Imitation System for Humanoid Robotics Heads

F. Cid, J.A. Prado, P. Manzano, P. Bustos and P. Núñez

***Abstract*—This paper presents a new system for recognition and imitation of a set of facial expressions using the visual information acquired by the robot. Besides, the proposed system detects and imitates the interlocutor's head pose and motion. The approach described in this paper is used for human-robot interaction (HRI), and it consists of two consecutive stages: i) a visual analysis of the human facial expression in order to estimate interlocutor's emotional state (i.e., happiness, sadness, anger, fear, neutral) using a Bayesian approach, which is achieved in real time; and ii) an estimate of the user's head pose and motion. This information updates the knowledge of the robot about the people in its field of view, and thus, allows the robot to use it for future actions and interactions. In this paper, both human facial expression and head motion are imitated by Muecas, a 12 degree of freedom (DOF) robotic head. This paper also introduces the concept of human and robot facial expression models, which are included inside of a new cognitive module that builds and updates selective representations of the robot and the agents in its environment for enhancing future HRI. Experimental results show the quality of the detection and imitation using different scenarios with Muecas.**

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

I. Introduction

Human Robot Interaction (HRI) is one of the most important tasks in social robotics. In the last decades, HRI has become an interesting research where different untrained users interact with robots in real scenarios. Most of the HRI methodologies use non-invasive techniques based on natural language (NL), in a similar to the way people interact in their daily life. Regarding this, verbal communication (speech, among others) or non-verbal communication (corporal language, gestures or facial expressiveness) have been successfully used for enhancing empathy, attention or understanding of social skills in a human-machine interaction [1], [2].

Social robots are usually designed in order to enhance the empathy and the attention of the HRI [3]. Thus, human shaped robots are typically used for decreasing the gap between the machine and the human communication styles. Besides, it allows the robot to adapt itself to the emotional state of the human interlocutor, which could be used for different purposes in a social affective communication. In order to have an efficient HRI, not only the robot shape is important, but also the knowledge of different elements of the human interlocutor state: pose in the environment, number of interlocutors in the scenario or emotional state, among others. To acquire this information, several techniques and methodologies have been studied and applied, such as facial expression recognition [15], skeletal modeling [17], use of corporal language [19] or speech recognition [20].

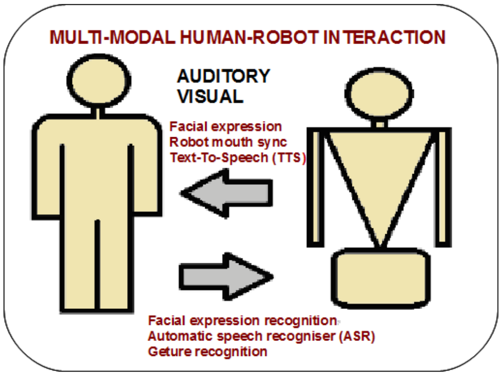

MULTI-MODAL HUMAN-ROBOT INTERACTION
AUDITORY
VISUAL
Facial expression
Robot mouth sync
Text-To-Speech (TTS)
Facial expression recognition
Automatic speech recogniser (ASR)
Geture recognition

Fig. 1. Multi-modal HRI is usually based on visual and auditory information.

Therefore, in order to interact with people, robots have to be able to perceive and share information with them using visual and auditory messages. Natural language, in conjunction with visual information is a very efficient method for an interaction paradigm with robots (see Fig. 1). On one hand, facial expression recognition provides an estimate of the interlocutor's emotional state through the understanding of visual information, providing support to the emotional responses of a robot inside a social dialog through audio media or visual aids creating a feedback for the content of the dialog [8]. In fact, interactive NL-based communication provides a fast feedback that is successfully used for handling errors and uncertainties. On the other hand, the human behavior imitation has been used for learning tasks and for enhancing the human-robot communication. Imitation of motions and emotions plays an important role in the cognitive development, and has been studied in the last year in social robotics [7], [6]. Both visual and auditory informations are used for mimicking human expressions as a mean of developing social and communication skills. Among social robots, the robotic head (e.g., Kismet[9], Saya [10] or WE-4RII [21]) mainly imitates facial expressions and corporal language, through the modification of the poses of different mechanical elements, such as eyes and mouth.

The imitation of corporal language not only depends on an accurate estimate of the user's pose, but also on tracking its motion. Most of the studies do not present solutions for uncontrolled environments because they need of a previous training with the user, or a high computational cost [11], [12]. The use of methods for estimating the pose and motions of the user's head allows other algorithms, such as the facial expression recognition algorithm, to obtain information to prevent errors in the detection or classification.

The proposed approach presents an imitation system which

F. Cid, P. Manzano, P. Bustos and P. Núñez are with University of Extremadura.
E-mail: fecidb@alumnos.unex.es
J.A. Prado is with University of Coimbra, Portugal
E-mail: jaugusto@isr.uc.pt

consists of two consecutive stages. First, a facial expression recognition system that allows detection and recognition of four different emotions (happiness, sadness, anger and fear) besides of the neutral state is presented. This system is based on a real-time Bayesian classifier where visual signal is analyzed in order to detect the expressivity of the interlocutor. The second part is a system that allows the robot to estimate the user's head pose and motion. The imitation system is developed and presented in this approach, where a robotic expressivity model is used as a bridge between the human expressivity and the final robotic head. This model is part of a new cognitive module that is able to build selective representations of itself, the environment and the agents in it. Finally, a set of experiments using a Muecas robotic head has been carried out in order to present and comment the results of the recognition and imitation systems.

This paper is organized as follows: In Section II, previous works in facial expression recognition, estimate of position and imitation systems are briefly described. Next, Section III presents the emotional state models associated to both interlocutors, robot and human, which are integrated inside the cognitive architecture for the proposed social robot. In Section IV, an overview of the proposed imitation system is presented. In Section VIII, experimental results are pointed out, and finally, Section IX describes the conclusions and future work of the presented approach.

II. Previous Works

To achieve affective Human-Robot Interaction, this paper primarily focuses on presenting different methodologies commonly used for facial expression recognition and imitation. In the automatic recognition of emotions is necessary multimodal, that is, it requires of verbal and non-verbal channels (face, gesture, body language), physiological signals or midterm activity modeling, among others [15], [23], [22]. One of the most significant works used by the scientific community in facial expression recognition using visual information is based on Paul Ekman's study [13], [14]. This author identifies and classifies the facial expressions through the study of different facial muscles in each expression, giving rise to the so-called Facial Action Coding System (FACS). The recognition of facial expressions is a very diversified field in its classification or detection methods, ranging from the use of active Appearance Models (AAM) [24], Support Vector Machines (SVM) [6], Gabor filter bank [25] and Dynamic Bayesian Network (DBN).

On the other side, several authors use robots in domestic environments with untrained users or people with disabilities [4], [5]. In these works, authors achieve a natural HRI through the generation of facial expressions by the robot with the goal of maintaining a level of empathy and emotional attachment to robots [3]. These facial expression and emotion generation methods differ in the amount of facial expression that is possible to generate by the robot due to physical constraint [21]. In robotic heads with human-like characteristics such as the robotic head used in this paper, different works provide solutions for emotion generation depending on their physical constraints [9].

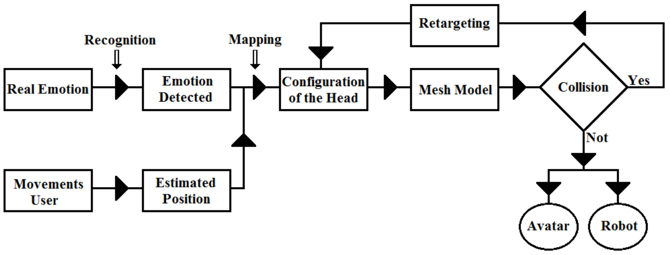


Fig. 2. Description of the human emotion module within the proposed cognitive architecture

Besides, the use of humanoid robotic heads in the HRI promotes the imitation of not only facial expressions as the position estimate for subsequent imitation of movements by the robotic head. In many studies this estimate depends on a previous training or marks [11], [12], showing poor performance on actual tests. Other studies using specific points (nose, mouth or eyes) [27], [26], presented better results but with a low stability.

Finally, robot's capability of imitating facial expressions and movements determines the design of the heads used in social robotics. Usually, imitation of facial expressions is achieved through mobile elements of the head (e.g., eyelids, eyebrows, eyes or mouth) [6], [21].

III. Emotional state modelling

The proposed approach is part of a new robotics cognitive architecture that builds selective representations (i.e. models) of the robot, the environment and the agents in it. This cognitive architecture performs internal simulations over these models to anticipate the outcome of future actions and interactions (e.g., safe navigation or path-planning, grasping of objects or more complex interactions) as shown in Fig. 2. The robotic head is represented by a mesh model used to avoid collisions between the models. Model-based representations of reality to help social robots achieve their tasks have been used in the last years with interesting results [17]. In order to achieve an affective HRI, non-contact interaction is modelled in the cognitive architecture, including movements, gesture and facial expression recognition, and detection of human emotional state. This last model is presented in this paper. Thus, human and robot emotional state models are similar and defined as: $M_{\{robot,human\}} = \{(m_0,p_0),(m_1,p_1),...(m_5,p_5)\}$, where m_i represents an emotional state $m_i = \{happy, sad, anger, fear, neutral\}$ for i= 1 to 5 and p_i the probability of this emotional state $0 \leq p_i \leq 1$ and $\sum_{p_i} = 1$. Both models M_{robot} and M_{human} will be updated once the facial expression and emotional state have been estimated by the proposed system.

IV. Imitation System

In this paper, an imitation system is presented. This system consists of two parts: a facial expression recognition system and estimate position and movements system, described in fig 3. The robot has a firewire camera in each eye, that allow it

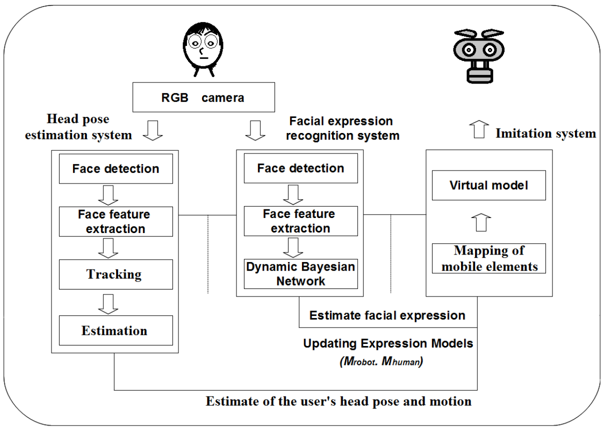

RGB camera
Head pose estimation system
Facial expression recognition system
Imitation system
Face detection
Face feature extraction
Tracking
Estimation
Face detection
Face feature extraction
Dynamic Bayesian Network
Virtual model
Mapping of mobile elements
Estimate facial expression
Updating Expression Models
(M_{robot}, M_{human})
Estimate of the user's head pose and motion

Fig. 3. Overview of the Imitation system proposed in this paper.

to obtain visual information for user detection. The imitation system can imitate the facial expression and movements of the user's head, through Muecas robotics head.

V. Facial Expression Recognition

In the first part, a real-time facial expressions recognition system is presented. This system will be integrated inside a cognitive architecture as a new module that provides a representation of the agent's and robot's emotional states. The proposed approach is described in Fig. 3. The robot acquires the information using a firewire camera inside the robot's eyes. This measurement is preprocessed in order to estimate the pose of the face in the robot's surrounding. Then, once the region of interest (i.e human face) is detected, the system extracts facial features for the subsequent classification task. This is achieved using a Dynamic Bayesian Network (DBN), allowing the robot to recognize the emotional state associated to the facial expression. In the next stage, the system updates the emotional state of the agents in the communication (self and interlocutor emotional state models), and finally, in the imitation system the facial expression is played by the Muecas robotic head's avatar.

A. Facial expression recognition system

In the design of the classifier, our own interpretation of FACS (Facial Action Coding System) is used, which drives a set of random variables different from those defined by other researchers. Each facial expression is composed by a specific set of Action Units (Fig. 4). Each of these Action Units is a distortion on the face induced by small muscular activity. Normally, a well determined set of face muscles is associated to a specific Action Unit, which can give the idea that all these basic distortions are independent. Nevertheless, some of these Action Units are antagonistic. One clear and understandable example is the case of two Action Units related with the lips corners: AU12 and AU15. When performing AU12, lips corners are pulled up. Contrary, when performing the AU15, the lip corners are pulled down. Therefore, the movements of the lip corners could be considered independent because they are performed by distinct muscle sets. However, when analyzed visually they are antagonistic and exclusive.

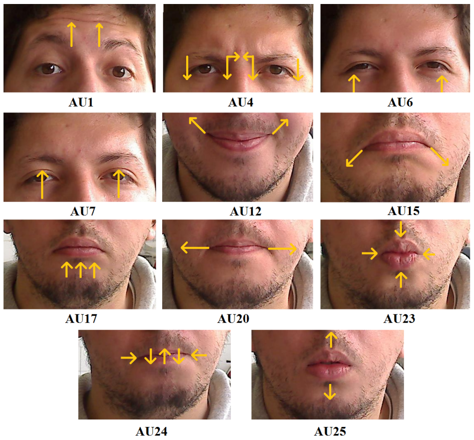


Fig. 4. Action Units (AUs)

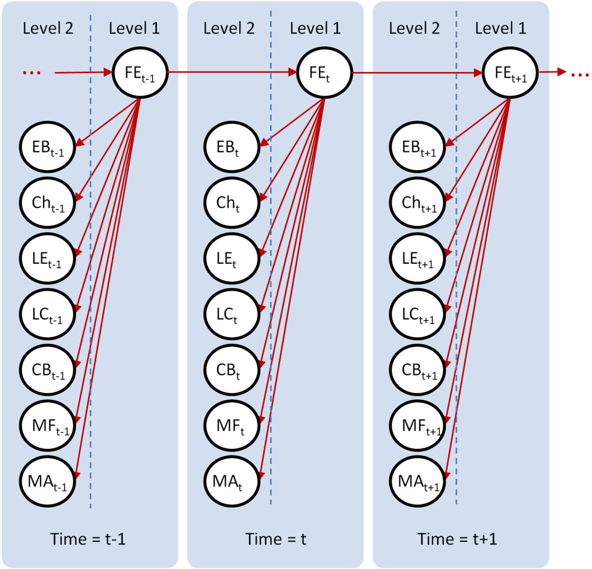


Fig. 5. Facial Expression Dynamic *Bayesian network,*three time intervals are shown ($t-1$, t, $t+1$).

The state space is assumed to be discrete, and in this case, hidden Markov models (HMM) can be applied. A hidden Markov model can be considered to be an instantiation of a dynamic Bayesian network and thus exact inference is feasible. Based in these principles, belief variables were defined and a dynamic Bayesian classifier of facial expressions was developed.

Facial Expression dynamic Bayesian network: The DBN takes advantage of the existing antagonism in some AUs to reduce the size of the dynamic Bayesian network. Thus, instead of using the 11 AUs as leafs for our DBN (Dynamic Bayesian Network), 7 variables are proposed. These variables group the related antagonist and exclusive Action Units. The two-level network structure is illustrated in figure 5. The time influence that characterizes this network as a dynamic Bayesian network is also represented in figure 5.

In the first level of the *dynamic Bayesian network*'s there is only one node. The global classification result obtained is provided by the belief variable associated to this node: $FE \in \{Anger, Fear, Sad, Happy, Neutral\}$, where the variable name stands for Facial Expression. Considering the structure of the dynamic *Bayesian network*, the variables in the second level have as parent this one in the first level: FE.

In the second level there are seven belief variables:

- $EB \in \{AU1, AU4, none\}$ is a belief variable related to the *Eye-Brows* movements. The events are directly related to the existence of AU1 and AU4.
- $Ch \in \{AU6, none\}$ is a belief variable which is related to *Cheeks* movements; more specifically, the events indicate if the cheeks are raised (AU6 is performed).
- $LE \in \{AU7, none\}$ is a belief variable which is related to the *Lower Eyelids* movements; AU7 is associated to lower eyelids set to up.
- $LC \in \{AU12, AU15, none\}$ is the belief variable related to the movements of the *Lips Corners*. When the corners did not perform any movement then the event *none* has a high probability. The event *AU12* has a big probability when the corners of the lips are pulled up. If the lip corners moves down the event *AU15* must have a big probability.
- $CB \in \{AU17, none\}$ is the belief variable collecting the probabilities related to the *Chin Boss* movements. The event *none* is related with the absence of any movement, while the event *AU17* had a great probability when the chin boss is pushed upwards.
- $MF \in \{AU20, AU23, none\}$ is the belief variable related to the Mouth's Form. The events *AU20* and *AU23* indicated, respectively, if the mouth is horizontally stretched or tightened.
- $MA \in \{AU24, AU25, none\}$ is the belief variable related to the Mouth's Aperture. The events *AU24* and *AU25* are related, respectively, with lips pressed together or with lips relaxed and parted.

The movements performed by the human in one area of the face can slightly affect muscles on other areas. However, this influence is very small and cannot be detected by the cameras of the robot. Thus, conditional independence among the 7 proposed variables was assumed.

The following equations illustrate the joint distribution associated to the Bayesian Facial Expressions Classifier.

$$\begin{aligned} &P(FE, EB, Ch, LE, LC, CB, MF, MA) = \\ &P(EB, Ch, LE, LC, CB, MF, MA|FE) * P(FE) = \\ &P(EB|FE) * P(Ch|FE) * P(LE|FE) * P(LC|FE) * \\ &P(CB|FE) * P(MF|FE) * P(MA|FE) * P(FE) \end{aligned} \tag{1}$$

The last equality is written assuming that the belief variables in the second level of the dynamic *Bayesian network* are independent.

From the joint distribution, the *posterior* can be obtained by the application of the Bayes rule as follows:

$$\begin{aligned} &P(FE|EB, Ch, LE, LC, CB, MF, MA) = \\ &P(EB|FE) * P(Ch|FE) * P(LE|FE) * P(LC|FE) * \\ &\ P(CB|FE) * P(MF|FE) * P(MA|FE) * P(FE) / \\ &\ P(EB, Ch, LE, LC, CB, MF, MA) \end{aligned} \tag{2}$$

From the Bayesian marginalization rule we can calculate:

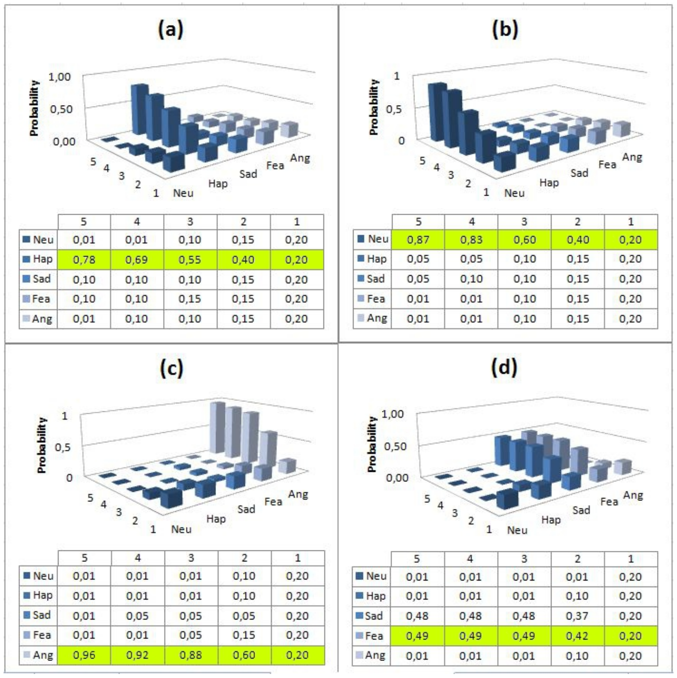
$$\begin{aligned} &P(EB, Ch, LE, LC, CB, MF, MA) = \\ &\sum_{FE} P(EB|FE) * P(Ch|FE) * P(LE|FE) * \\ &\quad P(LC|FE) * P(CB|FE) * P(MF|FE) * P(MA|FE) * P(FE) \end{aligned} \tag{3}$$

As a consequence of the dynamic properties of the network, convergence happens along 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 (Fig 6). 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 figure 6, camera grabbing was set to 5 fps (for the initial tests), therefore, the iteration axis represents the 5 utterances that happen in one second. The expression axis is the selected 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 happy, neutral and anger; the dynamic Bayesian network was capable of classifying the expected expression with a fast convergence. In (d), an example of ambiguity and misclassification is shown, where the expected result was sad but the result of classification was fear. In the presented example the obtained results are robust for the number of states of the system, as shown in Fig 6.

VI. User's head pose estimation

In the second stage of the proposed imitation system, a human head pose estimation system is presented. An overview of the approach is described in Fig. 7. The robot obtains the visual information from the firewire camera built in the robot's left eye. There is a first face-detection stage where the biggest visible human face is picked as target. Then, a set of key points is extracted from the image, mapped to a 3D surface and tracked along time in order to get a head pose estimation from each new frame. The human face detection is accomplished using Viola-Jones' cascade classifiers algorithm [18]. After a human face is detected, an alignment check is performed in order to guarantee that the face is looking straight to the camera, with no rotation over any of its three axis. Also, the region given by Viola-Jones's algorithm is clipped to keep only the "central" part of the face in order to avoid certain parts of the human face that may be difficult to track, such



(a)

	5	4	3	2	1
Neu	0,01	0,01	0,10	0,15	0,20
Hap	0,78	0,69	0,55	0,40	0,20
Sad	0,10	0,10	0,10	0,15	0,20
Fea	0,10	0,10	0,15	0,15	0,20
Ang	0,01	0,10	0,10	0,15	0,20

(b)

	5	4	3	2	1
Neu	0,87	0,83	0,60	0,40	0,20
Hap	0,05	0,05	0,10	0,15	0,20
Sad	0,05	0,10	0,10	0,15	0,20
Fea	0,01	0,01	0,10	0,15	0,20
Ang	0,01	0,01	0,10	0,15	0,20

(c)

	5	4	3	2	1
Neu	0,01	0,01	0,01	0,10	0,20
Hap	0,01	0,01	0,01	0,10	0,20
Sad	0,01	0,05	0,05	0,05	0,20
Fea	0,01	0,01	0,05	0,15	0,20
Ang	0,96	0,92	0,88	0,60	0,20

(d)

	5	4	3	2	1
Neu	0,01	0,01	0,01	0,01	0,20
Hap	0,01	0,01	0,01	0,10	0,20
Sad	0,48	0,48	0,48	0,37	0,20
Fea	0,49	0,49	0,49	0,42	0,20
Ang	0,01	0,01	0,01	0,10	0,20

Fig. 6. Results from facial expression classifier.

as beards, neck, ears or hair. Key point extraction is fulfilled using Good Features To Track algorithm, extracting the face's main corners. Once this first set of key points is extracted, each point is projected over the surface of a cylinder with a diameter equal to the image width. By doing this, a set of 3D points is created, which will be the main reference for the following pose estimation process. Thus, the face shape is being approximated to a cylinder. This may seem quite a rough approximation, but it has been proven to be good enough to work with any human test subject, instead of some more human-like projection surfaces. After these two initialization phases the initial key point set and the reference 3D model are built. Next, there will be one tracking phase performed by using Lukas-Kanade's optical flow algorithm, in which any not tracked key point will cause its correspondent 3D reference point to be dropped from the set. After this phase, if the percentage of lost key points has risen to a certain threshold the whole algorithm will be reset to avoid wrong estimations. If after the tracking phase the key point set is still big enough the pose estimation phase begins. The output of this phase is the pose estimation of the human head, meaning a translation vector T and a rotation matrix R. This estimation is performed using the POSIT algorithm. This algorithm calculates, after a number of iterations, which is the (R, T) transformation that, applied to the reference set of 3D points, will cause its projection to be as similar as possible to the current tracked key points.

VII. Imitation system in robotics head

Imitation is a key process in several social robotics applications as a means of developing social and communication skills (e.g learning or movement imitation in the context of HRI). In order to achieve a realistic imitation of the user,

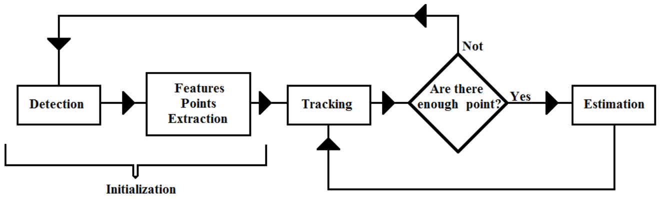


Fig. 7. Overview of the estimate pose and motion system proposed in this paper.

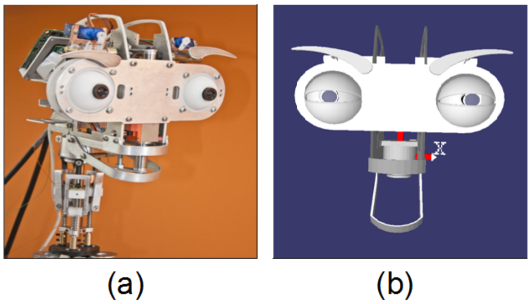


Fig. 8. a) 12DOF Robotics head Muecas; and b) Robotics Head Mueca's avatar.

this imitation system can recognize facial expressions, besides estimated position and movements of the head user. In most of facial expression mimicking approaches, visual and auditory information are used for achieving a multi-modal imitation system. In addition, it has been demonstrated that a more realistic communication derives from a robotic head with similar characteristics and movements to a human face [9]. Thus, a facial expression imitation system is described in this paper, where visual information is used to perform non-verbal communication in a more friendly and intuitive way using the Muecas robotic head (Fig 8a). with a graphical representation using virtual model (avatar) (Fig 8b). Besides, a mesh model is used for allowing the robot to be pro-active, by interpreting sensory information to predict the immediately relevant future inside the cognitive architecture.

1) Robotic head Muecas: The robotic head Muecas consists of 12 DOF and it has been designed by Iadex S.L in cooperation with RoboLab as a mean to transmit facial expressions and body lenguage for social robots [30][1]. One of the main goals in the design of Muecas was to imitate the movements and human emotional states according to the anatomy of the human head. Thus, for generation of facial expressions,the movement of the elements present in the recognition of facial expressions (e.g., eyes, eyebrows or mouth, among others) is similar to those of the human face, resulting in simplest and most natural imitations. Besides, for the imitation of the movements of the head user as: yaw, roll and pitch. The neck of the robot presents a combination of motors for imitating the human muscles, described in the fig 10.

Muecas has also its own virtual model, which consists of 16 DOF, with four degrees more than the real robotic

[1] For more information, you can visit www.robolab.unex.es

Emotion	AUs	Muecas' Component
Neutral	-	-
Happy	AU6-AU12-AU25	Eyebrows-Eyelids-Eyes-Mouth
Sad	AU1-AU4-AU15-AU17	Eyebrows-Eyelids-Eyes
Fear	AU1-AU4-AU20-AU25	Eyebrows-Eyelids-Mouth
Anger	AU4-AU7-AU17-AU23-AU24	Eyebrows-Eyelids

TABLE I
MOVEMENTS FOR THE ROBOTIC HEAD MUECAS' COMPONENTS ASSOCIATED TO THE EMOTION RECOGNITION

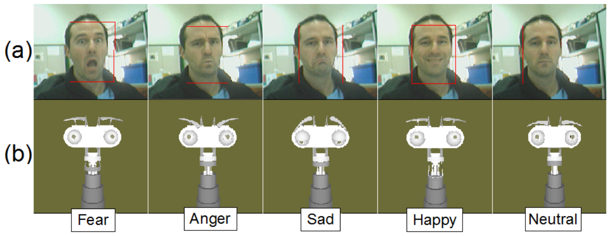


Fig. 9. a) Facial expression estimated by the recognition system; and b) Facial expressions imitated on the Muecas' avatar.

head (Eyelids). Besides, the mesh model of the robotic head is used as a bridge between the facial expression estimated by the system and the emotion reproduced by the robotic head, performing the necessary retargeting. That is, before generating facial expression and movements in the real robotic head, the system tries to generate all the cinematic chain of the mechanical motions and graphic representation of each imitated expression and movements of the head user through the avatar.

2) Facial Expression Generation: Facial expressions are detected and recognized using the recognition system described in Section V-A. Four different emotional states are estimated (i.e. Happiness, sadness, fear and anger), also a neutral state (i.e. no expression associated with an emotion). Fig. 9a illustrates the facial expressions estimated by the recognition system for different examples. These facial expressions are then mapped over the mesh model, modeling each one of the movements needed to generate the emotional state. Table I describes the set of mobile elements of the robotic head and the AUs for each emotion. In Fig. 9b the facial expressions generated by the mimicking system are illustrated using the virtual model of the robotic head.

3) Imitation of the user's head pose: The imitation system of the user's head motion is based on the detection and tracking system described in Section VI. Thus, the robot Muecas mimics the head pose according to the rotation and translation matrices (R, T) previously estimated by the system. In Fig. 11 the estimate of the user's head pose is shown. Four different examples are illustrated where different yaw, roll and pitch angles are modified. In fig. 10, the robotics head Muecas imitates the head pose and its motion. In order to generate the different movements of the head, the mapping over the mesh model prevents collisions and generate the cinematic chain of the mobile elements. Due to the mechanical constraint of the robot Muecas, in order to acquire the image data and after estimating and tracking the head pose, the motors of the robotic eyes dynamically changed the tilt and pan.

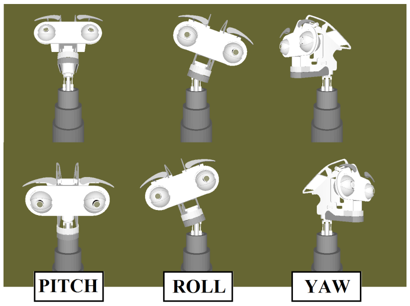


Fig. 10. Description of the robotics head movements using "Muecas" avatar.

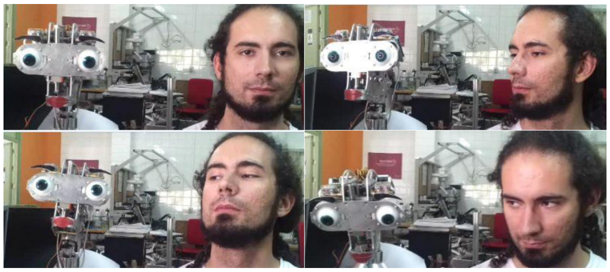


Fig. 11. Results of the User's head pose imitation system.

VIII. EXPERIMENTAL RESULTS

In this section, a set of tests has been performed in order to evaluate the effectiveness of the imitation system described in this paper. The software to control the system is built on top of the robotics framework *RoboComp* [28]. Making use of the components and tools it provides and its communication middleware, an easy to understand and efficient architecture has been developed.

The relationships between the different components used for the experimental setup have been drawn in Fig. 12. The main components of the proposed system are *MuecasemotionComp* and *Face3DTrackerComp*. They are connected, directly or indirectly, to the rest of the software components, such as: *camera* or *robotic head* among others (In the figure, not all components of the robotic head Muecas have been drawn to provide a simple explanation). The RGB image is provided by the *cameraComp* component, which sends it to *MuecasemotionComp* and *Face3DTrackerComp* components that estimate the facial expressions and pose of the interlocutor.

This components also updates: the movements of the robotic head, the robot and human emotional state models. that assigns the motion of each mobile element of the robotic head in order to generate a realistic facial expression or natural movements of the robotic neck. Then, *MuecasavatarComp* is used as a bridge between the imitation system (facial expression recognition and the estimated pose) and the robotic head Muecas (*MuecasComp*). Once the robotic head receives the positions of each mobile element, each motor commands are received and executed by its associated *dynamixelComp*.

Since the system was designed an implemented using component oriented design/programming, these components can

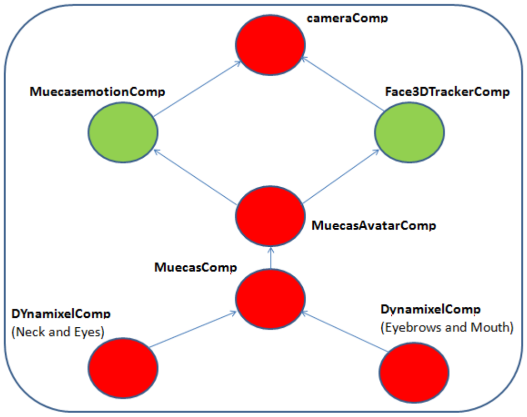


Fig. 12. Dependence Relationships between the different components used in the proposed approach.

Test	Percent of correctly detected facial expression (p_i)
Sad	74%
Happy	89%
Fear	95%
Anger	79%

TABLE II
ROBUSTNESS OF THE FACIAL EXPRESSION RECOGNITION SYSTEM

be easily used for other purposes, which is a very important feature in robotics development.

In order to evaluate the recognition and imitation of facial expressions, a set of experimental tests has been achieved in a real HRI scenario. A human interlocutor is located opposite to the robot, achieving different facial expressions (from sadness to happiness) in a continue mode. The proposed system is running on-line, acquiring and estimating the facial expressions in real-time (firewire camera acquires data at 25 fps). Then, the system updates the emotional state models (M_{robot}, M_{human}) and imitates the facial expression using Muecas in real-time too. These experiments are run 20 times with different interlocutors and generating different facial expressions. An example of the results are shown in the Table. II, where the evolution of each p_i is given. Robustness of the approach is given in Table II for the set of experiments achieved in the described scenario. As shown in Table II, the most of the facial expressions are correctly estimate.

The second part of the system are based on the estimate of the user's head pose and motion. For a correct evaluation of the system, a set of experimental tests were conducted to users with different ages and facial features. The tests consisted in the estimate and imitation of three basic movements of the user's head, such as: pitch, yaw and roll (i.e, movements around the pitch, yaw and roll axis, respectively). The movements were repeated 120 times using the robotics head Muecas, as is shown in Fig. 10. Finally, the tests demonstrated that movements, such as the pitch and roll movements, presented the best results with the robotics head. A summary of the results is illustrated in Table III.

Test	Percent of correctly estimate of pose
Pitch	80%
Roll	60%
Yaw	75%

TABLE III
ROBUSTNESS OF THE POSE AND MOTION ESTIMATION SYSTEM.

IX. CONCLUSION

In this paper, a imitation system for robotics head is presented. The imitation system consist of two parts: the first part is a system for recognition of the facial expression. First, a Dynamic Bayesian Network (DBN) structure has been used to classify facial expressions (happiness, sadness, anger, fear and neutral). This paper demonstrates the robustness of the solution for a common HRI scenario with different users and environmental conditions. Next, this facial expression is imitated in the robotic head Muecas, which has been designed for generating emotions. The full system has been incorporated in a social robot whose cognitive architecture has been pointed out in this paper. Thus, the robot and human emotional states are updated and tracked by the architecture in order to plan future actions and interactions. The second part estimated the motions and the pose of the user's head, allowing the robot to imitate the corporal language of the user and obtain the actual pose and orientation of the user's head.

Future works will be focused on a multi-modal interaction, where auditory information (e.g., speech or intensity) will be used in order to estimate the interlocutor's emotional state. This new module will be integrated in the architecture, taking into account the probabilities associated to each one of these emotional states. Besides, to achieve an affective HRI it would be interesting to study the empathy level of the presented solution in real scenarios with untrained interlocutors, and the use of more visual information about the user's body language to achieve a more natural behavior by the robot.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish Ministerio de Ciencia e Innovación TIN2011-27512-C05-04 and AIB2010PT-00149, and by the Junta de Extremadura projects IB10062. The authors also gratefully acknowledge support from Institute of Systems and Robotics at University of Coimbra (ISR-UC), Portuguese Foundation for Science and Technology (FCT) [SFRH/BD/60954/2009].

REFERENCES

[1] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll and L. Hall, "Caring for Agents and Agents that Care: Building Empathic Relations with Synthetic Agents", In *Third International Joint Conference on Autonomous Agents and Multiagents Systems*, Vol. 1, pp. 194-201, New York, USA, 2004.

[2] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive Robotics: The influence of robot gender on human behavior". In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2563-2568, October 2009.

[3] A. Tapus and M.j. Mataric, "Emulating Empathy in Socially Assistive Robotics", In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, Stanford, USA, March 2007.

[4] M.J. Matarić , J. Eriksson , D.J. Feil-Seifer1 and C.J. Winstein. “Socially assistive robotics for post-stroke rehabilitation”. In *Journal of NeuroEngineering and Rehabilitation*. 4:5. 2007.

[5] C. Jayawardena, I. H. Kuo, U. Unger, A. Igic, R. Wong, C. I. Watson, R. Q. Stafford, E. Broadbent, P. Tiwari, J. Warren, J. Sohn and B. A. MacDonald. “Deployment of a Service Robot to Help Older People”. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems* Taiwan, pp. 5990-5995, October 2010.

[6] 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.

[7] S. DiPaola, A. Arya, J. Chan, ”Simulating Face to Face Collaboration for Interactive Learning Systems”, In *In Proc: E-Learn 2005*, Vancouver, 2005.

[8] T. Chen, “Audio-Visual Integration in multimodal Communication”. In *IEEE Proceedings*, May, 1998.

[9] Kismet, Available at: *http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html*.

[10] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, “Development of the Face Robot SAYA for Rich Facial Expressions”. In *Proc. of 2006 SICE-ICASE International Joint Conference* Korea, pp. 5423-5428, October 2006.

[11] L. Guoyuan, Z. Hongbin, L.Hong, ”Affine Correspondence Based Head Pose Estimation for a Sequence of Images by Using a 3D Model”. In Proc. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR’04)* , pp. 632-637, 2004.

[12] D. DeCarlo, D. Metaxas, ”The Integration of Optical Flow and Deformable Models with Applications to Human Face Shape and Motion Estimation”, In Proc. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 231-238, 1996.

[13] P.Ekman, WV Friesen, JC Hager, ”Facial Action Coding System FACS”, the manual, 2002.

[14] P.Ekman, E. Rosenberg, ”What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)”, 2nd edn. Oxford Press, London.

[15] J.A. Prado, C. Simplcio, N.F. Lori and J. Dias, ”Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction”, In *International Journal of Social Robotics*, vol. 4, no. 1, pp. 29-51, December 2011.

[16] W. Zhiliang, L. Yaofeng, J. Xiao, “The research of the humanoid robot with facial expressions for emotional interaction”. In Proc. *First International Conference on Intelligent Networks and Intelligent Systems*, pp. 416-420. 2008.

[17] J.P. Bandera, “Vision-Based Gesture Recognition in a Robot Learning by Imitation Framework”, Ph.D. Thesis, University of Malaga, 2009.

[18] P. Viola, M. Jones, “Robust Real-time Object Detection”, In *Second International Workshop on statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling*, Canada, 2001.

[19] A. Aly and A. Tapus, “Speech to Head Gesture Mapping in Multimodal Human-Robot Interaction”, In *Proc. of the 5th European Conference on Mobile Robots ECMR 2011* Sweden, pp. 101-108, September 2011.

[20] C. Breazeal and L. Aryananda, ”Recognition of Affective Communicative Intent in Robot-Directed Speech”, Artificial Intelligence, pp. 83-104, 2002.

[21] M. Zecca, T. Chaminade, M. A. Umilta, K. Itoh, M. Saito, N. Endo, ”Emotional Expression Humanoid Robot WE-4RII- Evaluation of the perception of facial emotional expressions by using fMRI”, In *Robotics and Mechatronics Conference ROBOMEC2007*, Akita, Japan, pp. 2A1-O10, 2007.

[22] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann and S. Narayanan, ”Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information”, In *Proc. of ACM 6th International Conference on Multimodal Interfaces (ICMI 2004)*, 2004.

[23] Z. Zeng, M. Pantic, G. I. Roisman and T. Huang, ”A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions”, In textit IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 31, pp. 39 - 58, 2008.

[24] 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.

[25] 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.

[26] M. Gruendig, O. Hellwich, ” 3D Head Pose Estimation with Symmetry based Illumination Model in Low Resolution Video”. In Proc. *26th Symposium of the German Association for Pattern Recognition*, Germany, Vol. 3175, pp. 45-53, 2004.

[27] P. Fitzpatrick, ”Head pose estimation without manual initialization”, AI Lab, MIT. Cambridge, USA, 2000.

[28] L.J. Manso, P. Bachiller, P. Bustos, P. Nuñez, R. Cintas and L. Calderita. ”RoboComp: a Tool-based Robotics Framework”. In *Simulation, Modeling and Programming for Autonomous Robots* (SIMPAR). Pages 251-262. 2010.

[29] Open Source Computer Vision Library, Available at: *http://sourceforge.net/projects/opencvlibrary/*.

[30] Available at: *http://iadex.es*.