Engineering Applications of Artificial Intelligence 26 (2013) 467–477

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Facial expression recognition using tracked facial actions: Classifier performance analysis

Fadi Dornaika [a,b,*], Abdelmalik Moujahid [a], Bogdan Raducanu [c]

[a] *Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Spain*
[b] *IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*
[c] *Computer Vision center, Barcelona, Spain*

ARTICLE INFO

Article history:
Received 19 August 2011
Received in revised form
2 September 2012
Accepted 3 September 2012
Available online 29 September 2012

Keywords:
Visual face tracking
3D deformable models
Facial actions
Dynamic facial expression recognition
Human–computer interaction

ABSTRACT

In this paper, we address the analysis and recognition of facial expressions in continuous videos. More precisely, we study classifiers performance that exploit head pose independent temporal facial action parameters. These are provided by an appearance-based 3D face tracker that simultaneously provides the 3D head pose and facial actions. The use of such tracker makes the recognition pose- and texture-independent. Two different schemes are studied. The first scheme adopts a dynamic time warping technique for recognizing expressions where training data are given by temporal signatures associated with different universal facial expressions. The second scheme models temporal signatures associated with facial actions with fixed length feature vectors (observations), and uses some machine learning algorithms in order to recognize the displayed expression. Experiments quantified the performance of different schemes. These were carried out on CMU video sequences and home-made video sequences. The results show that the use of dimension reduction techniques on the extracted time series can improve the classification performance. Moreover, these experiments show that the best recognition rate can be above 90%.

1. Introduction

1.1. Overview

In recent times, there has been considerable technical progress within artificial intelligence in the field of computer vision to open the possibility of placing faces at the center of human–computer interaction (HCI). Facial expressions play an important role in recognition of human emotions. Psychologists postulate that facial expressions have a consistent and meaningful structure that can be backprojected in order to infer people inner affective state. Basic facial expressions typically recognized by psychologists are: happiness, sadness, fear, anger, disgust and surprise (Ekman, 1992). In the beginning, 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 (Fasel and Luettin, 2003; Pantic and Patras, 2006; Yeasin et al., 2006). The question of how to further exploit the results of the recognized facial expression actually motivates and fosters ongoing research in HCI, artificial intelligence and cognitive science. The field of 'emotional machines' (machines responsive to our emotions) is a vastly unexplored research domain with enormous potential.

A facial expression is formed by contracting or relaxing different facial muscles on human face which results in temporally deformed facial features like raising eyebrows and open mouth. The automated analysis of facial expressions is a challenging task because everyone's face is unique and interpersonal differences exist in how people perform facial expressions. Numerous methodologies have been proposed to solve this problem (Bartlett et al., 2006; Cheon and Kim, 2009; Naghsh-Nilchi and Roshanzamir, 2006; Sebe et al., 2007; Xiao et al., 2011; Zeng et al., 2009; Zhang et al., 2008).

1.2. Related works

In the past, a lot of effort was dedicated to recognize facial expression in still images (static recognition). For this purpose, many techniques have been applied: neural networks (Tian et al., 2001), Gabor wavelets (Bartlett et al., 2006) and active appearance models (AAM) (Sung and Kim, 2009). A very important limitation to the static strategy for facial expression recognition is the fact that still images usually capture the apex of the expression, i.e., the instant at which the indicators of emotion are most

* Corresponding author at: Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Manuel Lardizabal, 1, 20018 San Sebastian, Spain.
E-mail address: fdornaika@hotmail.fr (F. Dornaika).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.09.002

marked. Despite the fact that some of these techniques addressed non-apex expressions, their objective was to detect and recognize action units (e.g., Bartlett et al., 2006). In Zhang et al. (2012), the authors construct a sparse representation classifier (SRC). The effectiveness and robustness of the SRC method is investigated on clean and occluded facial expression images. Three typical facial features, i.e., the raw pixels, Gabor wavelets representation and local binary patterns (LBP) are extracted to evaluate the performance of the SRC method. In Moore and Bowden (2011), a sequential two stage approach is taken for pose classification and view dependent facial expression classification to investigate the effects of yaw variations from frontal to profile views. Local binary patterns (LBPs) and variations of LBPs as texture descriptors are investigated. Multi- class support vector machines are adopted to learn pose and pose dependent facial expression classifiers.

More recently, attention has been shifted particularly towards modeling dynamical facial expressions (Xiang et al., 2008; Robin et al., 2011). Recent research has shown that it is not just the particular facial expression, but also the associated dynamics that are important when attempting to decipher its meaning. The dynamics of facial expression can be defined as the intensity of the action units coupled with the timing of their formation. This is a very relevant observation, since for most of the communication act, people rather use 'subtle' facial expressions than showing deliberately exaggerated poses in order to convey their message. In Ambadar et al. (2005), the authors found that subtle expressions that were not identifiable in individual images suddenly became apparent when viewed in a video sequence.

Dynamical approaches can use shape deformations, texture dynamics (Yang et al., 2008) or a combination of them (Cheon and Kim, 2009). Dynamic classifiers try to capture the temporal pattern in the sequence of feature vectors related to each frame such as the hidden Markov models (HMMs) and dynamic Bayesian networks (Zhang and Ji, 2005). Cheon and Kim (2009) propose a dynamic recognition based on the differential active appearance model parameters. A sequence of input frames is fitted using the classical AAM, then a specific frame is selected as reference frame. The corresponding sequence of differential AAM parameters is recognized by computing the directed Hausdorff distance and the K nearest neighbor classifier. In Yeasin et al. (2006), a two-stage approach is used. Initially, a linear classification bank was applied and its output was fused to produce a characteristic signature for each universal facial expression. The signatures thus computed from the training data set were used to train discrete hidden Markov models to learn the underlying model for each facial expression. In Shan et al. (2006), the authors propose a Bayesian approach to modeling temporal transitions of facial expressions represented in a manifold. Xiang et al. (2008) propose a dynamic classifier that is based on building spatio-temporal model for each universal expression derived from Fourier transform. The recognition of unseen expressions uses Hausdorff distance in order to compute dissimilarity values for classification. Dornaika and Raducanu (2007) propose a dynamic classifier that is based on an analysis-synthesis scheme exploiting learned predictive models given by second order Markov models. Local binary patterns have been used for facial expression recognition in Shan et al. (2009) and Zhao and Pietikinen (2007).

Wu et al. (2010) explore Gabor motion energy (GME) filters as a biologically inspired representation for dynamic facial expressions. They show that GME filters outperform the Gabor energy filters, particularly on difficult low intensity expression discrimination. Huang et al. (2011) combine some extracted facial feature sets using confidence level strategy. Noting that for different facial components, the contributions to the expression recognition are different, they propose a method for automatically learning different weights to components via the multiple kernel learning. Meng et al. (2011) use two types of descriptors motion history histogram (MHH) and histogram of local binary patterns (LBP). Based on these two basic types of descriptors, two new dynamic facial expression features are proposed. Moore et al. (2010) uses weak classifiers are formed by assembling edge fragments with chamfer scores. An ensemble framework is presented with all-pairs binary classifiers. An error correcting support vector machine (SVM) is utilized for final classification.

1.3. Paper contribution

Automatic facial expression recognition from video sequences is a very challenging task. Indeed, one has to use several modules in sequence: face detection, model fitting, 3D face tracking, face deformation tracking before applying a classifier that can infer the type of the displayed expression. Therefore, the problems of face detection, 3D face tracking, and facial action tracking are out of the scope of the paper. For the completeness of presentation, our face recognition system is depicted in Fig. 1. We stress the fact that the focus of the paper is on the third stage, namely the dynamic facial expression recognition. The majority of the proposed dynamic facial expression techniques assume high resolution frontal facial images. However, very few works have been done in order to recognize facial expression in the presence of head motion in 3D space. Although Moore and Bowden (2011) studied facial expression recognition under different poses, it is a static method that infers the expression from one single snapshot.

In this paper, we focus on the dynamic facial expression recognition in the presence of head motion. The recognition follows the extraction and tracking of facial actions using our 3D face and facial action tracking system (Dornaika and Davoine, 2006). Adopting such a 3D face tracker will overcome two main disadvantages associated with many existing dynamic recognition schemes. First, the expression recognition will not depend on the texture appearance, and hence more flexibility is gained in the sense that the learned models are independent from texture appearances and their changes (texture independence). This a clear

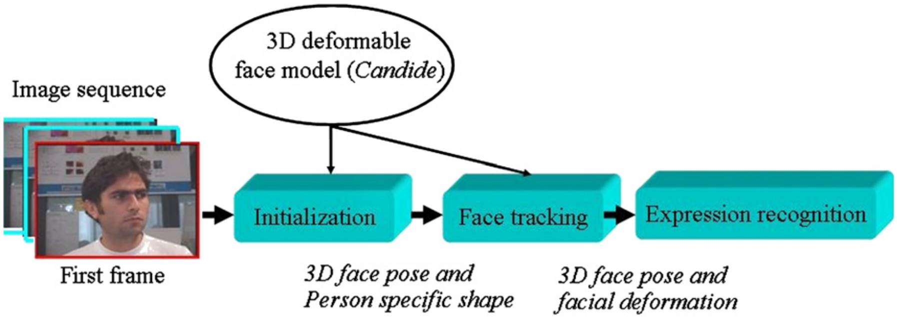

3D deformable face model (Candide)
Image sequence
First frame
Initialization
Face tracking
Expression recognition
3D face pose and Person specific shape
3D face pose and facial deformation

Fig. 1. Face recognition based on tracked facial deformation using the standard deformable face model *Candide.*

advantage over the methods relying on texture variations whose performance may be affected if significant noise affect the images. Second, since the tracked facial actions are associated with a generic 3D deformable face model (they are not expressed in the image plane), the facial expression recognition can be performed even in the presence of head motion (view independence).

The main contribution of the paper is the application and comparison of some machine learning schemes allowing the recognition of facial expressions from temporal facial actions (local facial deformations). More precisely, we explore two schemes that exploit facial action parameters estimated by our tracker (Dornaika and Davoine, 2006). The first scheme adopts a dynamic time warping technique for recognizing expressions where the training data are a set of signature examples associated with different universal facial expressions. The second scheme casts the dynamic recognition problem into a classification problem. It models temporal signatures associated with facial actions with fixed length feature vectors (observations), and uses some machine learning algorithms in order to recognize the displayed expression.

A related work can be found in Chakraborty et al. (2009). This work addresses emotion detection in high resolution images illustrating upright and frontal faces. The learning phase consists of three phases. First, three facial attributes (measured in image plane) are estimated using some image processing techniques. These facial attributes are mouth opening, eye opening, and eyebrow constriction. Then, every attribute measure is encoded into three distinct fuzzy set, each indicating the fuzzyness membership to a magnitude level (low, moderate, and high). A mapping from the fuzzified measurement space of facial attributes to the fuzzified emotion space is then constructed in order to recognize the emotion in test images. The main differences between our work and Chakraborty et al. (2009) are as follows: (i) our facial actions are directly linked to the standard facial action coding system (FACS), (ii) our retrieved facial actions are expressed in a local head coordinate system, which means that these actions can be retrieved even in the presence of head motions, and (iii) our work recognizes facial expressions by analyzing the temporal evolution of the facial action intensities, whereas Chakraborty et al. (2009) uses the average value of facial attribute over the images of the sequence, and (iv) our facial actions are retrieved in a more principled way based on a real-time tracker, whereas the facial attributes in Chakraborty et al. (2009) are retrieved using ad hoc techniques.

The rest of the paper is organized as follows. Section 2 describes the deformable 3D face model that we use to represent the face shape. Section 3 reviews our used face and facial action tracker. Section 4 describes the used two strategies for dynamic facial expression. Section 5 presents experimental results obtained with CMU subset as well as with some home-made video sequences. It also shows the performance of some classifiers. Section 6 concludes the paper.

2. Candide 3D model

Building a generic 3D face model is a challenging task. Indeed, such a model should account for the differences between specific human faces as well as between different facial expressions. This modeling was explored in the computer graphics, computer vision and model-based image coding communities. In our study, we use the 3D face model *Candide* (Ahlberg, 2002). This 3D deformable wireframe model was first developed for the purpose of model-based image coding. The 3D shape of this model is directly recorded in coordinate form, i.e., the 3D coordinates of the vertices. The theoretical 3D face model is given by the 3D coordinates of the vertices $\mathbf{P}_i$, $i=1,\dots,n$ 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

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{S}\boldsymbol{\tau}_\mathbf{s} + \mathbf{A}\boldsymbol{\tau}_\mathbf{a} \tag{1}$$

where $\overline{\mathbf{g}}$ is the standard shape of the model, and the columns of $\mathbf{S}$ and $\mathbf{A}$ are the shape and action units, respectively. A shape unit provides a way to deform the 3D wireframe such as to adapt the eye width, the head width, the eye separation distance, etc. Thus, the term $\mathbf{S}\boldsymbol{\tau}_\mathbf{s}$ accounts for shape variability (inter-person variability) while the term $\mathbf{A}\boldsymbol{\tau}_\mathbf{a}$ accounts for the facial action (intra-person variability). The shape and action variabilities can be approximated well enough for practical purposes by this linear relation. Also, we assume that the two kinds of variability are independent. In this study, we use 12 modes for the shape unit matrix and six modes for the action units matrix.

In Eq. (1), the 3D coordinates are expressed in a local coordinate system. However, one should relate the 3D coordinates to the image coordinate system (the 2D image coordinates). To this end, we adopt the weak perspective projection model. We neglect the perspective effects since the depth variation of the face can be considered as small compared to its absolute depth.[1]

For a given person, $\boldsymbol{\tau}_\mathbf{s}$ is constant. Estimating $\boldsymbol{\tau}_\mathbf{s}$ can be carried out using either feature-based or featureless approaches. In our recent work, we have shown that some components of the shape control vector can be automatically initialized with a featureless approach (Dornaika and Raducanu, 2010). The state of the 3D model is given by the 3D head pose (three rotations and three translations) and the control vector $\boldsymbol{\tau}_\mathbf{a}$. This is given by

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

where

- θ_x, θ_y, and θ_z represent the three angles associated with the 3D rotation between the 3D face model coordinate system (the user's face) and the camera coordinate system. In our case, the direction of the user's gaze is given by the two angles θ_x and θ_y.
- t_x, t_y, and t_z represent the three components of the 3D translation vector between the 3D face model coordinate system and the camera coordinate system.
- Each component of the vector $\boldsymbol{\tau}_\mathbf{a}$ represents the intensity of one facial action such as eyelid raiser, lip stretcher, eyebrow raiser, etc. This belongs to the interval [0,1] where the zero value corresponds to the neutral configuration (no deformation) and the one value corresponds to the maximum deformation. Without loss of generality, we have chosen the following action units: (1) jaw drop (action unit 26), lip stretcher (action unit 20), (3) lip corner depressor (action unit 15), (4) upper lip raiser (action unit 10), (5) eyebrow lowerer (action unit 4), (6) outer eyebrow raiser (Action Unit 2). The figures in parentheses depict the corresponding action unit in the FACS standard system.

  In our work, we suppose that these units are enough to cover most common facial actions (mouth and eyebrow movements). We point out that the vector $\boldsymbol{\tau}_\mathbf{a}$ encodes six local facial deformations caused by either facial expressions or spontaneous local facial motions.

3. 3D head pose and facial action tracking

The head and facial actions are tracked in 3D using the temporal face tracker developed by Dornaika and Davoine

[1] The perspective projection is the classical pin-hole camera model. The weak perspective projection can be seen as the zero approximation to the perspective projection.

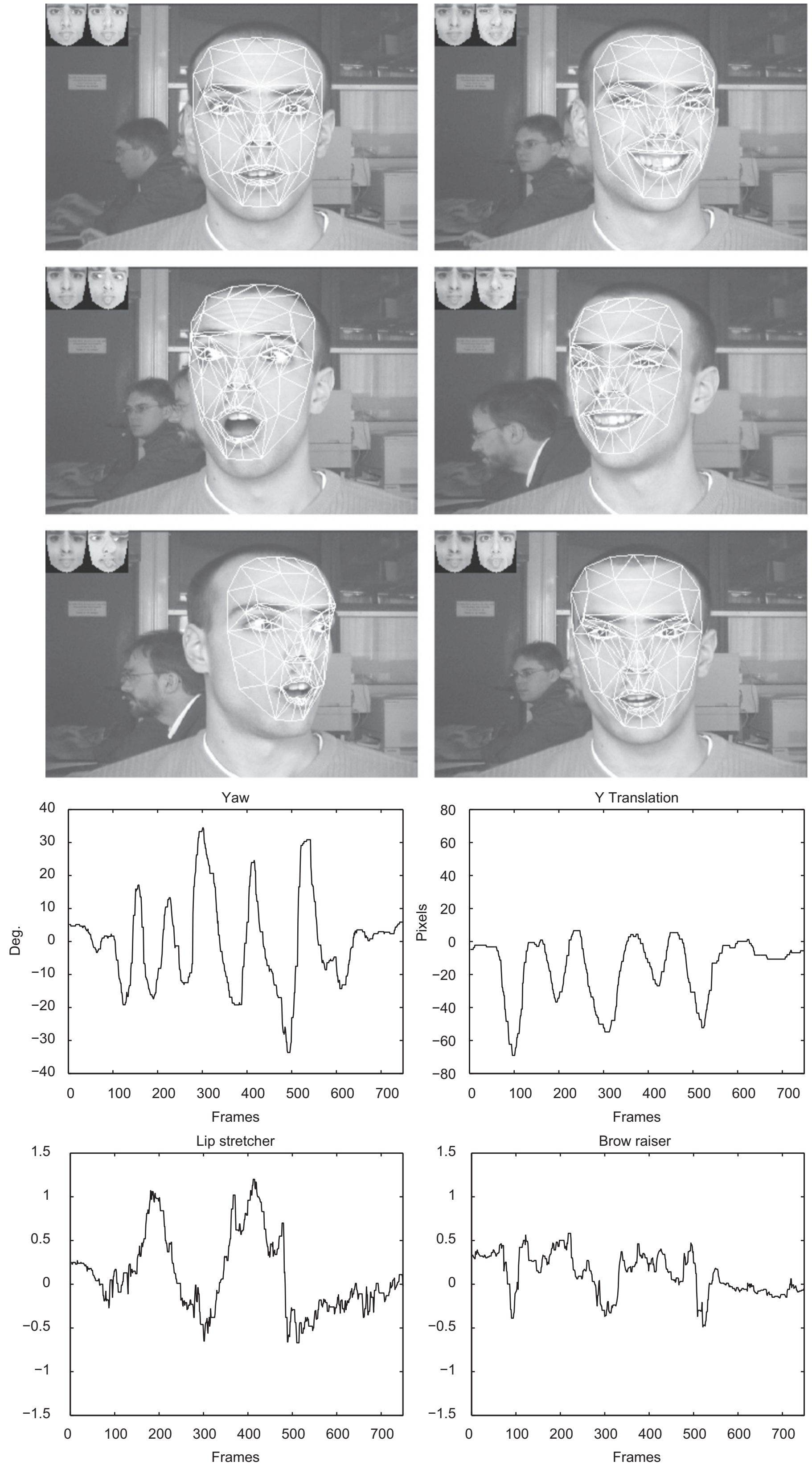


Fig. 2. Face and facial action tracking results using our appearance-based tracker. The upper part of the figure shows frames 63, 181, 282, 418, 492, and 683 in a 750-frame video sequence. The plots display the yaw angle (large value for frame 282), the vertical translation, the lip stretcher (large values for frames 181 and 418), and the brow raiser (large value for frame 282).

(2006). This appearance-based tracker aims at computing the 3D head pose and facial actions encapsulated in the vector **b**. The basic idea is to recover **b** by minimizing a distance between the incoming warped frame and the current appearance of the face. This minimization is carried out using a Gauss–Newton-like approach (Dornaika and Davoine, 2006). This tracker has two interesting features. First, the statistics of the appearance model are updated online. Second, the empirical gradient matrix is computed for each input frame. This scheme leads to a fast, efficient and robust tracking algorithm.

Fig. 2 displays the tracking results associated with eight frames of a 750-frame sequence featuring quite large pose variations as well as large facial actions. The sequence is of resolution 720×480 pixels. As can be seen, the 3D motion of the face as well as the facial actions associated with the mouth and the eyebrows are accurately recovered. The plots of this figure display the estimated value of the yaw angle, the vertical translation, the lip stretcher, and the brow raiser as a function of the frames of the sequence.

We stress the fact that the extracted facial actions are pose-independent. In other words, the used 3D face tracker provides the same value of the facial actions regardless of the face orientation with respect to the image/camera (see Fig. 3). On the other hand, most existing facial expression recognition methods compute/use facial features that are pose dependent. Due to the use of the *Candide 3* model, our facial expression recognition schemes described in the following section can be applied within a range of out-of-plane rotation from -50° to $+50^\circ$.

4. Dynamic facial expression recognition

4.1. First approach: dynamic time warping

4.1.1. Learning

In order to learn the spatio-temporal structures of the actions associated with facial expressions, we have used the following. Video sequences have been picked up from the CMU database (Kanade et al., 2000). These sequences depict five frontal view universal expressions (surprise, sadness, joy, disgust and anger). Each expression is performed by 21 different subjects, starting from the neutral one. Altogether we select 105 video sequences composed of around 15–20 frames each, that is, the average duration of each sequence is about half a second. The learning phase consists of estimating the facial action parameters $\boldsymbol{\tau}_{\mathbf{a}}$ (a 6-vector) associated with each training sequence, that is, the temporal sequence of the action parameters. In the sequel, the temporal sequence of $\boldsymbol{\tau}_{\mathbf{a}}(t_1), \boldsymbol{\tau}_{\mathbf{a}}(t_2), \ldots, \boldsymbol{\tau}_{\mathbf{a}}(T)$, is called “temporal signature”. This temporal signature encapsulates the facial deformation between time t_1 and time T.

Fig. 4 shows three video examples associated with the CMU database depicting surprise, anger, and joy expressions. The left frames illustrate the moderate magnitude of the expression. The right frames illustrate the apex of the expression. The training video sequences have an interesting property: all performed expressions go from the neutral expression to a high magnitude expression by going through a moderate magnitude around the middle of the sequence. Therefore, using the same training set we get two kinds of signatures: (i) an entire signature which models transitions from the neutral expression to a high magnitude expression, and (ii) a truncated trajectory (the second half part of a given signature) which models the transition from small/moderate magnitudes to high magnitudes. Note that the second kind of signatures can also model the variability of the action parameters for a given expression.

4.1.2. Recognition

In the recognition phase, the head pose and facial actions are recovered from the video sequence using the appearance-based tracker (Dornaika and Davoine, 2006). The basic idea of our proposed approach is that the expression class for a test sequence can be inferred from the extracted temporal signature by comparing it with learned/labeled signatures. The current facial expression is then recognized by computing a similarity or dissimilarity measure between the extracted facial actions $\boldsymbol{\tau}_{\mathbf{a}(t)}$ of the test sequence and those associated with each universal expression (obtained at the learning phase). This recognition scheme can be carried out either online or off-line. One can notice that a direct comparison between the estimated signatures and the learned ones is not feasible since there is no frame-to-frame correspondence between the tracked actions and the stored ones. To overcome this problem, we use dynamic time warping (DTW) technique (Keogh and Ratanamahatana, 2005) which allows temporal deformation of time series as they are matched against each other.

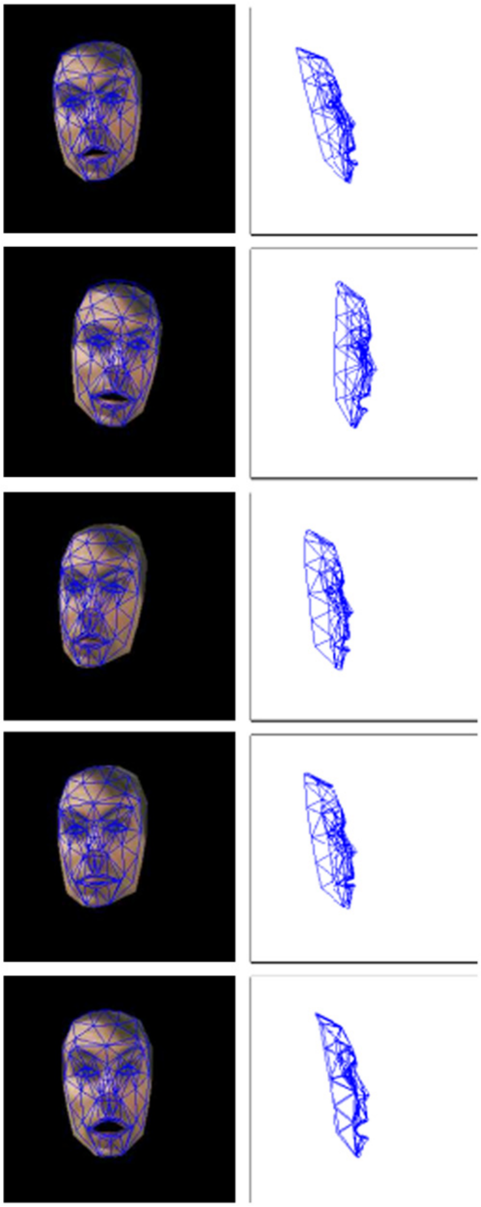
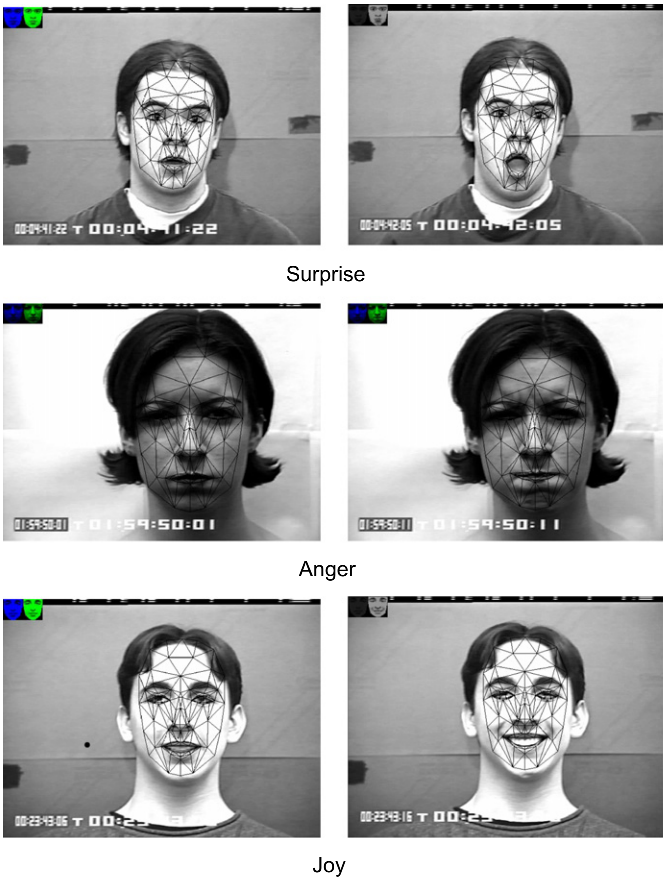


Fig. 3. Face recognition based on tracked facial actions using the standard deformable face model *Candide*. As can be seen, the lip lowerer facial action is invariant to head pose. In other words, the same intensity is obtained regardless of the face orientation with respect to the image/camera. On the other hand, most existing facial expression recognition methods compute/use facial features that are pose dependent.

Our proposed DTW-based classification scheme is illustrated in Fig. 5. We infer the facial expression associated with the current frame t by considering the estimated signature, i.e. the


Surprise

Anger

Joy

Fig. 4. Three video examples associated with the CMU database depicting surprise, anger, and joy expressions. The left frames illustrate the moderate magnitude of the expression. The right frames illustrate the apex of the expression.

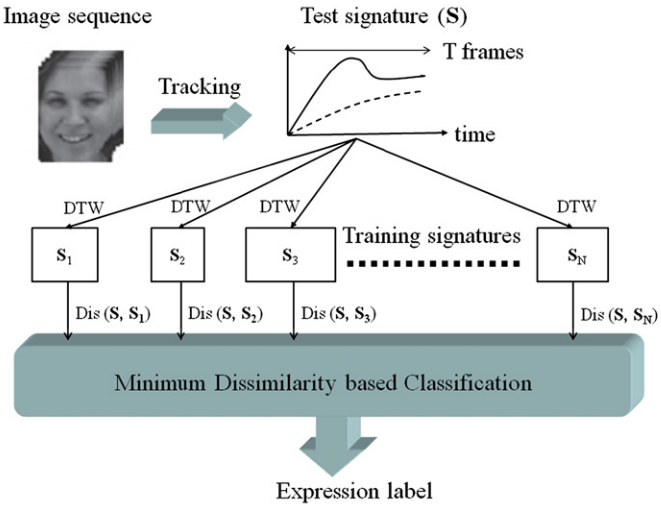


Fig. 5. Recognition phase of the proposed DTW-based classification. The new temporal signature (temporal facial actions) is matched with each training signature associated with the universal expressions and its dissimilarity score is calculated. It is finally classified into the expression class that corresponds to the minimum dissimilarity score. Notice that the temporal signatures (training and test) can have different duration.

sequence of vectors $\boldsymbol{\tau}_{\mathbf{a}(t)}$, within a temporal window of size T centered at the current frame t. In our tests, T is set to 9 frames. This signature is matched against the 105 training signatures (either the entire ones or the truncated ones) using DTW. For each training signature, the alignment performed by the DTW technique returns a dissimilarity measure between the test signature and the training one. A zero value indicates a perfect match between the two compared signatures and a high value indicates a mismatch. We propose two classification schemes. The first classification scheme stipulates that the smallest average dissimilarity decides the expression classification where the dissimilarity measures associated with a given universal expression are averaged over the whole training subjects. The second scheme is the nearest neighbor classifier, i.e., the smallest dissimilarity measure decides the expression classification. The proposed scheme accounts for the variability in duration since the DTW technique allows non-linear time scaling. The segmentation of the video is obtained by repeating the whole recognition scheme for every frame in the test video.

4.2. *Second approach: principal component analysis and linear discriminant analysis (PCA+LDA)*

As can be seen from the previous section, our first approach requires a matching in the sense of a dynamic time warping for every learned signature (temporal sequence of facial actions). Therefore, the CPU time of the recognition scheme based on the DTW technique will be proportional to the number of the subjects present in the database. Whenever this number is very large, the recognition scheme becomes computationally expensive. In this section, we show that the learned temporal signatures can be represented in a more compact form, namely with fixed length feature vectors. Thus, the problem of dynamic expression recognition which was mainly treated by dynamic time warping, hidden Markov models, or directed Hausdorff distance, can be carried out using any machine learning scheme since the learned examples (feature vectors) have the same dimension.

4.2.1. *Learning*

The learning phase is depicted in Fig. 6. Again, we use CMU database subset. In order to obtain temporal signatures having the same number of frames (duration), all signatures belonging to the same expression class are aligned using the DTW technique. Recall that this technique allows a frame-to-frame correspondence between two time series. Let $\mathbf{e}_i^j$ be the aligned signature i belonging to the expression class j. The example $\mathbf{e}_i^j$ is represented by a column vector of dimension $1 \times 6T$ and is obtained by simply concatenating the facial action 6-vectors $\boldsymbol{\tau}_{\mathbf{a}(t)}$:

$$\mathbf{e}_i^j = [\boldsymbol{\tau}_{\mathbf{a}(1)}; \boldsymbol{\tau}_{\mathbf{a}(2)}; \ldots; \boldsymbol{\tau}_{\mathbf{a}(T)}]$$

Note that T represents the duration of the aligned signatures which will be fixed for all examples. Thus, a nominal duration of 18 frames for the aligned signatures makes the dimension of all examples $\mathbf{e}_i^j$ (all i and j) equal to 108.

Applying a principal component analysis on the set of all training signatures yields the mean signature $\overline{\mathbf{e}}$ as well as the principal modes of variation. Any training signature $\mathbf{e}$ can be approximated by the principal modes using the q largest eigenvalues:

$$\mathbf{e} \simeq \overline{\mathbf{e}} + \mathbf{U}\mathbf{c} = \overline{\mathbf{e}} + \sum_{l=1}^{q} c_l \mathbf{U}_l$$

In our work, the number of principal modes is chosen such that the variability of the retained modes corresponds to 99% of the total variability. The vector $\mathbf{c}$ can be seen as a parametrization of any input signature, $\hat{\mathbf{e}}$, in the space spanned by the q basis vectors $\mathbf{U}_l$. The vector $\mathbf{c}$ is given by

$$\mathbf{c} = \mathbf{U}^T(\hat{\mathbf{e}} - \overline{\mathbf{e}}) \tag{3}$$

Thus, all training signatures $\mathbf{e}_i^j$ can now be represented by the vectors $\mathbf{c}_i^j$ (using (3)) on which a linear discriminant analysis

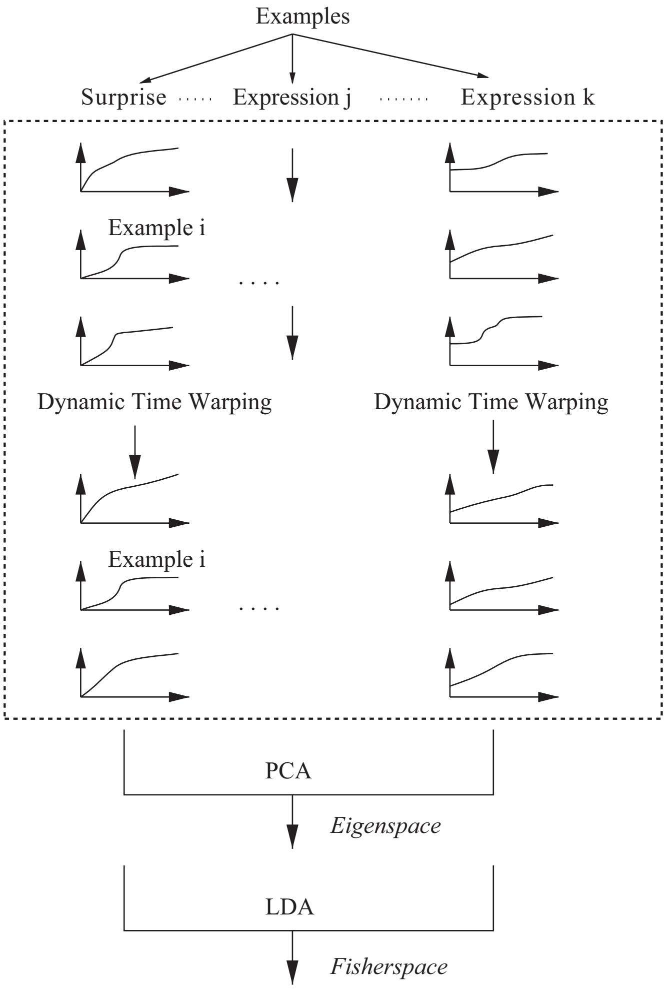


Fig. 6. The parameterized modeling of facial expressions using Eigenspace and Fisher space (PCA+LDA).

(LDA) can be applied. This gives Fisher space in which each training video sequence is represented by a vector of dimension $l-1$ where l is the number of expression classes.

It should be noticed that (i) PCA reduces the dimension of the signature and make them uncorrelated, and (ii) LDA enhances the discrimination among different classes (Yang and Yang, 2003).

4.2.2. Recognition

The recognition scheme follows the main steps of the learning stage. We infer the facial expression by considering the estimated facial actions returned by the temporal 3D face tracker. We consider the 1D vector $\mathbf{e}'$ (the concatenation of the facial actions $\tau_{\mathbf{a}(t)}$) within a temporal window of size T centered at the current frame t. Note that the value of T should be the same as in the learning stage. This vector is then mapped into a new vector using the learned PCA+LDA mapping. Two classification criteria are used: the Euclidean distance from the expression mean and the *Mahalanobis* distance from this mean.

5. Performance evaluation

In this section we provide quantitative evaluation of the proposed dynamic recognition schemes.

5.1. First approach: DTW

In order to quantify the recognition rate, in addition to the CMU video sequences, we have generated several test videos featuring the universal facial expressions. To this end, we have asked a volunteer student to perform each universal expression several times. The new subject was instructed to display the expression in a natural way, i.e. the displayed expressions were independent of any database.

The performance of the developed recognition scheme is evaluated by utilizing 52 home-made test videos. In other words, the signatures retrieved from the CMU video sequences will be considered as the training signatures, and the 52 signatures will be used for testing. Table 1 shows the confusion matrix associated with the 52 test signatures using the DTW technique (Fig. 5). We point out that the learned signatures were inferred from the CMU database while the used test videos were created at our laboratory. It is worthy to mention that the use of two different data sets for training and testing is, in general, more difficult than using a split of the same data set into training and testing data. Moreover, we stress the fact that the training signatures (retrieved from CMU image sequences) does not have the same duration (the duration of the training sequences varies between 10 and 27 frames). Furthermore, the video rate of the training sequences is 30 frames per second, whereas the rate of test sequences is 25 frames per second. We can conclude that the DTW technique was also useful for overcoming the video rate difference in video sequences.

As can be seen, the recognition rate of dynamic expressions was 100% for all basic expressions except for the disgust expression for which the recognition rate was 44%. The reason is that the disgust expression performed by our subject was very different from that performed by most of the CMU database subjects. This can be confirmed by Fig. 7. Therefore, for the above experiment, the overall recognition rate is 90.4%.

Table 2 shows the confusion matrix when the second classification rule was applied (nearest neighbor classifier). As can be seen, the recognition rate of the disgust expression becomes 77.8% and the overall recognition rate becomes 96.1%.

Table 1
Confusion matrix for the dynamic facial expression classifier using the DTW technique (the smallest average similarity). The learned signatures were inferred from the CMU database while the used test videos were created at our laboratory. The numbers in parenthesis indicate the number of test cases for each basic expression. The recognition rate of dynamical expressions was 100% for all basic expressions except for the disgust expression for which the recognition rate was 44%.

	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	5	0
Disg.	0	0	0	4	0
Ang.	0	0	0	0	10

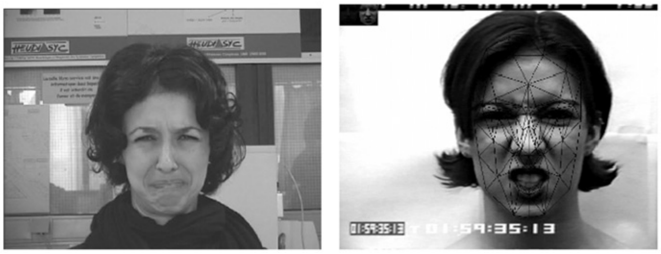
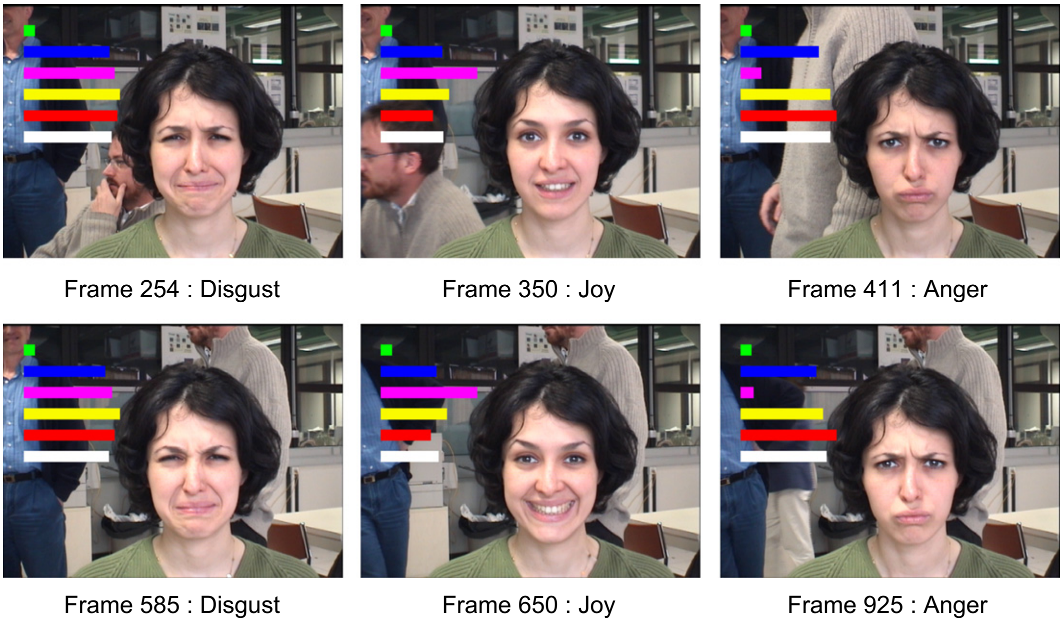


Fig. 7. The disgust expression performed by a CMU subject (right) and by our subject (left). Although both subjects claim that they are displaying a disgust expression, the mouth configurations are markedly different.

Table 2
Confusion matrix for the dynamical facial expression classifier using the DTW technique (the smallest similarity). The learned signatures were inferred from the CMU database while the used test videos were created at our laboratory. The numbers in parenthesis indicate the number of test cases for each basic expression. The recognition rate of dynamical expressions was 100% for all basic expressions except for the disgust expression for which the recognition rate was 77.8%.

	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	0	0
Disg.	0	0	0	7	0
Ang.	0	0	0	2	10


Frame 254 : Disgust

Frame 350 : Joy

Frame 411 : Anger

Frame 585 : Disgust

Frame 650 : Joy

Frame 925 : Anger

Fig. 8. DTW-based facial expression recognition associated with frames 254, 350, 411, 585, 650, and 925 in a 1000-frame video sequence. The annotated labels are automatically retrieved by selecting the expression that gave the minimum dissimilarity measure provided by the DTW technique (see the curves depicted in Fig. 9).

Fig. 8 shows the recognition results obtained with a 1000-frame test video (only six frames are shown). Fig. 9 illustrates the corresponding dissimilarity measure associated with every universal expression (returned by the DTW technique) as a function of time. As explained in the previous section, the recognition is performed by looking for the universal expression that has the smallest dissimilarity measure. The vertical bold arrows correspond to the frames depicted in Fig. 8. We can observe that even with low intensity facial actions the recognition scheme is still able to correctly recognize the displayed expressions. For example, we consider frame 585 in Fig. 8, this frame is recognized as a disgust expression. By visually inspecting the dissimilarity curve of each universal expression in Fig. 9, we can easily see that the disgust curve holds the lowest dissimilarity value for the frames preceding the frame 585 (for which the facial action intensity is relatively low). We can also observe that the DTW-based recognizer markedly and correctly labels the transition of the expressions when they start from the neutral configuration. This tends to confirm that even for subtle expressions, the dynamic scheme still able to correctly infer the displayed expression.

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 tracking of one frame is carried out in 50 ms.

5.2. Second approach: PCA+LDA

Table 3 shows the confusion matrix for the dynamical facial expression classifier using PCA+LDA mapping. The learned signatures were inferred from the CMU database while the used test videos were created at our laboratory. The recognition rate of dynamical expressions was 100% for all basic expressions except for the disgust expression for which the recognition rate was 55%. Therefore, for the above experiment, the overall recognition rate is 92.3%. Similar results have been obtained with the *Mahalanobis* distance. One can notice the slight improvement in the recognition rate over the classical recognition scheme based on the DTW technique adopting the first classification rule. Fig. 10 shows the recognition results for four frames.

We also proposed another simple scheme to segment frames into neutral and non-neutral ones. This scheme exploits an interesting property of the 3D deformable model. Indeed, an ideal neutral configuration of the face yields a zero vector for the vector $\boldsymbol{\tau}_{\mathbf{a}}$. Thus, a frame can be individually considered as a non-neutral expression if the sum of the absolute value of all components of $\boldsymbol{\tau}_{\mathbf{a}}$ is greater then a predefined threshold. Thus, the above classification using a sliding temporal window only occurs at non-neutral frames. This recognition scheme is illustrated in Fig. 11. Figs. 11 and 12 show the ability of our proposed approach to recognize facial expressions in the presence of head motions and non-frontal face views.[2]

On a 3.2 GHz PC, a non-optimized C code of the developed approach carries out the tracking and recognition in about 53 ms. The tracking of one frame is carried out in 50 ms.

[2] The used 3D face tracker works well as long as the out-of-plane rotation angle belongs to $[-50^{\circ}, +50^{\circ}]$.

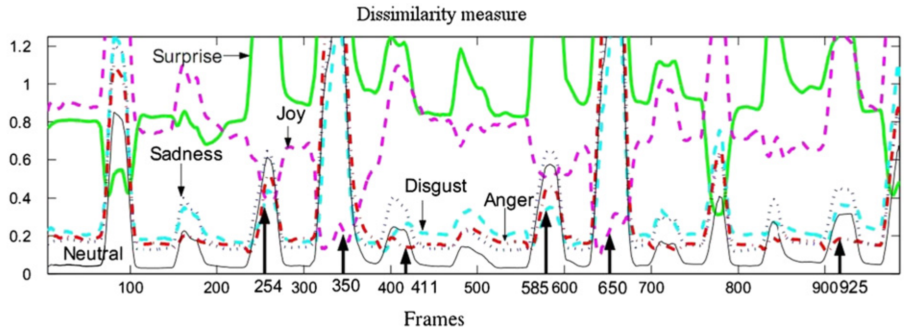
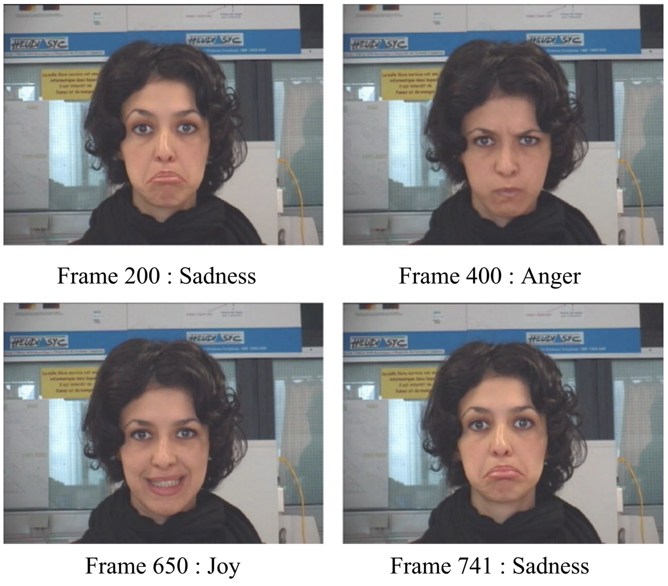


Fig. 9. Dissimilarity profiles of six facial expressions (obtained by DTW) for a 1000-frame video sequence. The vertical bold arrows correspond to the frames depicted in Fig. 8.

Table 3
Confusion matrix for the dynamical facial expression classifier using PCA+LDA mapping. The learned signatures were inferred from the CMU database while the used test videos were created at our laboratory. The numbers in parenthesis indicate the number of test cases for each basic expression. The recognition rate of dynamical expressions was 100% for all basic expressions except for the disgust expression for which the recognition rate was 55%.

	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	4	0
Disg.	0	0	0	5	0
Ang.	0	0	0	0	10


Frame 200 : Sadness

Frame 400 : Anger

Frame 650 : Joy

Frame 741 : Sadness

Fig. 10. PCA+LDA-based facial expression recognition associated with four frames of a test video sequence.

5.3. Method comparison

We also compared the recognition schemes with some state-of-the art classifiers. To this end, we used the leave one out cross validation (LOOCV) techniques. We used the same aligned temporal signatures associated with CMU data set.

Table 4 summarizes the recognition results obtained with the 105 video sequences using the LOOCV technique with the five classifiers: (i) support vector machines (SVM), (ii) K nearest neighbor (KNN), (iii) Naive Bayes (NB), (iv) Bayes network (BN), and (v) the proposed approach (PCA+LDA mapping) (Section 4.2). The first four methods were retrieved from WEKA library. The SVM classifier used is a polynomial of degree three. The KNN parameter was set to 1 ($K=1$). The rest of the parameters are set to their default values. We can observe that the best performance was obtained by the proposed approach and the SVM classifier.

We stress the fact that the first four classifiers have been applied on the original data (temporal signatures), while our proposed approach used the NN classifier on the embedded space represented by PCA+LDA, which is also a linear dimensionality reduction technique.

5.4. Static recognition versus dynamic recognition

In the previous section, we have described several methods for dynamic facial expression recognition. The extracted facial actions in video sequences can also be used in a static recognition fashion. This static recognition scheme uses the facial actions associated with only one single frame. In order to assess the benefit of using temporal information, we performed also the “static” facial expression recognition. In the static scheme, we considered the training frames in CMU subset that have moderate magnitudes. We then use the leave one out cross validation technique on these frames in order to asses the recognition accuracy. Table 5 summarizes the obtained accuracy for the dynamic and static schemes. We used two machine learning classification algorithms: (i) LDA followed by the nearest neighbor classifier and (ii) SVMs. We can observe that the dynamic recognition scheme has outperformed the static recognition scheme. It should be noticed that LDA can be applied directly to the static scheme since the dimension of feature vectors is small in that there is no need to apply the PCA pre-stage.

6. Conclusions and discussions

In this paper, we have addressed the analysis and recognition of facial expressions in continuous videos using tracked facial actions. We have introduced two different schemes that exploit facial actions estimated by an appearance-based 3D face tracker. The proposed schemes do not require tedious learning stages since they are not based on rawbrightness changes although the tracked facial actions are derived from them using an adaptive appearance tracker. We stress the fact that this is a not a contradiction with the claim that the two approaches are texture-independent. Indeed, the tracking of facial actions is carried out using online appearance models which dynamically learn the face appearance online. The proposed approaches have an additional advantage by which the facial expression recognition can be performed even when the face is in a non-frontal view. The proposed approaches take advantage of the spatio-temporal configuration of the facial actions. For both

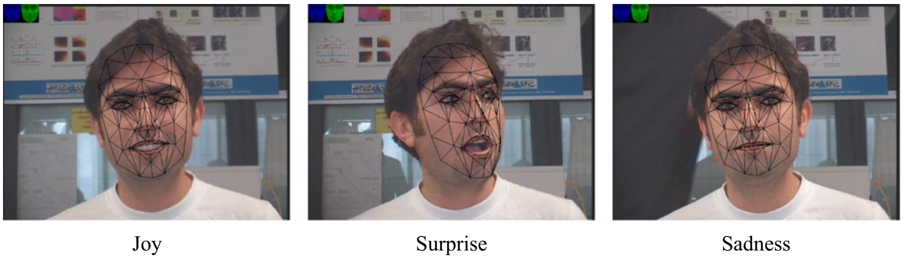


Fig. 11. Facial expression recognition associated with three frames of a test video. One can notice that the proposed scheme recognizes correctly the displayed expressions despite the head motion and the non-frontal face orientation.

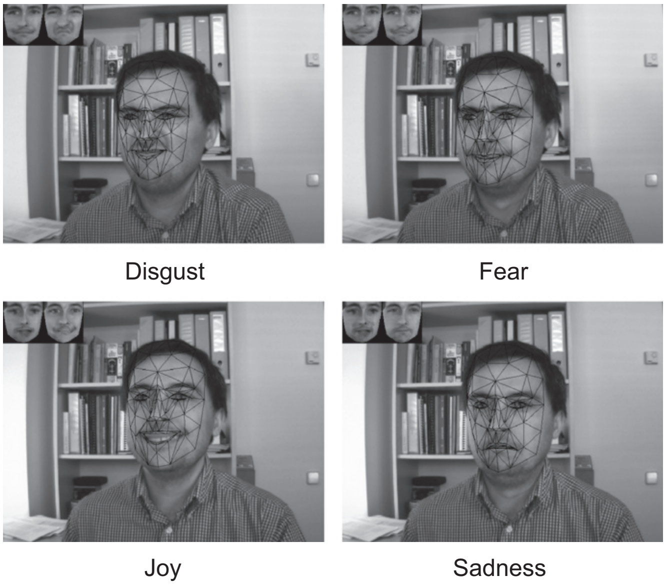


Fig. 12. Facial expression recognition associated with four frames of a test video. One can notice that the proposed scheme recognizes correctly the displayed expressions despite the head motion and the non-frontal face orientation.

Table 4
Overall classification results.

Method	Accuracy (%)
SVM	85.19
KNN	80.52
NB	70.91
BN	71.17
PCA+LDA (1 NN)	90.10

Table 5
Overall classification results for the dynamic and static classifiers.

Classifier type	LDA + NN (%)	SVM (%)
Static	80.00	82.80
Dynamic	90.10	85.57

proposed approaches, changes in either the video rate or the facial action duration do not affect the recognition accuracy this is due to the use of dynamic time warping technique which overcomes such non-linear time scale.

The proposed approaches, despite their flexibility, have recognition rates close to many sophisticated methods reported in the recent literature. The conducted experiments have shown that the mapping provided by the mapping PCA+LDA has provided better performance than the classifiers working on the raw facial action sequences. This can be explained by the fact that the PCA stage reduces noise and that the LDA stage enhances the discrimination between expressions.

Experiments have shown that accurate facial expression recognition can be obtained by only exploiting the tracked facial actions associated with the mouth and the eyebrows. There are several reasons that justify the selection of the six AUs: (1) These six units are associated with the mouth and eyebrows regions. These face parts are markedly affected by universal facial expressions. (2) Some subtle facial actions cannot be detected in real images where the face occupies a small region in the image (e.g., cheek raiser AU). (3) By including many actions units the 3D face and facial action tracker may become unsuitable for real-time applications. The current used appearance-based 3D face tracker adopts 12 unknown parameters for a given video frame (six degrees of freedom associated with the 3D head pose and the selected six action units).

It is worth noting that once a fixed length feature vector is computed from the time series representation of the extracted facial deformation, it is straightforward to use machine learning tools including the kernel techniques for the PCA and LDA which increase the discriminative power of the dimensionality reduction techniques.

Future work will be oriented towards non-linear dimensionality reduction techniques (kernel- and manifold-based methods) for facial expression representation, which are known for an increased discriminative power.

Acknowledgments

This work was supported by the Spanish Government under the Projects TIN2010-18856 and TIN2009-14404-C02.

References

Ahlberg, J., 2002. An active model for facial feature tracking. EURASIP J. Appl. Signal Process. 1 (6), 566–571.

Ambadar, Z., Schooler, J., Cohn, J., 2005. Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions. Psychol. Sci. 16 (5), 403–410.

Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J., 2006. Fully automatic facial action recognition in spontaneous behavior. In: IEEE International Conference on Automatic Face and Gesture Recognition, 2006.

Chakraborty, A., Konar, A., Chakraborty, U., Chatterjee, A., 2009. Emotion recognition from facial expressions and its control using fuzzy logic. IEEE Trans. Syst. Man Cybern.—Part A: Syst. Humans 39 (4), 726–743.

Cheon, Y., Kim, D., 2009. Natural facial expression recognition using differential-AAM and manifold learning. Pattern Recognition 42, 1340–1350.

Dornaika, F., Davoine, F., 2006. On appearance based face and facial action tracking. IEEE Trans. Circuits Syst. Video Technol. 16 (9).

Dornaika, F., Raducanu, B., 2007. Inferring facial expressions from videos: tool and application. Signal Process.: Image Commun. 22 (9), 769–784.

Dornaika, F., Raducanu, B., 2010. Person-specific face shape estimation under varying head pose from single snapshots. In: IEEE International Conference on Pattern Recognition.

Ekman, P., 1992. Facial expressions of emotions: an old controversy and new findings. Philos. Trans. R. Soc. London 335, 63–69.

Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: a survey. Pattern Recognition 36 (1), 259–275.

Huang, X., Zhao, G., Pietikinen, M., Zheng, W., 2011. Expression recognition in videos using a weighted component-based feature descriptor. In: Image Analysis, Lecture Notes on Computer Science, vol. 6688.

Kanade, T., Cohn, J., Tian, Y.L., 2000. Comprehensive database for facial expression analysis. In: International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March, pp. 46–53.

Keogh, E., Ratanamahatana, C.A., 2005. Exact indexing of dynamic time warping. Knowl. Inf. Syst. 7, 358–386.

Meng, H., Romera-Paredes, B., Bianchi-Berthouze, N., 2011. Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. In: IEEE International Conference on Face and Gesture Recognition—Workshop on Facial Expression Recognition and Analysis Challenge.

Moore, S., Bowden, R., 2011. Local binary patterns for multi-view facial expression recognition. Comput. Vision Image Understanding 115, 541–558.

Moore, S., Ong, E., Bowden, R., 2010. Facial expression recognition using spatio-temporal boosted discriminator classifiers. In: International Conference on Image Analysis and Recognition.

Naghsh-Nilchi, A.R., Roshanzamir, M., 2006. An efficient algorithm for motion detection based facial expression recognition using optical flow. Int. J. Eng. Appl. Sci. 2 (3), 141–146.

Pantic, M., Patras, I., 2006. Dynamics of facial expression: recognition of facial actions and their temporal segments form face profile images. IEEE Trans. Syst. Man Cybern. Part B 36 (2), 433–449.

Robin, T., Bierlairey, M., Cruz, J., 2011. Dynamic facial expression recognition with a discrete choice model. J. Choice Modelling 2 (1), 95–148.

Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S., 2007. Authentic facial expression analysis. Image Vision Comput. 25 (December), 1856–1863.

Shan, C., Gong, S., McOwan, P.W., 2006. Dynamic facial expression recognition using a bayesian temporal manifold model. In: Proceedings of the British Machine Vision Conference, vol. I, Edinburgh, UK, pp. 297–306.

Shan, C., Gong, S., McOwan, P.W., 2009. Facial expression recognition based on local binary patterns: a comprehensive study. Image Vision Comput. 27, 803–816.

Sung, J., Kim, D., 2009. Real-time facial expression recognition using STAAM and layered GDA classifier. Image Vision Comput. 27 (9), 1313–1325.

Tian, Y., Kanade, T., Cohn, J.F., 2001. Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. 23, 97–115.

Wu, T., Bartlett, M.S., Movellan, J.R., 2010. Facial expression recognition using Gabor motion energy filters. In: Computer Vision and Pattern Recognition Workshops (CVPRW).

Xiang, T., Leung, M.K.H., Cho, S.Y., 2008. Expression recognition using fuzzy spatio-temporal modeling. Pattern Recognition 41 (1), 204–216.

Xiao, R., Zhao, Q., Zhang, D., Shi, P., 2011. Facial expression recognition on multiple manifolds. Pattern Recognition 44, 107–116.

Yang, J., Yang, J., 2003. Why can LDA be performed in PCA transformed space? Pattern Recognition 36 (2), 563–566.

Yang, P., Liu, Q., Cui, X., Metaxas, D.N., 2008. Facial expression recognition using encoded dynamic features. In: Computer Vision and Pattern Recognition.

Yeasin, M., Bullot, B., Sharma, R., 2006. Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. Multimedia 8 (3), 500–508.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31 (1), 39–58.

Zhang, S., Zhao, X., Lei, B., 2012. Robust facial expression recognition via compressive sensing. Sensors 12, 3747–3761.

Zhang, Y., Ji, Q., 2005. Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. Pattern Anal. Mach. Intell. 27 (5), 699–714.

Zhang, Y., Ji, Q., Zhu, Z., Yi, B., 2008. Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters. IEEE Trans. Circuits Syst. Video Technol. 18 (10), 1383–1396.

Zhao, G., Pietikinen, M., 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 915–928.