Expressive face analysis and synthesis for visual interaction

Bouchra Abboud, Franck Davoine, Mô Dang

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

Abstract

In this paper, we address the task of extracting, from a natural image or a video sequence, parameters of an appearance model describing expressive human faces. The parameters contain the information needed to reproduce a natural looking synthetic expressive face, possibly with different expressions (for visual communication), as well as for facial expression interpretation (for man-to-machine interaction). Results illustrate how a given natural expressive face can be resynthesized and tracked in a video. In a second step, results reveal how to control and force new expressions on a natural face. Then, we show how a facial appearance model may be used for facial expression recognition.

1. Introduction

Nowadays, different research activities are oriented towards making man-machine interaction as natural as possible, based on everyday human communication means like speech, facial expressions and body gestures from both sides. Human to machine communication will apply through audio-video channels, and will be integrated and analysed coherently not only to perform low level tasks, like word recognition or eye movement tracking, but also high level interpretation and data fusion, like speech emotion and facial expression understanding. Machine to human communication, on the other hand, can be based on human-like audio-video feedbacks simulating a "person in the machine". The human machine will also be a humane machine, capable of providing consistent interaction with the multimodal stimuli which have been received and processed. The human machine will present a photo-realistic virtual face, a voice, a body and gestures. It will use its artificial senses to understand high level messages coming from the interacting human and as well as all its virtual actuators to provide acoustic-visual feedback, like facial expressions [1] for instance in order to convey information, through avatars or autonomous agents. Possible applications domains are tourism, cultural heritage, eCommerce, technology-enhanced learning, multimedia production, handicap, cognitive robots, vehicle on-board safety systems, etc.

As well as being used to recognize people, the face forms a source of many informative social signals [2]. Lip-reading helps us to understand speech. Gaze patterns can direct the attention and regulate the conversation as well as head gestures like nodding or frowning. Facial expressions do communicate information about other person feelings or mental states. Psychological studies have shown reasonably good recognition of a small number of basic emotional categories in nearly all cultures. These emotions include joy, sadness, anger, disgust, fear and surprise [3]. 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 [4, 5, 6, 7]. For instance, the working group on Synthetic-Natural Hybrid Coding (SNHC) of the MPEG-4 standard has retained these six expressions that can be specified by themselves or in combination with other face animation parameters (FAPs) to control artificial 3D head models.

In this paper, we describe the use of active facial appearance models for face analysis [8]. Such image-based parametric models are used to analyse and generate video-realistic faces, and differ from the physical 3D face models frequently used today in machine to man interaction systems. We will describe how such generative models, in which facial appearance classes are learned from examples, can be used for face tracking, facial expression synthesis and recognition.

2. Active facial appearance models

2.1. Model description

It has been shown that the active appearance model [8] 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. For each image of the training set, the shape is represented by a vector of manually selected landmark points coordinates and the texture is represented by a vector of grey level values inside the convex enveloppe of the landmark points. After having computed the mean shape $\bar{\mathbf{s}}$ and aligned all shapes of the training set on $\bar{\mathbf{s}}$ by means of similitude transforms, 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 synthesized 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, 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.

This work is supported by the French Incentive Concerted Action for Young Researchers (*ACI Jeunes, Ministère de la Recherche*) and the european Interface project (FP5 - IST) for funding.

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:

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \quad (3)$$

$$\mathbf{g}_i = \bar{\mathbf{g}} + Q_t \mathbf{c}_i \quad (4)$$

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.

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) and (4). 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 [9], 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. For this purpose, a set of training residual images are computed by displacing the appearance and pose parameters whithin allowable limits. These residuals are then used to compute matrices R_a and R_t establishing the linear relationships $\delta(\mathbf{c}) = -R_a\ \mathbf{r}(\mathbf{c}, \mathbf{p})$ and $\delta(\mathbf{p}) = -R_t\ \mathbf{r}(\mathbf{c}, \mathbf{p})$ between the parameter displacements and the corresponding residuals, so as to minimize $|\mathbf{r}((\mathbf{c}, \mathbf{p}) + \delta(\mathbf{c}, \mathbf{p}))|^2$. An iterative model refinement procedure [9] is then used to drive the appearance model towards the actual face in the image.

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 [10]. Each sequence of this database contains ten to twenty images, beginning with a neutral expression and ending with a high magnitude expression.

a b c

Figure 1: a: Target (original face). b: Model initialization. c: Iterative model refinement until convergence to the target face.

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 choosen manually and extracted from the video sequence.

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 shape-free texture vector $\mathbf{g}$ is composed of 3493 pixels. The active appearance search algorithm was tested on a previously unseen face with the mean shape and texture used as a first approximation (figure 1.b). The model output converges to the target face as shown in figure 1. Three other adaptations of a higher resolution appearance model computed on the same learning set are shown on figure 2.

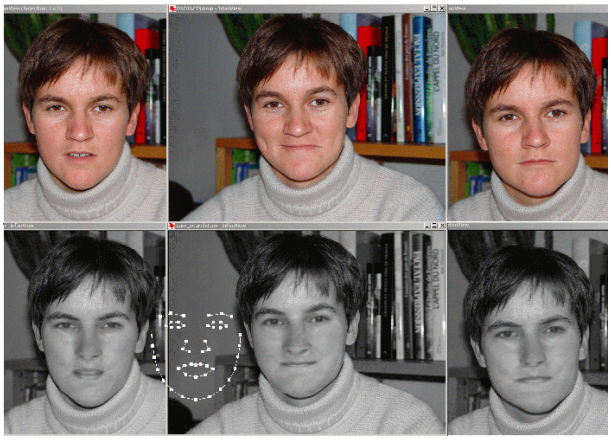


Figure 2: Upper row : original unseen faces. Lower row: reconstructed synthetic faces inserted onto the original ones (the shape of the facial mask represented by the model is illustrated by the set of white points).

3. Facial expression analysis and synthesis

3.1. Facial expression modeling

The aim of this section is to study a linear model, as it is proposed in [11, 12], correlating the appearance parameters to facial expression intensity according to:

$$\mathbf{c} = \mathbf{a}_{\mathbf{e}0} + \mathbf{a}_{\mathbf{e}1}\mathcal{I} + \boldsymbol{\varepsilon} \quad (5)$$

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 learned 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 $\mathbf{e}$:

$$\mathbf{c}_{\mathbf{e}}(\mathcal{I}) = \mathbf{a}_{\mathbf{e}0} + \mathbf{a}_{\mathbf{e}1}\mathcal{I} \quad (6)$$

Note that $\mathbf{c}_{\mathbf{e}}(\mathcal{I})$ is the same for all faces showing the same intensity of a given expression, and contains the information relative

to the person's expression. Hence, it is possible to modify the intensity of the synthesized expression by modifying the value of $\mathcal{I}$.

Given an unseen face with a determined expression (fig. 3a), it is possible to estimate the appearance parameter vector $\mathbf{c}_{op}$ that will synthesize an artificial face similar to this target face as described at the end of 2.1 (fig. 3b).

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 (6):

$$\mathcal{I}_{est} = \mathbf{a}_{\mathbf{e}1}^{+}(\mathbf{c}_{op} - \mathbf{a}_{\mathbf{e}0}) \tag{7}$$

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 $\mathbf{c}_{\mathbf{e}}(\mathcal{I}_{est})$, the latter being evaluated at the estimated expression intensity $\mathcal{I}_{est}$ using equation (6). This gives the identity vector $\mathbf{c}_{res}$:

$$\mathbf{c}_{res} = \mathbf{c}_{op} - \mathbf{c}_{\mathbf{e}}(\mathcal{I}_{est}) \tag{8}$$

Having this, it is possible to modify the facial expression intensity represented in vector $\mathbf{c}_{\mathbf{e}}(\mathcal{I}_{est})$ by modifying the $\mathcal{I}$ value in equation (6). In particular it is possible to filter out the expression by setting $\mathcal{I} = 0$.

$$\mathbf{c}_{\mathbf{e}}(0) = \mathbf{a}_{\mathbf{e}0} + \mathbf{a}_{\mathbf{e}1} \times 0 = \mathbf{a}_{\mathbf{e}0} \tag{9}$$

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 figure 3c:

$$\mathbf{c}_{neutral} = \mathbf{c}_{\mathbf{e}}(0) + \mathbf{c}_{res} \tag{10}$$

a b c

Figure 3: 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.

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 (5) 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.

$$\mathcal{I}'_{est} = \mathbf{a}_{\mathbf{e}'1}^{+}(\mathbf{c}_{neutral} - \mathbf{a}_{\mathbf{e}'0}) \tag{11}$$

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

$$\mathbf{c}_{\mathbf{e}'}(\mathcal{I}'_{est}) = \mathbf{a}_{\mathbf{e}'0} + \mathbf{a}_{\mathbf{e}'1}\mathcal{I}'_{est} \tag{12}$$

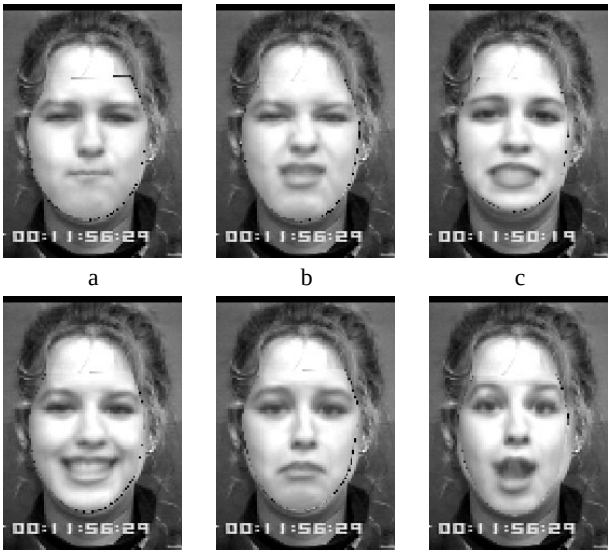
The new residual $\mathbf{c}_{res}$ is then given by:

$$\mathbf{c}_{res} = \mathbf{c}_{neutral} - \mathbf{c}_{\mathbf{e}'}(\mathcal{I}'_{est}) \tag{13}$$

The facial expression intensity represented in vector $\mathbf{c}_{\mathbf{e}'}(\mathcal{I}'_{est})$ can be controlled through the parameter $\mathcal{I}$ in equation (6). 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 $\mathbf{c}_{\mathbf{e}'}(\mathcal{I}'_{est})$ it is possible to modify the expression intensity shown on the target face from neutral to high magnitude as shown in figure 4.

$$\mathbf{c}_{intense} = \mathbf{c}_{\mathbf{e}'}(1) + \mathbf{c}_{res} \tag{14}$$



a b c

d e f

Figure 4: Generation of six synthetic expressions starting from the expression filtered face of figure 3.c. a: Anger. b: Disgust. c: Fear. d: Joy. e: Surprise. f: Sadness

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. The active appearance model is fit on each image of a video sequence using the previous model output in terms of appearance and pose parameters as a first approximation, as shown on figure 5. 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 (7), as well as the linearly predicted vector of appearance parameters $\mathbf{c}_{\mathbf{e}}(\mathcal{I}_{est})$ at this intensity.

Figure 5: Upper row : original surprised face. Lower row: synthetic expressive face inserted onto the original one. The appearance model may be used to track the face expression in a video.

For each facial expression, the $\mathbf{c_e}(\mathcal{I}_{est})$ 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 figure 6 showing the evolution of the first variation mode (first coefficient of $\mathbf{c_e}(\mathcal{I}_{est})$) over 15 video sequences showing a facial expression evolution from neutral to high magnitude joy (a) and disgust (b) of different faces.

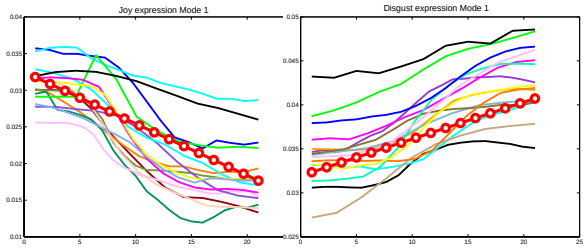


Figure 6: For joy and disgust, evolution of 1st mode of $\mathbf{c_e}(\mathcal{I}_{est})$ over each video sequence. Straight line: 1st mode of $\mathbf{c_e}(\mathcal{I})$, with $\mathcal{I}$ linearly varying from 0 to 1.

3.5. Facial expression recognition

To classify a new face represented by the parameter vector $\mathbf{c}_i$, we use a Linear Discriminant Analysis scheme, assuming that the expression classes have a common covariance matrix. We measure the squared Mahalanobis distance $d_M(\mathbf{c}_i, \bar{\mathbf{c}}_j) = (\mathbf{c}_i - \bar{\mathbf{c}}_j)^t \Sigma^{-1} (\mathbf{c}_i - \bar{\mathbf{c}}_j)$ from $\mathbf{c}_i$ to each of the j estimated mean vectors $\bar{\mathbf{c}}_j$, and assign $\mathbf{c}_i$ to the class of the nearest mean. $j \in$ [joy, fear, disgust, surprise, fear, sadness, neutral] and Σ is the estimated common covariance matrix of the training set.

4. Conclusion and perspectives

In this paper, we have presented different applications of Active Appearance Models for face analysis and synthesis. Such tool can 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.

	neut.	ang.	disg.	fea.	joy	surp.	sad.
neutral	38	1	1	4	0	0	8
anger	1	10	0	0	0	0	1
disgust	0	0	12	0	0	0	0
fear	0	0	1	8	0	0	0
joy	0	0	0	2	11	0	0
surprise	1	0	0	0	0	13	1
sadness	3	0	0	0	0	0	14

Table 1: Confusion matrix for the expression classifier, using 130 unknown test images. Globally, 81.5% of the images were correctly classified.

5. References

[1] P. Ekman, "Should we call it expression or communication?," *Innovations in Social Science Research*, vol. 10, no. 4, pp. 333–344, 1997.

[2] V. Bruce and A. Young, *In the eye of the beholder. The science of face perception*, University Press, Oxford, 1998.

[3] P. Ekman, *Facial Expressions*, chapter 16 of Handbook of Cognition and Emotion, T. Dalgleish and M. Power, John Wiley & Sons Ltd., 1999.

[4] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–988, October 1999.

[5] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, December 1999.

[6] 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.

[7] 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, December 2000.

[8] 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.

[9] 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.

[10] 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.

[11] T.F. Cootes, K. Walker, and C.J. Taylor, "View-based active appearance models," in *International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000, pp. 227–232.

[12] 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.