Facial Expression Recognition for HCI Applications

Fadi Dornaika
Institut Géographique National, France

Bogdan Raducanu
Computer Vision Center, Spain

INTRODUCTION

Facial expression plays an important role in cognition of human emotions (Fasel, 2003 & Yeasin, 2006). The recognition of facial expressions in **image sequences** with significant head movement is a challenging problem. It is required by many applications such as human-computer interaction and computer graphics animation (Cañamero, 2005 & Picard, 2001). To classify expressions in still images many techniques have been proposed such as Neural Nets (Tian, 2001), Gabor wavelets (Bartlett, 2004), and active appearance models (Sung, 2006). Recently, more attention has been given to modeling facial deformation in dynamic scenarios. Still image classifiers use feature vectors related to a single frame to perform classification. **Temporal classifiers** try to capture the temporal pattern in the sequence of feature vectors related to each frame such as the Hidden Markov Model based methods (Cohen, 2003, Black, 1997 & Rabiner, 1989) and Dynamic Bayesian Networks (Zhang, 2005). The main contributions of the paper are as follows. First, we propose an efficient recognition scheme based on the detection of **keyframes** in videos where the recognition is performed using a temporal classifier. Second, we use the proposed method for extending the human-machine interaction functionality of a robot whose response is generated according to the user's recognized facial expression.

Our proposed approach has several advantages. First, unlike most expression recognition systems that require a frontal view of the face, our system is view- and texture-independent. Second, its learning phase is simple compared to other techniques (e.g., the Hidden Markov Models and Active Appearance Models), that is, we only need to fit second-order Auto-Regressive models to sequences of facial actions. As a result, even when the imaging conditions change the learned Auto-Regressive models need not to be recomputed.

The rest of the paper is organized as follows. Section 2 summarizes our developed appearance-based 3D **face tracker** that we use to track the 3D **head pose** as well as the **facial actions**. Section 3 describes the proposed facial expression recognition based on the detection of keyframes. Section 4 provides some experimental results. Section 5 describes the proposed human-machine interaction application that is based on the developed facial expression recognition scheme.

SIMULTANEOUS HEAD AND FACIAL ACTION TRACKING

In our study, we use the *Candide* **3D face model** (Ahlberg, 2001). This 3D deformable wireframe model is given by the 3D coordinates of n vertices. Thus, the 3D shape can be fully described by the $3n$-vector $\boldsymbol{g}$ - the concatenation of the 3D coordinates of all vertices. The vector $\boldsymbol{g}$ can be written as:

$$g = g_s + A\tau_a \quad (1)$$

where $\boldsymbol{g}_s$ is the static shape of the model, τ_a is the facial action vector, and the columns of A are the Animation Units. In this study, we use six modes for the facial Animation Units (AUs) matrix A, that is, the dimension of τ_a is 6. These modes are all included in the *Candide* model package. We have chosen the six following AUs: lower lip depressor, lip stretcher, lip corner depressor, upper lip raiser, eyebrow lowerer and outer eyebrow raiser. A cornerstone problem in **facial expression** recognition is the ability to track the local **facial actions**/deformations. In our work, we track the head and facial actions using our **face tracker** (Dornaika & Davoine, 2006). This appearance-based tracker simultaneously computes the 3D **head pose** and the **facial actions** τ_a by minimizing a distance between

the incoming warped frame and the current appearance of the face. Since the **facial actions,** encoded by the vector τ_a, are highly correlated to the facial expressions, their time series representation can be utilized for inferring the facial expression in videos. This will be explained in the sequel.

EFFICIENT FACIAL EXPRESSION DETECTION AND RECOGNITION

In (Dornaika & Raducanu, 2006), we have proposed a **facial expression** recognition method that is based on the time-series representation of the tracked facial actions τ_a. An analysis-synthesis scheme based on learned auto-regressive models was proposed. In this paper, we introduce a process able to detect **keyframes** in videos. Once a keyframe is detected, the temporal recognition scheme described in (Dornaika & Raducanu, 2006) will be invoked on the detected keyframe. The proposed scheme has two advantages. First, the CPU time corresponding to the recognition part will be considerably reduced since only few keyframes are considered. Second, since a **keyframe** and its neighbor frames are characterizing the expression, the discrimination performance of the recognition scheme will be boosted. In our case, the keyframes are defined by the frames where the **facial actions** change abruptly. Thus, a keyframe can be detected by looking for a local positive maximum in the temporal derivatives of the facial actions. To this end, two entities will be computed from the sequence of facial actions τ_a that arrive in a sequential fashion: (i) the L_1 norm $\|\tau_a\|_1$, and (ii) the temporal derivative given by:

Figure 1. Efficient facial expression detection and recognition based on keyframes

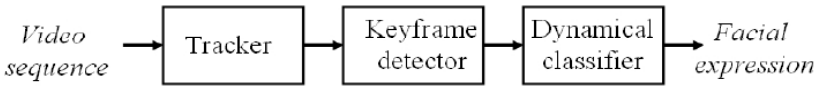
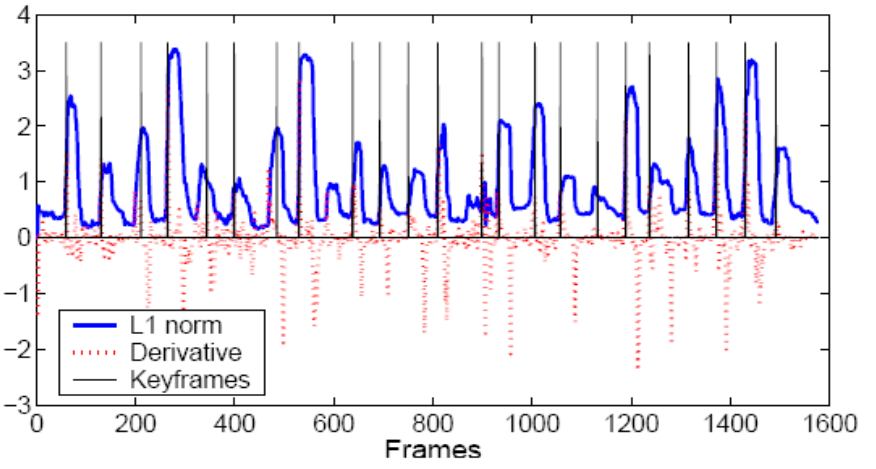


Figure 2. Keyframe detection and recognition applied on a 1600-frame sequence



$$D_t = \frac{\partial \|\tau_a\|_1}{\partial t} = \sum_{i=1}^{6} \frac{\partial \tau_{a(i)}}{\partial t} \quad (2)$$

In the above equation, we have used the fact that the facial actions are positive. Let W be the size of a temporal segment defining the temporal granulometry of the system. In other words, the system will detect and recognize at most one expression every W frames. In practice, W belongs to [0.5s, 1s]. The whole scheme is depicted in Figure 1.

In this figure, we can see that the system has three levels: the tracking level, the keyframe detection level, and the recognition level. The tracker provides the facial actions for every frame. Whenever the current video segment size reaches W frames, the keyframe detection is invoked to select a keyframe in the current segment if any. A given frame is considered as a **keyframe** if it meets three conditions: (1) the corresponding D_t is a positive local maximum (within the segment), (2) the corresponding norm $\|\tau_a\|_1$ is greater than a predefined threshold, (3) its far from the previous keyframe by at least W frames. Once a keyframe is found in the current segment, the dynamical classifier described in (Dornaika & Raducanu, 2006) will be invoked.

Figure 2 shows the results of applying the proposed detection scheme on a 1600-frame sequence containing 23 played expressions. Some images are shown in Figure 4. The solid curve corresponds to the norm $\|\tau_a\|_1$, the dotted curve to the derivative D_t and the vertical bars correspond to the detected keyframes. In this example, the value of W is set to 30 frames. As can be seen, out of 1600 frames only 23 keyframes will be processed by the expression classifier.

EXPERIMENTAL RESULTS

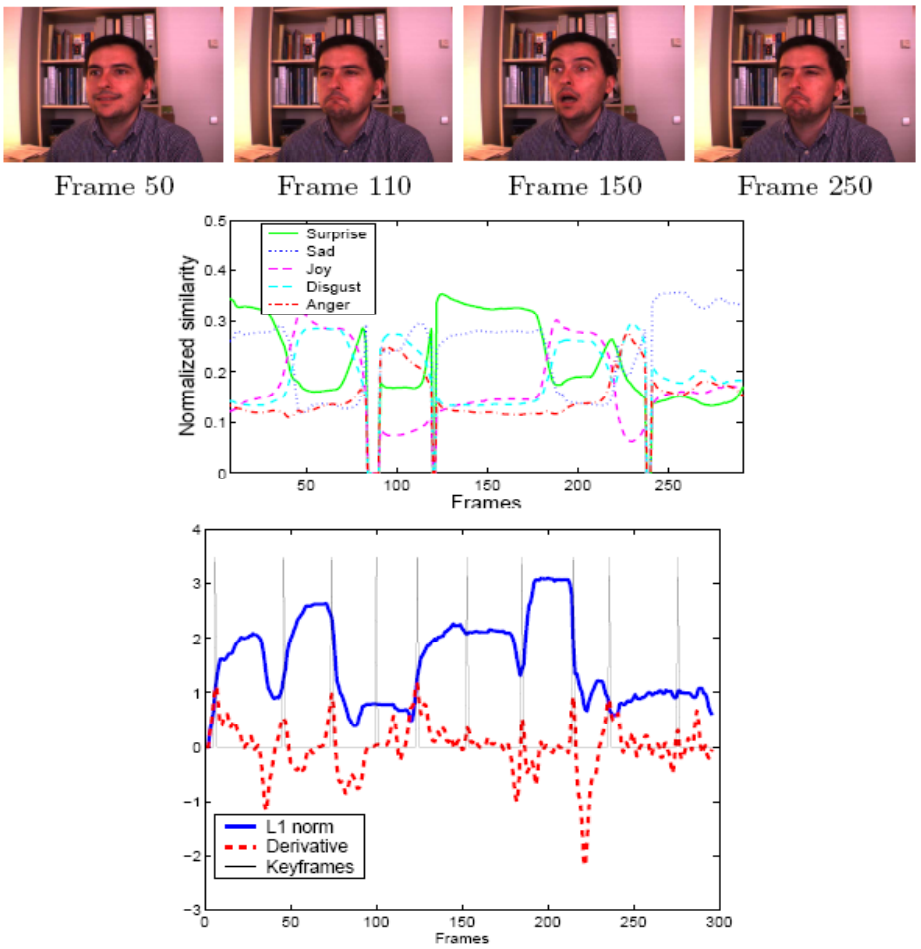
Recognition results: We used a 300-frame video sequence. For this sequence, we asked a subject to display several expressions arbitrarily (see Figure 3). The middle of this figure shows the normalized similarities associated with each universal expression where the recognition is performed for every frame in the sequence. As can be seen, the **temporal classifier** (Dornaika & Raducanu, 2006) has correctly detected the presence of the surprise, joy, and sadness expressions. Note that the mixture of expressions at transition is normal since the recognition is performed in a frame-wise manner. The lower part of this figure shows the results of applying the proposed keyframe detection scheme. On a 3.2 GHz PC, a non-optimized C code of the developed approach carries out the tracking and recognition in about 60 ms.

Performance study: In order to quantify the recognition rate, we have used 35 test videos retrieved from the CMU database. Table 1 shows the confusion matrix associated with the 35 test videos featuring 7 persons. As can be seen, although the recognition rate was good (80%), it is not equal to 100%. This can be explained by the fact that the expression dynamics are highly subject-dependent. Recall that the used auto-regressive models are built using data associated with one subject. Notice that the human 'ceiling' in correctly classifying facial expressions into the six basic emotions has been established at 91.7%.

Table 1. Confusion matrix for the facial expression classifier associated with 35 test videos (CMU data). The model is built using one unseen person

	Surprise (7)	Sadness (7)	Joy (7)	Disgust (7)	Anger (7)
Surprise	7	0	0	0	0
Sadness	0	7	0	5	0
Joy	0	0	7	0	0
Disgust	0	0	0	2	2
Anger	0	0	0	0	5

Figure 3. Top: Four frames (50, 110, 150, and 250) associated with a 300-frame test sequence. Middle: The similarity measure computed for each universal expression and for each non-neutral frame of the sequence-the framewise recognition. Bottom: The recognition based on keyframe detection.

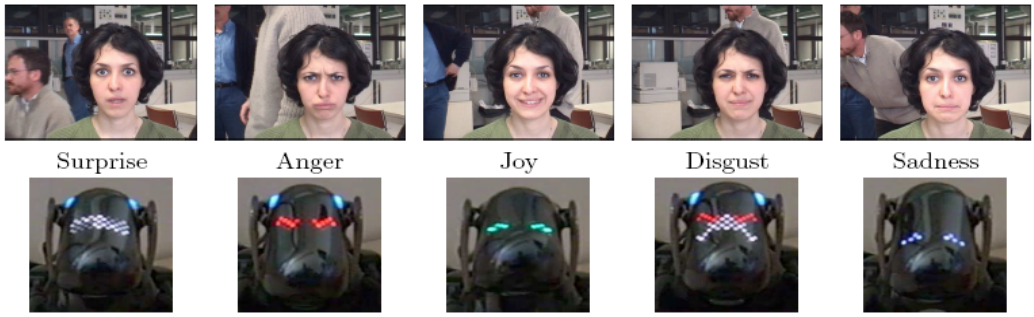


HUMAN-MACHINE INTERACTION

Interpreting non-verbal face gestures is used in a wide range of applications. An intelligent user-interface not only should interpret the face movements but also should interpret the user's emotional state (Breazeal, 2002). Knowing the emotional state of the user makes machines communicate and interact with humans in a natural way: intelligent entertaining systems for kids, interactive computers, intelligent sensors, social robots, to mention a few. In the sequel, we will show how our proposed technique lends itself nicely to such applications. Without loss of generality, we use the AIBO robot which has the advantage of being especially designed for Human Computer Interaction. The input to the system is a video stream capturing the user's face.

The AIBO robot: AIBO is a biologically-inspired robot and is able to show its emotions through an array of LEDs situated in the frontal part of the head. In addition to the LEDs' configuration, the robot response

Figure 4. Top: Some detected keyframes associated with the 1600-frame video. Middle: The recognized expression. Bottom: The corresponding robot's response.



contains some small head and body movements. From its concept design, AIBO's affective states are triggered by the Emotion Generator engine. This occurs as a response to its internal state representation, captured through multi-modal interaction (vision, audio and touch). For instance, it can display the 'happiness' feeling when it detects a face (through the vision system) or it hears a voice. But it does not possess a built-in system for vision-based automatic facial-expression recognition. For this reason, with the scheme proposed in this paper (see Section 3), we created an application for AIBO whose purpose is to enable it with this capability.

This application is a very simple one, in which the robot is just imitating the expression of a human subject. Usually, the response of the robot occurs slightly after the apex of the human expression. The results of this application were recorded in a 2 minute video which can be downloaded from the following address: http://www.cvc.uab.es/~bogdan/AIBO-emotions.avi. In order to be able to display simultaneously in the video the correspondence between subject's and robot's expressions, we put them side by side.

Figure 4 illustrates five detected keyframes from the 1600 frame video depicted in Figure 2. These are shown in correspondence with the robot's response. The middle row shows the recognized expression. The bottom row shows a snapshot of the robot head when it interacts with the detected and recognized expression.

CONCLUSION

This paper described a view- and texture-independent approach to facial expression analysis and recognition. The paper presented two contributions. First, we proposed an efficient facial expression recognition scheme based on the detection of keyframes in videos. Second, we applied the proposed method in a Human Computer Interaction scenario, in which an AIBO robot is mirroring the user's recognized facial expression.

ACKNOWLEDGMENT

This work has been partially supported by MCYT Grant TIN2006-15308-C02, Ministerio de Educación y Ciencia, Spain. Bogdan Raducanu is supported by the Ramon y Cajal research program, Ministerio de Educación y Ciencia, Spain. The authors thank Dr. Franck Davoine from CNRS, Compiegne, France, for providing the video sequence shown in Figure 4.

REFERENCES

Ahlberg, J. (2001). CANDIDE-3 – An Updated Parameterized Face. *Technical Report LiTH-ISY-R-2326*, Dept. of Electrical Engineering, Linköping University, Sweden.

Bartlett, M., Littleworth, G., Lainscsek, C., Fasel I. & Movellan, J. (2004). Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions. *Proc. of IEEE Conference on Systems, Man and Cybernetics*, Vol. I, The Hague, The Netherlands, pp.592-597.

Black, M.J. & Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23-48.

Breazeal, C. & Scassellati, B. (2002). Robots that Imitate Humans. *Trends in Cognitive Science*, Vol. 6, pp. 481-487.

Cañamero, L. & Gaussier, P. (2005). Emotion Understanding: Robots as Tools and Models. In *Emotional Development: Recent Research Advances*, pp. 235-258.

Cohen, I., Sebe, N., Garg, A., Chen, L. & Huang, T. (2003). Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. *Computer Vision and Image Understanding*, 91(1-2):160-187.

Dornaika, F. & Davoine, F. (2006). On Appearance Based Face and Facial Action Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(9):1107-1124.

Dornaika, F. & Raducanu, B. (2006). Recognizing Facial Expressions in Videos Using a Facial Action Analysis-Synthesis Scheme. Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. N/A, Australia.

Fasel, B. & Luettin, J. (2003). Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1):259-275.

Picard, R., Vyzas, E. & Healy, J. (2001) Toward Machine Emotional Intelligence: Analysis of Affective Psychological State. *IEEE Trasactions on Pattern Analysis and Machine Intelligence*, 23(10):1175-1191.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of IEEE*, 77(2):257-286.

Sung, J., Lee, S. & Kim, D. (2006). A Real-Time Facial Expression Recognition Using the STAAM. *Proc. of International Conference on Pattern Recognition*, Vol. I, pp. 275-278, Hong-Kong.

Tian, Y., Kanade T. & Cohn, J. (2001). Recognizing Action Units for Facial Expression Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, pp. 97-115.

Yeasin M., Bullot, B. & Sharma, R. (2006). Recognition of Facial Expressions and Measurement of Levels of Interest from Video. *IEEE Transactions on Multimedia* 8(3):500-508.

Zhang, Y. & Ji, Q. (2005). Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699-714.

KEY TERMS

3D Deformable Model: A model which is able to modify its shape while being acted upon by an external influence. In consequence, the relative position of any point on a deformable body can change.

Active Appearance Models (AAM): Computer Vision algorithm for matching a statistical model of object shape and appearance to a new image. The approach is widely used for matching and tracking faces.

AIBO: One of several types of robotic pets designed and manufactured by Sony. Able to walk, "see" its environment via camera, and recognize spoken commands, they are considered to be autonomous robots, since they are able to learn and mature based on external stimuli from their owner or environment, or from other AIBOs.

Autoregressive Models: Group of linear prediction formulas that attempt to predict the output of a system based on the previous outputs and inputs.

Facial Expression Recognition System: Computer-driven application for automatically identifying person's facial expression from a digital still or video image. It does that by comparing selected facial features in the live image and a facial database.

Hidden Markov Model (HMM): Statistical model in which the system being modeled is assumed to be

a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications.

Human–Computer Interaction (HCI): The study of interaction between people (users) and computers. It is an interdisciplinary subject, relating computer science with many other fields of study and research (Artificial Intelligence, Psychology, Computer Graphics, Design).

Social Robot: An autonomous robot that interacts and communicates with humans by following the social rules attached to its role. This definition implies that a social robot has a physical embodiment. A consequence of the previous statements is that a robot that only interacts and communicates with other robots would not be considered to be a social robot.

Wireframe Model: The representation of all surfaces of a three-dimensional object in outline form.