FACE APPEARANCE FACTORIZATION FOR EXPRESSION ANALYSIS AND SYNTHESIS

Bouchra Abboud, Franck Davoine

Heudiasyc Laboratory, CNRS, University of Technology of Compiègne.
BP 20529, 60205 COMPIEGNE Cedex, FRANCE.
E-mail: Bouchra.Abboud@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 factorization technique which decomposes the appearance parameters coding a natural image. This technique is then used to perform facial expression synthesis on unseen faces showing any undetermined facial expression, as well as facial expression recognition.

1. INTRODUCTION

Natural human-machine interaction is becoming an active and important research area. Adequate feedback 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. Furthermore, the human face constitutes a source of informative social signs which allow good communication expectation response.

Face configuration is mainly influenced by the interaction of two inherent major factors: identity and facial expression. Facial expression recognition works showed [1] that six basic emotional categories are universally recognized in nearly all cultures, 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.

This paper addresses the issue of appearance based face representation and appearance factorization for facial expression synthesis and recognition.

2. ACTIVE FACIAL APPEARANCE MODELS

We choose to represent faces using the active appearance model (AAM) [2] which is a powerful tool allowing to extract from any unknown target face, a set of appearance parameters coding a synthetic face similar to the target in terms of minimum texture error. AAM uses Principal Component Analysis to model both shape and texture variations seen in a training set according to:

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \quad \text{and} \quad \mathbf{g}_i = \bar{\mathbf{g}} + Q_t \mathbf{c}_i, \tag{1}$$

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 controlling the synthesized shape $\mathbf{s}_i$ and texture $\mathbf{g}_i$. $\bar{\mathbf{s}}$ and $\bar{\mathbf{g}}$ are the mean shape and texture computed on the aligned and normalized training faces.

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.

The active appearance model can automatically adjust parameters $\mathbf{c}$ and $\mathbf{p}$ to a target face 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] and returns parameters $\mathbf{c}_{op}$ and $\mathbf{p}_{op}$.

The appearance model is constructed 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 select 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 is chosen manually by extracting an intermediate frame from the video sequence. 37 other neutral faces are also added. The standard model is built using 50 shape modes, 170 texture modes and 120 appearance modes thus retaining 98 percent of the combined shape and texture variation. The shape-free texture vector $\mathbf{g_i}$ is composed of 5871 pixels and the shape vector $\mathbf{s_i}$ dimension is 106.

3. APPEARANCE FACTORIZATION

The AAM face representation described above allows automatic extraction of a set of appearance parameters from any unknown target face. The extracted parameters control simultaneously the reconstructed face shape and texture which contain information about the reconstructed face identity and facial expression. Hence, this representation might be used for facial expression recognition using classification of AAM parameters; and facial expression synthesis using direct control of AAM parameters through linear modelling [4].

However such an approach suffers from a major drawback. Indeed, *a priori* knowledge of the facial expression shown on a target face is required in order to perform new facial expression synthesis while keeping the target identity intact.

We wish to introduce a more general representation which allows to extract from any appearance vector a subset of parameters controlling exclusively facial expression independently of identity and without *a priori* knowledge of either expression or identity. Such a model would allow immediate expression synthesis on any unknown target face by replacing the extracted facial expression parameters with the parameters corresponding to the desired expression we wish to synthesize.

Similarly, extraction of the parameter subset exclusively controlling facial expression is expected to boost facial expression

recognition performance.

In this perspective, we choose to model the mapping from expression and identity parameters to natural faces using a bilinear factorization model.

Bilinear models are two-factor models with the property that their outputs are linear in either factors when the other is held constant. They provide rich factor interactions by allowing factors to modulate each other's contributions multiplicatively.

Tenenbaum and Freeman [5] model the interaction between face illumination and pose using bilinear models in order to perform face synthesis under novel illuminations as well as face pose recognition. Similarly Chuang *et al.* [6] use bilinear models to separate video data into expressive features and underlying content in order to perform facial expression synthesis on speaking faces. In a more general approach Vasilescu and Terzopoulos [7] propose multilinear analysis of faces to separate factors such as identity, viewpoint, illumination and expression from pixel grey level values. This representation is then used to perform face recognition in previously unseen viewpoint or under unknown illumination. Similarly, Wang and Ahuja [8] use multilinear modelling based on Higher Order Singular Value Decomposition (HOSVD) in conjunction with an automatic face appearance representation technique, in order to separate identity and facial expression from appearance parameters. This is achieved by decomposing a 3D observation tensor into a core tensor, a person subspace matrix, an expression subspace matrix and a facial feature subspace matrix. This decomposition is then used to perform facial expression synthesis and recognition of the seven basic facial expressions as well as face recognition. However, the proposed decomposition requires *a priori* knowledge of the facial expression shown on a target face in order to perform facial expression synthesis.

In this paper, we propose a more general bilinear factorization model which allows to separate identity and expression factors on any unknown target face showing an undetermined expression in order to perform facial expression synthesis and recognition. Results are compared to previous ones obtained by direct classification of AAM parameters.

4. BILINEAR MODEL LEARNING

Two types of bilinear models are described in this section, namely the symmetric bilinear model and the asymmetric bilinear model. The general symmetric model allows to represent the interaction between style and content factors for a given observation, whereas the simpler asymmetric model is style specific and requires one factor to be known in advance. Detailed model construction for both configurations is addressed below.

4.1. The bilinear symmetric model

A bilinear symmetric model represents the interaction between style $\mathbf{a^s}$ and content $\mathbf{b^c}$ factors for a given observation $\mathbf{y^{sc}}$ according to:

$$y_k^{\mathbf{sc}} = \mathbf{a^{sT}} \mathbf{w_k} \mathbf{b^c}, \tag{2}$$

where $y_k^{\mathbf{sc}}$ represents the k^{th} component of $\mathbf{y^{sc}}$ and $\mathbf{w_k}$ is a style and content independent matrix characterizing their interaction.

For a training set of $S \times C$ observations with S different styles and C different contents, the observation matrix is obtained by stacking the $S \times C$ observation vectors $\mathbf{y^{sc}}$ style-wise. We obtain the vector transpose of such a stacked matrix by permuting the stacking procedure to content-wise as shown in equation (3).

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y^{11}} & \cdots & \mathbf{y^{1C}} \\ \vdots & \ddots & \vdots \\ \mathbf{y^{S1}} & \cdots & \mathbf{y^{SC}} \end{bmatrix}, \quad \mathbf{Y^{VT}} = \begin{bmatrix} \mathbf{y^{11}} & \cdots & \mathbf{y^{1S}} \\ \vdots & \ddots & \vdots \\ \mathbf{y^{C1}} & \cdots & \mathbf{y^{CS}} \end{bmatrix} \tag{3}$$

The symmetric model can then be written in a compact matrix form:

$$\mathbf{Y} = \left[\mathbf{W^{VT}A}\right]^{\mathbf{VT}} \mathbf{B}, \tag{4}$$

where **A** and **B** represent the stacked style and content parameter matrices,

$$\mathbf{A} = \left[\mathbf{a^1} \ldots \mathbf{a^S}\right], \qquad \mathbf{B} = \left[\mathbf{b^1} \ldots \mathbf{b^C}\right], \tag{5}$$

and **W** is the stacked interaction weights matrix.

Training a bilinear symmetric model consists in learning style, content, and weight matrices **A**, **B** and **W** which minimize the total squared error over a training set between the actual and the reconstructed observations.

Let I and J be the dimensions of the style and content vectors. I is chosen to be equal to the number of styles shown in the training set ($I = S$) whereas it is recommended to choose J by looking for an elbow in the singular value spectrum of **Y** [5].

Least squares optimal values of **A** and **B** are iteratively estimated using singular value decomposition as follows:

1. $\mathbf{Y} = \mathbf{USV^T}$ and initialize **B** as the first J rows of $\mathbf{V^T}$.
2. $[\mathbf{YB^T}]^{\mathbf{VT}} = \mathbf{USV^T}$ and set **A** to be the first I rows of $\mathbf{V^T}$.
3. $[\mathbf{Y^{VT}A^T}]^{\mathbf{VT}} = \mathbf{USV^T}$ and update **B** to be the first J rows of $\mathbf{V^T}$.
4. Repeat until **A** and **B** are stable.

Upon convergence W is given by:

$$\mathbf{W} = \left[[\mathbf{YB^T}]^{\mathbf{VT}}\mathbf{A^T}\right]^{\mathbf{VT}}. \tag{6}$$

4.2. The bilinear asymmetric model

While the symmetric bilinear model decomposes an observation into a style component $\mathbf{a^s}$, a content component $\mathbf{b^c}$ and a style and content independent interaction component $\mathbf{w_k}$, the style specific asymmetric bilinear model decomposes observations into a content component $\mathbf{b^c}$ and a style specific linear mapping $\mathbf{W^s}$. For a given observation $\mathbf{y^{sc}}$ with a known style $''s''$, the bilinear asymmetric model is given by:

$$\mathbf{y^{sc}} = \mathbf{W^s b^c}. \tag{7}$$

Training the bilinear asymmetric model consists in learning the stacked content and weight matrices **B** and **W** ($\mathbf{Y} = \mathbf{WB}$) which minimize the total squared error between the actual and the reconstructed observations of a training set. The least square optimal factors are obtained by singular value decomposition of the training matrix: $\mathbf{Y} = \mathbf{USV^T}$. Then **W** is given by the first J column of **US** and **B** is given by the first J rows of $\mathbf{V^T}$.

4.3. Experimental setup

To build the bilinear model we extract from the CMU database [3] a training set containing 70 frontal face images of 10 different persons (contents) showing each of the seven basic facial expressions (styles). The observation matrix is built by stacking the appearance parameter vectors coding the training faces. Each column of the observation matrix $\mathbf{Y}$ contains the AAM appearance vectors of a specific person with different expressions whereas each row contains the appearance vectors of all the persons showing a specific expression.

We set the dimensionality of expression vectors to be equal to the number of expressions in the training set I=7 to allow maximum expressiveness, and the dimensionality of the identity vectors to be J=10 which corresponds to the maximum number of training identities and construct the bilinear symmetric and asymmetric models.

5. BILINEAR MODEL FITTING

Bilinear symmetric model fitting to an unknown target face consists in extracting from the optimal appearance parameters $\mathbf{c}_{op}$ (obtained by AAM search and coding this face), a subset of parameters exclusively coding expression $\mathbf{a^e}$ and a subset of parameters exclusively coding identity $\mathbf{b^i}$. The target face is of undetermined identity and undetermined known or unknown expression.

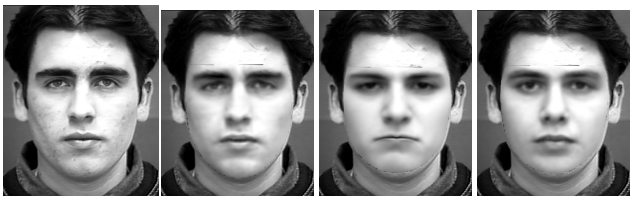
Adapting the symmetric bilinear model to a target face with undetermined identity and expression is an iterative procedure detailed below where $\mathbf{X}^+$ indicates the pseudo-inverse of $\mathbf{X}$:

1. Initialize $\mathbf{b^i}$ as the mean content (identity) vector of the training set
2. $\mathbf{a^e} = \left[[\mathbf{W}\mathbf{b^i}]^{\mathbf{VT}}\right]^+ \mathbf{c_{op}}$
3. $\mathbf{b^i} = \left[[\mathbf{W}^{\mathbf{VT}}\mathbf{a^e}]^{\mathbf{VT}}\right]^+ \mathbf{c_{op}}$
4. Repeat until $\mathbf{a^e}$ and $\mathbf{b^i}$ are stable.

The symmetric bilinear model fitting (eq. (4)) on a target face with undetermined identity and expression is shown in figure (1.c).

On the other hand, if the facial expression shown on the target face is *a priori* known, the extraction of the identity parameters $\mathbf{b}^i$ is immediate *via* simple least square fitting of the expression specific asymmetric bilinear model:

$$\mathbf{b^i} = [\mathbf{W^e}]^+ \mathbf{c_{op}} \tag{8}$$



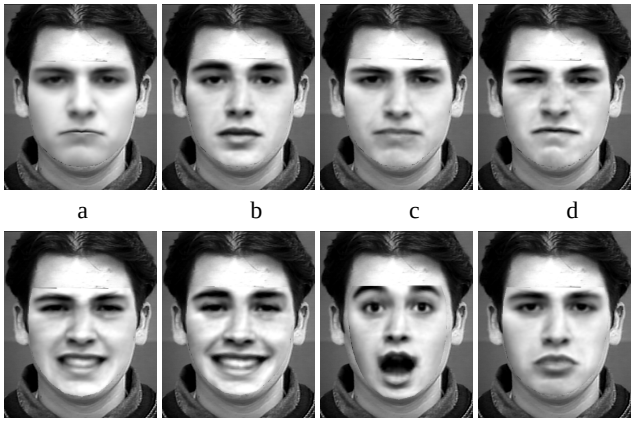
a b c d

Fig. 1. a: Unknown target face. b: AAM fitting. c: Symmetric bilinear model fitting d: Asymmetric bilinear model fitting. The difference in image quality between both bilinear models is discussed in section (6).

The asymmetric bilinear model fitting (eq. (7)) on a target face with undetermined identity and determined (neutral) expression is shown in figure (1.d).

6. FACIAL EXPRESSION SYNTHESIS

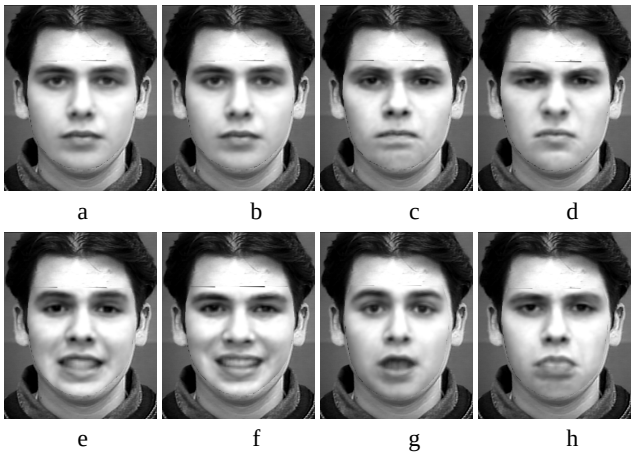
To perform facial expression synthesis on a unknown target face, with an undetermined identity and expression, represented by a set of appearance parameters $\mathbf{c}_{op}$, the bilinear symmetric model is first adapted to the target face. The corresponding expression $\mathbf{a^e}$ and identity $\mathbf{b^i}$ factors are then extracted according to the procedure described in section (5).



a b c d

e f g h

Fig. 2. Expression synthesis using the symmetric bilinear model. a: Symmetric bilinear model fitting. b: Neutral. c: Anger. d: Disgust. e: Fear. f: Joy. g: Surprise. h: Sadness.



a b c d

e f g h

Fig. 3. Expression synthesis using the asymmetric bilinear model. a: Asymmetric bilinear model fitting b: Neutral. c: Anger. d: Disgust. e: Fear. f: Joy. g: Surprise. h: Sadness.

To synthesize any novel expression e' while keeping identity intact an artificial appearance parameter is built by combining the extracted identity factor $\mathbf{b^i}$ with the desired expression factor learned from the training set $\mathbf{a^{e'}}$:

$$\mathbf{c_{synth}} = \left[\mathbf{W^{VT}a^{e'}}\right]^{\mathbf{VT}} \mathbf{b^i}. \quad (9)$$

Facial expression synthesis on the unknown target face of figure (1.a) without prior knowledge of the facial expression is shown on figure (2). Note that the symmetric model fitting (fig. (2.a)) shows a facial expression that differs from the actual target expression. This is expected since the expression is supposed to be initially undetermined and the similarity between the model output and target face increases when the correct target expression (neutral) is synthesized as shown on figure (2.b).

However, if the expression shown on the target face is determined, the asymmetric expression specific bilinear model is adjusted to the target face and the identity factor $\mathbf{b^i}$ is extracted. The artificial appearance parameter is then built by combining the extracted identity factor with the desired expression specific weights matrix ($\mathbf{c_{synth}} = \mathbf{W^{e'}b^i}$). Facial expression synthesis using the asymmetric model on the unknown neutral face of figure (1.a) is shown on figure (3).

7. FACIAL EXPRESSION RECOGNITION

An identity specific asymmetric bilinear model, with expression considered as content, is constructed using the same training set described in section (4.3) giving the identity specific weights matrices $\mathbf{W^i}$ and expression factors $\mathbf{B}$. To perform facial expression recognition, we use a test set of 112 unknown face image showing each of the seven basic facial expressions and run AAM optimization to extract the corresponding appearance parameters. The identity specific asymmetric model is then adjusted to the tested face assuming that its identity corresponds to the first training identity and the corresponding expression factor $\mathbf{b^e}$ is extracted. The Euclidian distance is then computed between the extracted expression factor and each of the training expression factors of matrix $\mathbf{B}$. The class yielding the minimum distance is selected. This experience is repeated for all the 10 training identities and finally the expression is attributed to the class with the maximum number of votes yielding a correct recognition rate of 82.14%.

The confusion matrix for asymmetric bilinear model based facial expression recognition is given in table (1).

	neut.	ang.	disg.	fea.	joy	surp.	sad.
neut.	31	1	2	0	0	2	5
ang.	1	8	1	0	0	0	0
disg.	1	0	5	0	0	0	1
fear	1	0	0	8	1	0	0
joy	0	0	0	2	17	0	0
surp.	1	0	0	0	0	14	0
sad.	1	0	0	0	0	0	9

Table 1. Confusion matrix for the asymmetric identity-specific factorization based expression classifier. The correct recognition rate for 112 unknown test images is 82.14%.

For comparison, we perform facial expression recognition by direct measurement of the Euclidian distance between the appearance parameters coding the tested face and the mean of the appearance parameters coding the training faces with a specific expression. The tested expression is attributed to the class with the nearest mean yielding a correct recognition rate of 75%. As expected, appearance factorization which allows the extraction of an identity-independent, expression-specific factor from any unknown target face, boosts facial expression recognition.

8. CONCLUSION AND PERSPECTIVES

We presented a technique for facial expression synthesis and recognition using a bilinear factorization method which allows to separate a vector of appearance parameters coding a face into an identity factor exclusively coding face identity and an expression factor exclusively coding facial expression.

The advantages of such a model compared to standard linear regression techniques described in [4] for facial expression synthesis, lies in the fact that no knowledge of the facial expression or face identity of a target face is *a priori* required. However, it is not possible to control the intensity of the synthesized facial expression which is essential to perform video synthesis. A possible solution could be the use of a multi-linear factorization model mapping face identity, facial expression, and facial expression intensity to appearance parameters.

A further extension of this work could be multi-linear factorization of faces allowing to extract illumination factors thus robustifying face and facial expression recognition and synthesis to miscellaneous illumination variations.

9. REFERENCES

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

[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] B. Abboud, F. Davoine, and M Dang, "Expressive face recognition and synthesis," in *IEEE workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, Madison, U.S.A., June 2003.

[5] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, pp. 1247–1283, 2000.

[6] E.S. Chuang, H. Deshpande, and C. Bregler, "Facial expression space learning," in *The Tenth Pacific Conference on Computer Graphics and Applications*, October 2002.

[7] M. A. Vasilescu and D. Terzopoulos, "Multilinear image analysis for facial recognition," in *International Conference on Pattern Recognition (ICPR'02)*, Quebec City, Canada, August 2002.

[8] H. Wang and N. Ahuja, "Facial expression decomposition," in *Proc. 9th Intern. Conf. on Computer Vision, ICCV'03*, Nice, France, September 2003.