MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
8-13

FACIAL EXPRESSION RECOGNITION IN CONTINUOUS VIDEOS USING LINEAR DISCRIMINANT ANALYSIS

Fadi Dornaika and Franck Davoine

HEUDIASYC Mixed Research Unit, CNRS/UTC
Compiègne University of Technology
60205 Compiègne Cedex, FRANCE
dornaika@cvc.uab.es
fdavoine@hds.utc.fr

ABSTRACT

In this paper, we address the recognition of facial expressions in continuous videos. We introduce a view- and texture-independent approach that exploits facial action parameters estimated by an appearance-based 3D tracker. We represent the learned facial actions associated with different facial expressions by time series. These time series are then efficiently and compactly represented in Eigenspace and Fisherspace for subsequent recognition. The developed approach is fast and can be used online. Experiments demonstrated the effectiveness of the developed method.

1. INTRODUCTION

Computational facial expression analysis is a challenging research topic in computer vision. It is required by many applications such as human-computer interaction and computer graphic animation. Facial expression recognition is an important part of the so-called Cognitive Vision (CV). Automatic facial expression recognition did not start until the 1990s. To classify expressions in still images many techniques have been proposed such as Neural Nets and Gabor wavelets [1]. Recently, more attention has been given to modelling facial deformation in dynamic scenarios. Still image classifiers use feature vectors related to a single frame to perform classification. Temporal classifiers try to capture the temporal pattern in the sequence of feature vectors related to each frame such as the Hidden Markov Models (HMM) based methods [2]. A survey on facial expression recognition methods can be found in [3].

Our work focuses on the design of classifiers used for performing the recognition following the extraction of facial actions using our 3D face and facial action tracker [4]. Note that tracking the face and the facial actions can be carried out using other appearance-based trackers. For example, the 3D Active Appearance Model tracker described in [5] can be used for such purposes.

In this paper, we introduce a view- and texture-independent approach that exploits the tracked facial actions. The developed approach is fast and can be used online. At the learning stage, training videos illustrating basic expressions are tracked. Then the computed facial actions are aligned in the time domain. These aligned trajectories (represented by one-dimensional vectors) are then used for building an eigensystem and Fisherspace (Principal Component Analysis plus Linear Discriminant Analysis). The modelling and recognition occur in the Fisherspace. One advantage of our proposed scheme is that the spatio-temporal structures of the facial expressions are represented in a compact form.

2. HEAD AND FACIAL ACTION TRACKING

In our study, we use the 3D face model *Candide* [6]. This 3D deformable wireframe model is given by the 3D coordinates of the vertices $\mathbf{P}_i, i = 1, \ldots, n$ where n is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$-vector **g** – the concatenation of the 3D coordinates of all vertices $\mathbf{P}_i$. The vector **g** can be written as:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{S}\,\tau_{\mathbf{s}} + \mathbf{A}\,\tau_{\mathbf{a}} \qquad (1)$$

where $\overline{\mathbf{g}}$ is the standard shape of the model, and the columns of **S** and **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 $\mathbf{S}\,\tau_{\mathbf{s}}$ accounts for shape variability (inter-person variability) while the term $\mathbf{A}\,\tau_{\mathbf{a}}$ accounts for the facial action (intra-person variability). In this study, we use 12 modes for the shape unit matrix and six modes for the action units matrix. Without loss of generality, we have chosen the following action units: 1) Lower lip depressor, 2) Lip stretcher, 3) Lip corner depressor, 4) Upper lip raiser, 5) Eyebrow lowerer, 6) Outer eyebrow raiser. Thus, for a given person, the state of the 3D model is given by the 3D head pose (three rotations and three translations) and the facial deformation/actions encoded by the control vector $\tau_{\mathbf{a}}$. This is given by the vector **b**:

$$\mathbf{b} \quad = \quad [\theta_x, \quad \theta_y, \quad \theta_z, \quad t_x, \quad t_y, \quad t_z, \quad {\tau_{\mathbf{a}}}^T \ ]^T \quad (2)$$

Tracking the head and facial actions is carried out using our tracker [4]. This appearance-based tracker aims at computing the 3D head pose and facial actions, i.e. the vector **b**, by minimizing a distance between the incoming warped frame and the current *shape-free* appearance of the 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. Figure 1 displays the tracking results associated with two video sequences featuring quite large pose variations as well as large facial actions. In this figure, for each video frame the *Candide* model is globally and locally deformed according to the computed vector **b**, and then projected onto the corresponding frame. In the sequel, we show that the estimated facial actions, encoded by the vector $\tau_{\mathbf{a}}$, can be utilized for recognizing the facial expression in continuous videos.

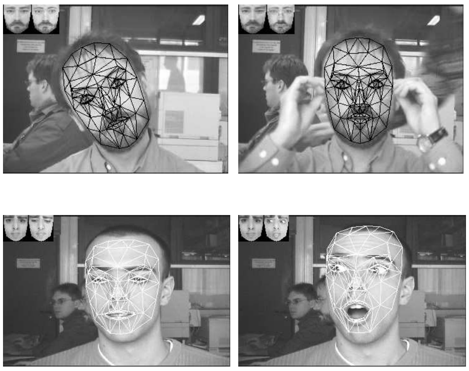
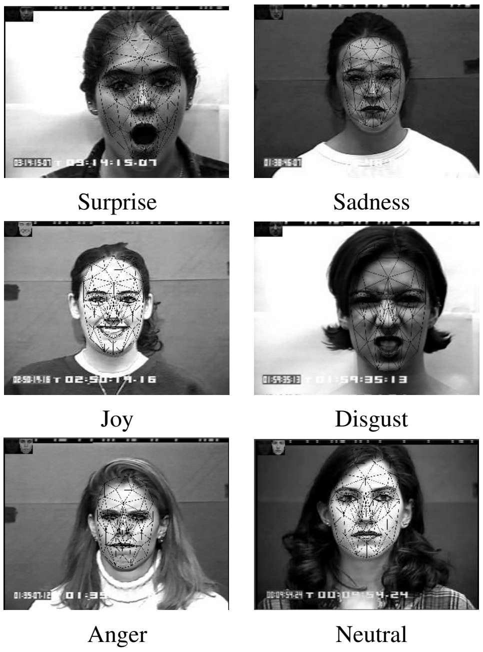


Fig. 1. Face and facial action tracking results using our appearance-based tracker [4] with two video sequences.

3. LEARNING AND MODELLING

The learning phase is split into two stages. In the first stage, continuous videos depicting different facial expressions are tracked and the retrieved facial actions are represented by time series. In the second stage, a compact model of each expression class is retrieved from these temporal representations.

The training video sequences have been picked up from the CMU database [7]. The used sequences depict five frontal view universal expressions (surprise, sadness, joy, disgust and anger). Each basic expression is performed by 7 different people. Altogether we use 35 video sequences composed of around 15 to 20 frames each. Each training video depicts the expression starting from a neutral configuration.



Surprise Sadness

Joy Disgust

Anger Neutral

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

Figure 2 shows six videos belonging to the CMU database. The first five images depict the adaptation/tracking results associated with the high magnitude of the five basic expressions.

The second stage is depicted in Figure 3. In order to obtain training trajectories with the same number of frames (duration) the trajectories belonging to the same expression class are aligned using the Dynamic Time Warping technique [8]. In this case, we pick up one training video arbitrarily and we align the remaining training videos with respect to the selected one.

Let $\mathbf{e}_i^j$ be the aligned trajectory i belonging to the expression class j where $i = 1 \ldots 7$ and $j = 1 \ldots 5$. The example $\mathbf{e}_i^j$ is represented by a column vector of dimension $6{\times}T$ and is obtained by simply concatenating the facial action 6-vectors $\tau_{\mathbf{a}(t)}$:

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

Note that T represents the duration of the aligned trajectories and is the same for all examples. In our study, a nominal duration of 18 frames for the aligned trajectories makes the dimension of all examples $\mathbf{e}_i^j$ (all i and j) equal to 108.

Applying a Principal Component Analysis (PCA) on the set of all training trajectories yields the mean trajectory $\bar{\mathbf{e}}$ as well as the principal modes of variation. Any training trajectory **e** can be approximated by the principal modes using the q largest eigenvalues:

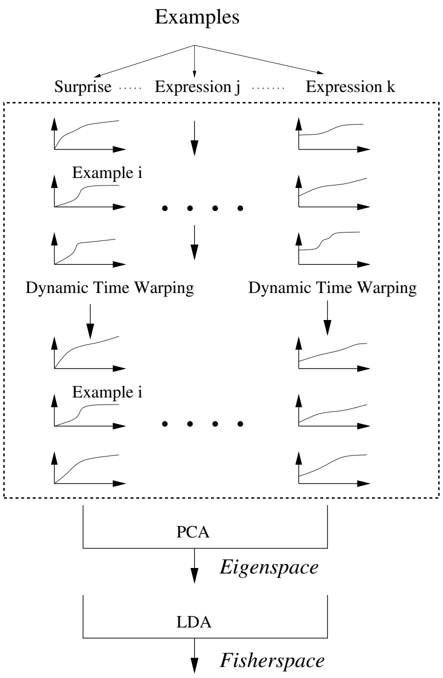


Fig. 3. The parameterized modelling of facial expressions using Eigenspace and Fisherspace.

$$\mathbf{e} \cong \bar{\mathbf{e}} + \mathbf{U}\,\mathbf{c} = \bar{\mathbf{e}} + \sum_{l=1}^{q} c_l\,\mathbf{u}_l \tag{3}$$

where $\mathbf{u}_l$ denote the q basis vectors (principal modes). 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 compact parametrization of any input trajectory, $\hat{\mathbf{e}}$, in the space spanned by the q basis vectors $\mathbf{u}_l$. The vector $\mathbf{c}$ is given by:

$$\mathbf{c} = \mathbf{U}^T\,(\hat{\mathbf{e}} - \bar{\mathbf{e}}) \tag{4}$$

Thus, all training trajectories $\mathbf{e}_i^j$ can now be represented by the vectors $\mathbf{c}_i^j$ using (4). Linear Discriminant Analysis can be applied on the computed vectors $\mathbf{c}_i^j$. This gives a new space (the Fisherspace) in which each training video sequence is represented by a point having $k-1$ coordinates where k is the number of expression classes. Eventually, each basic expression is represented by a cluster of points in the Fisherspace (each point represents a training video sequence), and the corresponding first and second moment can be easily computed.

Figure 4 illustrates the modelling results associated with the used 35 training videos/trajectories. It displays the second component versus the first one. In this space, each training video is represented by a point having 5 coordinates. Here, we have used six expression classes: the five basic expressions and the neutral expression. In order to obtain a realistic modelling of the neutral expression, we have used a set of sub-trajectories picked up arbitrarily from the original 35 training videos, i.e. we use the start phase of the expressions. In this figure, the ideal neutral trajectory (a sequence of zero vectors) is represented by a star.

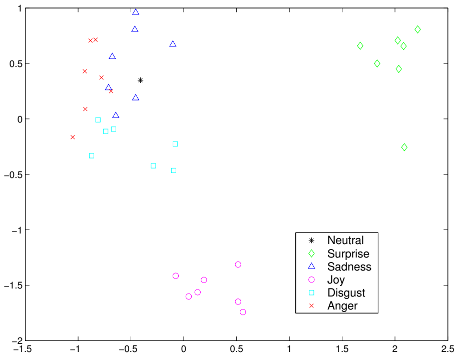


Fig. 4. The 35 training videos associated with the five basic facial expressions depicted in the Fisherspace. In this space, each training video is represented by a point having 5 coordinates. The plot displays the second component versus the first one. The ideal neutral trajectory (a sequence of zero vectors) is represented by a star.

4. FACIAL EXPRESSION 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 our tracker. We consider the one-dimensional 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 T should be the same as in the learning stage. This vector is then classified in the Fisherspace. Two classification criteria are used: the Euclidean distance from the expression mean and the *Mahalanobis* distance from this mean.

It is worthwhile noting that the computational cost of this recognition scheme does not depend on the number of examples since basic statistics have already been computed in the learning stage. Figure 5 shows the similarity measure associated with each expression class obtained with a 741-frame-long sequence. For each frame, the similarity is given by the Euclidean distance between the current trajectory and the expression mean, all expressed in the Fisherspace. Figure 6 shows the recognition results for frames 200, 400, 650, and 741 (these frames are marked with vertical arrows in figure 5). Similar results have been obtained with the *Mahalanobis* distance.

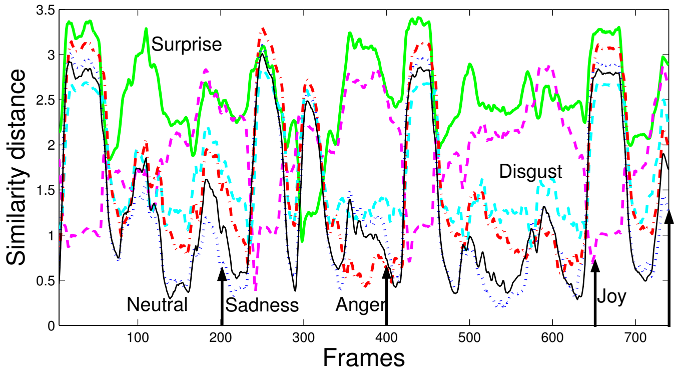
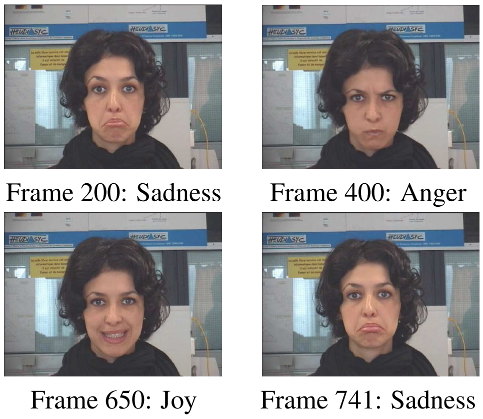


Fig. 5. The similarity measure for each expression class and for each frame of a 741-frame-length sequence.


Frame 200: Sadness

Frame 400: Anger

Frame 650: Joy

Frame 741: Sadness

Fig. 6. Facial expression recognition associated with frames 200, 400, 650, and 741. The associated similarity measures are shown with vertical arrows in Figure 5.

5. PERFORMANCE EVALUATION

In order to quantify the recognition rate, we have created videos featuring the basic facial expressions. To this end, we have asked a volunteer student to perform each basic expression several times in a relatively long sequence. The subject was instructed to display the expression in a natural way, i.e. the displayed expressions were independent of any database. Table 1 shows the confusion matrix obtained with the developed method. The learned models are inferred from the CMU database while the used test videos are created at our laboratory. The recognition rate of dynamical expressions was very good on tested video sequences for all basic expressions except for the disgust expression for which the recognition rate was 55%. 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 is confirmed by Figure 7. Therefore, for the above experiment, the overall recognition rate is 92.3%.

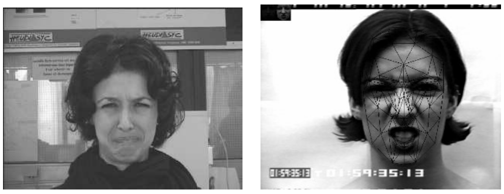


Fig. 7. The disgust expression performed by a CMU subject (Right) and by our subject (Left). Although both subjects claim that they are displaying a disgust expression, the mouth configurations are markedly different.

	Surp.	Sad.	Joy	Disg.	Ang.
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

Table 1. Confusion matrix obtained with the developed method. The learned models are inferred from the CMU database while the test videos are created at our laboratory.

6. REFERENCES

[1] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, “Machine learning methods for fully automatic recognition of facial expressions and facial actions,” in *IEEE Int. Conference on Systems, Man and Cybernetics*, 2004.

[2] I. Cohen, N. Sebe, A. Garg, L. Chen, and T.S. Huang, “Facial expression recognition from video sequences: Temporal and static modeling,” *Computer Vision and Image Understanding.*, vol. 91, no. 1-2, pp. 160–187, 2003.

[3] M. Pantic and L.J.M. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[4] F. Dornaika and F. Davoine, “Head and facial animation tracking using appearance-adaptive models and particle filters,” in *IEEE CVPR Workshop on Real-Time Vision for Human-Computer Interaction, Washington DC*, July 2004.

[5] I. Matthews and S. Baker, “Active appearance models revisited,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[6] J. Ahlberg, “CANDIDE-3 - an updated parametrized face,” Tech. Rep. LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden, 2001.

[7] T. Kanade, J. Cohn, and Y.L. Tian, “Comprehensive database for facial expression analysis,” in *International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000, pp. 46–53.

[8] D. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series,” in *AAAI-94 Workshop on Knowledge Discovery in Databases*, 1994.