Machine Vision and Applications (2013) 24:811–820
DOI 10.1007/s00138-012-0447-z

ORIGINAL PAPER

Texture-independent recognition of facial expressions in image snapshots and videos

Bogdan Raducanu · Fadi Dornaika

Received: 6 October 2011 / Revised: 3 August 2012 / Accepted: 13 August 2012 / Published online: 20 November 2012

Abstract This paper addresses the static and dynamic recognition of basic facial expressions. It has two main contributions. First, we introduce a view- and texture-independent scheme that exploits facial action parameters estimated by an appearance-based 3D face tracker. We represent the learned facial actions associated with different facial expressions by time series. Second, we compare this dynamic scheme with a static one based on analyzing individual snapshots and show that the former performs better than the latter. We provide evaluations of performance using three subspace learning techniques: linear discriminant analysis, non-parametric discriminant analysis and support vector machines.

Keywords Facial expression recognition · Subspace learning · Static and dynamic classifier · Human machine interaction

1 Introduction

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 to infer people inner affective state [14,15]. Basic facial expressions typically recognized by psychologists are: happiness, sadness, fear, anger, disgust and surprise [13]. In the beginning, facial expression analysis was essentially a research topic for psychologists. However, recent progresses in image processing and pattern recognition have significantly motivated research works on automatic facial expression recognition [16,20,36]. In the past, a lot of effort was dedicated to recognize facial expression in still images. For this purpose, many techniques have been applied: neural networks [32], Gabor wavelets [4] and active appearance models [30]. A very important limitation to this strategy 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 marked. In their daily life, people seldom show apex of their facial expression during normal communication with their counterparts, unless for very specific cases and for very brief periods of time.

More recently, attention has been shifted particularly towards modelling dynamical facial expressions [35,27]. This is because the differences between expressions are more powerfully modelled by dynamic transitions between different stages of an expression rather than their corresponding static key frames. 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 [3], the authors found that subtle expressions that were not identifiable in individual images suddenly became apparent when viewed in a video sequence.

Dynamical 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 [37]. In [5], parametric 2D flow models

B. Raducanu (✉)
Computer Vision Center, Edifici "O", Campus UAB,
08193 Bellaterra, Barcelona, Spain
e-mail: bogdan@cvc.uab.es

F. Dornaika
University of the Basque Country UPV/EHU,
Paseo Manuel de Lardizabal, 1, 20018 San Sebastian, Spain

F. Dornaika
IKERBASQUE, Basque Foundation for Science,
48011 Bilbao, Spain
e-mail: fadi_dornaika@ehu.es

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. In [36], 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 HMMs to learn the underlying model for each facial expression. In [29], the authors propose a Bayesian approach to modelling temporal transitions of facial expressions represented in a manifold. However, the fact that the method relies heavily on the gray level of the image can be a serious limitation. In [34], the authors explore Gabor motion energy filters (GME) 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. In [18], the authors 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. In [21], the authors use two types of descriptors motion history histogram (MHH) and histogram of local binary patterns (LBPs). LBP was applied to each frame of the video and was used to capture local textural patterns. Based on these two basic types of descriptors, two new dynamic facial expression features are proposed. In [23], the authors 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. In [38], the authors construct a sparse representation classifier (SRC). The effectiveness and robustness of the SRC method are investigated on clean and occluded facial expression images. Three typical facial features, i.e., the raw pixels, Gabor wavelets representation and LBPs are extracted to evaluate the performance of the SRC method. In [22], 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. LBPs and variations of LBPs as texture descriptors are investigated. Multi-class SVMs are adopted to learn pose and pose-dependent facial expression classifiers.

As can be seen, most of the proposed expression recognition schemes require a frontal view of the face. Moreover, most of them rely on the use of image raw brightness changes. 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 [8,25,26] as well as training of social robots [6,7].

This paper introduces a novel scheme for dynamic facial expression recognition that is based on the appearance-based 3D face tracker [11]. It has two main contributions. First, we introduce a view- and texture-independent scheme that exploits facial action parameters estimated by an appearance-based 3D face tracker. We represent the learned facial actions associated with different facial expressions by time series. Second, we compare this dynamic scheme with a static one based on analyzing individual snapshots and show that the former performs better than the latter. We provide evaluations of performance using three subspace learning techniques: linear discriminant analysis (LDA), non-parametric discriminant analysis (NDA) and SVMs.

Compared to existing dynamical facial expression methods, our proposed approach (first contribution) 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—invariant geometrical parameters. Third, its learning phase is simple compared to other techniques (e.g., the HMM). As a result, even when the imaging conditions change, the learned expression dynamics need not to be recomputed. In this work, we compare the proposed approach for dynamic facial expression against individual frame-based recognition methods. This comparison shows a clear superiority in terms of recognition rates and robustness.

The most related works are our previous works [10] and [12]. These works utilize the intensities of facial actions for the dynamic facial expression recognition. However, [10] proposes a dynamic classifier based on a brute force matching of temporal trajectories. [12] proposes a dynamic classifier that is not based on examples. It is based on an analysis–synthesis scheme exploiting learned predictive models (second order Markov models). The current paper provides a substantial novelty since it addresses dynamic and static recognition schemes with moderate and high magnitude facial expressions. Moreover, the current work utilizes and compares several subspace learning techniques for both the dynamic and static recognition frameworks such as LDA, NDA and SVM.

The rest of the paper is organized as follows. Section 2 briefly presents the used 3D face and facial action tracker. Section 3 describes the proposed recognition scheme. In Sect. 4 we report some experimental results and method comparisons. Finally, in Sect. 5, we present our conclusions and some guidelines for future work.

2 3D facial dynamics extraction

2.1 A deformable 3D face model

In our work, we use the 3D face model *Candide* [2]. This 3D deformable wireframe model was first developed for the purpose of model-based image coding and computer animation. The 3D shape of this wireframe model is directly recorded in coordinate form. It is given by the coordinates of the 3D 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 $\mathbf{g}$ is 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, $\boldsymbol{\tau}_\mathbf{a}$ is the animation control vector, and the columns of $\mathbf{A}$ are the animation units (AUs). The static shape is constant for a given person. In this study, we use six modes for the facial AUs matrix $\mathbf{A}$. We have chosen the following AUs or facial actions: lower lip depressor, lip stretcher, lip corner depressor, upper lip raiser, eyebrow lowerer, and outer eyebrow raiser. These facial actions are enough to cover most common facial animations. Moreover, they are essential for conveying emotions. Thus, for every frame in the video, the state of the 3D wireframe model is given by the 3D head pose parameters (three rotations and three translations) and the internal face animation control vector $\boldsymbol{\tau}_\mathbf{a}$. This is given by the 12-dimensional vector $\mathbf{b}$:

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

where:

- θ_x, θ_y, and θ_z represent the three angles associated with the 3D rotation between the 3D face model coordinate system and the camera coordinate system.
- 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. 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. In the sequel, the word "facial action" will refer to the facial action intensity.

2.2 Simultaneous face and facial action tracking

In order to recover the facial expression, one has to compute the facial actions encoded by the vector $\boldsymbol{\tau}_\mathbf{a}$ which encapsulates the facial deformation. Since our recognition scheme is view independent, these facial actions together with the 3D head pose should be simultaneously estimated. In other words, the objective is to compute the state vector $\mathbf{b}$ for every video frame.

For this purpose, we use the tracker based on Online Appearance Models [11]. This appearance-based tracker aims at computing the 3D head pose and the facial actions, i.e. the vector $\mathbf{b}$, by minimizing a distance between the incoming warped frame and the current *shape-free* appearance of the face (a geometrically normalized frontal face). This minimization is carried out using a gradient descent method. The statistics of the *shape-free* appearance as well as the gradient matrix are updated every frame. This scheme leads to a fast and robust tracking algorithm. The remaining of this section will describe the main features of the used 3D face tracker [11]. These features are also illustrated in Fig. 1 which shows the model fitting for every video frame.

Given a video sequence depicting a moving head/face, we would like to recover, for each frame, the 3D head pose and the facial actions encoded by the control vector $\boldsymbol{\tau}_\boldsymbol{a}$. In other words, we would like to estimate the vector $\mathbf{b}_t$ (Eq. 2) at time t given all the observed data until time t, denoted $\mathbf{y}_{1:t} \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$. In a tracking context, the model parameters associated with the current frame will be handed over to the next frame.

For each input frame $\mathbf{y}_t$, the observation is simply the warped texture patch (the shape-free image) associated with the geometric parameters $\mathbf{b}_t$. We use the HAT symbol for the tracked parameters and textures. For a given frame t, $\hat{\mathbf{b}}_t$ represents the computed geometric parameters and $\hat{\mathbf{x}}_t$ the corresponding shape-free patch (shown in the upper part of Fig. 1), i.e.,

$$\hat{\mathbf{x}}_t = \mathbf{x}(\hat{\mathbf{b}}_t) = \mathcal{W}(\mathbf{y}_t, \hat{\mathbf{b}}_t) \tag{3}$$

where $\mathcal{W}$ is a piece-wise affine transform that maps the input image underneath the projected 3D mesh to the shape-free face image.

The model of face appearance (shape-free face images) is given by a multivariate Gaussian with a diagonal covariance matrix $\boldsymbol{\Sigma}$. The choice of a Gaussian distribution is motivated by the fact that this kind of distribution provides simple and general model for additive noises. In other words, this multivariate Gaussian is the distribution of the facial patches $\hat{\mathbf{x}}_t$. Let $\boldsymbol{\mu}$ be the Gaussian center and $\boldsymbol{\sigma}$ the vector containing the square root of the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are d-vectors (d is the size of the shape-free face image $\mathbf{x}$).

The estimation of $\hat{\mathbf{b}}_t$ from the sequence of images will adopt the following steps (illustrated in Fig. 1):

Face pose and facial action estimation using Online Appearance Models:

1. Acquire a new image $\mathbf{y}_t$.
2. Set the initial solution for $\hat{\mathbf{b}}_t$ as $\hat{\mathbf{b}}_t = \hat{\mathbf{b}}_{t-1}$ (the solution estimated at the previous image).

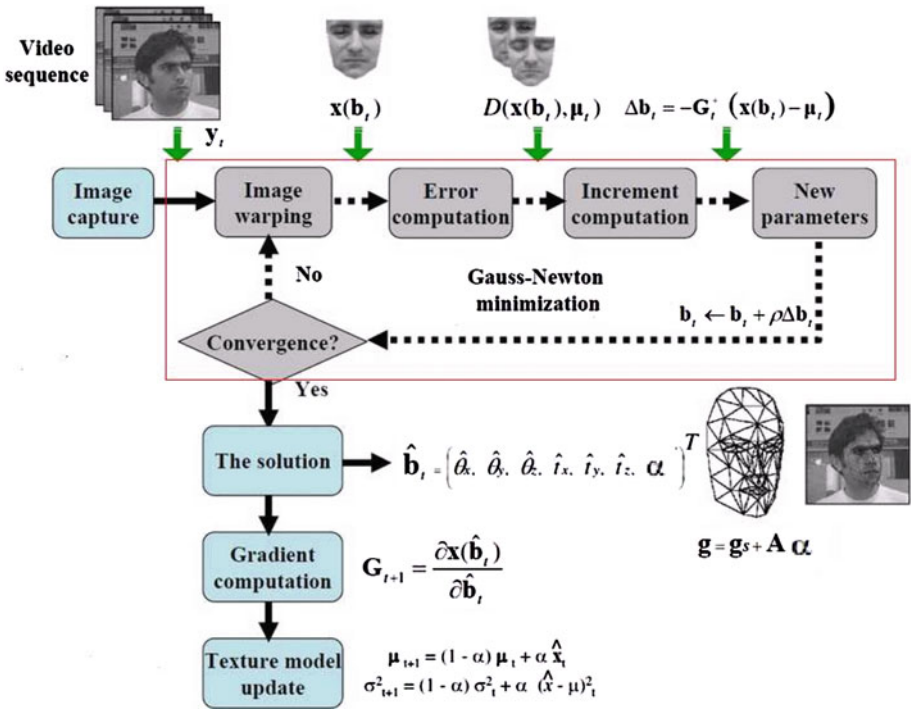


Fig. 1 The fitting algorithm used for tracking the 3D face pose and facial actions in a video sequence. See Sect. 2.2 for a more detailed explanation. All depicted processing steps are used for every frame in the video sequence

3. Iteratively update the solution $\hat{\mathbf{b}}_t$ by minimizing the error between the current warped texture (based on $\mathbf{y}_t$) and the current texture model (shape free textures). A Gauss–Newton like minimization is used. The error is set to the *Mahalanobis* distance between the warped texture and the current face texture model (Eq. 4). This minimization process is illustrated in the nested loop (Fig. 1).

$$\min_{\mathbf{b}_t} e(\mathbf{b}_t) = \min_{\mathbf{b}_t} D(\mathbf{x}(\mathbf{b}_t), \boldsymbol{\mu}_t) = \sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \quad (4)$$

4. At convergence, use the current solution $\hat{\mathbf{b}}_t$ for computing the current shape-free facial image $\mathbf{x}(\hat{\mathbf{b}}_t)$.
5. Use the estimated shape-free facial image $\mathbf{x}(\hat{\mathbf{b}}_t)$ for updating the texture model (the mean and the covariance of the multivariate Gaussian) using the following equations:

$$\mu_{i_{(t+1)}} = (1-\alpha)\,\mu_{i_{(t)}} + \alpha\,\hat{x}_{i_{(t)}} \quad (5)$$

$$\sigma^2_{i_{(t+1)}} = (1-\alpha)\,\sigma^2_{i_{(t)}} + \alpha\,(\hat{x}_{i_{(t)}} - \mu_{i_{(t)}})^2 \quad (6)$$

where α is a positive scalar controlling how fast the past observations are forgotten.

6. Compute the current gradient matrix $\frac{\partial \mathbf{x}}{\partial \mathbf{b}}$ by the finite difference method. This gradient matrix encodes the texture variation with respect to all geometrical parameters (face pose and the facial actions).
7. Hand over the current face pose and facial actions, the gradient matrix and the face texture model to the next frame. Set $t \leftarrow t+1$. Go to step 1.

3 Facial expression recognition

3.1 Learning

In order to learn the spatio-temporal structures of the facial actions associated with facial expressions, we have used a simple supervised learning scheme that consists in two stages. In the first stage, video sequences depicting different facial expressions are tracked using the appearance-based tracker. The retrieved facial actions $\boldsymbol{\tau}_\mathbf{a}$ are represented by time series. In other words, an example (expression going from neutral to apex) is encoded by a sequence of facial actions $\boldsymbol{\tau}_{\mathbf{a}(1)}, \ldots, \boldsymbol{\tau}_{\mathbf{a}(T)}$. In the second stage, in order to get the same dimension for all training examples, all facial action sequences are registered in the time domain using the dynamic time warping (DTW) technique [24]. DTW is a well-known technique to find an optimal alignment between two given (time dependent) sequences under certain restrictions. An illustration of DTW is given in Fig. 2. This temporal alignment is needed in order to subsequently use machine learning tools that require the observations to have the same dimension. The motivation behind this process, resides in the fact that the generation of a facial expression, in terms

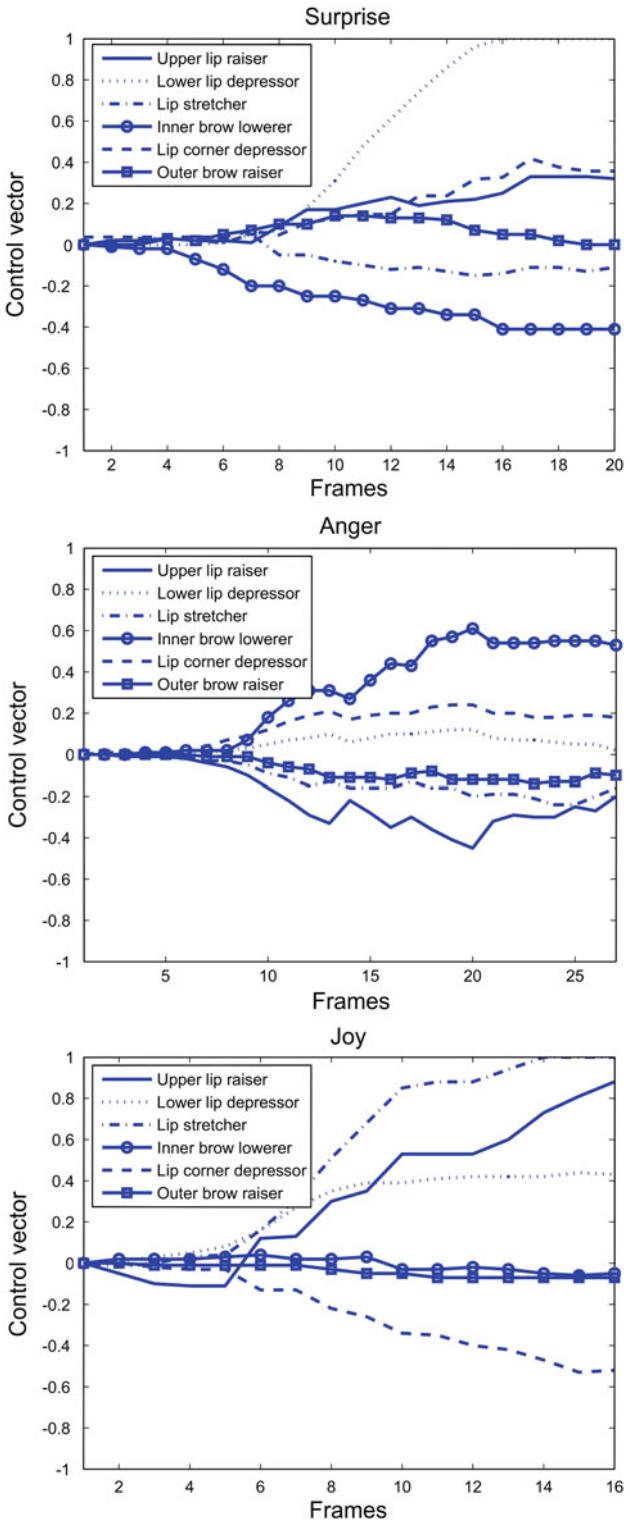


Fig. 2 Three examples (sequences) of learned facial action parameters as a function of time

of speed and intensity, is person-specific and furthermore, it is different for the same person, depending on the context. For this reason, we have to make sure that all observations we acquire are conveniently resized in the time domain for their computational analysis. For example, a given example (expression) is always represented by a feature vector obtained by concatenating the vectors $\boldsymbol{\tau}_{\mathbf{a}}$ belonging to the registered temporal sequence.

Video sequences have been picked up from the CMU database [19]. These sequences depict five frontal view universal expressions (surprise, sadness, joy, disgust and anger). Each expression is performed by seven different subjects, starting from the neutral one. Altogether, we select 35 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-element vector) associated with the frames of each training sequence, that is, the temporal trajectories of the action parameters. Figure 3 shows the retrieved facial action parameters associated with three sequences: surprise, anger, and joy. 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 trajectories: (1) an entire trajectory which models transitions from the neutral expression to a high magnitude expression, and (2) a truncated trajectory (the second half part of a given trajectory) which models the transition from small/moderate magnitudes (half apex of the expression) to high magnitudes (apex of the expression). Figure 4 shows the half apex and apex facial configurations for three expressions: surprise, anger, and joy. In the final stage of the learning, all training trajectories are aligned in the time domain using the DTW technique by fixing a nominal duration for a facial expression. In our experiments, this nominal duration is set to 18 frames.

3.2 Recognition

In the recognition phase, the 3D head pose and facial actions are estimated online from the video sequence using the appearance-based face and facial action tracker. We infer the facial expression associated with the current frame t by considering the estimated trajectory, i.e. the sequence of vectors $\boldsymbol{\tau}_{\mathbf{a}(t)}$ within a temporal window of size 18 centered at the current frame t. This trajectory (feature vector) is then classified using classical classification techniques that rely on the learned examples. We have used three different classification methods: (1) LDA, (2) NDA [17], and (3) SVMs with a radial basis function [9].

It is worth noting that the static recognition scheme will use the facial actions associated with only one single frame. In this case, the training examples correspond to the apex of the expression or to its half apex.

Fig. 3 Three video examples associated with the CMU database depicting surprise, anger, and joy expressions. The *left frames* illustrate the half apex of the expression. The *right frames* illustrate the apex of the expression

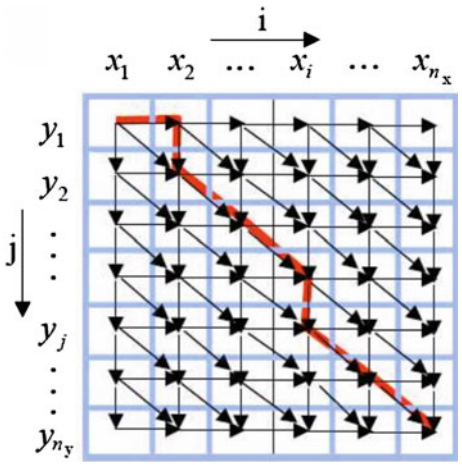


Fig. 4 The DTW aims at aligning two multivariate signals having different number of samples. The matching of these two signals reduces to finding an optimal path (*red path*) that minimizes a global cost, which is carried out by dynamic programming [24]

4 Experimental results and method comparisons

4.1 Recognition using a small data set

In our experiments, we used a subset from the CMU facial expression database, containing seven persons who are displaying five expressions: surprise, sadness, joy, disgust and anger. For dynamical facial expression recognition evaluation, we used the truncated trajectories, i.e., the temporal sequence containing nine frames, with the first frame representing a "subtle" facial expression (corresponding more or less with a "half apex" state, see the left column of Fig. 4) and the last one corresponding to the apex state of the facial expression (see the right column of Fig. 4). We decided to remove in our analysis the first few frames (from initial, "neutral" state to "half-apex") since we found them irrelevant for the purposes of the current study. The results reported in this section are based on the "leave-one-out" cross-validation

Table 1 Overall classification results for K-NN based on LDA

Classifier type	$K = 1$ (%)	$K = 3$ (%)	$K = 5$ (%)
Dynamic	94.28	88.57	82.86
Static (apex)	91.42	91.43	88.57
Static (half-apex)	85.71	82.85	80.00

Table 2 Overall classification results for K-NN based on NDA

Classifier type	$K = 1$ (%)	$K = 3$ (%)	K=5 (%)
Dynamic	88.57	88.57	85.71
Static (apex)	85.71	88.57	91.42
Static (half-apex)	82.85	80.00	80.00

Table 3 SVM: overall recognition rate for the dynamic classifier

c ($g = 1/54$)	$g/2$ (%)	g (%)	2 g (%)
1	91.43	91.43	91.43
5	91.43	94.28	97.14
10	97.14	97.14	97.14
50	97.14	100.00	97.14
100	100.00	97.14	97.14
500	97.14	97.14	97.14
1,000	97.14	97.14	97.14

strategy. Several subspace learning techniques have been tested: LDA and NDA. For LDA and NDA, the classification was based on the K-Nearest Neighbor rule (K-NN = 1, 3, 5), meanwhile SVM classifier has been applied on raw data.

In order to assess the benefit of using temporal information, we performed also the "static" facial expression recognition, by comparing frame-wise the instances corresponding to half-apex and apex states, respectively.

The results for K-NN, based on LDA and NDA representations, are reported in Tables 1 and 2, respectively. The SVM results for the dynamic classifier are reported in Table 3. The kernel used was a radial basis function. Thus, the SVM had two parameters to tune 'c' (cost) and 'g' (gamma). In this case, we wanted to see how the variation of parameters 'c' and 'g' affect the recognition performance. We considered seven values for 'c' and three for 'gamma'. Table 4 shows the confusion matrix associated with a given "leave-one-out" for the dynamic classifier using SVM. Since we noticed that 'gamma' does not have a significant impact on the classification results, for the study of frame-wise case, we set this parameter to its default value ($1/dim(featurevector) = 1/54$) and considered only different values for the 'c' parameter.

As can be seen, in Table 3, by increasing the value of c, the recognition rate obtained by the leave-one-out cross-validation increases, then it reaches a maximum for several values of c (it may slightly decrease). This behavior can be

Table 4 Confusion matrix for the dynamic classifier based on SVM

	Surprise	Sadness	Joy	Disgust	Anger
Surprise (7)	7	0	0	0	0
Sadness (7)	0	6	0	1	0
Joy (7)	0	0	7	0	0
Disgust (7)	0	0	0	7	0
Anger (7)	0	1	1	1	4

The results correspond to the case when $c = 1$ and g= 0.0185

explained by the fact that the size of the whole dataset is relatively small. As found by many researchers, the best value of c is not predictable and depends on the problem at hand and on the data used. Note that the parameter c controls the trade-off between a hard margin SVM (all training samples are correctly classified) and a soft margin SVM (some misclassified training samples are allowed). If c tends to infinity, the soft margin SVM tends to the hard margin SVM for separable data.

The results for the static classifier based on SVM are presented in Table 5. At the same time, we present in Table 6, a comparison between our approach and other state of the art methods: AAM + LDA [1], LBP [28], Gabor [31] and LBP [33]. It can be appreciated that our approach outperforms the existing ones. The value which appears in Table 6 has been obtained by averaging the recognition rates corresponding to the apex, from Table 5.

To conclude this part of the experimental results, we could say that, in general, the dynamic recognition scheme has outperformed all static recognition schemes. Moreover, we found out that the SVM clearly outperforms K-NN in classification accuracy.

Besides the experiments described above, we performed also a cross-check validation. In the first experiment, we trained the static classifier with the frames corresponding to half-apex expression and use the apex frames for test. We refer to this case as 'minor' static classifier. In another set of experiments, we trained the classifier with the apex frames and test it using the half-apex frames ('major' static classifier). The classification results for K-NN (based on LDA and NDA) and SVM are presented in the Tables 7, 8 and 9, respectively.

In conclusion, we could observe that the 'minor' static classifier has comparable results to the static "half apex" classifier. This means that a learning based on data featuring half apex expressions will have very good generalization capabilities since the tests with both kinds of data (half-apex and apex expressions) have a high recognition rate. Also, one can notice that the recognition rate of the 'minor' static classifier is higher than that of the 'major' static classifier.

This result may have very practical implications assuming that training data contain non-apex expressions, specially for

Table 5 SVM: overall recognition rate for the static classifier

Static type	$c = 1$ (%)	$c = 5$ (%)	$c = 10$ (%)	$c = 50$ (%)	$c = 100$ (%)	$c = 500$ (%)	$c = 1{,}000$ (%)
Apex	82.86	97.14	100.00	94.28	94.28	94.28	94.28
Half-apex	82.85	82.85	85.71	94.28	94.28	94.28	91.43

Table 6 Comparison of different approaches for the static classifier

Method	Overall results (%)
Our approach	**93.9**
AAM + LDA [1]	82.9
LBP [28]	92.1
Gabor [31]	92.2
LBP [33]	55.6

Table 7 K-NN results based on LDA: cross-check validation results for the static classifier

Static classifier	$K = 1$ (%)	$K = 3$ (%)	$K = 5$ (%)
Minor	82.85	85.71	85.71
Major	57.14	65.71	62.85

Minor: train with half-apex frames and test with apex. Major: train with apex frames and test with half-apex

real-world applications. In human–computer interaction scenarios, for instance, we are interested in quantifying human reaction based on its natural behavior. For this reason, we have to acquire and process data "online" without any external intervention. In this context, it is highly unlikely to capture automatically a person's apex of the facial expression. Most of the time we are tempted to show more "subtle" versions of our expressions and when we indeed show apex, this is in very specific situations and for very brief periods of time.

4.2 Recognition using a large data set

In order to evaluate the performance of the proposed recognition schemes with a large data set, we proceed as follows. We generate 350 synthetic sequences from the existing 35 sequences by exploiting the convexity principle that states that each example can be approximated by a given linear combination of two real examples. This process is commonly used in machine learning for enlarging the size of the training data set.

In our implementation, a synthetic sequence is reconstructed by a random linear combination of two real sequences chosen at random. The sequence is then perturbed by a Gaussian noise. This will give a data set of 350 synthetic sequences. For training we used the original CMU sequences, and for test we used the synthetic ones. The recognition results associated with K-NN (based on LDA and NDA) and SVM are summarized in Table 10.[1] As can be seen, the recognition rates are slightly better than those obtained with small data sets.

Table 8 K-NN results based on NDA: cross-check validation results for the static classifier

Static classifier	$K = 1$ (%)	$K = 3$ (%)	$K = 5$ (%)
Minor	94.28	88.57	85.71
Major	65.71	62.66	60.00

Minor: train with half-apex frames and test with apex. Major: train with apex frames and test with half-apex

4.3 Discussion

The proposed scheme does not require tedious learning stages since it is 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 not a contradiction with the claim that the approach is texture independent. Indeed, the tracking of facial actions is carried out using Online Appearance Models which dynamically learn the face appearance online. Due to the use of a deformable 3D model, our approach has an additional advantage by which the facial expression recognition can be performed even when the face is in a non-frontal view. Another advantage comes from the dynamic nature of the scheme: it exploits the spatio-temporal configuration of the facial actions. For this reason, changes in either the video rate or the facial action duration do not affect the recognition accuracy this is due to the use of DTW technique which overcomes such non-linear time scale.

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); and (3) by including more actions units, the 3D face and facial action tracker may become not suited for real-time applications. The

[1] Several parameters relative to K-NN and SVM have been tested, but only the best results are shown.

Table 9 SVM: cross-check validation results for the static classifier

Static type	$c = 1$	$c = 5$	$c = 10$	$c = 50$	$c = 100$	$c = 500$	$c = 1{,}000$
Minor	80.00	80.00	85.71	85.71	82.86	80.00	82.86
Major	48.57	60.00	51.43	45.71	48.57	48.57	48.57

Minor: train with half-apex frames and test with apex. Major: train with apex frames and test with half-apex

Table 10 Recognition rate using K-NN (based on LDA and NDA) and SVM applied to the large data set

Subspace learning type	Recognition rate (%)
K-NN (LDA)	90.57
K-NN (NDA)	91.71
SVM	98.00

The training was based on the CMU data and the synthetic data was used for test

currently 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).

5 Conclusions and future work

In this paper, we addressed the dynamic facial expression recognition in videos. We introduced a view and texture-independent scheme that exploits facial action parameters estimated by an appearance-based 3D face tracker. We represented the corresponding learned facial actions associated with different facial expressions by time series. In order to show even better the benefits of employing a dynamic classifier, we compared it with static classifiers, built on the half-apex and apex frames of the corresponding facial expressions. We also showed that by only using half-apex frames to train the static classifiers, we still get very reliable predictions about the real facial expression (tests were done with apex and half-apex frames).

By including the retrieval of the facial actions by the real-time face tracker the task of recognizing one frame (static recognition) can take about 60 ms on current PCs. It is slightly more than 60 ms for the dynamic recognition. It is worth noting that despite the fact that the recognition methods are fast enough, a real-time performance is not required since on average humans display a dynamic expression in about half a second. In consequence, any physical system based on our frameworks will be rapid enough to be used without missing the detection of dynamic facial expressions displayed in videos. Due to the use of the *Candid 3* model, our facial expression recognition schemes can be applied within a range of out-of-plane face rotation from −50 to +50 degrees.

In the future, we want to further explore the results obtained in this paper by focusing on two directions: using facial actions as a hint to assess persons' level of interest during an event and trying to discriminate between a fake and a genuine facial expression.

Acknowledgments B. Raducanu is supported by the projects TIN2009-14404-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018), Ministerio de Ciencia e Innovacion, Spain. This work is supported in part by the Spanish Government under the project TIN2010-18856.

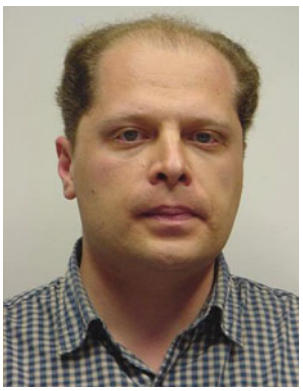
References

1. Abboud, B., Davoine, F.: Facial expression recognition and synthesis based on appearance model. Signal Process. Image Commun. **19**(8), 723–740 (2004)
2. Ahlberg, J.: Model-based coding: extraction, coding and evaluation of face model parameters. Ph.D. Thesis, Department of Electrical Engineering, Linköping University, Sweden (2002)
3. Ambadar, Z., Schooler, J., Cohn, J.: Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions. Psychol. Sci. **16**(5), 403–410 (2005)
4. Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, vol. I, pp. 592–597. The Hague, The Netherlands (2004)
5. Black, M.J., Yacoob, Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. Int. J. Comp. Vis. **25**(1), 23–48 (1997)
6. Breazeal, C.: Robot in society: friend or appliance? In: Proceedings of Workshop on Emotion-Based Agent Architectures. Seattle (1999)
7. Breazeal, C.: Sociable machines: expressive social exchange between humans and robots. PhD Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, USA (2000)
8. Cañamero, L., Gaussier, P.: Emotion understanding: robots as tools and models. In: Nadel, J., Muir, D. (eds.) Emotional Development: Recent Research Advances, pp. 235–258. Oxford Univesity Press, Oxford (2005)
9. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
10. Dornaika, F., Davoine, F.: View and texture-independent facial expression recognition in videos using dynamic programming. In: Proceedings of IEEE International Conference on Image Processing, Genova, Italy (2005)
11. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. IEEE Trans. Circuits Syst. Video Technol. **16**(9), 1107–1124 (2006)

12. Dornaika, F., Raducanu, B.: Inferring facial expressions from videos: tool and application. Signal Process. Image Commun. **22**(9), 769–784 (2007)
13. Ekman, P.: Facial expressions of emotions: an old controversy and new findings. Philos. Trans. Royal Soc. Lond. **B335**, 63–69 (1992)
14. Ekman, P.: Facial expression and emotion. Am. Psychol. **48**(4), 384–392 (1993)
15. Ekman, P., Davidson, R.: The Nature of Emotion: Fundamental Questions. Oxford University Press, New York (1994)
16. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recogn. **36**(1), 259–275 (2003)
17. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, Boston (1990)
18. Huang, X., Zhao, G., Pietikäinen, M., Zheng, W.: Expression recognition in videos using a weighted component-based feature descriptor. In: Heyden, A., Kahl, F. (eds.) Image Analysis. Springer, Berlin (2011)
19. Kanade, T., Cohn, J., Tian, Y.L.: Comprehensive database for facial expression analysis. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53. Grenoble, France (2000)
20. Kim, Y., Lee, S., Kim, S., Park, G.: A fully automatic system recognizing human facial expressions. In: Negoita, M., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based Intelligent Information and Engineering Systems. Springer, Berlin (2004)
21. Meng, H., Romera-Paredes, B., Bianchi-Berthouze, N.: Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. In: Proceedings of IEEE International Conference on Face and Gesture Recognition—Workshop on Facial Expression Recognition and Analysis Challenge, Santa Barbara, 25 March 2011
22. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. Comp. Vis. Image Underst. **115**, 541–558 (2011)
23. Moore, S., Ong, E., Bowden, R.: Facial expression recognition using spatiotemporal boosted discriminator classifiers. In: Campilho A., Kamel, M. (eds.) International Conference on Image Analysis and Recognition, Springer, Berlin (2010)
24. Müller, M.: Information Retrieval for Music and Motion. Springer, Berlin (2007)
25. Pantic, M.: Affective Computing. In: Pagani, M. (ed.) Encyclopedia of Multimedia Technology and Networking vol. I, pp. 8–14. Idea Group Publishing, USA (2005)
26. Picard, R.W., Vyzas, E., Healy, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1175–1191 (2001)
27. Robin, T., Bierlairey, M., Cruz, J.: Dynamic facial expression recognition with a discrete choice model. J. Choice Model. **2**(1), 95–148 (2011)
28. Shan, C., Gong, S., McOwan, P.W.: Robust facial expression recognition using local binary patterns. In: Proceedings of International Conference on Image Processing, vol. II, pp. 370–373. Genoa, Italy (2005)
29. Shan, C., Gong, S., McOwan, P.W.: Dynamic facial expression recognition using a bayesian temporal manifold model. In: Proceedings of British Machine Vision Conference, vol. I, pp. 297–306. Edinburgh, UK (2006)
30. Sung, J., Lee, S., Kim D.: A real-time facial expression recognition using the STAAM. In: Proceedings of International Conference on Pattern Recognition, vol. I, pp. 275–278. Hong Kong, PR China (2006)
31. Tian, Y.: Evaluation of face resolution for expression analysis. In: Proceedings of IEEE Workshop on Face Processing in Video. Washington, DC, USA (2004)
32. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 97–115 (2001)
33. Valstar, M.F., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: Proceedings of IEEE International Conference on Face and Gesture Recognition—Workshop on Facial Expression Recognition and Analysis Challenge, Santa Barbara, 25 March 2011
34. Wu, T., Bartlett, M.S., Movellan, J.R.: Facial expression recognition using gabor motion energy filters. In: Computer Vision and Pattern Recognition Workshops (CVPRW). San Francisco, USA (2010)
35. Xiang, T., Leung, M.K.H., Cho, S.Y.: Expression recognition using fuzzy spatio-temporal modeling. Pattern Recogn. **41**(1), 204–216 (2008)
36. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. Multimed. **8**(3), 500–508 (2006)
37. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 699–714 (2005)
38. Zhang, S., Zhao, X., Lei, B.: Robust facial expression recognition via compressive sensing. Sensors **12**, 3747–3761 (2012)

Author Biographies

Bogdan Raducanu received his B.Sc. degree in computer science from the University "Politehnica" of Bucharest, Bucharest, Romania, in 1995 and the Ph.D. degree "Cum Laude" from the University of the Basque Country, Bilbao, Spain, in 2001. Currently, he is a senior researcher at the Computer Vision Center in Barcelona, Spain. His research interests are computer vision, pattern recognition, machine learning, artificial intelligence, social computing and human–robot interaction. He is the author or co-author of about 70 publications in international conferences and journals. In 2010, he was the leading Guest Editor of Image and Vision Computing journal for a special issue on 'Online Pattern Recognition'.



Fadi Dornaika received his Ph.D. in signal, image, and speech processing from Institut National Polytechnique de Grenoble, France, in 1995. He is currently an Ikerbasque research professor at the University of the Basque Country. He has published more than 150 papers in the field of computer vision. His research concerns geometrical and statistical modelling with focus on 3D object pose, real-time visual servoing, calibration of visual sensors, cooperative stereo-motion, image registration, facial gesture tracking, and facial expression recognition.