A Novel Real Time Facial Expression Recognition system based on Candide-3 Reconstruction Model

P. Romero, F. Cid and P. Núñez

***Abstract*—In the last decade, affective Human-Robot Interaction has become an interesting topic for researching. Facial expressions are rich sources of information about affective behaviour and have been commonly used for emotion recognition. In this paper, a novel facial expression recognition algorithm using RGB-D information is proposed. The algorithm achieves in real-time the detection and extraction of a set of both, invariant and independent facial features based on the well-known Candide-3 reconstruction model. Experimental results showing the effectiveness and robustness of the approach are described and compared to previous related works.**

***Index Terms*—Social Robot, Facial Expression Recognition, Human-Robot-Interaction.**

I. INTRODUCTION

The days of the robots entering into the daily life are coming soon. In these real scenarios, robots must be programmed to interact and communicate with humans or other robots by following social behaviors and rules. In this respect, as with human-human interaction, affective states are one of the most critical elements in a natural human-robot interaction (HRI). Thus, knowing and understanding these human emotions helps social robots in adapting their communication in real time, improving and enriching the interaction [1]. This kind of HRI is usually known as affective HRI, which has become a central part in the field of social robotics in the last years.

Determining the affective state of a human in a real interaction is not an easy task. Most of the affective HRI techniques use only one information channel, called *mode*, in order to recognize human emotions. Among the typical modes are speech, body language or facial expressions, as shown in Fig. 1. This last one is the richest and most powerful source of affective information and therefore, the facial expressiveness has been mainly used as input data in classical emotion recognition systems. These systems are commonly based on facial muscle deformations, as was described in the Facial Action Coding System (FACS) proposed by Ekman *et al*. [2]. FACS describes facial movements as groups of *Action Units* (AU) and is considered to be the most commonly used standard that describes facial expressions.

This paper describes a new real-time facial expression recognition system using a RGB-D sensor that relies on the Candide-3 reconstruction model described in [3]. Contrary to other approaches, the **main contribution** of this work is to develop a low computational cost algorithm for extracting a set of independent and antagonistic facial features from the Candide-3 face model, which constitute the input of a Dynamic Bayesian Network (DBN) used as classifier [4]. Four different emotions, as well as a non-emotional state, are robustly detected using this Bayesian approach (happiness, sadness, anger, fear and neutral).

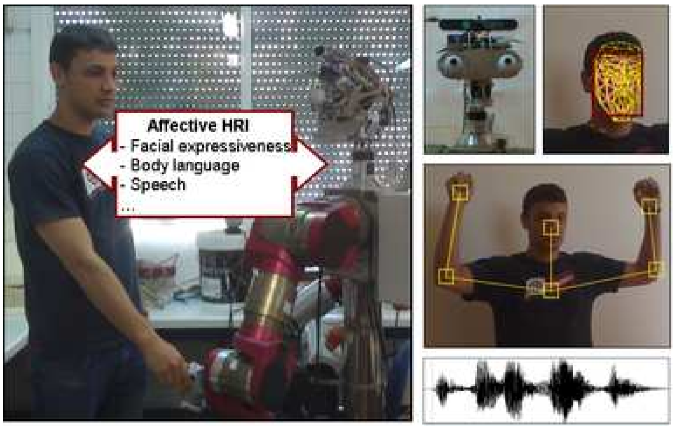


Fig. 1. Multimodal affective Human Robot Interaction. Facial expressiveness, body gestures and speech are usually used for an affective interaction.

This paper is organized as follows: After discussing known approaches to facial expression recognition in Section II, Section III presents an overview of the proposed system. In Section IV, the experimental results are outlined, and finally, Section V describes the conclusions and future works of the presented approach.

II. RELATED WORKS

Automatic emotion recognition systems have been studied in the last decade by several research groups. Current state-of-the-art systems are multimodal, that is, they estimate emotions using different modes: such as the face [8], [9], body gestures [19], speech [11] or physiological signals, among other [12], [13]. Facial expressiveness is the essential mean of transmitting emotions, and thus, the majority of the approaches are based on the analysis of facial expressions using visual information. An interesting and updated review is presented in [14]. In these works, a large part of the studies are based on the Facial Action Coding System (FACS) proposed by Ekman *et al.* [2], which classify the set of all the facial expressions into only six prototypical expressions conveying the basic emotions. FACS proposes that each facial expression is composed by a specific set of *Action Units* (AUs), which represent a distortion on the face induced by small muscular activity.

Implementing facial expression recognition systems is not an easy task. The main problems are related to the detection,

P. Romero, F. Cid and P. Núñez are member of Robotics and Artificial Vision Lab. *Robolab* Group, University of Extremadura, Spain.
E-mail: promerog@alumnos.unex.es; fecidb@alumnos.unex.es; pnuntru@unex.es

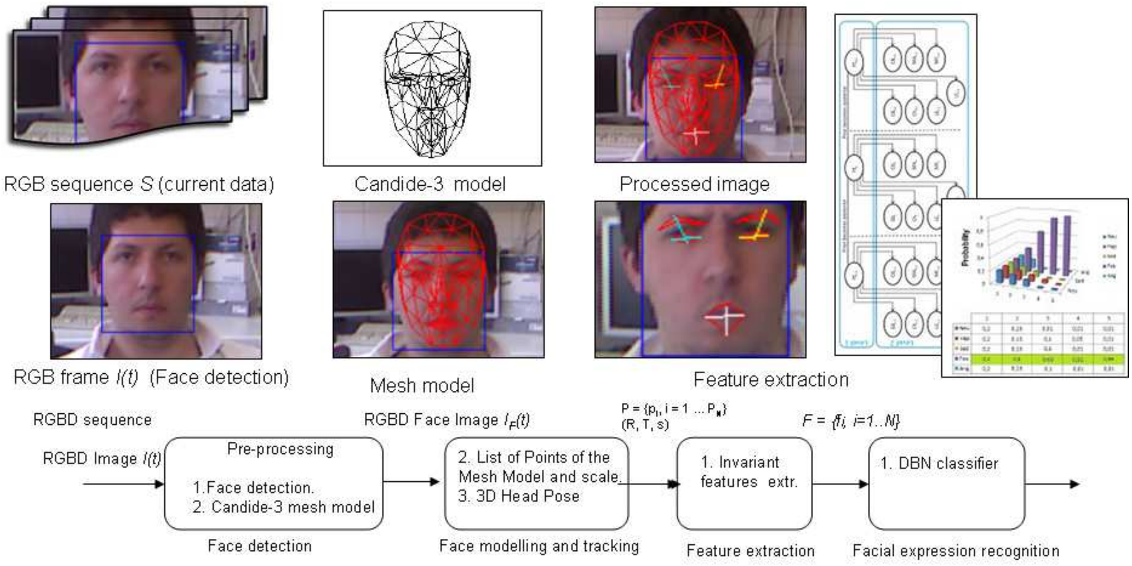


Fig. 2. Overview of the proposed system

extraction and later classification of the user's features, which is the basis of the most of the emotion recognition systems. Current literature describes different systems and methods that try to solve these problems. In [10], the authors use Active Appearance Models (AAM) and a Dynamic Bayessian Network (DBN) classifier for facial emotion recognition in image sequences. Support Vector Machines (SVM) in conjunction with a nonlinear mass-spring model are used in [15] for facial emotion recognition and imitation. Gabor filter banks are also used for extracting features from the image, such as was described in [8]. Finally, other facial features, which are directly related to AUs, are extracted from RGB images and later classified into a set of five basic emotions using a DBN [6],[4]. The proposed work is a modified version of these previous works, where features are now extracted not from the RGB image, but using RGBD information and the Candide-3 mesh model.

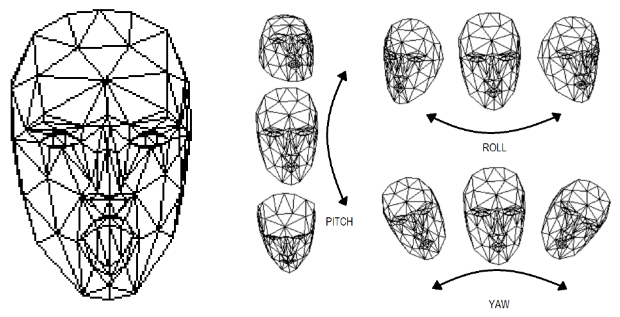
Several authors use RGB cameras to obtain a sequence of images as the basis of their emotion recognition algorithms [15], [10]. Using only RGB images for detecting facial features is a hard problem. Light conditions, shadows or noises are typical situations where these systems present mistakes to detect features and/or the user's head pose. Recently, new sensors that combine RGB and Depth information have become very popular. Microsoft Kinect sensor, for instance, is a low-cost RGB-D sensor that simultaneously provides both real-time, depth and color images. At present, there are some studies that show the advantages of using depth in the extraction of facial features, such as Li et al.'s work [16]. In this work, the authors use the Kinect sensor in order to recognize faces under challenging conditions. The proposal described in this work is focused on detecting invariant facial features using a face mesh model. This idea is not new, for instance, in [17], [18] the authors use similar models to extract facial features. However, in these works other more expensive sensors are used, in contrast to the low-cost RGB-D sensor used in this paper.

Finally, the use of RGB-D sensors within social robotics is increasing in the last years due to its low cost and its reasonable performance. These sensors allow to analyze the user features and their environment in real time, allowing the robot to efficiently interact with users, objects and other robots. An assessment of the Kinect sensor has been described in [21]. In this work, a set of tests under different scenarios is achieved, demonstrating that advantages of the Kinect sensor for HRI applications.

III. Automatic Facial Recognition System

In this paper, a novel facial expression recognition system is presented. This work is based on the Candide-3 reconstruction mesh model, which is used for detecting and extracting facial features. An overview of the proposed method is shown in Fig. 2, which flows from left to right. The system acquires a RGBD sequence in real-time and uses the *Microsoft Kinect SDK* for face tracking [20]. This algorithm uses the widely used Candide-3 deformable face model. The outputs of the face tracking algorithm implemented by *Microsoft SDK* are the head's pose (rotation and traslation matrices, and a scale factor) and a list of points associated to the Candide-3 mesh model. Through this information, it is possible to extract the user's facial features and its position in relation to the Kinect sensor. These features and their time evolution are used as input to a Dynamic Bayesian Network, which classifies them into an emotion (Happiness, Sadness, Fear, Anger and Neutral).

The proposed system is based on a client-server model using the framework RoboComp [5]. The head pose and the list of points of the face model are transferred through a middleware to a client computer connected to the local network. This same client computer processes these data for extracting the user's facial features and classifying into a basic emotion. The



(a) (b)

Fig. 3. Data transferred: a) Candide-3 Model; b) The Positions of rotation and traslation of user's head.

proposal presented in this paper is explained in the following subsections.

A. Data Adquisition

The facial expression recognition system proposed in this paper uses RGB-D information acquired by Microsoft Kinect sensor. This RGB-D camera simultaneously provides two different images, a depth image and a color image (640x480) at real-time frame rates. Kinect sensor is optimized to work from a close range as 0.5m to 3m, which allows to use it for typical Human Robot Interaction.

B. Face tracking algorithm

Let S being a sequence of RGBD image acquired by the robot in a real interaction, and let $I(t)$ being a frame of this sequence at instant of time t. *Microsoft Kinect SDK*, by means of its *Face Tracking* algorithm, is used for detecting and tracking the human face in real-time. The outputs of the algorithm are the head's pose, H^t, and the list of points associated to the Candide-3 mesh model, P^t.

Candide-3 is a face deformable model based on a low number of polygons, controlled by global and local Action Units (AUs) [3]. It consists of 113 interconnected nodes forming 168 triangles and depicts a face in neutral posture and expression. In Fig. 3, the outputs of the SDK face tracking algorithm is illustrated. Fig. 3(a) draws the candide-3 model. A list of interconnected points $P^t = \{p_i | i = 1...113\}$ is obtained by the algorithm. Fig. 3(b) illustrates the head pose estimate by the system, which is described by $H^t = \{R, T, s\}$, being R and T the rotation and translation of the user respect to the sensor, and s a scale factor.

C. Communication between components

As was afore-mentioned, the proposed system uses a client-server model. First, the head's pose in conjunction with the Candide-3 mesh model are obtained through a computer with the *Microsoft Kinect SDK* (running in Windows). Then, this computer performs the data transfer through the *Ice* middleware to other RoboComp components running in Linux. A detailed description of the process is described in the following stages:

- First, the data acquired by the RGB-D sensor is processed by the *SDK face tracking* algorithm. H^t and P^t are the outputs of the algorithm.
- A communication module that allows transfering this information to other computers and processes has been implemented, through the *Internet Communications Engine* middleware (Ice).
- Finally, a client RoboComp component (running in Linux) receives the information from the middleware. Through the retrived information, the Candide-3 face model is reconstructed, and the data of the face position are processed.

In summary, the data transferred through the middleware to other computers for processing are:

- RGBD Sensor : The RGBD image information, as well as its width, height and number of bits per pixel, from the Kinect sensor.
- 2D mesh and point : A list of 2D points obtained from the mesh, based on Candide-3 deformable model.
- 3D head Pose : The rotation and traslation matrices, and a scale factor of the user's head.

D. Invariant Facial Feature Extraction

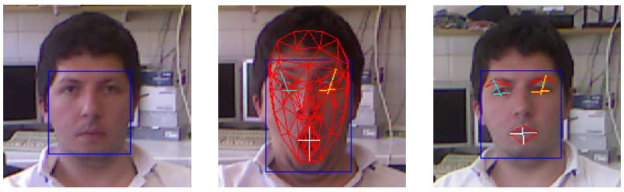
Facial feature extraction algorithm proposed in this work is based on the analysis of the Candide-3 mesh model. The algorithm extracts a set of m relevant features from the retrieved Candide-3 mesh model, F^I=$\{f_i^I | i = 1...m\}$, which are directly related to the Action Units (AUs) described in the Facial Action Coding System (FACS) [2]. The AUs used in the proposal possess independent and antagonistic qualities, as shown in Fig. 4 (*e.g.,* AU24 and AU25 are related to the distortions caused by the aperture of the mouth).

The facial features used in this work are obtained directly from P^t, the mesh model of the face. These features f^I are calculated using the Euclidean distances between different points in the mesh: i) the upper contour of the eyebrows and lower edge of the eyes, d_{eb}; ii) lip corners, d_{lc}; iii) upper contour and lower edge of the lips (mouth's aperture), d_{ma}; and iv) cheek, d_{ch}. In order to become independent of the limitations as the scale or distance from the user to the sensor, all these features are normalized by using the values extracted in the non-emotional state (*Neutral*). In Fig. 5(a), an input RGBD image is acquired by the sensor. Fig. 5(b) shows the Candide-3 model displayed over the face. This image is obtained by using the Ice middleware described in Section III-C. Through the mesh model, a set of facial features are easily detected as illustrated in Fig. 5(c), labeled as: EB d_{eb} (yellow), Ch d_{ch} (yellow), LC d_{lc} (white) and MA d_{ma} (white). The main limitations of the proposed facial features extraction subsystem (*e.g.,* the limit distance in order to detect features) are due to the own limitations of the *SDK face tracking* algorithm.

E. Dynamic Bayesian Network

The Dynamic Bayesian Network DBN implemented in this system, uses antagonistic properties in some AUs to increase

Fig. 4. Action Units (AUs).



(a) (b) (c)

Fig. 5. facial feature extraction based on Candide-3 model: a) Face detection ; b) Mesh model of the face; and c) Features extracted in the mesh model of the image.

the performance and reduce the number of variables to be considered in the network. The proposed system uses only 11 Action Units, to reduce and optimize the information processing. These 11 AUs will be grouped according to antagonistic and exclusive properties into only 7 variables. The two-level network structure is illustred in Fig. 6. The time influence that characterizes this networks as a dynamic bayesian networks is also represented in Fig. 6.

The first level of the DBN contains the belief variable *FE*, that representing the emotional state resulting from the classifier. Each facial expression is associated with a possible emotional states of the user such as: $FE_{[Neutral]}$, $FE_{[Anger]}$, $FE_{[Sad]}$, $FE_{[Happy]}$ and $FE_{[Fear]}$. The variables in the second level have as parent this one in the first level: *FE*.

The second level consists of 7 belief varibles:

- EB: $\{AU1, AU4, none\}$; stands for *Eye-Brows*.
- Ch: $\{AU6, none\}$; stands for *Cheeks*
- LE: $\{AU7, none\}$; stands for *Lower Eyelids*
- LC: $\{AU12, AU15, none\}$ stands for *Lips Corners*
- CB: $\{AU17, none\}$ stands for *Chin Boss*
- MF: $\{AU20, AU23, none\}$ stands for *Mouth's Form*
- MA: $\{AU24, AU25, none\}$ stands for *Mouth's Aperture*

The DBN required to be supplemented with learning data from the values of these 7 variables in the second level for each emotional state. Usually, in each new probability a threshold is used to find the matches. But, in this case to avoid the extant gaps, a pre-processing stage is done before the learning stage, fitting a Gaussian distribution to the data. In the process of detection of the AUs, the mesh model performs a tracking of the muscles of the face. These movements don't affect significantly to other nearby muscles, allowing these 7 variables to be independent given the facial expression *FE*.

The data *D* obtained in the facial feature extraction algorithm proposed in this paper, have the following set up:

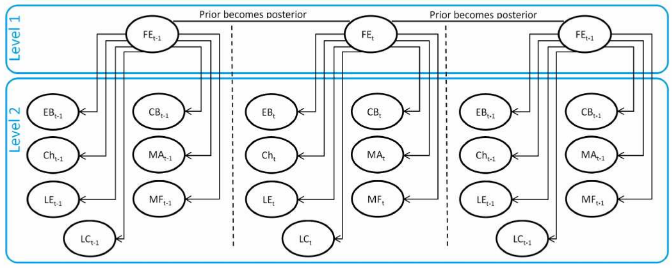


Fig. 6. Dynamic Bayesian Network, three time intervals are shown.

$$D = ((x_1, y_1)...(x_n, y_n)), x_i \in \mathbb{R}^d, y_i \in \mathbb{R} \quad (1)$$

Consider that y_1 to y_5 are the five possible emotional states ($FE_{[neutral]}, FE_{[happy]}, FE_{[sad]}, FE_{[fear]}, FE_{[anger]}$); 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

$$X_i \sim N(prior^T x_i, \sigma^2) \quad (2)$$

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

Finally, by using the Bayes's rule, we have the posterior equation:

$$P(FE|x_m) = \frac{\prod_1^n P(x_i|FE) * P(FE)}{P(x_m)}, \quad (3)$$

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

$$P(x_m) = \sum_{FE} \prod_1^k P(x_i|FE) * P(FE), \quad (4)$$

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

This model includes a dynamic convergence property over time. The resultant histogram from the previous frame is passed as prior knowledge for the current frame. The maximum number of frames for convergence has been limited to 5. If the convergence reaches a 80% threshold before 5 frames, the classification is considered complete. If not it keeps converging up to the fifth frame. If the fifth frame is reached and 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 classification result. The threshold is used as a control measure for the classification errors generated in the detection of the Action Units (AUs).

In Fig. 7, the iteration axis illustrates the 5 utterances that happen to each results. The expression axis draws the scope of possible expressions. Notice that the sum of probability at each iteration among the five possible expressions is always 1. In examples (a), (b) and (c), respectively, inputs were given for neutral, fear, happy, sad and anger; the Dynamic Bayesian Network was capable of classifying the expected expression

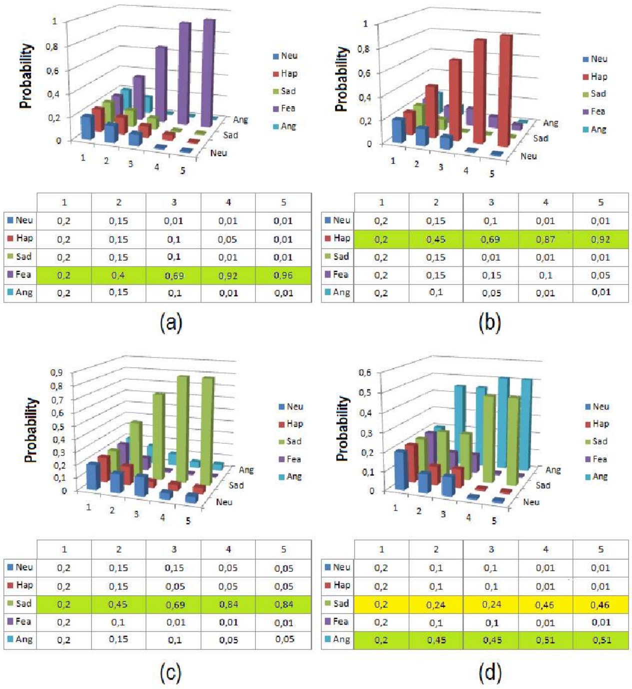


Fig. 7. Results from the classified for different facial expressions.

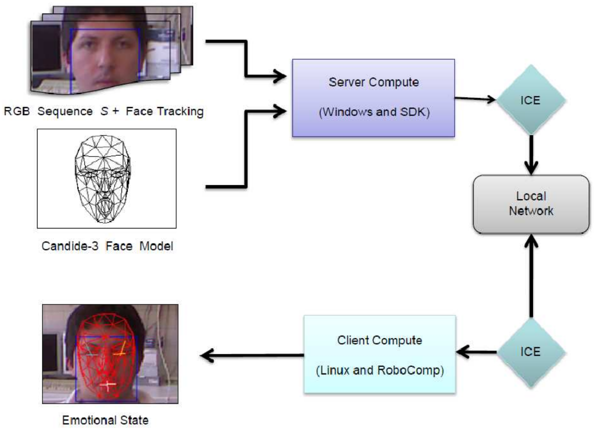


Fig. 8. Architecture of the system in the tests

with a fast convergence. In the Fig 7(d), an example of ambiguity is illustrated, As shown in the figure, the difference between emotional states: Sad (yellow) and Anger (green), is limited. For this example, the expected result exceeds the threshold, being a correct detection.

IV. Experimental Results

In this section, a set of test has been achieved in order to evaluate the performance of the facial expression recognition system. These tests are divided into two different modalities of evaluation: In the first modality, it will conduct a series of tests with different types of users to evaluate the performance, robustness and effectiveness of the proposed system. The second modality will perform a comparative study of the proposed algorithm with the previous facial expression recognition system [6].

Test	Percentage of correctly detected facial expressions (r)
Sad	90%
Happy	98%
Fear	95%
Anger	95%
Neutral	92%

TABLE I
Robustness of the facial expression recognition system

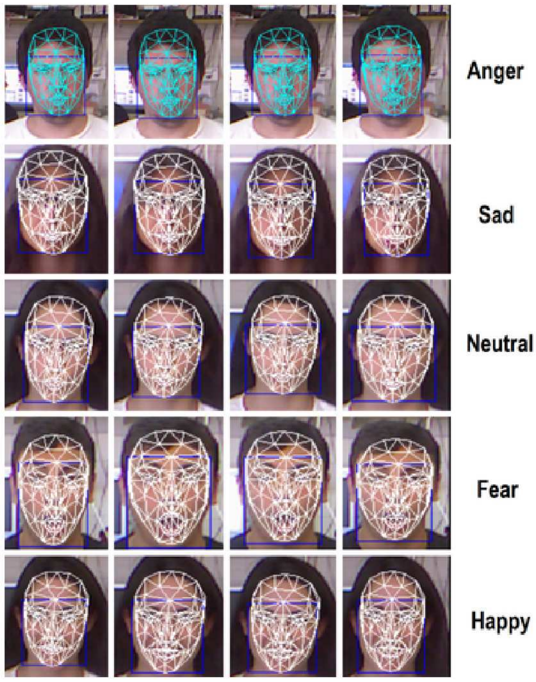


Fig. 9. Users with different gender and facial features in the tests

A. Evaluation

The first test is based on the estimate of the robustness of the facial expression recognition system. For this test, a server with an Intel(R) Core(TM) i3 CPU running Windows and the SDK, was used to transfer the Candide-3 face model and 3D user pose data, to another computer on the local network, using the *Ice* middleware. The client computer on which the tests were performed has a 2.8 Ghz Intel(R) Core(TM) i7 CPU and 4Gb RAM running Linux. This computer receives data from the server, using the framework Robocomp that controls the facial expression recognition system. The architecture of the proposed system is illustrated in Fig. 8.

To test the robustness and homogeneity of the results, a group of 20 users with different gender and facial features has been used (as illustrated in Fig. 9). For each test the user must perform 5 random sequences of facial expressions. In table I, the percentages of correctly detected facial expression (r) are shown.

B. Comparative Study

The second test is achieved in order to compare the system proposed in this paper with the method described in [6]. In order to evaluate these two methods, they have been used

Test	Percentage of improvement in the detection of FE (p_i)
Sad	15%
Happy	10%
Fear	4%
Anger	16%
Neutral	9%

TABLE II
PERCENTAGE OF IMPROVEMENT OF THE FACIAL EXPRESSION RECOGNITION SYSTEM

under similar conditions in each sequence (*e.g.,* users, lighting conditions and distance from the user to the sensor). In this second test a group of 15 users with different gender and facial features, must perform 5 random sequences of facial expressions with each method. The results of this comparison are shown in Table II, where the percentages of improvement in the detection of each facial expression p_i are illustrated. The proposed system in this work shows significant improvements over the previous system [6] in each emotional state, especially in the emotional states associated with high activation levels (i.e., happy, anger or sad). This is because the states with a high level of activation requiring high muscular activity in the face, resulting in a remarkable deformation in the mesh model [7]. Besides the above-mentioned aspects of this new system, it is also important to mention an improvement in the user detection and features extraction. Besides, it prevents detection errors. In summary, the main improvements of this new system in relation to the previous system are the minimum amount data on the trainer, the better detection of the neutral state among others emotional states and the elimination of errors in the user detection.

V. CONCLUSION

In this paper, a real-time facial expression recognition system has been presented. This proposed system is based on the use of the Candide-3 face model, which allows the extraction of a set of facial features, as well as rotations and translations of the user's face, from a RGBD image. Through the face model, the standardized variables associated to the AUs of the user are obtained, allowing real-time recognition of each facial expression with different factors, such as: lighting conditions, gender, unusual facial features (like injuries or scars), among others. In relation to the previous method [6], the results of the comparative study show better performance for the proposed system in uncontrolled environments, in addition to an improved capability for detecting and tracking of the user's head, which eliminates the errors of detection in these type of system.

Future work will focus on the extraction of more facial features, which would allow to recognize a higher number of emotional states associated with these features. Besides, it is planned to integrate the proposed system into a multimodal system that allows analyzing not only facial expressions, but also the user's body language.

ACKNOWLEDGMENT

This work has been partially supported by the MICINN Project TIN2012-38079-C03-01.

REFERENCES

[1] R. W. Picard, "Affective Computing". MIT Press, pp. 88-91, 2000.
[2] P. Ekman, WV Friesen, JC Hager, "Facial Action Coding System FACS", The manual, 2002.
[3] Jrgen Ahlberg, "CANDIDE-3 - an updated parameterized face", Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linkping University, Sweden, 2001.
[4] 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.
[5] 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.
[6] F. Cid, J.A. Prado, P. Bustos and P. Núñez. "Development of a Facial Expression Recognition and Imitation Method for Affective HRI", In Workshop of Physical Agents, pp. 119-124, 2012.
[7] G. Caridakis, K. Karpouzis, M. Wallace, L. Kessous, N. Amir. "Multimodal user's affective state analysis in naturalistic interaction", In Journal of Multimodal User Interfaces 3(1-2), pp. 49-66, 2010.
[8] H.-B. Deng, L.-W. Jin, L.-X. Zhen and J.-C. Huang. "A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA", International Journal of Information Technology. Vol.11, no.11. pp. 86-96, 2005.
[9] T. Fukuda, M.-J. Jung, M. Nakashima, F. Arai and Y. Hasegawa, "Facial Expressive Robotic Head System for Human-Robot Comunication and Its Application in Home Environment", In *Proc. of the IEEE* , Vol. 92, No. 11, pp. 1851-1865, 2004.
[10] K.-E. Ko, K.-B. Sim, "Development of a Facial Emotion Recognition Method based on combining AAM with DBN", in *International Conference on Cyberworlds 2010*, pp. 87-91 ,2010.
[11] T. Chen, "Audio-Visual Integration in multimodal Communication". In *IEEE Proceedings*, May, 1998.
[12] R. Cowie and R. Cornelius. "Describing the emotional states that are expressed in speech", Speech Communication 40, pp. 5-32, 2003.
[13] 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.
[14] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art", Tech Report, College of Computing, Georgia Institute of Technology, 2012.
[15] S. S. Ge, C. Wang, C.C. Hang, "Facial Expression Imitation in Human Robot Interaction", In *Proc. of the 17 IEEE International Symposium on Robot and Human Interactive Communication*, Germany, pp. 213-218, 2008.
[16] B.Y.L. Li, A.S. Mian, W. Liu and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise", In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, IEEE Computer Society 2013, ISBN 978-1-4673-5053-2, USA, pp.186-192,2013.
[17] M. Hernandez, J. Choi and G. Medioni, "Laser Scan Quality 3-D Face Modeling Using a Low-COST DEPTH CAMERA", In *20th European Signal Processing Conference (EUSIPCO 2012)*, pp. 1995-1999, Romania, 2012.
[18] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," In *ISMAR 2011, 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 127-136, USA, 2011.
[19] L. Kessous, G. Castellano, G. Caridakis. "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis", In Journal on Multimodal User Interfaces, Vol. 3, No. 1. (2010), pp. 33-48.
[20] Face Tracking, http://msdn.microsoft.com, last visited: 07.06.2013
[21] L.V. Calderita, L. Manso, P. Bustos and P. Nez. "Assessment of PrimeSense RGB-D camera for robotic applications", In Workshop on Physical Agents, 2012.