Available online at www.sciencedirect.com
ScienceDirect
Signal Processing: *Image Communication* 22 (2007) 769–784
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Inferring facial expressions from videos: Tool and application ☆

Fadi Dornaika[a,*], Bogdan Raducanu[b]

[a]*Institut Géographique National, 94165 Saint-Mandé, France*
[b]*Computer Vision Center, 08193 Bellaterra, Barcelona, Spain*

Received 8 December 2006; received in revised form 15 June 2007; accepted 19 June 2007

Abstract

In this paper, we propose a novel approach for facial expression analysis and recognition. The main contributions of the paper are as follows. First, we propose a temporal recognition scheme that classifies a given image in an unseen video into one of the universal facial expression categories using an analysis–synthesis scheme. The proposed approach relies on tracked facial actions provided by a real-time face tracker. Second, we propose an efficient recognition scheme based on the detection of keyframes in videos. Third, we use the proposed method for extending the human–machine interaction functionality of the AIBO robot. More precisely, the robot is displaying an emotional state in response to the user's recognized facial expression. Experiments using unseen videos demonstrated the effectiveness of the developed methods.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Facial expression recognition; Temporal classifiers; Keyframes; Human machine interaction; AIBO robot

1. Introduction

Facial expression plays an important role in cognition of human emotions. Basic facial expressions typically recognized by psychologists are happiness, sadness, fear, anger, disgust and surprise [13]. In the past, facial expression analysis was essentially a research topic for psychologists. However, recent progresses in image processing and pattern recognition have motivated significantly research works on automatic facial expression recognition [15,17,26].

The recognition of facial expressions in image sequences with significant head motion is a challenging problem. It is required by many applications such as human–computer interaction and computer graphics animation [7,20,21].

To classify expressions in still images many techniques have been proposed such as Neural Nets [25], Gabor wavelets [2], and active appearance models [24]. The still images usually capture the apex of the expression, i.e., the instant at which the indicators of emotion are most marked. 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 (or dynamical classifiers) try to capture the temporal pattern in the sequence of feature vectors related to each frame such as the Hidden Markov Model-based

☆This work was supported by the MEC project TIN2006-15308-C02 and The Ramón y Cajal Program.
*Corresponding author.
E-mail addresses: fadi.dornaika@ign.fr (F. Dornaika), bogdan@cvc.uab.es (B. Raducanu).

0923-5965/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2007.06.006

methods [8,4,23] and Dynamic Bayesian Networks [27]. In [4], parametric 2D flow models associated with the whole face as well as with the mouth, eyebrows, and eyes are first estimated. Then, mid-level predicates are inferred from these parameters. Finally, universal facial expressions are detected and recognized using the estimated predicates.

The main contributions of the paper are as follows. First, we propose a temporal recognition scheme that classifies a given image in an unseen video into one of the universal facial expression categories using an analysis–synthesis scheme. Second, we propose an efficient recognition scheme based on the detection of keyframes in videos. Third, 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.

We propose a novel and flexible scheme for facial expression recognition that is based on an appearance-based 3D face tracker. Our developed approach enables facial expression recognition using an analysis–synthesis scheme based on auto-regressive models. Although auto-regressive models have been widely used in the tasks of 2D tracking and synthesis, to the best of our knowledge they have not been used for facial expression recognition. The proposed approach proceeds as follows. First, a tracker provides the time-varying facial actions related to the lips and the eyebrows. Second, using learned auto-regressive models (each universal expression has a model) the facial actions are then temporally synthesized. Then similarity measures between the synthesized trajectories and the actual ones will decide the expression.

Compared to existing temporal facial expression methods our proposed approach has several advantages. First, unlike most expression recognition systems that require a frontal view of the face, our system is view independent since the used tracker simultaneously provides the 3D head pose and the facial actions. Second, it is texture independent since the recognition scheme relies only on the estimated facial actions—geometrical parameters. Third, 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. Fourth, the computational load of the proposed approach is low. 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. This section also describes an efficient recognition scheme 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.

2. Simultaneous head and facial action tracking

In our study, we use the 3D face model *Candide* [1]. This 3D deformable wireframe model is given by the 3D coordinates of the vertices $\mathbf{P}_i, i = 1, \ldots, n$ where n is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$-vector $\mathbf{g}$—the concatenation of the 3D coordinates of all vertices $\mathbf{P}_i$. The vector $\mathbf{g}$ can be written as

$$\mathbf{g} = \mathbf{g}_s + \mathbf{A}\boldsymbol{\tau}_\mathbf{a}, \tag{1}$$

where $\mathbf{g}_s$ is the static shape of the model, $\boldsymbol{\tau}_\mathbf{a}$ the facial action vector, and the columns of $\mathbf{A}$ are the animation units (AUs). In this study, we use six modes for the facial AUs matrix $\mathbf{A}$, that is, the dimension of $\boldsymbol{\tau}_\mathbf{a}$ 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. These AUs are enough to cover most common facial animations (mouth and eyebrow movements). Moreover, they are essential for conveying emotions. Many studies have shown that image regions associated with the mouth and the eyebrows are the most informative regions about the facial expression (e.g., [3,4]).

Thus, the state of the 3D model is given by the 3D head pose (three rotations and three translations) and the vector $\boldsymbol{\tau}_\mathbf{a}$. This is given by the 12-vector $\mathbf{b}$:

$$\mathbf{b} = [\theta_x,\ \theta_y,\ \theta_z,\ t_x,\ t_y,\ t_z,\ \boldsymbol{\tau}_\mathbf{a}^\mathrm{T}]^\mathrm{T}. \tag{2}$$

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 tracker [10]. This appearance-based tracker simultaneously computes the 3D head pose and the facial actions encapsulated in the vector $\mathbf{b}$ by minimizing a distance between the incoming warped frame and the current appearance of the face. This minimization is carried out using a

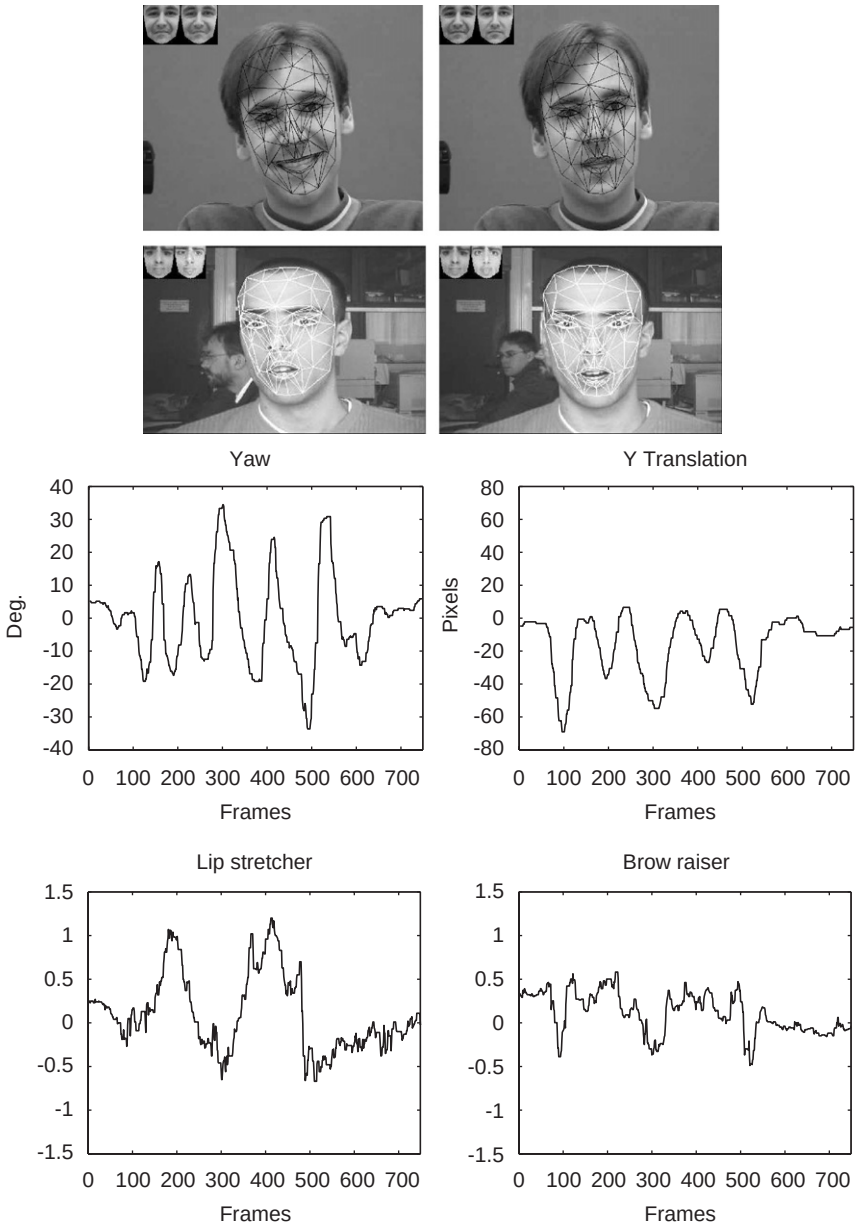

Fig. 1. *Top*: Simultaneous head and facial action tracking results associated with two video sequences. *Bottom*: The yaw angle, the vertical translation, the lip stretcher, and the eye brow raiser associated with the second video sequence.

Gauss–Newton method. The statistics of the appearance as well as the gradient matrix are updated every frame. This scheme leads to a fast and robust tracking algorithm. Fig. 1 shows the tracking results associated with two video sequences. The first video[1] consists of 1000 frames, and depicts a subject engaged in conversation with another person. The second video consists of 750 frames, and depicts a subject featuring quite large head pose variations as well as large facial actions. The bottom of this figure shows the estimated value of the yaw angle, the vertical translation, the lip stretcher, and the brow raiser associated with the second sequence. Since the facial actions $\boldsymbol{\tau}_{\mathbf{a}}$ 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. We stress the fact that since these actions are independent from

[1] www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/.

the 3D head pose, our proposed facial expression recognition method will be view independent.

3. Approach

In this section, we introduce a recognition scheme based on the dynamics of facial expressions. We argue that the facial expression can be inferred from the temporal representation of the tracked facial actions. For this purpose, we use continuous dynamical systems described by the facial action parameters $\boldsymbol{\tau}_{\mathbf{a}_{(t)}}$.

Corresponding to each universal expression there is a dynamical model, supposed to be a second-order Markov model. It is a Gaussian auto-regressive process (ARP) defined by

$$\boldsymbol{\tau}_{\mathbf{a}_{(t)}} = \mathbf{A}_1\boldsymbol{\tau}_{\mathbf{a}_{(t-1)}} + \mathbf{A}_2\boldsymbol{\tau}_{\mathbf{a}_{(t-2)}} + \mathbf{d} + \mathbf{B}\mathbf{w}_{(t)} \quad (3)$$

in which $\mathbf{w}_{(t)}$ is a vector of 6 (6 is the dimension of $\boldsymbol{\tau}_{\mathbf{a}_{(t)}}$) independent random $\mathcal{N}(0,1)$ variables. The parameters of the model are: (i) deterministic parameters $\mathbf{A}_1, \mathbf{A}_2$, and $\mathbf{d}$ and (ii) stochastic parameters $\mathbf{B}$. It is worthwhile noting that the above model can be used for predicting the process from the previous two values. The predicted value at time t obeys a multivariate Gaussian centered at the deterministic value of (3) with $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ being its covariance matrix [5,18,19]. The reason of using second-order Markov models is twofold. First, these models are easy to estimate. Second, they are able to model complex motions. For example, these models have been used in [5] for learning the 2D motion dynamics of talking lips, beating hearts, and writing fingers.

3.1. Learning

Given a training sequence $\boldsymbol{\tau}_{\mathbf{a}_{(1)}}, \ldots, \boldsymbol{\tau}_{\mathbf{a}_{(T)}}$, with $T > 2$, belonging to the same universal expression, it is well known that a maximum likelihood estimator provides a closed-form solution for the parameters $\mathbf{A}_1, \mathbf{A}_2, \mathbf{d}$, and $\mathbf{B}$.

We have built a second-order auto-regressive model for each universal expression. We have used two different training sets. The first training set was provided by the CMU data [16]. This first set consists of 35 short videos depicting five universal expressions: surprise, sadness, joy, disgust, and anger. Each universal expression was performed by seven persons (see Fig. 2). Each short video depicts a face going from the neutral configuration to the apex configuration. The average length of these videos is about 18 frames. The first frame depicts a neutral face and the last frame depicts the apex configuration.

Notice that the corresponding auto-regressive models were computed by concatenating the facial actions associated with the seven persons.

The second data set consists of five 30 second videos. Each video sequence contains several cycles depicting a given universal expression. All these training videos have depicted a student who simulated the expressions. These training videos have been acquired in our laboratory with a

Fig. 2. Six training videos from the CMU database. The first five images depict the high magnitude of the five basic expressions together with the fitted 3D deformable model.

high-quality camera under daylight. Their resolution is 640×480. Their frame rate is 25 frames per second. We stress the fact the illumination conditions as well as the camera used are not required to be the same for training and testing. The reason is twofold. First, the tracker is based on online appearance. Second, the expression classification does not use facial textures, it only deals with tracked facial actions. The goal of these training videos is to compute the auto-regressive models.

Fig. 3 shows the tracked facial actions associated with five training 30 second videos.

It should be noticed that neutral expressions can present slight deformations even when the face of the subject seems expressionless. However, we have not used auto-regressive models for modeling the dynamics of these slight deformations. Instead we use the $L1$ norm of the vector $\boldsymbol{\tau}_{\mathbf{a}_{(t)}}$ to decide if the corresponding current frame depicts a neutral expression or not.

3.2. Recognition

In our previous works [11,12], we have shown that analysis–synthesis schemes based on learned auto-regressive models can be used for facial expression recognition.

In our work [12], we infer the facial expression in videos by considering the vectors $\boldsymbol{\tau}_{\mathbf{a}_{(t)}}$ within a temporal window of size T centered at the current frame t. These vectors are provided by the 3D face tracker. The expression for frame t is recognized using the following analysis–synthesis scheme. This is a two-step approach. In the first step, we locally synthesize the facial actions, $\hat{\boldsymbol{\tau}}_{\mathbf{a}(i)}$ within the temporal window using all auto-regressive models and the actual tracked facial actions. In the second step, the model providing the most similar facial action trajectory to the actual one will decide the classification.

For the synthesis purpose, we utilize (3). Recall that second-order auto-regressive models require two initial frames. Thus, in our synthesis process, we generate two synthesized facial action trajectories for each expression category according to the choice of these initial values. The first trajectory is generated by invoking (3) for each synthesized frame where the two initial frames are set to their corresponding actual values. The second trajectory is generated by invoking (3) where the two initial frames are set to the local average value of the actual trajectory. In order to get stable synthesized trajectories, Eq. (3) has been used with the random noise $\mathbf{Bw}_{(t)}$ set to zero.

Let $\mathbf{r}$ be the actual facial action trajectory—the concatenation of the tracked facial actions $\boldsymbol{\tau}_{\mathbf{a}_{(t)}}$ within the temporal window of size T. The dimension of $\mathbf{r}$ is $6T$. Let $\mathbf{s}$ be a synthesized facial action trajectory—a $6T$-vector. Since we have five auto-regressive models and two synthesis schemes then we have 10 synthesized trajectories $\mathbf{s}_{kl}, k = 1,\ldots,5, l = 1,2$ (recall that we have two synthesized trajectories for each universal expression).

Comparing the actual trajectory $\mathbf{r}$ with a synthesized one $\mathbf{s}_{kl}$ can be carried out using the cosine of the angle between the two vectors:

$$d_{kl} = \frac{\mathbf{r}^{\mathrm{T}}\,\mathbf{s}_{kl}}{|\mathbf{r}|\,|\mathbf{s}_{kl}|}. \tag{4}$$

Let $d_k = \max(d_{kl}), l = 1,2$. In other words, we only retain the synthesized trajectory that is the most consistent with the actual tracked one. Therefore, the most probable universal expression depicted in the current frame will be given by

$$k = \arg\max_{k}(d_k), \quad k = 1,\ldots,5.$$

The above similarity measures can be normalized. For example, the following normalization can be used:

$$p_k = \frac{\mathrm{e}^{d_k}}{\sum_{j=1}^{5}\mathrm{e}^{d_j}}.$$

Fig. 4 illustrates the analysis–synthesis scheme associated with the lower lip depressor parameter as a function of time (13 frames). Each graph corresponds to a given auto-regressive model. The solid curve corresponds to the actual facial parameter which corresponds to a tracked surprise transition. The solid curve is the same for all graphs. The dashed and dotted curves correspond to the synthesized parameter using the learned auto-regressive models: the dashed ones correspond to the case where the initial conditions are set to the local average while the dotted ones to the case where the initial conditions are set to two actual values. In this real example, one can easily see how the surprise auto-regressive-based synthesized trajectories are very similar to the actual trajectory (top-left).

Based on experimental observations we found that the minimum time required for passing from the neutral expression to the apex expression is about half a second. Therefore, the neighborhood

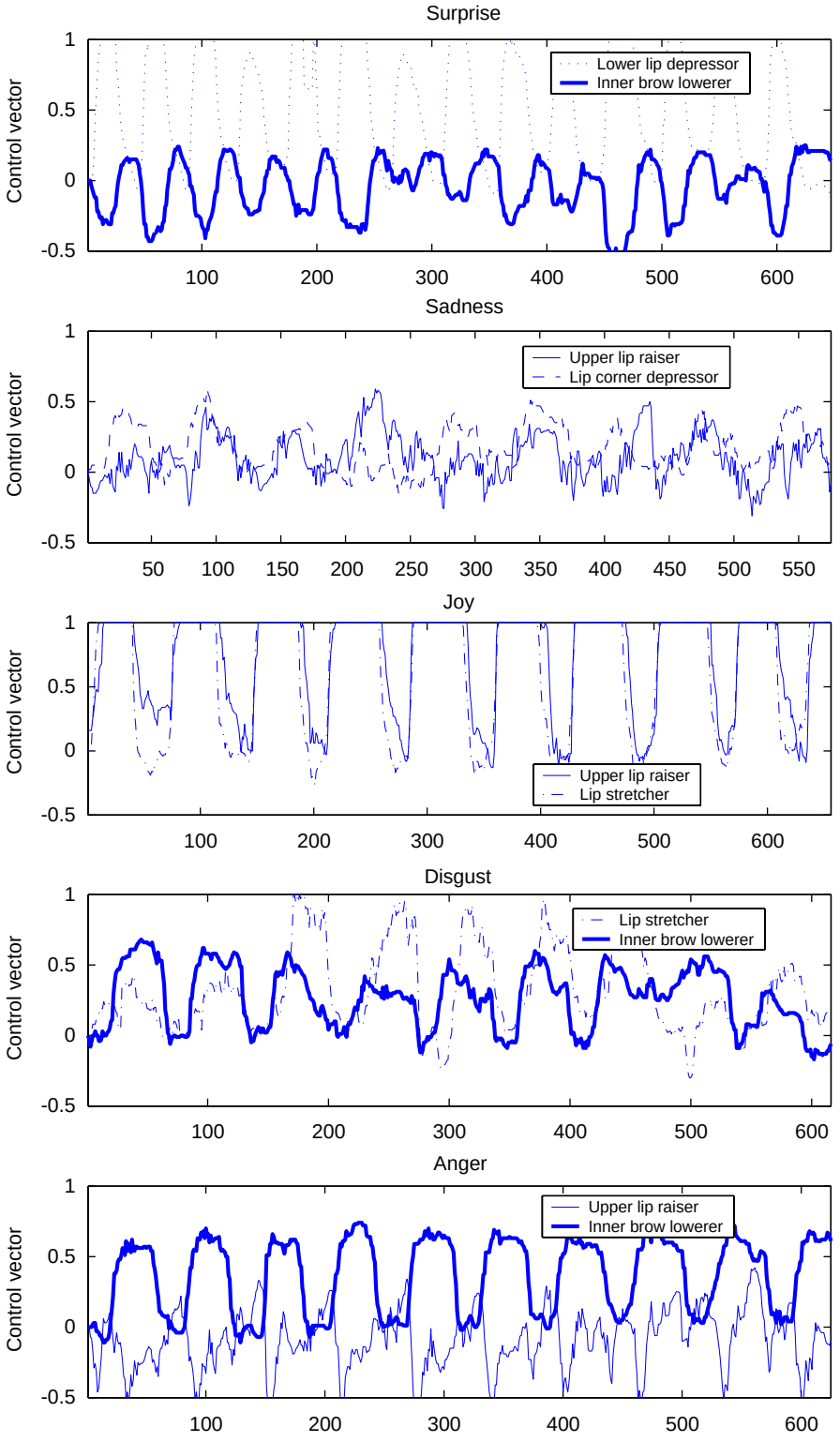

Fig. 3. The tracked facial actions, $\tau_{\mathbf{a}_{(t)}}$, associated with five training videos. For a given plot, only two components are shown.

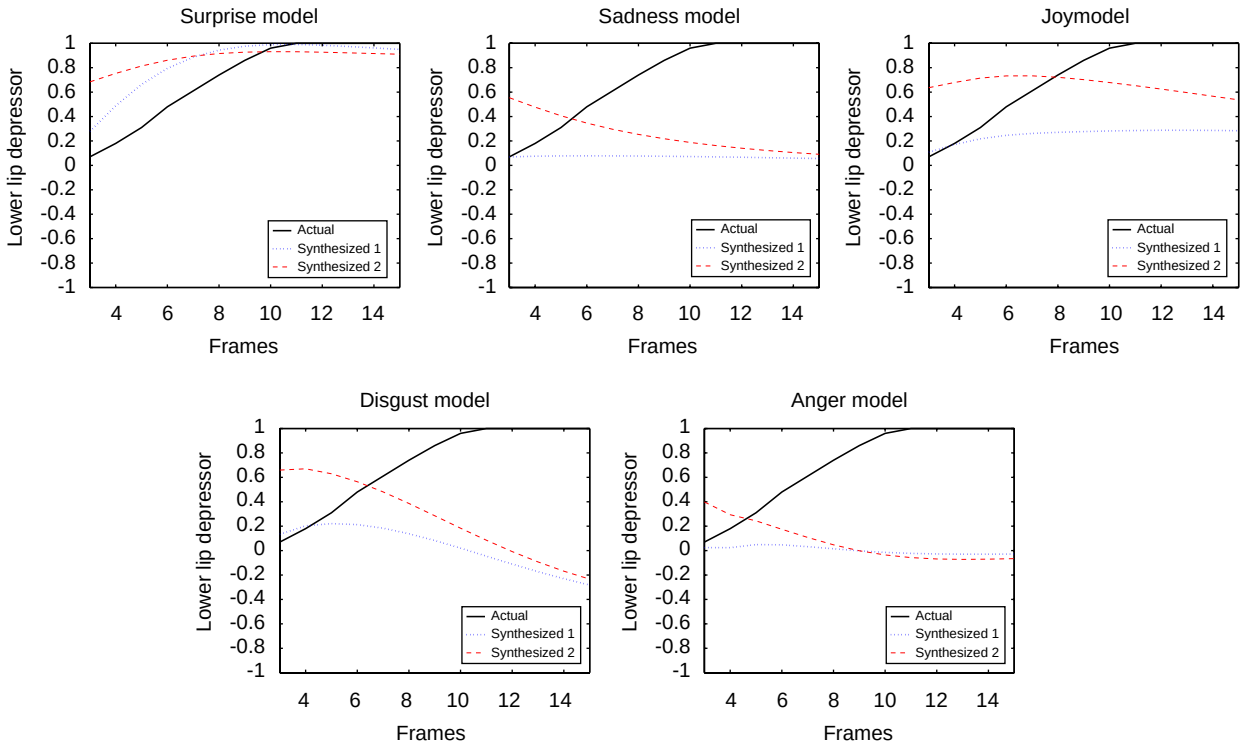

Fig. 4. This example illustrates the analysis–synthesis scheme associated with the lower lip depressor parameter as a function of time (13 frames). Each graph corresponds to a given auto-regressive model. The actual parameter (solid curve) which corresponds to a tracked surprise transition is the same for all graphs. In this example, one can easily see how the synthesized trajectories are very similar to the actual one when the surprise auto-regressive model is used (top-left).

size T should be greater than or equal to 0.5 s. If the frame rate of the video is 25, we can conclude that T should be greater than or equal to 12.5 frames. Our results were obtained with T being set to 15 frames. We have observed that a size which is slightly greater than 15 frames will not affect the recognition results. However, a very small T may affect the recognition results.

3.3. *Efficient detection and recognition*

In many cases, the framewise recognition is not required. In other words, the system is required to provide the expression whenever it appears on the user's face. In this section, we propose a fast and simple scheme able to detect and recognize facial expression in continuous videos without having to apply the above recognition scheme to every frame in the video. Instead, the above scheme is applied only on keyframes. This scheme has two advantages. First, the CPU time corresponding to the recognition part will be considerably reduced. Second, since a keyframe and its surrounding 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. More precisely, the keyframes will correspond to a sudden increase in the facial actions, which usually occurs in a neutral-to-apex transition. Recall that the tracker should process all frames in order to provide the time-varying facial actions. Obviously, we adopt a heuristic definition for the keyframe. Using this definition, this keyframe is forced to be a frame at which several facial actions change significantly.

Therefore, a keyframe can be detected by looking for a local positive maximum in the derivatives of the facial actions. To this end, two entities will be computed from the sequence of facial actions $\boldsymbol{\tau}_{\mathbf{a}}$ that arrive in a sequential fashion. The $L1$ norm $\|\boldsymbol{\tau}_{\mathbf{a}}\|_1$ and the the first derivative $\partial\boldsymbol{\tau}_{\mathbf{a}}/\partial t$. The ith component of this vector $\partial\tau_{\mathbf{a}(i)}/\partial t$ is given using the values associated with

four frames

$$\frac{\partial \tau_{\mathbf{a}(i)}}{\partial t} = 2\left(\tau_{\mathbf{a}(i)_{(t+1)}} - \tau_{\mathbf{a}(i)_{(t-1)}}\right) + \tau_{\mathbf{a}(i)_{(t+2)}} - \tau_{\mathbf{a}(i)_{(t-2)}}, \tag{5}$$

where the subscript i stands for the ith component of the facial action vector $\boldsymbol{\tau}_{\mathbf{a}}$. Since we are interested in detecting the largest variation in the neutral-to-apex transition, we use the temporal derivative of $\|\boldsymbol{\tau}_{\mathbf{a}}\|_1$:

$$D_t = \frac{\partial \|\boldsymbol{\tau}_{\mathbf{a}}\|_1}{\partial t} \tag{6}$$

$$= \sum_{i=1}^{6} \frac{\partial \tau_{\mathbf{a}(i)}}{\partial t}. \tag{7}$$

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. The parameter W controls the rate of the recognition outputs. It does not intervene in the classification process. If one is only interested in detecting and recognizing all the subject's facial expressions, W should be small in order not to skip any displayed expression. In this case, the minimum value should correspond to the duration of the fastest expression (in our case, this was set to 15 frames $\approx$ 0.5 s). On the other hand, when a machine or a robot should react online according to the subject's facial expression W should be large so that the machine can achieve the actions before receiving a new command. In this case, skipping some expressions is allowed.

The whole scheme is depicted in Fig. 5. 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 and 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 $\|\boldsymbol{\tau}_{\mathbf{a}}\|_1$ is greater than a predefined threshold, and (3) it is far from the previous keyframe by at least W frames. Once a keyframe is found in the current segment, the dynamical classifier described in the previous section will be invoked, that is, the temporal window will be centered on the detected keyframe. Adopting the above three conditions is justified by the following facts. The first one is to make sure that the chosen frame corresponds well to a significant facial deformation. The second one is to make sure that the chosen keyframe does not correspond to a small facial deformation (excluding quasi-neutral frames at which the derivatives can be local maxima). The third one is to avoid a large number of detected expressions per unit of time. Every facial action $\boldsymbol{\tau}_{\mathbf{a}(i)}$ is normalized: a zero value corresponds to a neutral configuration and a one value corresponds to a maximum deformation. Thus, the $L1$ norm of the vector encoding these facial actions can be used to decide whether the current frame is a neutral frame or not. In our implementation we used a threshold of 1. A small threshold may lead to the detection of many keyframes since the derivatives have local maxima even for very small facial deformations. On the other hand, many keyframes cannot be detected if one adopts a large threshold. Based on experimental measurements, the $L1$ norm associated with the apex configurations for all universal expressions is between 2 and 3.5.

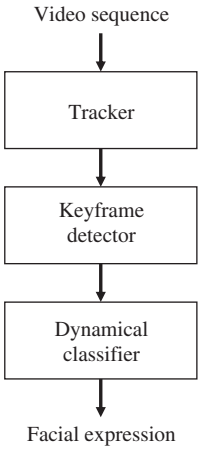

Fig. 5. Efficient detection and recognition based on keyframes.

Fig. 6 shows the results of applying the proposed detection scheme on a 1600-frame sequence containing 23 played expressions. For this 1600-frame test video, we asked our subject to adopt arbitrarily different facial gestures and expressions for an arbitrary duration and in an arbitrary order. In this video, there is always a neutral face between two expressions. The solid curve corresponds to the norm $\|\boldsymbol{\tau}_{\mathbf{a}}\|_1$, 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. Fig. 7 shows the results of applying the

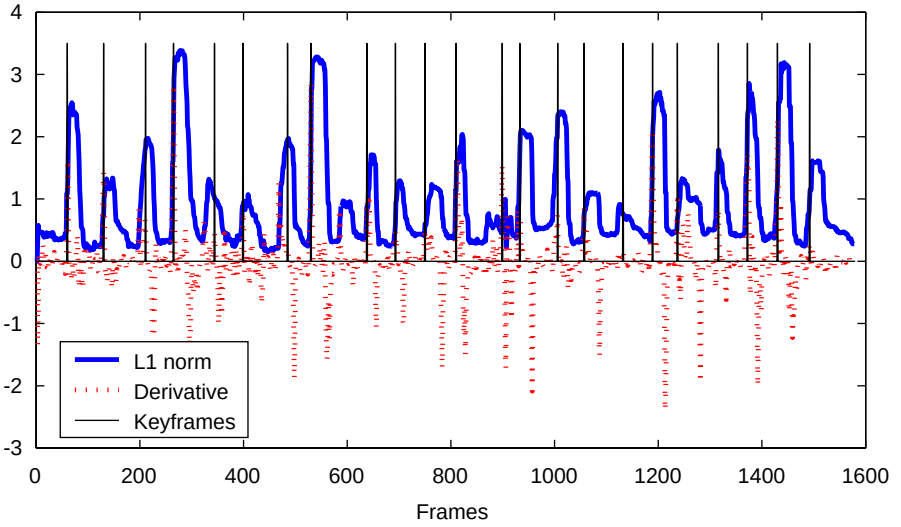

Fig. 6. Keyframe detection and recognition applied on a 1600-frame sequence.

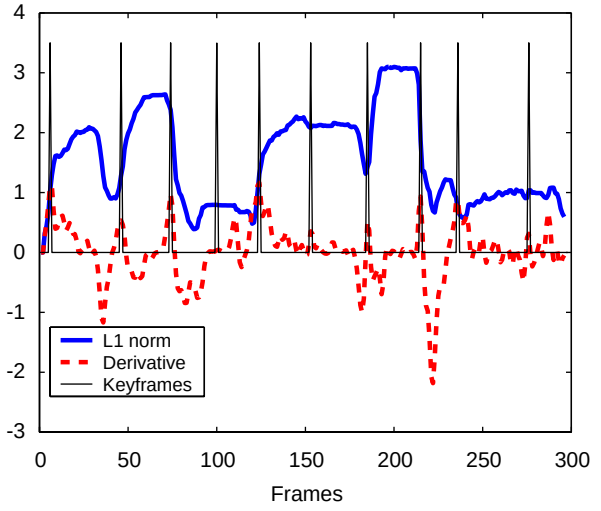

Fig. 7. Keyframe detection and recognition applied on a 300-frame sequence.

proposed scheme on a 300-frame sequence. The video contained nine arbitrary displayed expressions performed by a researcher. Some of the displayed expressions has a long duration coupled with a facial deformation. This explains why the detected keyframes are 10.

The decrease in the CPU is better illustrated in the example shown in Fig. 6. This figure shows 1600 frames with 23 detected keyframes. Let t_c be the CPU time associated with the recognition task (the temporal classifier described in Section 3.2). If we use a framewise recognition scheme then the CPU time for processing the whole video will be $1600t_c$. Should we use the efficient recognition scheme based on keyframes, the same CPU time becomes $23t_c$. So the decrease in the CPU time spent for recognition is $\frac{1600}{23} = 69.5$. In other words, the efficient recognition is 69.5 times faster then the frame-wise recognition scheme. Note that even when the framewise recognition scheme skips all neutral frames the decrease in the CPU time will remain significant.

This efficient detection scheme will be used by the application we will introduce in Section 5.

4. Experimental results

4.1. Recognition

Our first set of experiments was performed on three video sequences. Each test video was acquired by a different camera and depicted a series of facial expressions performed by an unseen subject. In other words, the subject in each test video was different from those used for learning the auto-regressive models. The three videos were performed by one Ph.D. student and two researchers. They contain 10, 9, and 4 expressions, respectively. Since they are for testing the classifier, the produced expressions are used as they are. We label the displayed expressions according to the subject's claim.

In the first experiment, we have used a 748-frame test sequence. Eight frames of this sequence are

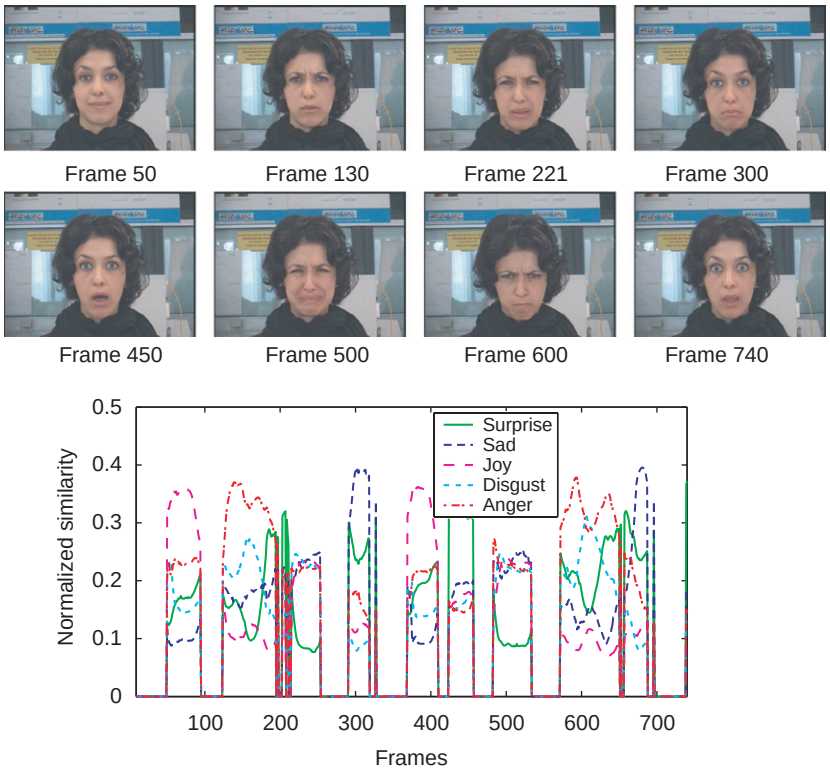

Fig. 8. *Top*: Eight frames associated with a 748-frame test sequence. *Bottom*: The similarity measure computed for each universal expression and for each non-neutral frame of the sequence.

shown in Fig. 8. The number of the displayed expressions in this original video is 10. The bottom of this figure shows the normalized similarities associated with each universal expression obtained with the sequence using a sliding temporal window of 15 frames. The used auto-regressive models were built with the CMU data. By inspecting the original video we have found that all displayed expressions were correctly classified by the developed approach (Section 3) except for the disgust expression for which the approach provides a mixture of three expressions (see the similarity curves at frames 200 and 500). Note that the temporal window size should be greater than or equal to the minimum time needed by an expression to go from the neutral configuration to a perceived expression.

In the second experiment, we used a 300-frame video sequence. For this sequence, we asked a subject to display several expressions arbitrarily (see Fig. 9). The bottom of this figure shows the normalized similarities associated with each universal expression. In this case, the auto-regressive models were built with the second training data set (the five 30 second videos). As can be seen, the algorithm 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 framewise manner.

One can notice that the images 110 and 250 are almost the same but the similarity measures are not the same even though the maximum of these measures is indicating a sadness expression for both frames. The reason for this is that the similarity measure for frame 110 is based on frames 103–117 (the temporal window size T is set to 15). In a similar manner, the similarity measure for frame 250 is based on frames 243–257. We compared the 15 frames centered at frame 110 with the 15 frames centered at frame 250. In the first case, we found that the lower lip moved upwards (frames 103–110), whereas in the second case (frames 243–250) the lower lip was motionless. Thus, even though frames 110 and 240 are the same, their seven preceding frames are not the same. This results in different similarity measures.

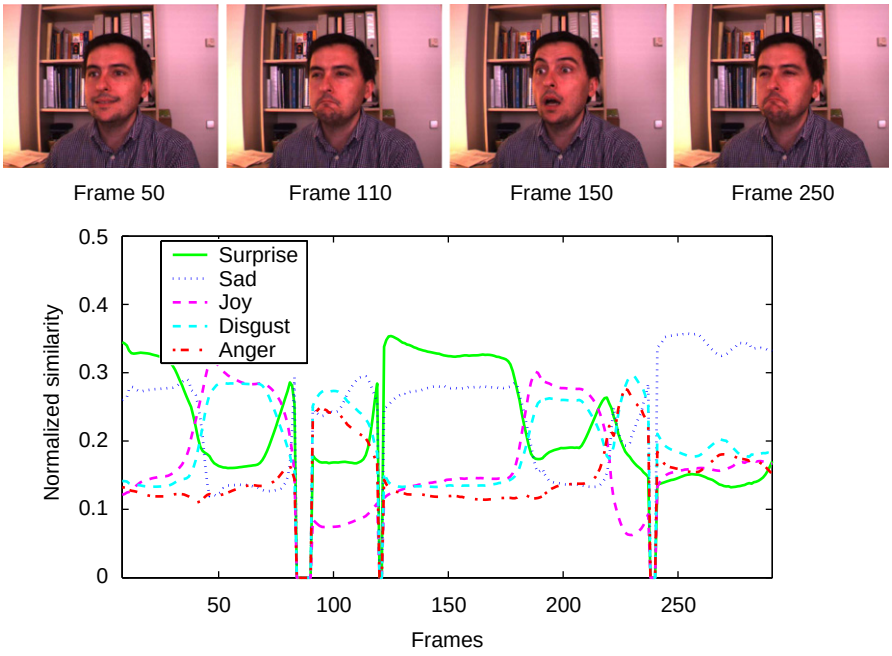

Fig. 9. *Top*: Four frames (50, 110, 150, and 250) associated with a 300-frame test sequence. *Bottom*: The similarity measure computed for each universal expression and for each non-neutral frame of the sequence.

Fig. 10. Three frames associated to the third test sequence. The recognition results are indicated in parentheses.

In the third experiment, we have used a 325-frame video sequence. Fig. 10 shows the recognition results associated with this video.

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.

The proposed approach does not require a frontal face since the facial actions and the 3D head pose are simultaneously tracked. However, there are some limitations that are inherited from the 3D face tracker. Indeed, the proposed method builds on our 3D face tracker [10]. This developed tracker can easily track out-of-plane face rotations (yaw and pitch angles) belonging to the interval $[-45°, 45°]$. There is no limitation on the roll angle since the face will be visible regardless of the value of this angle. We have processed two videos depicting out-of-plane face motions: the first one is shown in Fig. 9 and the second one is shown in Fig. 16.

4.2. Performance study

In order to quantify the recognition rate, we have used the 35 CMU videos for testing using the auto-regressive models built with the second training data set. Table 1(a) shows the confusion matrix associated with the 35 test videos illustrating seven persons. Table 1(b) shows the same confusion matrix obtained with the method proposed in [9]. This method is based on the dynamic time warping technique.

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

Table 1
Confusion matrices for the facial expression classifier associated with 35 test videos (CMU database)

	Surp. (7)	Sad. (7)	Joy (7)	Disg. (7)	Ang. (7)
(a) *Auto-regressive models* (*proposed approach*)					
Surp.	7	0	0	0	0
Sad.	0	7	0	5	0
Joy	0	0	7	0	0
Disg.	0	0	0	2	2
Ang.	0	0	0	0	5
(b) *Dynamic time warping technique*					
Surp.	7	0	0	0	0
Sad.	0	7	0	5	0
Joy	0	0	6	0	0
Disg.	0	0	1	2	4
Ang.	0	0	0	0	3

The training data are associated with one unseen person. (a) Illustrates the confusion matrix obtained with the proposed approach. (b) Illustrates the confusion matrix obtained with a dynamic programming approach.

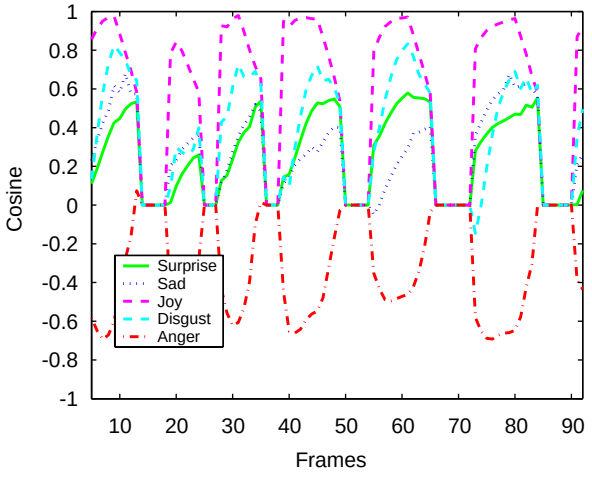

Fig. 11. The cosine angle associated with seven short video illustrating joy expressions performed by seven different persons (the videos are concatenated into one single video). As can be seen, the maximum of the cosine correctly indicates the joy expression.

could be highly subject dependent. Recall that the used auto-regressive models are built using data associated with one single person. Notice that the human ceiling in correctly classifying facial expressions into the six basic emotions has been established at 91.7% by Ekman and Friesen [14].

Fig. 11 summarizes the joy test data (CMU data) used for the confusion matrix computation. This figure shows the value of the cosine as defined by (4) for seven test videos concatenated into one single sequence.

In another set of experiments, we have used the model built with the CMU data (seven persons) and then asked two unseen persons to play the universal expressions. Tables 2 and 3 show the confusion matrices associated with the two persons, respectively. By combining the results associated with these two tables, we can see that out of 101 played expressions there were 14 misclassified expressions leading to a recognition rate of 86.14%. If every person is considered separately then the recognition rate will be 98.1% for the first person and 73.5% for the second one. Since facial expressions and dynamics are subject dependent, it is not surprising to get different recognition results even when the same learned auto-regressive models are used.

Table 2
Confusion matrix associated with an unseen person's videos

	Surp. (14)	Sad. (9)	Joy (10)	Disg. (9)	Ang. (10)
Surp.	14	0	0	0	0
Sad.	0	9	0	0	0
Joy	0	0	10	1	0
Disg.	0	0	0	8	0
Ang.	0	0	0	0	10

Table 3
Confusion matrix associated with an unseen person's videos

	Surp. (10)	Sad. (9)	Joy (9)	Disg. (10)	Ang. (11)
Surp.	10	0	0	0	0
Sad.	0	9	0	0	11
Joy	0	0	9	2	0
Disg.	0	0	0	8	0
Ang.	0	0	0	0	0

5. Application

Interpreting non-verbal face gestures is used in a wide range of applications. An intelligent user-interface not only should interpret the face motions but also should interpret the user's emotional state [6,22]. 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. Although the presented application is independent of the technique, it might be very useful for many practitioners working in the domain of human–computer interaction. Without loss of generality, we use the AIBO robot which has the advantage of being especially designed for human–computer interaction (see Fig. 12). The experimental setup is depicted in Fig. 13. The input to the system is a video stream capturing the user's face.

Fig. 12. The AIBO robot.

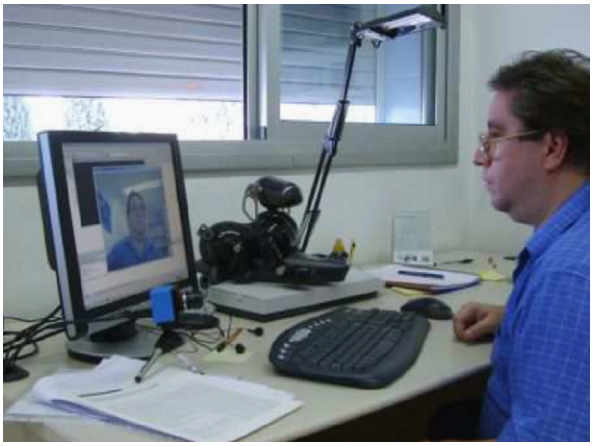

Fig. 13. The experimental setup.

5.1. The AIBO robot

AIBO is a biologically inspired robot and is the flagship of a whole generation (social robotics). Its name can be interpreted in two ways: one is to see it as an acronym for 'Artificial Intelligent RoBOt'; on the other hand, its name in japanese means 'pal', 'companion'. Created initially for entertainment purposes only, it was rapidly adopted by the scientific community which saw it as a very powerful 'toolbox' to test and to develop different theories related with the field of social robotics (like cognitive learning, affective computing, etc.). A very important characteristic is that it possess an 'innate' sense of curiosity. In consequence, its behavioral patterns will develop as it learns and grows. It matures through a continuous interaction with the environment and the people it cohabitates with. For this reason, each AIBO is unique. Its human-like communication system is implemented through series of instincts and senses (affection, movement, touch, hearing, sight and balance senses). One of the most crucial instincts is the 'survival' instinct. Whenever it feels the battery level is below a certain value it starts searching the recharging station. Another one is represented by its ability to display emotional states.

AIBO is able to show its emotions through an array of LEDs situated in the frontal part of the head. These are depicted in Fig. 14, and are shown in correspondence with the six universal expressions. Notice that the blue lights that appear, in certain images, on each part of the head, are blinking LEDs whose meaning is to inform that

Fig. 14. AIBO is able to show its emotions through an array of LEDs situated in the frontal part of the head. The figure illustrates the LEDs' configuration for each universal expression.

the robot is remotely controlled.[2] This is a built-in feature and cannot be turned off. In addition to the LEDs' configuration, the robot response contains some small head and body motion.

5.2. Results

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 multimodal 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 3.3), 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. In other words, we wanted to see its reaction according to the emotional state displayed by a person. 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: www.cvc.uab.es/~bogdan/AIBO-emotions.avi. In order to be able to display simultaneously in the video the correspondence between person's and robot's expressions, we put them side by side. In this case only, we analyzed offline the content of the video and commands with the facial expression code were sent to the robot.

Fig. 15 illustrates nine detected keyframes from the 1600 frame video depicted in Fig. 6. These are shown in correspondence with the robot's response. The middle column shows the recognized expression. The right column shows a snapshot of the robot head when it interacts with the detected and recognized expression.

6. Conclusion

This paper described a view- and texture-independent approach to facial expression analysis and

[2]AIBO can function in two modes: autonomous and remote controlled. The application described in this paper, was built using the remote framework (RFW) programming environment (based on C++ libraries), which works on a client-server architecture over a wireless connection between a PC and the AIBO.

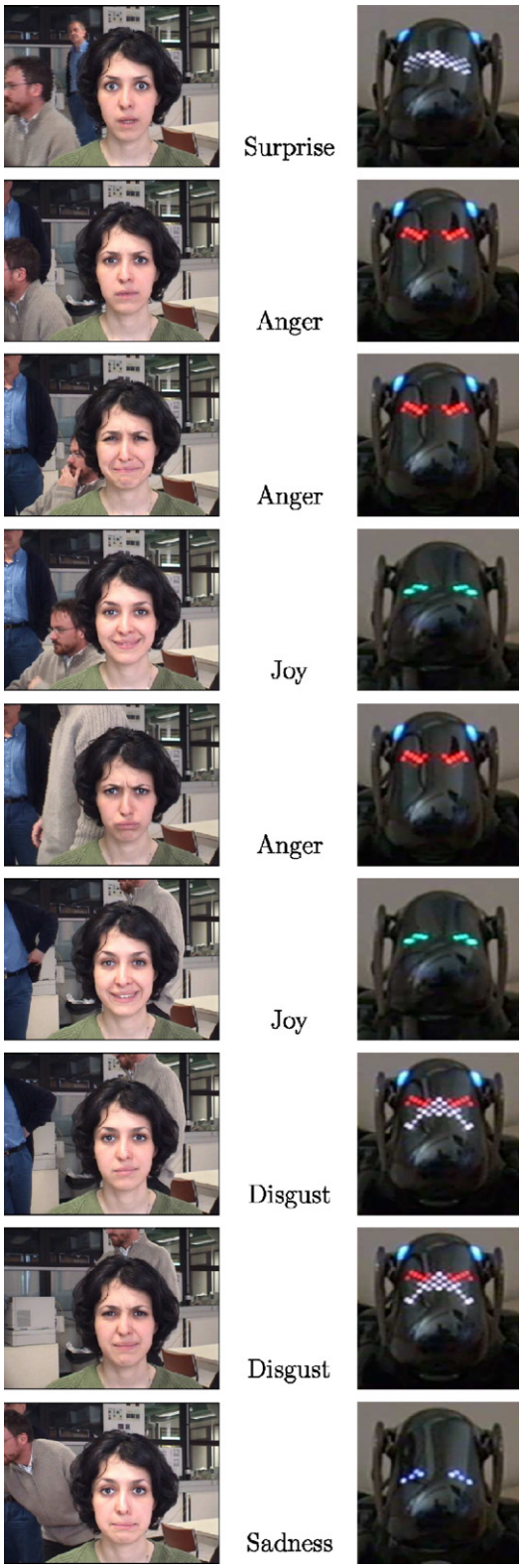

Fig. 15. *Left column*: Some detected keyframes associated with the 1600-frame original video. *Middle column*: The recognized expression. *Right column*: The corresponding robot's response.

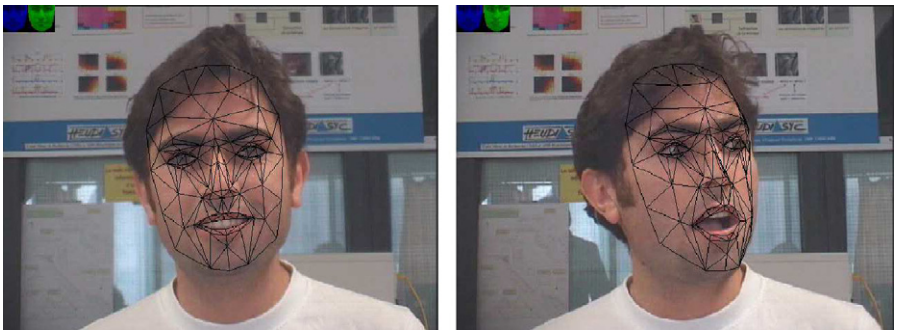

Fig. 16. Facial expression recognition associated with a video depicting out-of-plane face motions.

recognition. The paper presented three contributions. First, we proposed a temporal recognition scheme that classifies a given image in an unseen video into one of the universal facial expression categories using an analysis–synthesis scheme. The proposed approach relies on tracked facial actions provided by a real-time face tracker. Second, we propose an efficient recognition scheme based on the detection of keyframes in videos. Third, we applied the proposed method in a human–computer interaction scenario, in which an AIBO robot is mirroring the user's recognized facial expression. Future work may investigate the use of the on-board camera for recognizing the users' facial expressions.

The detection of keyframes is very robust since it utilizes geometrical constraints on normalized facial actions. However, the classifier has some limitations that are directly related to the ability of the learned models to be as universal as possible. One possible solution is to construct one auto-regressive model per subject and per expression. At run time, the synthesis scheme can use all available auto-regressive models that are associated with a given universal expression. In our study, we have assumed that a quasi-neutral expression occurs between two consecutive facial expressions. However, in practice, should two facial expressions occur immediately one after the other, one can expect that the system can capture the keyframe corresponding to the second displayed expression. Indeed, two conditions can be satisfied for detecting the corresponding keyframe. (1) One or more facial action (the total number is 6) will increase since there is a transition from one apex configuration to another apex configuration. (2) There is a significant facial deformation.

It is worth noting that the keyframes are defined using a heuristic based on the tracked facial actions. Since the final goal of the system is to detect and recognize the facial expression, it does not matter where the detected keyframe is located in the time domain. A detected keyframe simply indicates that there is a significant facial deformation which requires further processing by the expression classifier. In general, there is no one to one mapping between the detected keyframes and the displayed expressions since the latter ones may have arbitrary durations and modes.

In our study, we tracked facial actions associated with the mouth and the eyebrows only. Many studies have shown that image regions associated with the mouth and the eyebrows are the most informative regions about the facial expressions. Certainly, the configuration of the eye openings is affected by the surprise and joy expressions. However, the movements of the mouth and the eyebrows are more informative than those associated with the eye openings.

We believe that the strength of our approach is better exploited when the same person or dynamics are used. This fact is consistent with many researchers' findings that stipulate that temporal expression classifiers are very accurate when they deal with the same person.

References

[1] J. Ahlberg, CANDIDE-3—an updated parametrized face, Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden, 2001.

[2] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, J. Movellan, Machine learning methods for fully automatic recognition of facial expressions and facial actions, in: IEEE International Conference on Systems, Man and Cybernetics, vol. I, The Hague, The Netherlands, October 2004, pp. 592–597.

[3] J.N. Bassili, Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face, J. Pers. Social Psychol. 37 (1979) 2049–2059.

[4] M.J. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, Int. J. Comput. Vision 25 (1) (1997) 23–48.
[5] A. Blake, M. Isard, Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion, Springer, New York, 2000.
[6] C. Breazeal, B. Scassellati, Robots that imitate humans, Trends Cognitive Sci. 6 (2002) 481–487.
[7] L. Cañamero, P. Gaussier, in: Emotion understanding: robots as tools and models, Emotional Development: Recent Research Advances, 2005, pp. 235–258.
[8] I. Cohen, N. Sebe, A. Garg, L. Chen, T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling, Comput. Vision Image Understanding 91 (1–2) (2003) 160–187.
[9] F. Dornaika, F. Davoine, View- and texture-independent facial expression recognition in continuous videos using dynamic programming, in: IEEE International Conference on Image Processing, 2005.
[10] F. Dornaika, F. Davoine, On appearance based face and facial action tracking, IEEE Trans. Circuits Syst. Video Technol. 16 (9) (September 2006) 1107–1124.
[11] F. Dornaika, F. Davoine, Recognizing facial expressions in videos using autoregressive models, in: IEEE International Conference on Pattern Recognition, Hong Kong, 2006, pp. 520–523.
[12] F. Dornaika, B. Raducanu, Recognizing facial expressions in videos using a facial action analysis–synthesis scheme, in: IEEE International Conference on Advanced Video and Signal based Surveillance, Australia, 2006.
[13] P. Ekman, Facial expressions of emotion: an old controversy and new findings, Philos. Trans. Roy. Soc. London B 335 (1992) 63–69.
[14] P. Ekman, W. Friesen, Pictures of Facial Affect, Consulting Psychologists Press, Palo Alto, CA, USA, 1976.
[15] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (1) (2003) 259–275.
[16] T. Kanade, J. Cohn, Y.L. Tian, Comprehensive database for facial expression analysis, in: International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000, pp. 46–53.
[17] Y. Kim, S. Lee, S. Kim, G. Park, A fully automatic system recognizing human facial expressions, in: Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, vol. 3215, 2004, pp. 203–209.
[18] L. Ljung, System Identification: Theory for the User, second ed., Prentice-Hall International, UK, 1987.
[19] B. North, A. Blake, M. Isard, J. Rittscher, Learning and classification of complex dynamics, IEEE Trans. Pattern Anal. Mach. Intell. 22 (9) (2000) 1016–1034.
[20] M. Pantic, in: Encyclopedia of Multimedia Technology and Networking, Affective Computing, vol. 1, 2005, pp. 8–14.
[21] R.W. Picard, E. Vyzas, J. Healy, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.
[22] H. Prendinger, J. Mori, M. Ishizuka, Recognizing, modeling, and responding to users' affective states, in: Proceedings of the 10th International Conference on User Modeling (UM-05), Springer Lecture Notes in Artificial Intelligence, vol. 3538, 2005, pp. 60–69.
[23] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[24] J. Sung, S. Lee, D. Kim, A real-time facial expression recognition using the STAAM, in: International Conference on Pattern Recognition, vol. I, Hong Kong, 2006, pp. 275–278.
[25] Y. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 97–115.
[26] M. Yeasin, B. Bullot, R. Sharma, Recognition of facial expressions and measurement of levels of interest from video, IEEE Trans. Multimedia 8 (3) (2006) 500–508.
[27] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 699–714.