STATISTICAL MODELING FOR FACIAL EXPRESSION ANALYSIS AND SYNTHESIS

Bouchra Abboud, Franck Davoine, Mô Dang

Heudiasyc Laboratory, CNRS, University of Technology of Compiègne.
BP 20529, 60205 COMPIEGNE Cedex, FRANCE.
E-mail: Franck.Davoine@hds.utc.fr

ABSTRACT

Facial expression interpretation, recognition and analysis is a key issue in visual communication and man to machine interaction. In this paper, we present a technique for extracting appearance parameters from a natural image or video sequence, which allow reproduction of natural looking expressive synthetic faces. This technique was used to perform face synthesis and tracking in video sequences as well as facial expression recognition and control.

1. INTRODUCTION

Natural human-machine interaction is becoming an active and important research area. Adequate feedbacks like speech, facial expression and body gestures are essential components of such interaction since these communicative events satisfy certain communication expectations in human-human interaction. The human face comprises major information about identity and emotion. It constitutes a source of many informative social signs and allows good communication expectation response.

Regarding emotions, previous works showed that in nearly all cultures, facial expressions of six basic emotional categories are universally recognized, namely: joy, sadness, anger, disgust, fear and surprise. Several other emotions and many combinations of emotions have been studied but remain unconfirmed as universally distinguishable. Thus, most of the research up to now has been oriented towards detecting these six universal expressions.

In this paper we address the issue of face modeling, video realistic face generation and facial expression synthesis and recognition using statistical active facial appearance models [1].

Thanks to the French Incentive Concerted Action for Young Researchers (*ACI Jeunes, Ministère de la Recherche*) and the european Interface project (FP5 - IST) for funding.

2. ACTIVE FACIAL APPEARANCE MODELS

2.1. Model description

It has been shown that the active appearance model [1] is a powerful tool for face synthesis and tracking. It uses Principal Component Analysis to model both shape and texture variations seen in a training set of visual objects. After having computed the mean shape $\bar{\mathbf{s}}$ and aligned all shapes from the training set by means of a Procrustes analysis, the statistical shape model is given by:

$$\mathbf{s}_i = \bar{\mathbf{s}} + \Phi_s \mathbf{b}_{si} \tag{1}$$

where $\mathbf{s}_i$ is the synthesized shape, Φ_s is a truncated matrix describing the principal modes of shape variations in the training set and $\mathbf{b}_{si}$ is a vector that controls the shape.

It is then possible to warp textures from the training set of faces onto the mean shape $\bar{\mathbf{s}}$ in order to obtain shape-free textures. Similarly, after computing the mean shape-free texture $\bar{\mathbf{g}}$ and normalizing all textures from the training set relatively to $\bar{\mathbf{g}}$ by scaling and offset of luminance values, the statistical texture model is given by:

$$\mathbf{g}_i = \bar{\mathbf{g}} + \Phi_t \mathbf{b}_{ti} \tag{2}$$

where $\mathbf{g}_i$ is the synthesized shape-free texture, Φ_t is a truncated matrix describing the principal modes of texture variations in the training set and $\mathbf{b}_{ti}$ is a vector that controls the synthesized shape-free texture.

By combining the training shape and texture vectors $\mathbf{b}_{si}$ and $\mathbf{b}_{ti}$ and applying further PCA the statistical appearance model is given by:

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \quad \text{and} \quad \mathbf{g}_i = \bar{\mathbf{g}} + Q_t \mathbf{c}_i \tag{3}$$

where Q_s and Q_t are truncated matrices describing the principal modes of combined appearance variations in the training set, and $\mathbf{c}_i$ is a vector of appearance parameters simultaneously controling both shape and texture [1].

[1] Note that only one PCA instead of three could have been used to compute the model, as explained and tested in section 4

a b c

Fig. 1. a: Target (original face); b: model initialization; c: after iterative refinement, the model has converged to the target face.

Given the parameter vector $\mathbf{c}_i$, the corresponding shape $\mathbf{s}_i$ and shape-free texture $\mathbf{g}_i$ can be computed respectively using equations (3). The reconstructed shape-free texture is then warped onto the reconstructed shape in order to obtain the full appearance of a face. Furthermore, in order to allow pose displacement of the model, it is necessary to add to the appearance parameter vector $\mathbf{c}_i$ a pose parameter vector $\mathbf{p}_i$ allowing control of scale, orientation and position of the synthesized face.

While a couple of appearance parameter vector $\mathbf{c}$ and pose parameter vector $\mathbf{p}$ represents a face, the active appearance model can automatically adjust those parameters to a target face [2], by minimizing a residual image $\mathbf{r}(\mathbf{c}, \mathbf{p})$ which is the texture difference between the synthesized face and the corresponding mask of the image it covers. The optimization scheme used here is based on the first order Taylor expansion described in [2].

In the following, the appearance and pose parameters obtained by this optimization procedure will be denoted respectively as $\mathbf{c}_{op}$ and $\mathbf{p}_{op}$.

2.2. Experimental setup

The appearance model is built using the CMU expressive face database [3]. Each sequence of this database contains ten to twenty images, beginning with a neutral expression and ending with a high magnitude expression. We selected 338 frontal still face images composed of 26 neutral expression faces, 26 moderate and 26 high magnitude *anger, disgust, fear, joy, surprise and sadness* expressions. Each moderate expression has been chosen manually by extracting an intermediate frame from the video sequence.

In order to build the model, a shape $\mathbf{s}$ of 57 landmarks is manually positioned on each of the 338 images, yielding shape-free texture vectors $\mathbf{g}$ of 3493 pixels. The model is built using 50 shape modes, 150 texture modes and 40 appearance modes: the vector $\mathbf{c}$ is composed of 40 components, that retain 98 percent of the combined shape and texture variation of the training set of faces. The model appears thus very compact: assuming that each of the 40 components of $\mathbf{c}$ is quantized on average with 10 bits, a compression ratio of about $\frac{40\times10}{4000 \text{ pixels}\times 8\text{bpp}} = 1:80$ is achieved.

The active appearance search algorithm is illustrated on a previously unseen face (Fig. 1a), with the mean shape and texture used as a first approximation (Fig. 1b). The model output converges to the target face as shown in Fig. 1c.

3. FACIAL EXPRESSION ANALYSIS / SYNTHESIS

3.1. Facial expression modeling

The aim of this section is to study a linear model, as it is proposed in [4], correlating the appearance parameters to facial expression intensity according to:

$$\mathbf{c} = \mathbf{a}_{\mathbf{e}0} + \mathbf{a}_{\mathbf{e}1}\mathcal{I} + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathcal{I}$ is a scalar varying from $\mathcal{I} = 0$ to indicate neutral expression to $\mathcal{I} = 1$ to indicate a high magnitude expression and $\boldsymbol{\varepsilon}$ is the approximation error. $\mathbf{a}_{\mathbf{e}0}$ and $\mathbf{a}_{\mathbf{e}1}$ are coefficient vectors, learnt for each facial expression ($\mathbf{e}$ is joy, fear, disgust, surprise, fear, sadness or neutral) by linear regression over the training set. The linear regression is performed using 3 control points for each expression namely neutral expression ($\mathcal{I} = 0$), moderate expression ($\mathcal{I} = 0.5$) and high magnitude expression ($\mathcal{I} = 1$).

3.2. Facial expression filtering

Once the coefficient vectors $\mathbf{a}_{\mathbf{e}0}$ and $\mathbf{a}_{\mathbf{e}1}$ have been learnt for a given expression $\mathbf{e}$, the linear model can be used to predict an artificial vector of appearance parameters $\mathbf{c}_{\mathbf{e}}(\mathcal{I})$ for a given intensity $\mathcal{I}$ of the expression $\mathbf{e}$:

$$\mathbf{c}_{\mathbf{e}}(\mathcal{I}) = \mathbf{a}_{\mathbf{e}0} + \mathbf{a}_{\mathbf{e}1}\mathcal{I}. \tag{5}$$

Note that a given intensity $\mathcal{I}$ of the expression $\mathbf{e}$ generates an unique value of the vector $\mathbf{c}_{\mathbf{e}}(\mathcal{I})$: the latter encodes thus an average appearance of this expression intensity, independently of any particular person. This means that the information about *identity* is contained in the residual $\boldsymbol{\varepsilon}$ in equation 4. Therefore, in order to synthesize a new expression intensity $\mathcal{I}'$ for a given person, the residual for this person has to be added to the average appearance $\mathbf{c}_{\mathbf{e}}(\mathcal{I}')$ of this new intensity. The procedure for doing this is detailed below.

Starting from an unseen face with a given expression (Fig. 2a), an appearance parameter vector $\mathbf{c}_{op}$ is first estimated as described at the end of 2.1, so that $\mathbf{c}_{op}$ synthesizes an artificial face similar to this target face (Fig. 2b).

Having a priori knowledge of the facial expression $\mathbf{e}$ represented on the target face, it is then possible to estimate the intensity of this expression by inverting equation (5):

$$\mathcal{I}_{est} = \mathbf{a}_{\mathbf{e}1}^{+}(\mathbf{c}_{op} - \mathbf{a}_{\mathbf{e}0}) \tag{6}$$

where $\mathbf{a}_{\mathbf{e}1}^{+}$ is the pseudo inverse of $\mathbf{a}_{\mathbf{e}1}$. The information relative to the person's identity will then be retrieved by filtering out the expression information contained in vector

a b c

Fig. 2. a: Target (original face). b: Reconstructed face using $\mathbf{c_{op}}$ obtained by iterative model adjustment to target face. c: Neutral expression obtained by canceling joy intensity.

$\mathbf{c_e}(\mathcal{I}_{est})$, the latter being evaluated at the estimated expression intensity $\mathcal{I}_{est}$ using equation (5). This gives the identity vector $\mathbf{c}_{res}$:

$$\mathbf{c}_{res} = \mathbf{c}_{op} - \mathbf{c_e}(\mathcal{I}_{est}). \quad (7)$$

Having this, it is possible to modify the facial expression intensity represented in vector $\mathbf{c_e}(\mathcal{I}_{est})$ by modifying the $\mathcal{I}$ value in equation (5). In particular it is possible to cancel the expression by setting $\mathcal{I} = 0$.

$$\mathbf{c_e}(0) = \mathbf{a}_{\mathbf{e}0} + \mathbf{a}_{\mathbf{e}1} \times 0 = \mathbf{a}_{\mathbf{e}0} \quad (8)$$

Then, by adding the identity vector $\mathbf{c}_{res}$ to the corrected expression, it is possible to modify the expression intensity shown on the target face from high magnitude to neutral as shown in Fig. 2c:

$$\mathbf{c}_{neutral} = \mathbf{c_e}(0) + \mathbf{c}_{res}. \quad (9)$$

3.3. Facial expression synthesis

Starting from the artificially generated neutral expression of the target face, it is possible to artificially generate any desired expression $\mathbf{e}'$ by applying the same method described in 3.2.

It is assumed that the linear model (4) for the new expression $\mathbf{e}'$ has been learnt on the training set, giving the corresponding $\mathbf{a}_{\mathbf{e}'0}$ and $\mathbf{a}_{\mathbf{e}'1}$ parameters. In the procedure described above, the appearance parameters describing the target face $\mathbf{c}_{op}$ will now be replaced by $\mathbf{c}_{neutral}$. It is then possible to estimate the intensity of the desired expression on the artificial neutral face. This value should be close to zero.

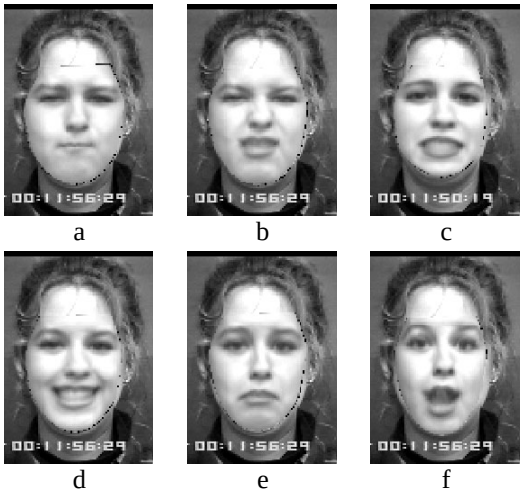
$$\mathcal{I}'_{est} = \mathbf{a}_{\mathbf{e}'1}^{+}(\mathbf{c}_{neutral} - \mathbf{a}_{\mathbf{e}'0}) \quad (10)$$

The estimated expression information vector at the $\mathcal{I}'_{est}$ intensity of the desired expression is given by:

$$\mathbf{c_{e'}}(\mathcal{I}'_{est}) = \mathbf{a}_{\mathbf{e}'0} + \mathbf{a}_{\mathbf{e}'1}\mathcal{I}'_{est} \quad (11)$$

The new residual $\mathbf{c}_{res}$ is then given by:

$$\mathbf{c}_{res} = \mathbf{c}_{neutral} - \mathbf{c_{e'}}(\mathcal{I}'_{est}) \quad (12)$$



a b c

d e f

Fig. 3. Generation of six synthetic expressions starting from the expression filtered face of figure 2.c. a: Anger. b: Disgust. c: Fear. d: Joy. e: Sadness. f: Surprise.

The facial expression intensity represented in vector $\mathbf{c_{e'}}(\mathcal{I}'_{est})$ can be controlled through the parameter $\mathcal{I}$ in equation (5). In particular it is possible to generate a high magnitude expression parameter estimation by setting $\mathcal{I} = 1$.

Then, by adding the identity vector $\mathbf{c}_{res}$ to the corrected expression estimation vector $\mathbf{c_{e'}}(\mathcal{I}'_{est})$, it is possible to modify the expression intensity shown on the target face from neutral to high magnitude as shown in figure 3.

$$\mathbf{c}_{intense} = \mathbf{c_{e'}}(1) + \mathbf{c}_{res}. \quad (13)$$

3.4. Evolution of the expression over a video sequence

In order to analyze the temporal behaviour of the linear model obtained in section 3.1, a series of experiments has been performed on a set of 15 videos representing different persons showing a facial expression gradually evolving from neutral to high magnitude.

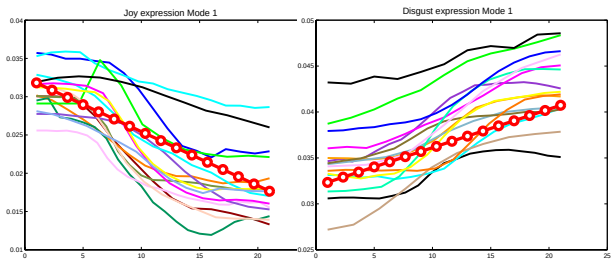


Fig. 4. For joy and disgust, evolution of 1st mode of $\mathbf{c_e}(\mathcal{I}_{est})$ over each video sequence. Straight line: 1st mode of $\mathbf{c_e}(\mathcal{I})$, with $\mathcal{I}$ linearly varying from 0 to 1.

The active appearance model is fit on each image of a

video sequence, using the previous output of the model as a first approximation of appearance and pose parameters.

At each step of the video sequence, the obtained $\mathbf{c}_{op}$ parameter vector allows to estimate the facial expression intensity $\mathcal{I}_{est}$ using equation (6), as well as the linearly predicted vector of appearance parameters $\mathbf{c}_\mathbf{e}(\mathcal{I}_{est})$ at this intensity. For each facial expression, the $\mathbf{c}_\mathbf{e}(\mathcal{I}_{est})$ parameters will follow a well defined trajectory whose behavior can be linearly approximated by the linear model computed in section 3.1. This is illustrated in Fig. 4 showing the evolution of the first variation mode (first coefficient of $\mathbf{c}_\mathbf{e}(\mathcal{I}_{est})$) over 15 video sequences; on each sequence, the face displays an expression going from neutral to high magnitude, respectively for joy (left plot) and disgust (right plot).

4. FACIAL EXPRESSION RECOGNITION

To classify a new face represented by the parameter vector $\mathbf{c}_i$, we use a Linear Discriminant Analysis scheme [5], assuming that the expression classes have a common covariance matrix. We measure the squared Euclidian distance $d_M^{lda}(\mathbf{c}_i^{lda}, \bar{\mathbf{c}}_j^{lda}) = (\mathbf{c}_i^{lda} - \bar{\mathbf{c}}_j^{lda})^t(\mathbf{c}_i^{lda} - \bar{\mathbf{c}}_j^{lda})$ where $\mathbf{c}_i^{lda}$ is the projection of the $\mathbf{c}_i$ vector on the LDA space and $\bar{\mathbf{c}}_j^{lda}$ is the j estimated mean vector in the LDA space. Finally we assign $\mathbf{c}_i^{lda}$ to the class of the nearest mean. Results are shown in table 1.

	neut.	ang.	disg.	fea.	joy	surp.	sad.
neutral	31	6	1	1	0	0	7
anger	2	14	1	0	0	1	1
disgust	0	2	16	0	1	0	0
fear	0	1	0	16	2	0	0
joy	1	1	1	0	25	0	0
surprise	1	0	1	2	2	21	2
sadness	3	0	0	1	0	0	15

Table 1. Confusion matrix for the classical model expression classifier, using 179 unknown test images. Globally, 77% of the images were correctly classified.

We also tested an alternative modelling approach of the shape and texture variations seen in the training set by using only one PCA. This approach consists in concatenating the aligned training shape $\mathbf{s}_i$ and shape free texture vector $\mathbf{g}_i$ in one appearance vector $\mathbf{b}_{sti} = ((\mathbf{W}_s\mathbf{s}_i)^t, \mathbf{g}_i^t)^t$. $\mathbf{W}_s$ is a suitably chosen weight matrix allowing to balance shape and texture variations. Applying PCA to the latter gives the one PCA appearance model (14) where $\mathbf{Q}_{st}$ is a truncated matrix describing the principal modes of $\mathbf{b}_{sti}$ variations, and $\mathbf{c}_{st}$ is a vector of appearance parameters controlling shape and texture. In practice we choose to keep 40 modes (96% of the observed variation) for the appearance vector $\mathbf{c}_{st}$.

$$\begin{pmatrix} \mathbf{W}_s\mathbf{s}_i \\ \mathbf{g}_i \end{pmatrix} = \mathbf{b}_{sti} = \bar{\mathbf{b}}_{st} + \mathbf{Q}_{st}\mathbf{c}_{st} \qquad (14)$$

The same optimisation procedure used in section 2.1 will be used to determine the optimal $\mathbf{c}_{st_{op}}$ parameters allowing to adjust the appearance model to an unknown target face. We classify a new face represented by the parameter vector $\mathbf{c}_{st_i}$ to the closest class, considering the squared Euclidian distance in the LDA space. Results are shown in table 2.

	neut.	ang.	disg.	fea.	joy	surp.	sad.
neutral	34	5	1	0	0	0	6
anger	1	15	3	0	0	1	0
disgust	0	2	16	0	1	0	0
fear	0	0	1	16	2	0	0
joy	1	0	1	0	25	0	0
surprise	1	0	0	1	2	23	2
sadness	4	1	0	1	0	0	13

Table 2. Confusion matrix for the one PCA model expression classifier, using 179 unknown test images. Globally, 79% of the images were correctly classified.

5. CONCLUSION AND PERSPECTIVES

In this paper, we present applications of Active Appearance Models for face analysis and synthesis, that could be useful for human to machine and machine to human interaction. Other classification approaches are currently being investigated, based on linear or non-linear schemes. In addition to the greylevel texture, multiresolution Gabor responses are also exploited to perform facial expression recognition.

6. REFERENCES

[1] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.

[2] T.F. Cootes and P. Kittipanya-ngam, "Comparing variations on the active appearance model algorithm," in *British Machine Vision Conference*, Cardiff University, September 2002, pp. 837–846.

[3] 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.

[4] H. Kang, T.F. Cootes, and C.J. Taylor, "Face expression detection and synthesis using statistical models of appearance," in *Measuring Behavior*, Amsterdam, The Netherlands, August 2002, pp. 126–128.

[5] S. Dubuisson, F. Davoine, and M. Masson, "A solution for facial expression representation and recognition," *Signal Processing: Image Communication*, vol. 17, no. 9, pp. 657–673, October 2002.