Video-based facial animation with detailed appearance texture*

WANG Jin[1,2**], SUN Hanqiu[3] and PENG Qunsheng[2]

(1. Department of Computer Science & Technology, Xuzhou Normal University, Xuzhou 221009, China; 2. State Key Laboratory of CAD & CG, Zhejiang University, Zhejiang 310027, China; 3. Department of Computer Science & Engineering, CUHK Shatin, N.T., Hong Kong)

Received May 16, 2005; revised September 13, 2005

Abstract Facial shape transformation described by facial animation parameters (FAPs) involves the dynamic movement or deformation of eyes, brows, mouth, and lips, while detailed facial appearance concerns the facial textures such as creases, wrinkles, etc. Video-based facial animation exhibits not only facial shape transformation but also detailed appearance updates. In this paper, a novel algorithm for effectively extracting FAPs from video is proposed. Our system adopts the ICA-enforced direct appearance model (DAM) to track faces from video sequences; and then, FAPs are extracted from every frame of the video based on an extended model of Wincandidate 3.1. Facial appearance details are transformed from each frame by mapping an expression ratio image to the original image. We adopt wavelet to synthesize expressive details by combining the low-frequency signals of the original face and high-frequency signals of the expressive face from each frame of the video. Experimental results show that our proposed algorithm is suitable for reproducing realistic, expressive facial animations.

Keywords: image based rendering, facial animation, texture transform.

In graphics, image processing and computer vision, performance driven facial animation has been an interesting research topic, and has been widely used in entertainment industry, virtual human, artificial life, on-line gaming, teleconference, wireless presence, and so on. Research work in this field was focused on three main parts: face detection and tracking, FAPs (facial animation parameters) extraction, texture mapping and driving 3D model.

Early efforts in face detection and tracking can be dated back to the 1970s[1]. Over the past decades there were lots of research works addressing several important aspects of face modeling and animation[2,3]. The difficulties in reproducing life-like characters in facial animation brought the performance driven approach, in which tracked human actors control the animation. No wonder that accurately tracking facial feature points or edges is important to maintain consistent and life-like quality of animation. Techniques used for tracking face can be generally classified into feature-based tracking and statistical training-based approaches. Representative techniques of the former include snakes[4] and optical flow. Snakes are widely used to track intentionally marked facial features. Many systems apply tracked snakes coupled with underlying muscle mechanisms to drive facial animation[5,6]. Muscle contraction parameters are estimated according to the tracked facial feature displacements in video sequences. Because the tracking error accumulates over image sequences, the method may lose the tracking contour. Colored markers can aid in tracking facial expression. However, marks on the face are in general intrusive and impractical. On the other hand, optical flow[7] performs natural feature tracking and therefore avoids the need for intentional marks on the face[8,9]. This approach also has limited applications because the complete and accurate knowledge of face is not used.

Statistical training-based approaches employ a training procedure that converts the image space to a multi-dimensional feature space. Techniques used to derive this feature space include principal component analysis (PCA) and factor analysis (FA). Active appearance model (AAM)[10—12] is a typical example of statistical training-based approaches. The constraints adopted for feature space definition can be classified into three types: (1) structure information constraints, (2) color information constraints, (3) texture information. And statistical training-based approaches make full use of the above three types of

* Supported by the Major State Basic Research Development Program of China (Grant No. 2002 CB312101) and National Natural Science Foundation of China (Grant No. 60403038)

** To whom correspondence should be addressed. E-mail: jwang@cad.zju.edu.cn

constraints. Based on the tracked 2D feature motions, FAPs can be extracted and used to directly drive facial animation[13,14].

Our work presented in this paper addresses the following three aspects: (1) Video-based face tracking: taking into account the complex correlation between components in video tracking, we adopt ICA (independent component analysis) to deal with components statistically independent and the precision is increased by about 7%. (2) FAPs extraction based on the 2D tracked results: we propose a novel algorithm to effectively and robustly extract FAPs. (3) Facial expressive mapping: we adopt a method to transform facial appearance details from the source character (e.g. excessive wrinkles, lentigines, fleck in the video), while preserving the illumination effect of the original image.

1 Face detection and tracking

The concept of AAM was firstly introduced by Coots et al.[10], and since then AAM has attracted a lot of attentions[15]. AAM is a powerful model for face alignment, recognition and synthesis[16,17]. It adopts the subspace analysis techniques to model both shape variation and texture variation, and sets up the correlations between them. However, the analysis conducted by Hou et al.[18] on mutual dependencies of shape, texture and appearance parameters in the AAM feature space models shows that there exist some admissible appearances that cannot be modeled and hence cannot be investigated by AAM. Thus, they presented a direct appearance model (DAM) to cope with this problem, which employs the texture information to predict the shape and update the estimates of position and appearance. DAM improves the convergence and accuracy significantly.

Our algorithm is based on DAM, but enforces ICA on DAM to resolve the correlation between components, which frequently causes problems for precision. The ICA embedded in DAM can also reduce the dimensions of feature space and thus increase the tracking speed.

1.1 Active appearance model (AAM)

Assume that a training set is given as $W = \{(S_0, T_0)\}$, where shape $S_0 = ((x_1, y_1), \cdots, (x_k, y_k)) \in \mathbb{R}^{2K}$ is a sequence of K points on the 2D image plane, and the texture T_0 is the patch of image pixels enclosed by S_0. Let $\bar{s}$ be the mean shape which is modeled by k principal modes learned from the training shapes using PCA. Let $\bar{t}$ be the mean texture which is obtained after the shapes are aligned to the tangent space of $\bar{s}$. An instance of shape can then be represented as

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{\Phi}_s \boldsymbol{b}_s, \tag{1}$$

where $\boldsymbol{b}_s$ is a vector in the low dimensional shape subspace; $\boldsymbol{\Phi}_s$ is the matrix consisting of k principal orthogonal modes of variation in $\{S_0\}$, obtained from the training set.

Similarly, the PCA texture model can be expressed as $\boldsymbol{t} = \bar{\boldsymbol{t}} + \boldsymbol{\Phi}_t \boldsymbol{b}_t$; the appearance of each example is a concatenated vector $\boldsymbol{b}_{st}$

$$\boldsymbol{b}_{st} = \begin{pmatrix} w_s \boldsymbol{b}_s \\ \boldsymbol{b}_t \end{pmatrix} = \begin{bmatrix} w_s \boldsymbol{\Phi}_s^{\mathrm{T}} (\boldsymbol{s} - \bar{\boldsymbol{s}}) \\ \boldsymbol{\Phi}_t^{\mathrm{T}} (\boldsymbol{t} - \bar{\boldsymbol{t}}) \end{bmatrix}. \tag{2}$$

Then,

$$\boldsymbol{b} = \bar{\boldsymbol{b}} + \boldsymbol{\Phi}_{st} \boldsymbol{b}_{st}, \tag{3}$$

where $\boldsymbol{\Phi}_{st}$ is the matrix consisting of principal orthogonal vectors of the variation in $\{\boldsymbol{b}\}$ for all training samples$\}$. The object instance, $(\boldsymbol{b}, \boldsymbol{t})$, is synthesized by warping the pixel intensities of $\boldsymbol{t}$ onto the geometry of shape $\boldsymbol{s}$. In AAM, the residual vectors between the model and image, $\delta \boldsymbol{t} = \boldsymbol{t}_{\text{model}} - \boldsymbol{t}_{\text{image}}$, are regressed against the known displacement vectors, $\delta \boldsymbol{b}_{st}$, using the principal components regression: $\delta \boldsymbol{b}_{st} = \boldsymbol{\Phi}_{st} \delta \boldsymbol{t}$. Embedded into an iterative updating scheme, this has been proven to be a very efficient way of matching these models with novel images.

1.2 Direct appearance model (DAM)

Because mapping from the texture space to the shape space is many-to-one, the shape parameters should be determined completely by texture parameters but not *vice versa*. DAM consists of a shape model, a texture model and a prediction model. Unlike AAM's crucial idea of combining shape and texture information into an appearance model, DAM predicts the shape parameters with the texture parameters. Based on an assumption concerning a linear relationship between shape and texture, the prediction function is given by

$$\boldsymbol{s} = \boldsymbol{R}\boldsymbol{t} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{\varepsilon}$ is the error vector and $\boldsymbol{R}$ is a projection matrix, and its objective cost function is defined as

$$C(\boldsymbol{R}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \mathrm{tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}). \tag{5}$$

Then, its minimum cost solution is

$$\boldsymbol{R}^* = E(\boldsymbol{s}\boldsymbol{t}^{\mathrm{T}})[E(\boldsymbol{t}\boldsymbol{t}^{\mathrm{T}})]^{-1}. \tag{6}$$

During searching, we employ $\delta \boldsymbol{T}$'s principal compo-

nents, $\delta T'$, to predict the position displacement

$$\delta s = R_s \delta T', \quad (7)$$

where R_s is the prediction matrix learned from linear regression upon $\{\delta s, \delta T'\}$.

1.3 Independent component analysis (ICA)

PCA belongs to unsupervised learning algorithms which discover significant features in the input data without a teacher. If the source data conforms to Gaussian distribution, PCA can extract the feature space accurately. However, few samples adhere to the Gaussian distribution exactly. ICA[19] is an alternative approach which allows its components to be as statistically independent as possible. Since the relativity of components is a very sensitive factor for tracking face accurately, we apply ICA to the set of principle components driven from PCA to improve the independency with the same dimensions as PCA[20]. Experiments on static images showed that the error rate can be reduced about 7% by performing ICA on PCA (Fig. 1).

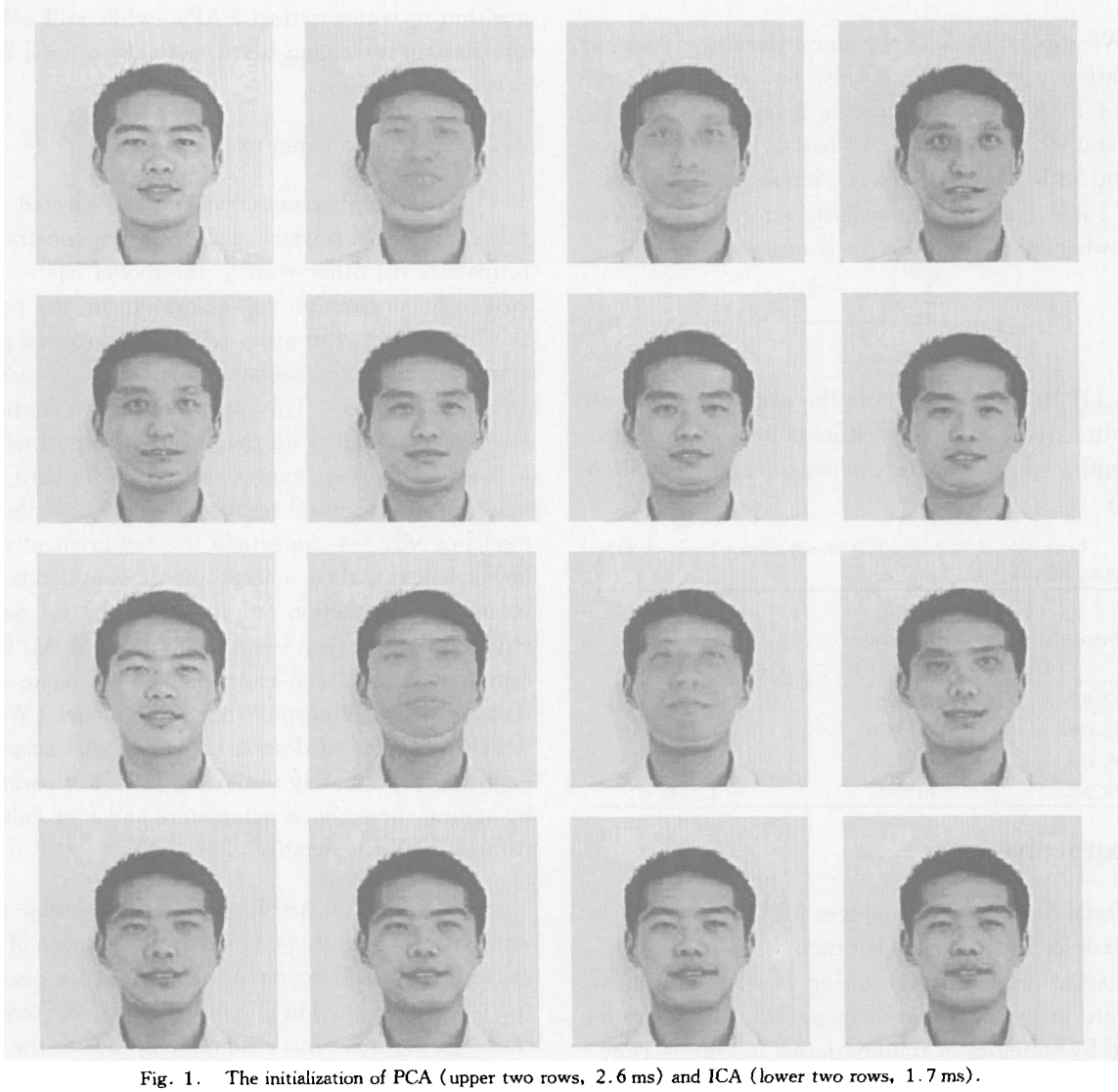

Fig. 1. The initialization of PCA (upper two rows, 2.6 ms) and ICA (lower two rows, 1.7 ms).

Assuming that the principal components are included in Eq. (3) after performing PCA, our approach adopts ICA to find an $m \times m$ matrix W as described by Eq. (8). Note that when U is a matrix of independent components, W is proper. In order to calculate W, we conduct an iterative procedure

$$\Delta W = \eta\left(I + \frac{1}{3m}(I - 2\psi(U))U^{\mathrm{T}}\right)W, \quad (8)$$

where η is a learning parameter, I is an identity matrix; a sub-variable of $Y = \psi(U)$ is defined as

$$y_{ij} = 1/(1 + e^{-u_{ij}}). \quad (9)$$

In our experiments, η is set to be 0.2, and the initial

value of every sub-variable of $\boldsymbol{W}$ is set to be a random value between 0.0 and 0.01.

For the experimental purpose, we chose grayscale face images with size of 320×240 as training samples and selected input component in grayscale with gradient of 5×5. There are 260 samples in our training dataset containing typical expressions and their rotations for up, down, left, and right within a small angle ($< 30°$). Appearance details contain mainly forehead wrinkles, brows wrinkles, jaw wrinkles, crow's-feet, and nose wing wrinkles.

We applied PCA to the above training dataset at four different significance levels, and ICA to the results of PCA. The significance levels 92%, 95%, 98% and 99.5% produced 14 basis, 19 basis, 26 basis, and 38 basis, respectively. We tested 50 face images. Table 1 shows the analysis error for PCA and ICA, where the analysis error is measured by

$$\varepsilon = \frac{\sum_m | l_m - \hat{l}_k |}{\sum_m \hat{l}_m}. \qquad (10)$$

In Eq.(10), l_m and $\hat{l}_k$ denote the length of the m-th edge after triangulation of feature points, $\hat{l}_k$ is the exact length, while l_m is the evaluated length with PCA or ICA.

Table 1. Error rate of PCA and ICA at significance levels of 92%, 95%, 98%, and 99.5%

Significance level	PCA Error samples/ All samples	ICA Error samples/ All samples
92.0%	28.33%	21.16%
95.0%	14.08%	8.55%
98.0%	11.87%	4.76%
99.5%	10.54%	2.13%

2 Extracting FAPs

Facial animation parameters (FAPs) are widely used to describe the shape deformation and the associated feature displacements during facial expressions. FAPs are in general derived using 2D feature points tracked by statistically training-based methods. However, this method is not robust and sometimes may produce singularity expressions (e.g. skew eyes, malformed mouth) in driving animation, for the important facial structure-constraint information is lost.

Instead of tracking 2D points to get FAPs to control the movement of the 3D wire-frame model, Ahlberg[14] suggested a robust technique for extracting FAPs, which applies feature space analysis to the FAPs of MEPG4. However, the correlation between components of FAPs is a sensitive factor affecting the performance of statistical training-based method. If the 3D parameterized model has complex correlation, the processing will be slow. It seems that the method is suitable for the simplified 3D parameterized model only. Unfortunately, with the simplified model, it is impossible to describe complex facial expressions accurately.

Next, we will propose a robust and efficient algorithm to better extract FAPs, while still adhere to the statistical training-based methods to track 2D features in real-time.

2.1 Selecting a parameterized model

The facial parameterized model selected should reflect the facial physiological structure constraint information. In other words, the model has to satisfy two basic constraints: a) the movement of a point on the face is related to many relevant expression parameters; b) an expression parameter can move many relative 3D points. This many-to-many relationship must be consistent with the physical transformation of a face during face expression. An awkward model may lead to unnatural results. We find that the facial model of MEPG4 can satisfy this requirement nicely. Nevertheless, when a large set of parameters is involved, its extraction from video is by no means a trivial task. We then select a sub-model of MPEG4 to represent major facial expressions. We make an extension to the Wincandidate3.1[21] model (WCM), which is capable of describing some facial shapes and expressions accurately with a simple set of parameters by adding three shape parameters and four animation parameters for generality.

According to AAM, after triangulation of the face with the feature points, the distribution of triangle centers should represent the five sense organs on the face properly. The 3D mesh of WCM, however, cannot finely meet this requirement, due to the sparse distribution of feature points in some regions. To improve the results of face detecting and tracking, we revise the original WCM model by inserting some supplemental points acting as the mediate points (see the blue points in Fig.2 (b)). We refer this extent model as IWCM. By applying IWCM to the 3D mesh face model, an original template can be created with ten paths (Fig.2). The ten paths are now all closed,

with the mouth being a hole. Our IWCM model has 84 parameters in total. To further improve the efficiency and robustness, we select only 19 parameters among them with which most of the typical facial expressions can be described, including upper lip raising, jaw drop, eyebrow lowering, lip corner depressing, outer brow raising, eye closing, and nose wrinkle.

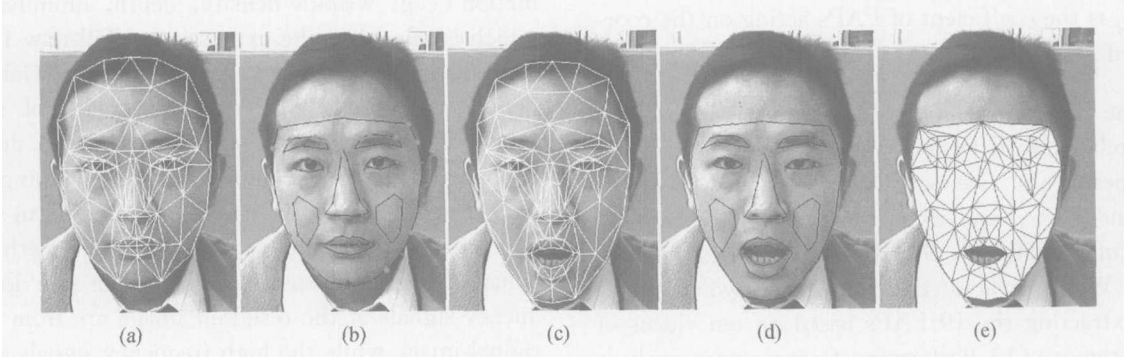

Fig. 2. Mesh of WCM contains feature points. (a) Original mesh of WCM; (b) generated shape with ten paths, blue points are the vertices added in IMCM; (c) mesh with open mouth expression; (d) the shape corresponding to (c); (e) the result after triangulation of feature points.

2.2 Extracting FAPs

Based on the multistage approach for camera calibration[22], we calculate separately the exterior and interior camera parameters during camera calibration. We first calibrate the following 5 interior parameters by using the general pin-hole camera model, including: (u, v), the coordinates of the principle point; (f_x, f_y), the scale factors of x and y axes in image space; n, the parameter describing the skew of the two image axes. The method is similar to Zhang[23], which makes use of a planar calibration pattern. To improve the tracking accuracy, we took the mean value of 20 experiments with 1000 frames each time. Our experiments used a desktop camera, and estimated the interior parameters only once regarding a stable circumstance.

We take the frontal neutral expression at the first frame. There are two approaches for estimating the exterior parameters: 1) Using two consecutive frames f_k, f_{k+1} (k means the kth frame) directly; 2) capitalizing on the first frame, which includes two steps: first, we estimate the movement of f_k, f_{k+1} to the first frame separately, then we obtain the relative movement between f_k and f_{k+1}. The first approach is efficient, but may produce accumulated errors. We adopt the second approach, in which the first frame is only calculated once. When tracking a video, the initialization of DAM is performed in the first frame, then the convergent positions of the feature points are adopted as initial locations in the next frame and likewise for the consecutive frames.

The exterior parameters and the displacement of the feature points are calculated consequently. They are optimized alternatively until attaining the required precision. There are six exterior parameters (α, β, γ, $\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{Z}_t$) and 19 expression unit parameters to be estimated. We use conjugate gradient method to solve the optimization problem, and the target functions about a pair of 3D points $p(x^i, y^i, z^i)$ in mesh and 2D points $q(\boldsymbol{I}_u^i, \boldsymbol{I}_v^i)$ in image are as follows:

$$w\begin{bmatrix} I_u^i \\ I_v^i \\ 1 \end{bmatrix} = \begin{bmatrix} af_x + dn + gu & bf_x + en + hu & cf_x + sn + gu & pf_x + qn + ru \\ df_y + gv & ef_y + hv & sf_y + iv & qf_y + rv \\ g & h & i & r \end{bmatrix} \begin{bmatrix} \psi(x^i) \\ \psi(y^i) \\ \psi(z^i) \\ 1 \end{bmatrix}, \tag{11}$$

where w is an arbitrary scalar, $i \in [0, 71]$, and

$$\begin{bmatrix} a & d & g \\ b & e & h \\ c & s & i \end{bmatrix} = \begin{bmatrix} \cos\gamma\cos\alpha - \sin\alpha\sin\beta\sin\gamma & -\sin\alpha\cos\beta & \cos\alpha\sin\gamma + \sin\alpha\sin\beta\cos\gamma \\ \cos\gamma\sin\alpha + \cos\alpha\sin\beta\sin\gamma & \cos\alpha\cos\beta & \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma \\ -\cos\beta\sin\gamma & \sin\beta & \cos\beta\cos\gamma \end{bmatrix}, \tag{12}$$

$$[p \quad q \quad r] = [-aX_t - bY - cZ_t \quad -dX_t - eY - sZ \quad -gX_t - hY - iZ], \tag{13}$$

$$\psi(l) = \sum_{j=0}^{m} \sum_{Q=0}^{N_j} k_{jQ}^{l}, \qquad (14)$$

where $m = 18$, N_j is the number of vertices on the 3D mesh affected by the jth unit animation parameter; k_{jQ}^{l} is the coefficient of FAPs acting on the coordinate of l ($l \in (x, y, z)$).

The constraint dataset includes 72 pairs 2D-3D points related with 19 unit animation parameters. In our experiments, we set the significance levels as 96% and 98%, and extracted FAPs with training dataset of 30, 50, 70, 90, and 110 samples, respectively. We compared the following two methods while extracting the 19 FAPs based on ten videos of 31458 frames: (1) Performing feature space analysis to 2D feature points, extracting FAPs based on 2D tracked results; (2) performing feature space analysis directly to the 3D face model. Fig. 3 shows the results.

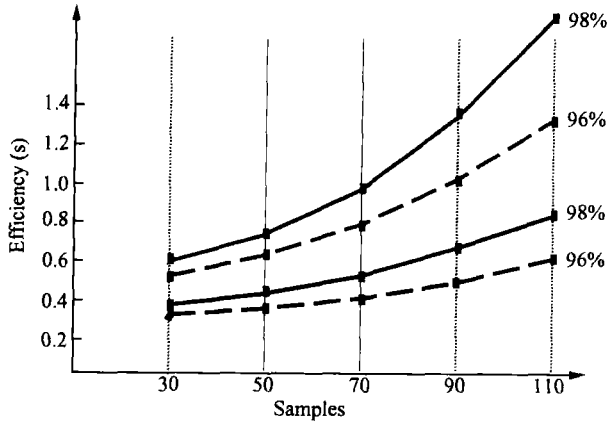

Fig. 3. Efficiency of extracting FAPs. Solid lines are produced by method (2), and dash lines are produced by method (1).

3 Transforming appearance details

Note that the effects of FAPs do not cover the personalized appearance details of a face such as forehead wrinkle, nose wing wrinkles, etc. We model appearance details by two steps: a) alignment of the major facial features such as nose, eyes, mouth, jaw, eyebrows; b) transforming appearance details information (e.g. wrinkle density, depth, illumination), which is related to the expressions. Till now few results have been reported upon this issue. Galton[24] generated multiple photographic images of several faces after aligning the eye positions. Later development adopted the technique of image warping[25] to map the component face images onto the mean shape. Tiddeman[26] presented a wavelet-based method for transforming facial texture, in which the low-frequency signals of the resultant image are from the original image while the high frequency signals are extracted from another image. Liu[27] synthesized realistic expressive facial expressions by mapping an expression ratio image (ERI) to another face image. Liu's method requires two reference images and is sensitive to the direction of the light.

We have developd a novel technique to transform texture details, which can keep the character of original face, and the desired expression details transformed from another image with the illumination effects of the original face (Fig. 4). Given an original face image A with frontal neutral expression and another face B with special expression, we derive a low-pass filter from a cubic B-Spline (H) and a high-pass filter from a real Gabor (G). For an original image X, the first layer pyramid decomposition is defined as

$$f(X) = X' + X'_x + X'_y, \qquad (15)$$

where X' represents the low frequency signals and X'_x, X'_y the image signals of high frequency. Note that the low frequency signals include the illumination information. Similarly, the second layer decomposition is expressed by

$$f(X') = X'' + X''_x + X''_y. \qquad (16)$$

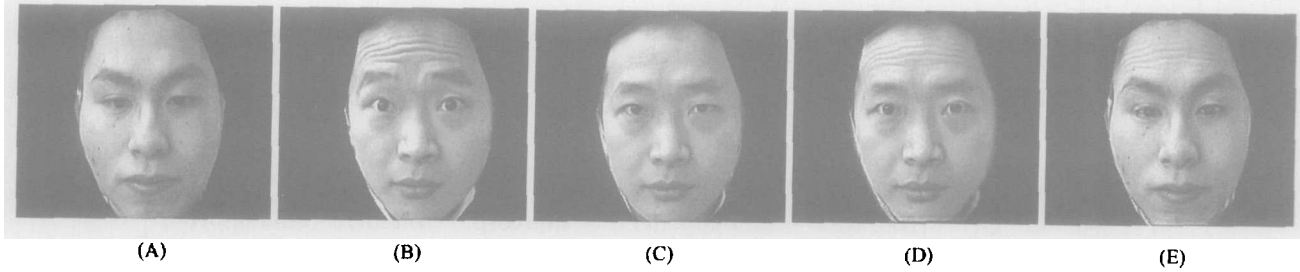

Fig. 4. The procedure of transforming appearance details. (A) Source neutral expression face; (B) source face with appearance detail; (C) the third face: medi-face; (D) transforming details from B to C; (E) expression mapping with $A * (D/C)$.

Now, collapsing the pyramid by $(AB)' = A''\Theta B_x'' \Theta B_y''$, we can generate the mixed image $(AB)'$, where Θ indicates a pyramid synthesis operator. Similarly, by $(AB)'\Theta B_x'\Theta B_y'$, we can reproduce the final transforming expression details. The approach may face a problem in some cases as shown in Fig.5, in which some flecks on the original image A are lost due to low-pass filtering. To address this problem, we introduce medi-face C, which corresponds to the neutral face of B. First, we produce a texture image D by transforming details of B to C with special ratio t. The equations are

$$(BC)' = (1 - t)C''\Theta t(B_x'\Theta B_y'), \quad (17)$$

and

$$D = (1 - t)(BC)'\Theta t(B_x'\Theta B_y'). \quad (18)$$

The coefficient $t(0<t<1)$ is used to control the expression extent (Fig.6). With C and D, an expression ratio image (ERI) $\boldsymbol{R}$ is generated. After mapping $\boldsymbol{R}$ to A, the final expressive face E can be generated, which maintains the original illumination of A with appearance transforms details from B.

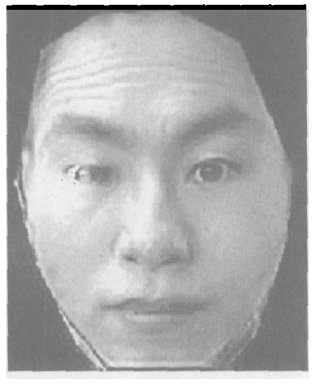

Fig. 5. Transforming wrinkle by wavelet filter.

We conducted three experiments on expression details transfer. The texture details are the forehead wrinkles in the bottom images of Fig. 6. Figs. 6, 7, 8 show more examples.

4 Video based face animation

Some work has been done on video-based facial animation[28], including simple parameter models (e.g. the Water' muscle model[29]). Previous work mostly relied on some auxiliary facilities, such as markers, to track the feature points. The produced expression was relatively simple. In our system, facial animation is driven in three steps: 1) creating a personalized 3D model M; 2) extracting the FAPs of another face from a video (see Section 2); 3) applying the FAPs to M to drive facial animation. In this section, we will address Steps 1) and 3) in detail.

4.1 Creating individual 3D model

We adopt the orthogonal image method[30] (Fig.9) to generate the simplified individualized parametric model. We obtain the 2D estimated result $\pounds$ of a frontal neutral expression image first. Some offset (e.g, at the eye socket, nose wing) may appear. We then need to adjust those inexact FAPs of $\pounds$. Based on $\pounds$, we can extract the facial shape parameters (FSPs), similar to extracting FAPs. Regardless of rotation, this result has high precision. We then utilize profile face to adjust Z unit parameters (ZUPs). In our IWCM, only five ZUPs (nose lifting, checks lifting, eyes lifting, mouth lifting, and eyebrow lifting) require interactive adjustment. The Z coordinate of other vertices can be fit by

$$f(p) = \sum_i c^i \boldsymbol{\Phi}(\| p - p^i \|), \quad (19)$$

where p^i is a group of 3D mesh points and u^i is their corresponding 2D displacement. We can obtain the coefficients of c^i by $u^i = f(p^i)$. We select RBF as

$$\boldsymbol{\Phi}(r) = e^{\frac{-r}{64}}. \quad (20)$$

The constraint equation is

$$\begin{cases} u^i = f(p^i), \\ \sum_i c^i = 0, \\ \sum_i c^i \cdot p^{i^{\mathrm{T}}} = 0. \end{cases} \quad (21)$$

For the texture mapping, we adopt bilinear interpolation.

4.2 Driving facial animation

Driving facial animation in our case is to drive a 3D facial model with expressions mimicing the video samples. This can be achieved conveniently with our extracted FAPs. Warping 3D model, and transforming the appearance details are the two key operations of expression cloning.

Note that a single animation parameter may shift multi-vertices, and the displacement of one vertex is the accumulated effects of several animation unit parameters. This mutual relation has been effectively simulated by our proposed IWCM. With this model, we can easily transform the shape of five sense organs by applying FAPs extracted from the video.

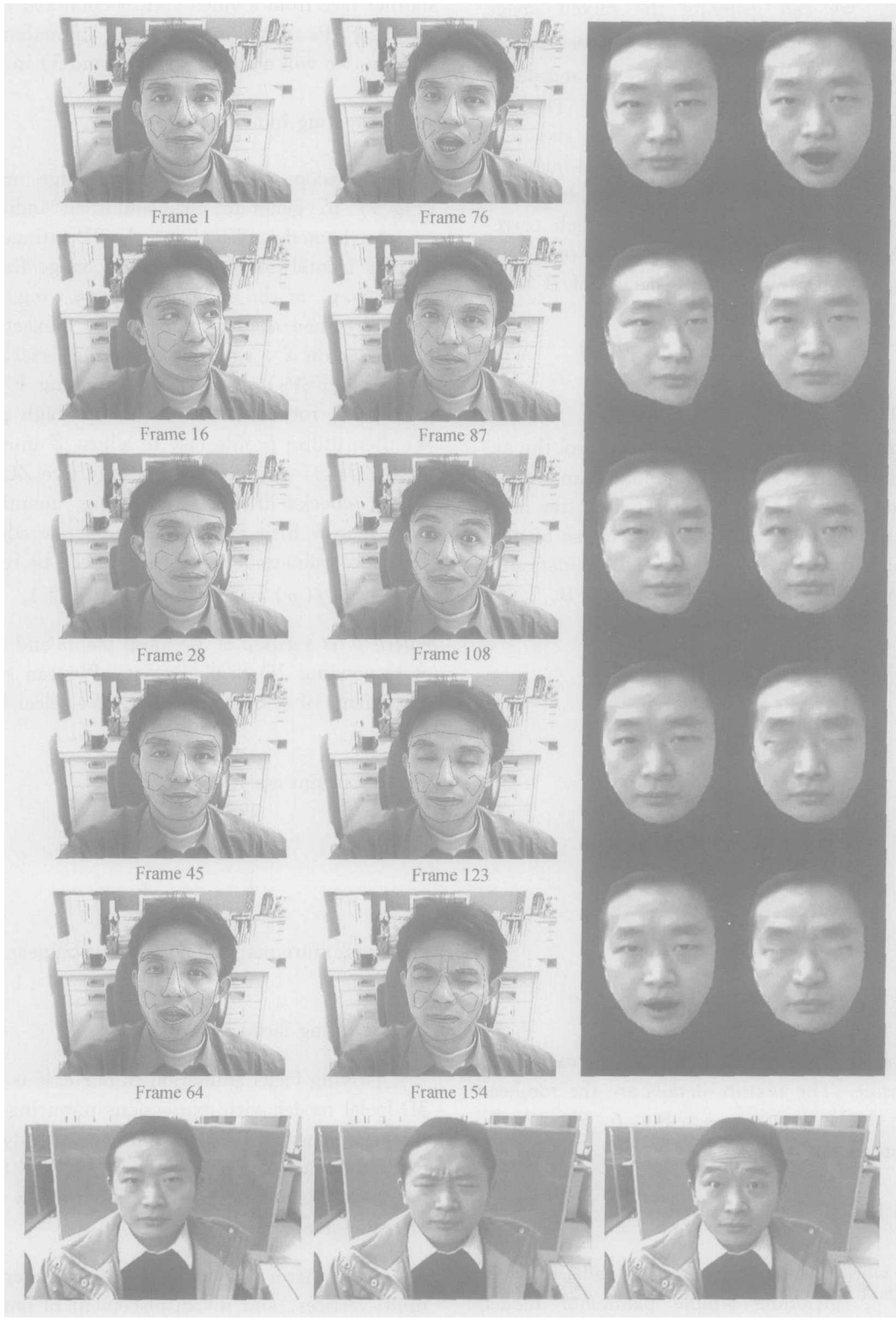

Fig. 6. Tracking a video (the left two wlumns) and extracting FAPs to drive facial animation (the right two columns) including acts of rotation, opening mouth, raising brows, closing eyes, and wrinkling nose; the lower row is the front neutral face used as texture, the sample of forehead wrinkle and brows wrinkle, respectively.

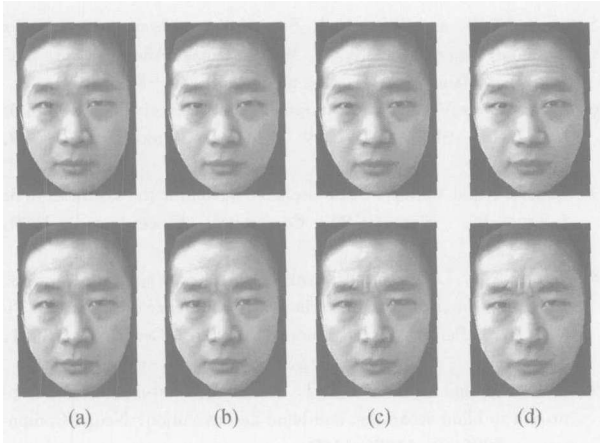

Fig. 7. Transform expressive details with various ratios of (a) 0.2; (b) 0.4; (c) 0.6; (d) 0.8.

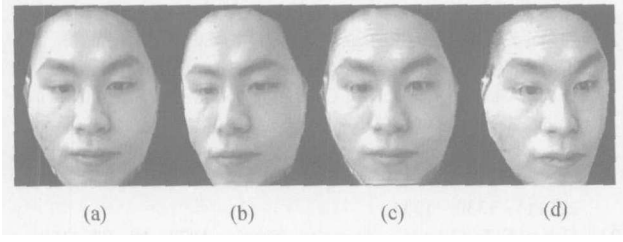

Fig. 8. (a) Original frontal face with neutral expression; (b), (c) and (d) transform expression details with ratio $t = 0.2, 0.4, 0.6$, respectively, when raising brows.

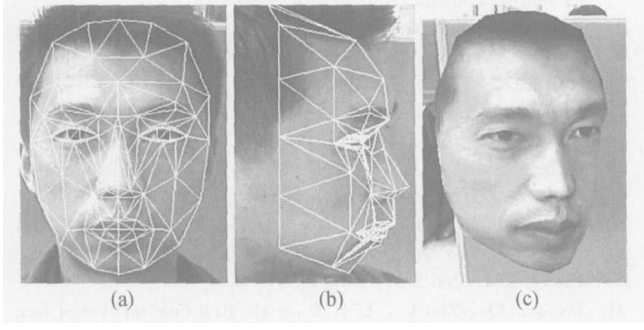

Fig. 9. (a) The frontal neutral expression; (b) the profile neutral expression face; (c) 3D textured face created by (a) and (b).

Texture details include the distribution of the global wrinkles, and the shape of each wrinkle. After warping the facial mesh, a reasonable distribution of wrinkles should be incorporated. However, modeling the shape of each individual wrinkle is difficult. Only a few papers try to express the details by the geometric model. Based on the study of image processing, the formulation of wrinkle details can be apprehended from macrocosmic, by pseudo-imitation to produce the acceptable visual effect. Our idea is also based on macrocosmic imitation (Fig. 8). The shape of a wrinkle is controlled by FAPs. In Eqs. (17) and (18), the value of parameter t is determined by the coefficients of FAPs. When multiple unit expression parameters act in the same region, we combine them by simple linear combination. Our method of expression transformation not only is simple, but also can produce satisfactory visual appearance. We establish a library of texture-detail samples to store some representative expressions. Their coefficients are set to be 1.0 when generating a typical expression. Accordingly, these coefficients are set to be 0.0 for the neutral expression. If the sample in library is identical with the face to be driven, the appearance details of the personalized face should be adopted; otherwise, the texture details are cloned to drive a new face.

5 Discussion and conclusion

Expressions, illumination and variant-views are the three key factors affecting the results of face tracking. In this paper, we mainly focus on extracting FAPs and transforming appearance details. For training and tracking, we adopt the gradient of a 5×5 neighborhood as input component, which can greatly improve illumination independence. Nowadays, techniques including sampling under various illuminations are available, and can be selectively incorporated into our system. Our current system can be extended conveniently to track the same faces from multi-views. We may establish 5 sample libraries about five views (-90—-60, -60—-30, -30—30, 30—60, 60—90),[31] which are stored according to the rotation angle, and compute the in-between angles of rotation by interpolation.

We have proposed and developed an effective system for video-based facial animation with appearance details cloning. Our system includes four processes: detecting and tracking face, extracting FAPs, facial appearance transfer, and driving facial expression animation. In comparison with others approaches, our main contributions in this work are as follows: (1) we have embedded ICA with DAM to produce basis vectors that are statistically independent, which increases the detecting and tracking precision by about 7%; (2) we have proposed a novel algorithm to extract FAPs rapidly and robustly, based on the 2D tracked results; (3) we have developed an effective method to transfer appearance details, while holding the original character of original face, and keeping the original illumination. In addition, expression extents are determined by FAPs.

Acknowledgements The authors would like to thank Dr. Mikkel B. Stegmann for his helpful discussions, thank Mr. Wu Yuan, Tang Feng, Zhou Weihua, and Prof. Bao Hujun for their suggestions and support.

References

1 Sakai T., Nagao M. and Kanade T. Computer analysis and classification of photographs of human faces. In: Proc. First USA-Japan Computer Conference, 1972, 2—7.
2 Guenter B., Grimm C., Wood D. et al. Making faces, In: Siggraph Proceedings 1998, 1998, 3(9): 55—66.
3 Pighin F., Hecker J., Lischinski D. et al. Synthesizing realistic facial expressions from photographs. In: Siggraph Proceedings, 1998, 75—84.
4 Kass M., Witkin A. and Terzopoulos D. Snakes: Active contour models. International Journal of Computer Vision, 1987, 1(4): 321—331.
5 Magnenat-Thalmann N., Cazedevals A. and Thalmann D. Modelling facial communication between an animator and a synthetic actor in real time. In: Proc. Modeling in Computer Graphics. Genova, Italy, June 1993, 387—396.
6 Terzopouos D. and Szeliski R. Tracking with Kalman snakes. In: Active Vision, MIT Press, 1992, 3—20.
7 Horn B.K.P. and Schunck B.G. Determining optical flow. Artificial Intelligence, 1981, 17: 185—203.
8 Essa I.A., Basu S., Darrell T. et al. Modeling, tracking and interactive animation of faces and heads using input from video. In: Proceedings of Computer Animation Conference, Geneva, Switzerland, IEEE Computer Society Press, June 1996, 172—179.
9 Essa I. A., Darrell T. and Pentland A. Tracking facial motion. In: Proceedings of the IEEE Workshop on Non-rigid and Articulate Motion. Austin, Texas, November, 1994, 36—42.
10 Cootes T., Edwards G. J. and Taylor C. J. Active appearance models. In: the 5th Proc. European Conference on Computer Vision, 1998, 484—498.
11 Coots T.F., Edwards G.J. and Taylor C.J. Active appearance models. In: ECCV98, 1998, 2, 484—498.
12 Edwards G.J., Coots T.F. and Yaylor C.J. Interpreting face images using active appearance models. In: Proc. International Conference on Automatic and Gesture Recognition, Japan, 1998, 300—305.
13 Essa I. A., Basu S., Darrell T. et al. Modeling, tracking and interactive animation of faces and heads using input from video. In: Proceedings of Computer Animation Conference, Geneva, Switzerland: IEEE Computer Society Press, 1996, June, 123—129.
14 Ahlberg J. MPEG-4 Facial Animation: The Standard, Implementation and Applications. New York: John Wiley & Sons, 2002, 103—112.
15 Cootes T.F. and Tayor C.J. Statistical models of appearance for computer vision. Draft report, Wolfson Image Analysis Unit, University of Manchechester, December 2000, 84—86.
16 Blanz V. andVetter T. A morphable model for the synthesis of 3d faces. In: SIGGRAPH' 99 Conference Proceedings, 1999, 187—194.
17 Blanz V. and Vetter T. A morphable model for the synthesis of 3d faces. In: Siggraph'99 Conference Proceedings, 1999, 187—194.
18 Hou X.W., Li S.Z. and Zhang H.J. Direct appearance models. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Hawaii, December, 2001, 828—833.
19 Bell A.J. and Sejnowski T.J. An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 1995, 6: 1129—1159.
20 Hyvarinen A. Survey on independent component analysis. Neural Computing Surveys, 1999, 2, 94—128.
21 Ahlberg J. Candide-3-an updated parameterized face, Report No. LiTH-ISY-R_2326, Dept. of EE, Linkoping University, Sweden, January 2001. http: //www.icg.isg.isy.liu.se/candide.
22 Zhang Z. A new multistage approach to motion and structure estimation by gradually enforcing geometric constraints. In: The 3rd Proc. Asian Conference on Computer Vision (ACCV'98), Hong Kong, January 1998, 567—574.
23 Zhang Z.Y. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330—1334.
24 Galton F.J. Composite portraits. Nature, 1878, 18: 97—100.
25 Ruprecht D. and Muller H. Image warping with scattered data interpolation. IEEE Computer Graphics and Applications, 1995, 15(2): 37—43.
26 Tiddeman B., Burt D. and Perrett D. Prototyping and transforming facial textures for perception research. IEEE Computer Graphics and Applications, 2001, 21: 42—50.
27 Liu Z.C. Shan Y. and Zhang Z.Y. Expressive expression mapping with ratio images. In: Siggraph' 01 Conference Proceedings, 2001, 271—276.
28 Parke F.I. Computer generated animation of faces. In: Proc. ACM Annual Conf., 1972, 451—457.
29 Waters K. and Frisbie J. A coordinated muscle model for speech animation. Graphics Interface, 1995, 163—170.
30 Horace H.S.I. and Yin L.J. Constructing 3D individualized head model from two orthogonal views. The Visual Computer, 1996. 12(5): 254—266.
31 Zhang Z.Q., Zhu L., Li S.Z. et al. Real-time multi-view face detection. In: Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition. Washington, DC, USA, May, 2002, 20—21.