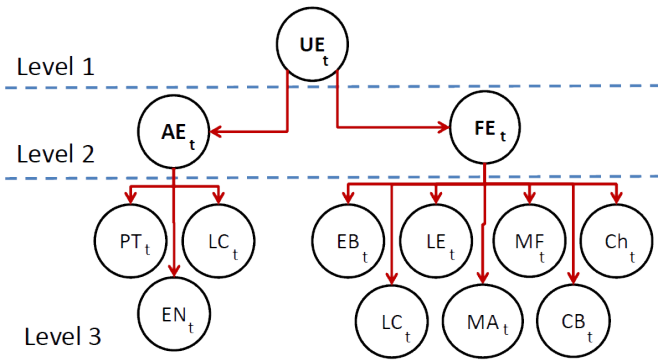


Fig. 7. Emotion Dynamic *Bayesian network, three levels are shown.*

the intensity of emotions. Moreover, each of the elements of the second level has as its father the node AE. In order to perform the estimation, the network first needs to be filled with learning information about the features in each emotional state. Henceforth, similarly to the facial expression network, Bayes' rule is once again used for inference given learning, as:

$$P(AE|x_a) = \frac{\prod_1^n P(x_i|AE) * P(AE)}{P(x_a)} \tag{11}$$

Where x_a is the most recently acquired auditory information, given by the variables in the second level (PT, EN, TE). Using the Bayesian marginalization rule, we can compute:

$$P(x_a) = \sum_{AE} \prod_1^k P(x_i|AE) * P(AE) \tag{12}$$

being $k = 3$, the number of variables in the network.

C. Multimodal fusion for emotion recognition

In this section, the system analyzes the information obtained by both modalities. Through the use of a multimodal system, it is possible to eliminate errors in the detection or classification, by checking the results with those obtained through another modality. The results of both modalities are used as input to a third bayesian network of three levels, giving rise to the estimate of the emotional state of the user, in a similar way to humans.

The last DBN classifies the results of the two previous DBN, to get the emotional state of the user in the conversation. This third DBN has a structure of three levels that meets all the nodes of the two previous networks, where the node U_E is the parent of the nodes of the second level FE and AE. For the nodes in the first and second level: U_E, FE and AE, there are only five possible results (Neutral, Happiness, Fear, Anger, Sadness). For the variables of the third level, 7 variables belong to the node FE that estimated the emotions through facial expressions, and 3 variables belong to the node AE that estimate the emotional states through the speech, as shown in Fig. 7.

To estimate the emotional state on the basis of these two results, the joint distribution asociated to the Bayesian Fusion is used, by the following equation:

$$P(x_c) = \sum_{U_E} \prod_1^j P(x_i|U_E) * P(U_E) \tag{13}$$

From the joint distribution, using the Bayes' rule:

$$P(U_E|x_c) = \frac{\prod_1^n P(x_i|U_E) * P(U_E)}{P(x_c)} \tag{14}$$

Through this third bayesian network, the emotional state of the user is estimated in this multimodal system, giving rise to a robust, real-time result.

Time Control: One of the main problems of the modal systems is related to the synchronization between the different blocks that make up the recognizer. The detection of a facial expression does not have to coincide in time with a corporal expression, and much less with the emotional information that can be extracted from the speech analysis. For this reason, a control block of time that synchronizes the results of each of the Bayesian networks are established, but instead of giving the output of each subsystem an estimate of human emotion in the same instant of time, the proposed system uses the emotion recognizer based on the analysis of facial expression as the dominant mode in the entire system [2]. Thus, only when there is audio information or bodily expressions during the interaction, the time control is the module for which these data should be merged into the final dynamic classifier.

In the Fig. 8 shows the behavior in time of the proposed system. Where, the predominant system corresponds to the output of recognizer of facial expressions. Only when there is audio data results are merged into a single output of multimodal system.

IV. Experimental Results

In this section, a set of tests evaluates the performance of the proposed system, through the evaluation of each of the two modalities separately and together. In the two first tests, the visual and auditory modalities are tested separatedly. In the last one, the whole set of the two modalities working together is examined.

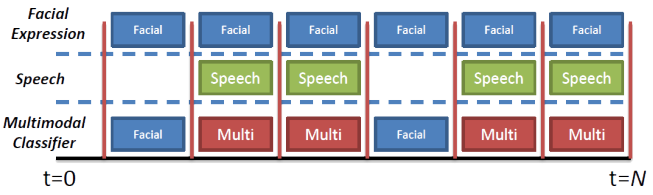


Fig. 8. Operation in the time of the proposed system.

The algorithms presented in this paper were developed in C ++, and the benchmark tests were performed on a computer with a 2.8 GHz Intel(R) Core(TM) i7 CPU and 4 Gb RAM running Linux. The software to control the system is built on the top of the robotics framework RoboComp [19]. For each test, two groups of evaluation were used: The first group is composed of visual and auditory information from users was provided by the Surrey Audio-Visual Expressed Emotion ($SAVEE$) database [20], through video and audio files. Meanwhile, a second group is comprised of 40 users with different gender and facial features has been used (The tests related to the audio, were performed in: [12]). This last group represents the non-trained users between adolescents and adults who were volunteers for these experiments in real time, within the laboratory ROBOLAB from the University of Extremadura. The visual and auditory information of these non-trained users was stored in our own database, through 10 tests or sequences for each user. So in each test, the user shall represent each of the emotional states in random order determined by the user. The relevant information of the users is analyzed through audio and video files for each test.

The information for each user group is presented in audio and video files, which are used as the main entrance of the system for each evaluation. In order to differentiate between each group and to avoid confusion, was appointed the first group as ”Database”, and the second group as ”Volunteers”. Finally, the evaluation of the system is divided into the following tests:

1) Recognition from Facial Expression: The first test checks the performance of the system for the recognition of facial expressions. This system detects and extracts features to estimate the emotional states of the user. Therefore, during the test all the visual information is provided by the two database mentioned above, through video files of users with different facial features, to verify the correct detection and classification of this system. The results related to each group of users show important differences, associated to the quantities of errors in the classification, as shown in Table IV and V. These differences between the data for each group is due to the reliability and the conditions for the acquisition of information in each case. On the one hand, the database ”SAVEE” provides best information in a controlled scenario, with high-quality images, trained users and a controlled lighting (without natural light). On the other hand, the information gained from the volunteers is related to non-trained users (mainly students and adults), in a non-controlled environment with natural light and ambient noise. Besides, in the Table VIII detail the errors in this test in 3 groups: $Misclassification$, $Ambiguous$ (between two emotional states with low recognition rates.) and $Under\ threshold$ (not exceed the minimum threshold in a state).

Test P_{FE}	a	b	c	d	e	Errors
a. Sadness	**90%**	0%	0%	0%	3%	7%
b. Happiness	0%	**97%**	0%	0%	0%	3%
c. Fear	1%	2%	**93%**	0%	0%	4%
d. Anger	2%	0%	0%	**94%**	0%	4%
e. Neutral	3%	0%	0%	0%	**95%**	2%

TABLE IV
RECOGNITION FROM FACIAL EXPRESSION (BASED ON SELECTED VOLUNTEERS FOR THESE TESTS)

Test P_{FE}	a	b	c	d	e	Errors
a. Sadness	**96%**	0%	0%	0%	1%	3%
b. Happiness	0%	**98%**	0%	0%	0%	2%
c. Fear	1%	3%	**89%**	0%	0%	7%
d. Anger	2%	0%	0%	**93%**	0%	5%
e. Neutral	1%	0%	0%	0%	**97%**	2%

TABLE V
RECOGNITION FROM FACIAL EXPRESSION (BASED ON THE DATABASE $SAVEE$)

2) Recognition from speech: The second test checks the robustness of the emotion recognition system from speech, analyzing the sentences in the audio signal to estimate the emotional states of the user. For this second test all the auditory information from users was provided by the two database mentioned above, to analyze and verify the performance of the system. The results of this second test are seen in the Table VI and VII, where the emotions with less intensity showed better results in the recognition from speech. However, the results of the recognition of emotions based on speech presents a lower performance in comparison with the recognition of emotions based on facial expressions. Since the auditory information can have factors that affect the final results, such as: dialect (cultural), personality (psychological), nervousness (psychological), among others. These factors have consequences that can be analyzed in the Table VI and VII, which despite having a number of differences in the methods of recording and the quality of the data acquired by the volunteers of this experiment and the database ”SAVEE”, exhibit similar results. The details of the errors are illustrated in the Table VIII.

3) Errors in the tests: The main errors described in Table VIII, are associated with the classification and estimation of the emotional states of the user through the auditory information of both groups. These failures in the classification is due to the inconsistency of the data that cause ambiguities between several states and errors of the classifier, called ”misclassification” . In relation to the visual information, the facial expressions show more errors in the classification process as ”misclassification” and ”under the threshold”, due

Test P_{AE}	a	b	c	d	e	Errors
a. Sadness	**87**%	0%	0%	0%	2%	11%
b. Happiness	0%	**71**%	5%	0%	0%	24%
c. Fear	2%	2%	**67**%	5%	0%	24%
d. Anger	2%	0%	5%	**78**%	0%	15%
e. Neutral	5%	0%	0%	0%	**82**%	13%

TABLE VI

RECOGNITION FROM SPEECH (BASED ON SELECTED VOLUNTEERS FOR THESE TESTS)

Test P_{AE}	a	b	c	d	e	Errors
a. Sadness	**83**%	0%	0%	0%	4%	13%
b. Happiness	0%	**76**%	3%	0%	0%	21%
c. Fear	0%	3%	**81**%	4%	0%	12%
d. Anger	1%	0%	6%	**67**%	0%	26%
e. Neutral	4%	0%	0%	0%	**89**%	7%

TABLE VII

RECOGNITION FROM SPEECH (BASED ON THE DATABASE $SAVEE$)

to certain unusual behavior of the user or facial distortions that are not associated with a specific emotion.

4) Emotion recognition from facial expressions and speech: Finally, to demonstrate the benefits of the proposed multimodal system, a third test was performed to verify the multimodal system based on the bayesian network, which analyzes the information from the two methods described above. The last system allows for the correction of detection errors and reduces the uncertainty in the classification. The results of the last test are seen in the Table IX.

Table IX shows the differences between the results of the two databases. Where, the advantage of the database ($SAVEE$) are based on good performance in the predominant system. However, the benefits of this type of systems are clear. Given that removes a range of problems associated with errors of classification, ambiguity, among others.

V. CONCLUSIONS

In this paper, an emotion recognition system through a multimodal approach is presented. This system recognizes emotional states through two modalities, the visual and the auditory. The visual modality uses observed information to recognize facial expressions. It is based on the use of contours and Gabor filters. This allows the recognition of patterns in the user's face, even with interference by the light and a wide range of different users. The second modality, auditory, recognizes emotion related items from the audio, using a sentence from a conversation to determine the emotional state of the user, getting emotional information in a similar way as a human would. The results show the improvements of the multimodal approach against those systems based on a single modality. Multimodal solutions show a significant reduction of detection and classification errors.

Future work will focus on using more precise information, through the use of new features based on the spectrum and a microphone array. This would improve the information

Errors	Misclassification	Ambiguous	Under threshold
Volunteers (P_{FE})	2%	1%	1%
Volunteers (P_{AE})	11%	2%	5%
Database (P_{FE})	2%	1%	2%
Database (P_{AE})	13%	1%	2%

TABLE VIII

DETAILS OF THE ERRORS IN THE TESTS

Test	Sadness	Happiness	Fear	Anger	Neutral
Volunteers P_{UE}	93%	97%	93%	**95**%	97%
Database P_{UE}	**98**%	**99**%	**91**%	95%	**98**%

TABLE IX

ROBUSTNESS OF THE EMOTION RECOGNITION SYSTEM (P: PERCENTAGE OF CORRECTLY DETECTED EMOTION)

obtained by the system, allowing the use of other types of information in the multimodal system as: body language, among other.

ACKNOWLEDGMENT

This work has been partially supported by the MICINN Project TIN2012-38079-C03-01, and by the Institute of Electrical and Electronics of the Universidad Austral de Chile.

REFERENCES

[1] R. W. Picard, "Affective Computing". MIT Press, pp. 88-91, 2000.
[2] N. Sebe, I. Cohen, T. Gevers, T. S. Huang, "Multimodal Approaches for Emotion Recognition: A Survey", In *Internet Imaging VI, SPIE'05*, USA, 2005.
[3] J. A. Prado, C. Simplício, N. F. Lori, J. Dias. "Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction", In International Journal of Social Robotics, Vol.4, Issue 1, pp. 29-51,2012.
[4] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art", Tech Report, College of Computing, Georgia Institute of Technology, 2012.
[5] P. Ekman, WV Friesen, JC Hager, "Facial Action Coding System FACS", The manual, 2002.
[6] Z. Zeng, M. Pantic, G. I. Roisman and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions", In *IEEE Transactions on Pattern Analysis and Machine Intelligence*,Vol. 31, pp. 39-58, 2008.
[7] L. Kessous, G. Castellano and G. Caridakis, "Multimodal Emotion Recognition in Speech-based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis", Journal on Multimodal User Interfaces, Vol. 3, No. 1, pp. 33-48, 2010.
[8] A. Jaimes and N. Sebe, "Multimodal Human Computer Interaction: A survey", In *IEEE International Workshop on Human Computer Interaction in conjuntion whith ICCV 2005*, beijing, China, 2005.
[9] P. Viola and M. J. Jones,"Robust real-time face detection", In International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004.
[10] X. Tan and B. Triggs. "Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions", IEEE Transactions on Image Processing, Vol. 19, pp. 1635-1650, 2007.
[11] R. Cowie and R. Cornelius. "Describing the emotional states that are expressed in speech", Speech Communication 40, pp. 5-32, 2003.
[12] F. Cid, J. Moreno, P. Bustos and P. Núñez. "Muecas: A Multi-Sensor Robotic Head for Affective Human Robot Interaction and Imitation". In Sensors 2014, 14(5), pp. 7711-7737, 2014.
[13] Sound eXchange, SoX. Available online: http://sox.sourceforge.net/
[14] D. Childers, D. Skinner and R. Kemerait. "The Cepstrum: A Guide to Processing". Proc. of the IEEE, Vol.65, No.10, pp. 1428-1443, 1977.
[15] A. Nogueiras, J. Marino, A. Moreno and A. Bonafonte. "Speech emotion recognition using hidden markov models". In European Conf. on Speech Communication and Technology (Eurospeech 01), 2001.

[16] T. Chen, "Audio-Visual Integration in multimodal Communication". In *IEEE Proceedings*, May, 1998.

[17] B. Schuller, G. Rigoll and M. Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture". In Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I–577–I–580, 2004.

[18] Noll, M. "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate". In Proceedings of the Symposium on Computer Processing in Communications, Vol. XIX, Polytechnic Press: Brooklyn, New York, pp. 779-797, 1970.

[19] L. Manso, P. Bachiller, P. Bustos, P. Núñez, R. Cintas and L.Calderita. "RoboComp: a Tool-based Robotics Framework", In Proceedings, SIMPAR Second International Conference on Simulation, Modeling and Programming for Autonomous Robots. pp 251-262. 2010.

[20] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition", In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, DOI 10.4018/978-1-61520-919-4, chapter 17, pp. 398-423, July 2010.