IMITATION OF FACIAL EXPRESSIONS FOR LEARNING EMOTIONS IN SOCIAL ROBOTICS

Valéria de C. Santos*, Sergio R. Díaz-Miguel Coca†, Giampaolo L. Libralon‡, Roseli A. F. Romero*

*Institute of Mathematics and Computer Sciences
University of São Paulo
São Carlos, SP, Brazil

†Departamento de Ingenería Electronica
Universidad Politécnica de Madrid, Madrid, Spain

‡Federal Institute of São Paulo
São Carlos, SP, Brazil

Emails: valeriac@icmc.usp.br, sdcoca@gmail.com, glibralon@ifsp.edu.br, rafrance@icmc.usp.br

Abstract— Social robots must be able to interact, communicate, understand and relate to humans in a natural way. Although many social robots have been developed successfully, there are still many limitations to overcome. Important advances are needed in the development of mechanisms that allow more realistic interactions and that regulate the relationship between robots and humans. One way to make more realistic interactions is through facial expressions of emotion. In this context, this work provides ability for imitation of facial expressions of emotion to a virtual robotic head, in order to allow more realistic and lasting interactions with humans. For such, learning by imitation is used, in which the robotic head mimics facial expressions made by a user during social interaction. The imitation learning was performed by artificial neural networks. Facial expressions considered in this work are: neutral, happiness, anger, surprise and sadness.

Keywords— Artificial emotion, imitation learning, social robotics, artificial neural networks, facial expressions

1 Introduction

Social robots are agents capable of recognizing other robots or humans and interact with them. Therefore, they must have perceptions, be able to interpret the environment, learn and relate to human naturally (Breazeal, 2002).

There are scientific and practical motivations for the development of social robots. From the scientific point of view, it is possible to learn about the social nature of human beings during the process of development of socially intelligent robots (Breazeal, 2002), (Webb, 2000). From the practical point of view, with the progress in social robotics research and development and the potential abilities of these robots, it is expected a growing use of them as an aid for humans in completing an increasing number of tasks. The social robots are particularly important in situations where they must interact with humans in order to solve specific issues, or in situations where they could be used as a persuading machine. In this context, some researchers have employed social robots as educational tools or as means of interaction during therapy with autistic children (Robins et al., 2004), (Björne and Balkenius, 2005).

A biologically inspired approach has been recently used in order to develop robots that are able to imitate or emulate the social or intelligent behavior found in living entities. Biologically inspired projects are based on neuroscience and biologic theories, including anthropology, psychology, etiology and sociology, among others. These theories have been widely used to lead projects on motivational, motor, cognitive, and behavioral systems of robots (Breazeal, 2000; Dautenhahn et al., 2002).

Among the biologically inspired approaches is the imitation learning. Humans and animals use imitation as a mechanism to gain knowledge. Imitation learning is explored in order to make an unexperienced agent able to learn by observation of an experienced agent during execution of a task. This approach can be seen as a collaborative learning based on interaction between the human and the agent (Bombini et al., 2009).

Most social robots have limited perception, cognition and behavioral abilities as opposed to human beings. Building robots with the ability to socialize, i.e., robots that are able to develop social skills including empathy and understanding of their surrounding world, is a complex task.

One way to contribute to robot socialization is to provide them with capabilities for facial expression of emotions, thus turning interaction with humans more realistic. This way, the main goal of this work is to provide a virtual robotic head with abilities of facial expression of emotions learned from the ones expressed by a human. Each of the imitated facial expressions will be associated to one of the five basic emotions (disgust, fear, happiness, sadness and surprise) present in the sci-

entific literature. The experiments show that the system works well in real interactions with users.

This document is organized in the following way: Section 2 presents the proposed system; Section 3 presents the results of the experimental evaluation and Section 4 presents the conclusion and future work.

2 System for Imitation of Facial Expression of Emotions

The imitation system has been conceived in two distinct modules: Automatic Facial Features Extraction (AFFE) and Facial Expressions Recognition (FER). In Figure 1, it is showed the system's architecture. The images received from the camera are analised by AFFE module that extracts the facial features. AFFE send these features to FER module which by its turn determines expressive features in order to make a robotic head to reproduce the same expression presented to the system.

The AFFE module processes the image, searches for a human face and analyses it to generate an array of features. This array is built based on points and textures which describe the features of face expression in image.

On the other hand, the FER module is responsible for receiving the array of features generated, recognizing the expression and realizing a suitable mapping between the features extracted from human face and the expressive ability of the robotic head.

Next, the details of each module are presented.

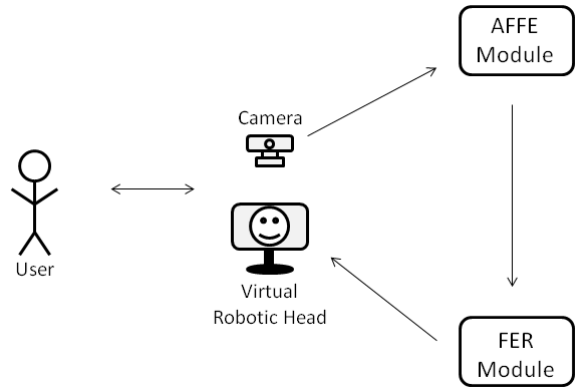

Figure 1: Architecture of the system for imitation of facial expressions of emotions.

2.1 *Automatic Facial Features Extraction Module*

The extraction system used in this module was provided by (Saragih et al., 2011). This system was developed as an optimization strategy for local experts-based deformable model fitting. Deformable model fitting is the problem of registering a parametrized model for an image so that its frames correspond to the locations that form the object of interest. Starting from a base reference model, which is comprised of several frames, the algorithm tries to align the elements of the face according to the frames that are present in the reference model. This problem is considered complex because it involves high dimensional optimization where the appearance of the object can vary a lot between instances of the object due to light conditions, noise in the image, resolution and intrinsic sources of variability.

This approach is based on the mean-shift algorithm, proposed by (Fukunaga and Hostetler, 1975). The difference is that this approach is applied for every landmark simultaneously, and imposes a global prior over their joint motion.

The algorithm starts by computing the set of candidates of locations for every landmark in the reference model by searching in a rectangular region and performing a non-parametric estimation of the map of responses. Then, the linearized model is found and the mean shift vectors are computed. These vectors point towards the direction of maximum growth of the local density function. Finally, the parameters of vector are updated.

2.2 *Facial Expressions of Emotions Recognition and Generation Module*

Valerie, the original name of the virtual robotic head used in this work, was first used on the work from (Gockley et al., 2005).

In this work, five expressions of emotions were defined for Valerie to express: neutral, happiness, anger, surprise, and sadness, as shown, respectively, in Figure 2. In order to define each expression, they were assigned different values to five parameters: jaw, smile, brows, sneer and pout. It can be noted that it is possible to modify the intensity of those parameters but it is not possible to modify the face points to exhibit new gestures. For instance, the smile can be intensified or diminished but it is not possible to move the mouth aside. Due to this limitations in Valerie's expressive abilities, we are going to consider only five basic expressions rather than the usual six that are found in the literature.

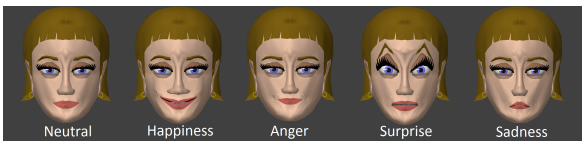

Figure 2: Facial expressions defined for Valerie.

With the points provided by the AFFE mod-

ule, slopes formed between eyebrows, mouth and nose's points were calculated. The points considered are shown in Figure 3, where LE is left eyebrow, RE is right eyebrow, UN is upper nose, LN is lower nose, LM is left mouth and RM is right mouth. Then, the slopes of lines formed by the following points are calculated: a_{1t} (LE1 and UN); a_{2t} (LE2 and UN); a_{3t} (LE3 and UN); a_{4t} (RE1 and UN); a_{5t} (RE2 and UN); a_{6t} (RE3 and UN); a_{7t} (LM and LN); a_{8t} (RM and LN), where t represents the time. These values are normalized in the interval [0,1].

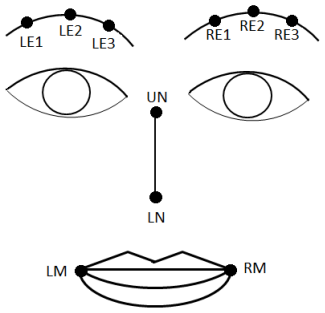

Figure 3: Points considered to calculate the angles from the AFFE module.

This database was used in a MLP network for learning phase and a TLFN (Focused Time Lagged Feedforward Network) network for interaction phase to prevent Valerie from moving brusquely by adding previous output values to the current value.

Here, it is proposed a different combination of the TLFN inputs. Instead to consider only the p inputs delayed in time, we decide to consider the t inputs pondered by respective weights, aiming to have a memory of the past inputs. So, the $x(t)$ input pattern is a combination of the previous $x(t-1), ..., x(1)$ values presented before, according to:

$$x(t) = \alpha^{(t-1)}x(1) + \alpha^{(t-2)}(1-\alpha)x(2) + \ldots + \alpha(1-\alpha)x(t-1) + (1-\alpha)x(t) \quad (1)$$

According to the literature, it is possible to perform imitation with both approaches: through mapping of a human features to a robot (Ito and Tani, 2004; Gotoh et al., 2007), and through classification of the user features and later generation of the matching expression of the robot (Boucenna et al., 2010; Ge et al., 2008), making the system independent of the robot in use. In this way, two MLP networks were implemented for two ways of learning, namely, MLP I and MLP II, respectively. Both networks have the same architecture: 8 neurons on the input layer, 15 on the middle layer and 5 on the output. The input signals correspond to the 8 angles obtained from the user's face features. The amount of neurons in the middle layer was defined empirically by choosing the one with the minimum error. The output values have a different function on each network as follows.

In the first network, MLP I, the outputs correspond with the parameters that will define Valerie's face expression. In order to know which emotion that output represents, another MLP network was trained, named Evaluation network. Evaluation network has five neurons in the output layer. The inputs are the parameter of the mentioned MLP I. The output will correspond with one of the five predefined emotions: neutral, happiness, anger, surprise and sadness. Those outputs were represented by n-dimensional vectors with canonical base. More specifically, vector $(10000)^T$, identifies the first emotion (i.e: neutral), vector: $(01000)^T$ identifies the second emotion (i.e: happiness), and so on.

In the second network, MLP II, where the classification of the expression of emotion was done for the later generation of Valerie's expression, the outputs correspond to the five defined expressions. In the same way as in the Evaluation network, those outputs were represented with a 5-dimensional vector with canonical base.

Each video correspond to one facial expression of emotion of one person. As the facial features were being extracted during the video, the outputs were being registered.

A cross validation method (Larson, 1931) was used to validate the network. Cross validation is an statistical method for algorithm validation and comparison, where the data is divided in two parts: one is used for training and the other for test/validation. During a typical cross validation, the training and testing sets have to be crossed upon several iterations, so that all of the data samples have a chance to be validated. The basic form of cross validation is the k-fold cross validation, where the data is partitioned in k equally sized sets. Then, k iterations of training and test are executed so that during each iteration a different set of data is taken for test while the other $k-1$ are used for training. Other forms of cross validation are special cases of k-fold cross validation or involve repeated iterations of it. The 10-fold cross-validation is the most commonly used where 9 sets are used as training on each iteration and one for test. In the 5x2-fold cross-validation (Refaeilzadeh et al., 2009), the 2-fold cross validation, is executed five times, resulting in ten values of accuracy.

With the database obtained from AFFE module, we also performed experiments using Support Vector Machines - SVMs (Steinwart and Christmann, 2008) - as a classifier, for comparison purposes. SVMs are learning algorithms based on the theory of statistical learning, following the principle of Structural Risk Minimization (SRM). The high generalization capacity obtained by SVMs re-

Table 1: Confusion Matrix - MLP I

	Predicted class				
Actual class	Neutral	Happiness	Anger	Surprise	Sadness
Neutral	**8**	0	1	0	1
Happiness	0	**10**	0	0	0
Anger	0	0	**10**	0	0
Surprise	0	0	0	**10**	0
Sadness	0	0	0	0	**10**

Table 2: Confusion Matrix - MLP II

	Predicted class				
Actual class	Neutral	Happiness	Anger	Surprise	Sadness
Neutral	**10**	0	0	0	0
Happiness	0	**10**	0	0	0
Anger	0	0	**10**	0	0
Surprise	1	0	0	**9**	0
Sadness	0	0	0	0	**10**

sults from the use of the statistical learning theory, principle presented in the decade of 60 and 70 by Vapnik and Chernovenkis (Vapnik and Ya, 1971).

3 Experiments and Results

The data that was used for training and test were extracted from videos of 10 people, each of them expressing the five basic emotions, summing up a total of 50 videos. With the data that was extracted, 6 folders were created, each of them containing 5 frames of each video, summing up a total of 250 data instances in each folder. Distributing the training in that way, the training data as well as the test have information of the features of every person showing every emotion.

To validate the performance of the neural networks, it was used the 5x2 cross-validation method, summing up 10 runs. The Table 1 illustrates the confusion matrix obtained by one of the runs that were performed over MLP I. The accuracy obtained on the tests generated an average hit rate of 92,4%. This high rate is due to the knowledge that the network acquired from the facial features of all of the people in the training sets. It can be noted a mistaken classification between sadness and neutral expressions. Apart from that, anger was also classified as neutral twice. The other expressions, happiness, anger, and surprise, were correctly classified in every case. Figure 4 shows instances of imitation of facial expression obtained with the training of MLP I and the output of the TLFN network. Looking at the figure it can be understood the difficulty of discerning between the expressions of sadness and neutral.

The accuracy obtained on the 10 runs with the MLP II generated an average hit rate of 97,6%. Table 2 illustrates the confusion matrix obtained by one of the runs that were performed over MLP II. The network was able to match every single expression more accurately than MLP I network. It was noted that the most common errors were made between sadness and neutral, anger and neutral, sadness and neutral. Figure 5 shows instances of imitation of facial expressions obtained by training MLP II and the outputs of TLFN. It is the most similar imitation obtained.

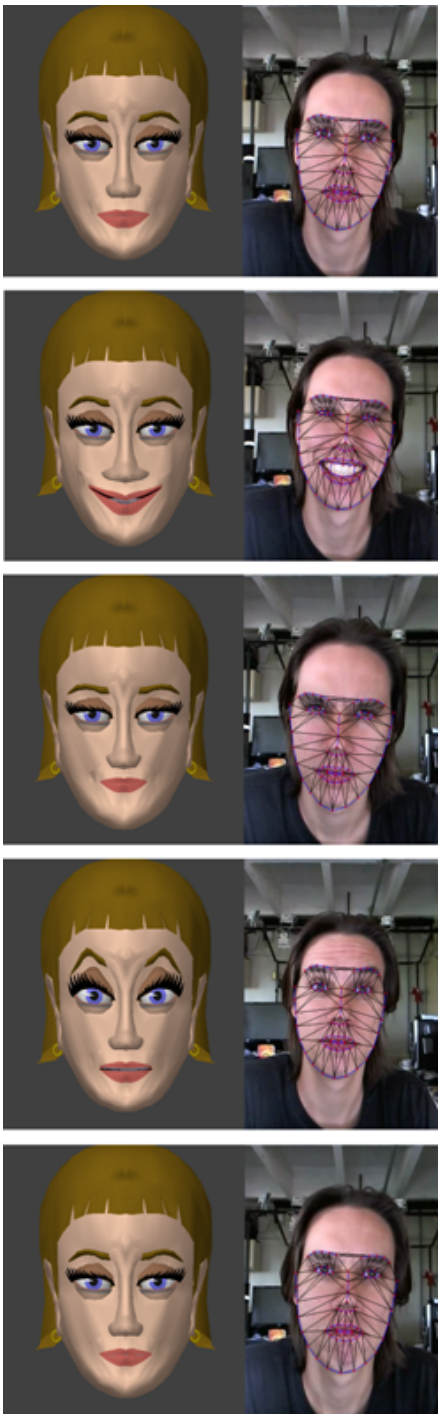

Figure 4: Facial expressions imitation from an user during social interaction, obtained by MLP I learnig. From up to down, the expressions are neutral, happiness, anger, surprise and sadness.

For training the SVMs, the Polynomial, Gaussian and Linear kernels were explored. Different values of kernel dependent parameters were also investigated. As mentioned, the 5x2-fold cross-validation was used to evaluate the method. The accuracy obtained on the 10 runs generated an average hit rate of 96,0%. Table 3 illustrates the confusion matrix obtained by one of the runs that were performed over SVMs.

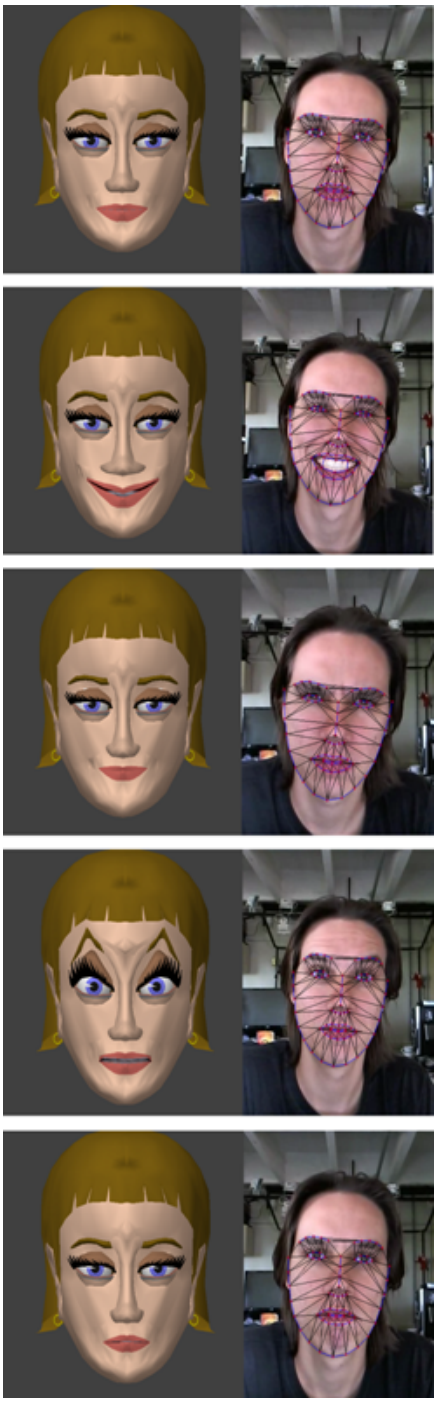

Figure 5: Facial expressions imitation from an user during social interaction, obtained by MLP II learning. From up to down, the expressions are neutral, happiness, anger, surprise and sadness.

The mean hit rates obtained by SVMs was quite similar to the ones obtained by the neural network MLP II and better than the ones from MLP I. It was noted that the most common errors were made between emotions anger and surprise.

4 Conclusions

One way to make the human-robot interaction more realistic is to provide the robot with abilities for facial expression of emotions. The goal of this work was to provide a virtual robotic head with the ability of imitation of facial expressions of emotions from the expressions exercised by a human during social interaction. The imitation process involves two steps: recognition of the user's expression, and generation of the matching robot expression. For that end, it was developed a system composed of two different modules: the automatic facial feature extraction module and the facial expressions of emotions recognition and generation module.

Table 3: Confusion Matrix - SVM

	Predicted class				
Actual class	Neutral	Happiness	Anger	Surprise	Sadness
Neutral	**10**	0	0	0	0
Happiness	0	**10**	0	0	0
Anger	0	0	**10**	0	0
Surprise	0	0	2	**8**	0
Sadness	0	0	0	0	**10**

For the automatic facial features extraction module, the extraction system developed in (Saragih et al., 2011) was used. In order to develop the facial expressions of emotions recognition and generation module, it was used an MLP neural network for the classification pattern learning and a TLFN network for the expressions generation in the virtual robotic head during user interaction. Furthermore, experiments with SVM were done.

The proposed system was analyzed using a cross-validation techniques. For the data sets generation, they were aggregated information from every user and every expression for training and test, leading to very satisfying results.

This system is expected to be used during real interactions upon which it is not possible to obtain information of the facial features for the network to learn from every user. This way, it will be investigated a way for the neural network to learn about the facial expressions in a more user-independent manner. One approach would be to use a bigger amount of data from a wider range of users for training with more facial features diversity, such as color or lips natural shape. Furthermore, it can be altered the intensity of the light were the videos are being recorded as well as user face position.

Acknowledges

The authors would like to thank FAPESP (process 2010/04325-3), CNPq and EMECW program for their financial support.

References

Björne, P. and Balkenius, C. (2005). A model of attentional impairments in autism: first steps toward a computational theory, *Cognitive Systems Research* **6**(3): 193–204.

Bombini, G., Mauro, G. D., Basile, N., Ferilli, T. M. A. and Esposito, S. (2009). Relational learning by imitation, *Agent and Multi-Agent Systems: Technologies and Applications* pp. 273–282.

Boucenna, S., Gaussier, P., Andry, P. and Hafemeister, L. (2010). Imitation as a communication tool for online facial expression learning and recognition, *International Conference on Intelligent Robots and Systems* .

Breazeal, C. (2000). *Sociable Machines: Expressive Social Interaction Between Human and Robots*, PhD thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA.

Breazeal, C. (2002). *Designing Sociable Robots*, The MIT Press.

Dautenhahn, K., Ogden, B. and Quick, T. (2002). From embodied to socially embedded agents-implications for interactionaware robots, *Cognitive Systems Research* **3**(3).

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition, *Information Theory, IEEE Transactions on* **21**(1): 32–40.

Ge, S. S., Wang, C. and Hang, C. C. (2008). Facial expression imitation in human robot interaction, *17th IEEE International Symposium on Robot and Human Interactive Communication*, Munich, Germany, pp. 213–218.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C. and Wang, J. (2005). Designing robots for long-term social interaction, *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, pp. 1338–1343.

Gotoh, M., Kanoh, M., Kato, S. and Itoh, H. (2007). A neural-based approach to facial expression mapping between human and robot, *Proceedings of the 11th international conference, KES 2007 and XVII Italian workshop on neural networks conference on Knowledge-based intelligent information and engineering systems: Part III*, Springer-Verlag, Berlin, Heidelberg, pp. 194–201.

Ito, M. and Tani, J. (2004). On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system, *Adaptive Behavior* **12**(2): 93–115.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation, *Journal of Educational Psychology* **22**: 45–55.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009). Cross-validation, *in* L. Liu and M. T. Ozsu (eds), *Encyclopedia of Database Systems*, Springer US, pp. 532–538.

Robins, B., Dickerson, P., Stribling, P. and Dautenhahn (2004). Robot-mediated joint attention in children with autism: A case study in robot-human interaction, *Interaction Studies* **5**(2): 161–198.

Saragih, J., Lucey, S. and Cohn, J. (2011). Deformable Model Fitting by Regularized Landmark Mean-Shift, *International Journal of Computer Vision* **91**(2): 200–215.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*, 1st edn, Springer Publishing Company, Incorporated.

Vapnik, V. N. and Ya (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities, *Theory of Probability and its Applications* **16**(2): 264–280.

Webb, B. (2000). What does robotics offer animal behaviour?, *Animal Behaviour* **60**(5): 545–558.