TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 08/17 pp51-57
Volume 12, Number 1, February 2007

Personalized Multi-View Face Animation with Lifelike Textures*

LIU Yanghua (柳杨华), XU Guangyou (徐光祐)**

Key Laboratory on Pervasive Computing (Tsinghua University) of the Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract: Realistic personalized face animation mainly depends on a picture-perfect appearance and natural head rotation. This paper describes a face model for generation of novel view facial textures with various realistic expressions and poses. The model is achieved from corpora of a talking person using machine learning techniques. In face modeling, the facial texture variation is expressed by a multi-view facial texture space model, with the facial shape variation represented by a compact 3-D point distribution model (PDM). The facial texture space and the shape space are connected by bridging 2-D mesh structures. Levenberg-Marquardt optimization is employed for fine model fitting. Animation trajectory is trained for smooth and continuous image sequences. The test results show that this approach can achieve a vivid talking face sequence in various views. Moreover, the animation complexity is significantly reduced by the vector representation.

Key words: face animation; point distribution model (PDM); texture; multi-view

Introduction

Personalized face animations need to generate realistic and natural faces. Many human head movements and subtle nuances of facial expressions are very difficult to model. The difficulties are mainly related to realism, which is how to build a generative model that can be used to analyze and synthesize photorealistic facial textures, and naturalness, which is how the face model animates multiple facial views. Existing face animation technologies can be roughly divided into two classes: 3-D model-based face animation technologies and image-based face animation technologies.

The complicated textures on an animated face are very difficult to attain in realistic face animations. The intricate and sensitive facial textures are rooted in the intricate facial motions and expressions, such as the obscuring of the teeth by the lip movements, texture diversification caused by tongue movements, texture wrinkles of the cheeks caused by expressions and illumination changes during pronunciation. A vivid animation image should correctly exhibit those visual textures of lips, teeth, and tongue. 3-D model-based face animation approaches require very sophisticated models for very complicated mouth textures. For generic 3-D face models, various approaches[1-3] have been proposed to parameterize the facial geometry and texture in 3-D face models. By manipulating the face model parameters over time, 3-D model-based approaches can easily generate pose-variable expressional faces. Researchers have tried many methods to improve the visual realism of modeled face animations, such as laser scanners and computerize texture mapping techniques[1-3] to acquire the required precise 3-D information. The clip-and-paste approach has also been used in 3-D model-based face animations to deal with subtle texture deformations of parts of the face such as the mouth and eyes, but for the 3-D modeling, complexity and the rendering latency affects the animation realism.

Received: 2005-09-14; revised: 2006-02-20

* Supported by the National Natural Science Foundation of China (No. 60673189)

** To whom correspondence should be addressed.
E-mail: xgy-dcs@mail.tsinghua.edu.cn; Tel: 86-10-62785483

Image-based approaches[4-7] can effectively improve visual effects of face animations by choosing and training a set of images to model facial objects without apparent 3-D information. Beier and Neely[4] interpolated and meta-morphed between two facial images by providing natural feature-based specifications and interactions, resulting in non-video realistic animations. Ezzat and Poggio[6] built a 2-D facial animation system based on a set of phoneme images using optical flow arithmetic. The speech driven facial animation system generates pronouncing image sequences by interpolating between successive phoneme images, but the image sequence is not flowing enough. Ezzat et al.[7] later presented another face animation approach with the capability to generate personal pronouncing video by assembling intricate mouth texture image from a small number of prototype images. In the system, the basic pixel flow and pixel appearance vectors are represented in a multi-dimensional fashion in the multi-dimensional morphable model (MMM). The effectiveness of the method has been demonstrated by Black et al.[8] who showed that the method is such a robust statistical framework which models possible variations as a probability mixture of causes that allows complex models of illumination variations and iconic change such as mouth movements. However, these types of image-based methods are only able to generate fixed-pose facial animations.

In image-based approaches, one of the major difficulties in facial animations is to express facial views with changes in the shape and texture of the face as the pose changes. Image-based approaches can achieve reasonable facial muscular and skin tissue textures, but fall short in multi-pose facial animations. Those approaches cannot easily predict a facial pixel's displacement resulting from changes in facial expressions or rotation and also cannot accurately estimate the obstructions which inevitably occur in convex and concave parts of the face. Human expressions always involve head movement, so pose changes are important for natural facial animations. Though many approaches have been proposed for multiple facial view in 2-D[9,10], the results still have various limitations. Researchers have tried to improve active shape model[10] and active appearance model[9]; however, these models are limited since they model nonlinear multi-view facial motions with linear models, can be computationally intensive[10] or have narrow restricted views[9]. View-based techniques defined in eigen space or the support vector machine model of the face space[11] used arbitrary divisions of the facial space so they tend to be susceptible to distortion when facial pose changes.

This paper presents a multi-view face model that generates realistic pose-variable facial animations with acceptable computational costs. The model parameterizes the facial shape by means of a 3-D point distribution model (PDM)[12,13] with the facial texture represented by a set of selected prototype image textures in a compact training texture vector space. Pose-variable realistic animations, especially around the mouth, can be generated by images collection and machine learning techniques. 2-D triangle mesh structures serve to bridge the shape space and the texture space. Instead of using linear optical flow to represent facial shape transformations as in the MMM model[7], the present model uses a flexible 3-D face geometry PDM, a sparse set of facial landmarks and meshes without real 3-D textures to efficiently represent facial motions by relating the landmarks over a collection of sample shapes to realize pose-variability. The obstructions are naturally identified in the 3-D information. Additionally, the PDM provides topology-reserved relationships between facial movements by compactly and flexibly describing non-rigid facial motions with the texture distortion in the images achieved by mesh warping. When the deformed 3-D geometry facial model is projected into a new view with depth obstruction estimation, the 2-D mesh structure of the animated face shape is obtained and the multi-pose animation image is acquired by texture warping between meshes. Compared to texture warping pixel by pixel, texture mesh warping is more efficient and stable, and has similar visual effects.

1 Training Data Capture and Pre-Processing

1.1 Training data capture

A variety of human audiovisual corpuses were collected for training image data using two synchronized and calibrated cameras from front and half-profile (about 45°) view. The talking head had no other expressions except pronunciation. The analysis focused on modeling deformations and textures of the

talking-related mouth area, which has the richest textures, the most complicated motions, and the most intricate obstructions in the human face (Fig. 1). The two images that were captured at the same time by the two cameras constituted an image pair which was used to reconstruct the 3-D facial geometry based on binocular vision[14] for the PDM construction.

Fig. 1 Left: frontal view; middle: half-profile view; upper right: mask used to separate non-rigid and rigid parts of the face; lower right: non-rigid part of facial image after pre-processing

The corpuses involved 200 Chinese short sentences or phases, which had 1318 syllables and 1897 triphonemes, covering all possible pronouncing mouth appearances. The recording took about 17 min.

1.2 Pre-processing

Since the talking head has no other expressions except for pronunciation, the face was divided into a non-rigid part and a rigid part by masking the image as shown in Fig. 1. The rigid part of the face only rotates and the images were synthesized using image warping. The animation of the non-rigid part was more difficult and that was the emphasis of this work.

The images were normalized to compensate for unexpected residual head movement to retain only the mouth motions because the human subject inescapably had small head movements even though she tried to keep her head still. The residual head movements were analyzed using 4 corresponding points between the chosen reference image and the current image[7], with the small head movement approximated as perspective motions in the facial surface plane which had only 8 degrees of freedom.

2 Model

The facial image representation was decomposed into shape representation and texture representation.

A set of prototype images $\{\boldsymbol{I}_i\}_{i=1}^{2N}$ was chosen to represent the facial appearance in the texture space. These images should cover all possible textures. The images normally include a reference image $\boldsymbol{I}_1$, which has no expressions or motion.

3-D facial geometry PDM was used to represent the shape deformation by eigenvectors in the PDM space. Optical flow was used for the shape representation of fixed pose facial animations since the shape deformation can be approximated as linear for fixed pose animation so they can be modeled by linear concatenation of the optical flow. Linear concatenation does not work with pose changes.

The 3-D facial geometry model and the texture space were connected by constructing 2-D triangular mesh elements, for the frontal view $\{\boldsymbol{S}_i^1\}_{i=1}^{M_1}$ and for the half-profile view $\{\boldsymbol{S}_i^2\}_{i=1}^{M_2}$ (M_1 and M_2 denote the number of triangular mesh elements for the frontal and half-profile views), corresponding to the triangular mesh used for the 3-D facial geometry model $\{\boldsymbol{S}_i\}_{i=1}^{M}$ (M denotes the number of triangular mesh elements in the 3-D facial model). The 2-D elements were obtained by projecting the 3-D model to each view after the obstruction estimate based on the vertices depth variations. The 2-D elements were also the basis for synthesizing the facial animation image when texture warping was applied.

2.1 Texture space

The texture space is made up of a set of prototype images with vector representation used to represent the mouth area appearance.

The prototype images were selected from the training data images. Appropriate frontal prototype images (denoted by view1) and the corresponding half-profile prototype images (denoted by view2) were well selected for the set.

The set of prototype images should construct or at least overlay orthogonal axes in the image texture space. Principle component analysis (PCA) was performed to analyze the set $\left\{\boldsymbol{I}_{j,\text{view1}}\right\}_{j=1}^{N_\text{T}}$ (N_T denotes the total number of frontal training images and $\boldsymbol{I}_{j,\text{view1}}$ denotes the j-th frontal view image) which resulted from the masking, with each image having a size of 250×180 pixels. 14 principle basis vectors and the covariance matrix were obtained with a deviation of 95%.

The basis vector in the PCA space did not exactly correspond to the images in the training data, so the images were selected using *k*-means clustering, which is performed on a low dimension set $\{\boldsymbol{p}_{j,\text{view1}}\}_{j=1}^{N_\text{T}}$ by computing the Mahalanobis distance metric between $\boldsymbol{p}_{m,\text{view1}}$ and $\boldsymbol{p}_{n,view1}$.

$$d(\boldsymbol{p}_{m,\text{view1}}, \boldsymbol{p}_{n,\text{view1}}) = (\boldsymbol{p}_{m,\text{view1}} - \boldsymbol{p}_{n,view1})^\text{T} \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{p}_{m,\text{view1}} - \boldsymbol{p}_{n,\text{view1}}) \quad (1)$$

where $\boldsymbol{\Sigma}$ is the original PCA eigen matrix.

When an image's clustering center did not overlap any image, the nearest image was chosen to replace that image. Each image $\boldsymbol{I}_{j,\text{view1}}$ was then represented by PCA vectors as $\boldsymbol{p}_{j,\text{view1}}$.

$$\min\ (\sum_{i=1}^{N} |\boldsymbol{I}'_{i,\text{view1}} - \boldsymbol{I}_{i,\text{view1}}|) \quad (2)$$

31 images were selected by the *k*-mean clustering. With the reference image, a total of N=32 prototype frontal images make up the set $\{\boldsymbol{I}_{i,\text{view1}}\}_{i=1}^{N}$. The corresponding half-profile prototype image set was $\{\boldsymbol{I}_{i,\text{view2}}\}_{i=1}^{N}$. These 2$N$ prototype images made up the texture space axis.

Given an image $\boldsymbol{I}^\text{synth}$ and its texture parameters $\boldsymbol{T}^\text{synth} = \{\boldsymbol{t}_i^\text{synth}\}_{i=1}^{2N}$ ($\sum_{i}^{2N} \boldsymbol{t}_i^\text{synth} = 1$), an animation image can be synthesized by

$$\boldsymbol{I}^\text{synth} = \sum_{i=1}^{N} \boldsymbol{t}_i^\text{synth} \boldsymbol{I}_{i,\text{view1}} + \sum_{i=N+1}^{2N} \boldsymbol{t}_i^\text{synth} \boldsymbol{I}_{i-N,\text{view2}} \quad (3)$$

2.2 3-D facial geometry PDM

2.2.1 Definition of 3-D facial geometry PDM

The 3-D point distribution model is a shape space represented by a set of number-fixed 3-D vertices[13]. The model can efficiently describe a typical shape distribution and the allowed variability by statistic techniques and vector representations. Triangular mesh is added to 3-D PDM to create a topology structure fixed 3-D PDM geometry model. Assume a 3-D shape vector

$$\boldsymbol{V}_i = \{x_{i1}, y_{i1}, z_{i1}, x_{i2}, y_{i2}, z_{i2}, \cdots, x_{iL}, y_{iL}, z_{iL}\}^\text{T} \quad (4)$$

where $\{x_{ik}, y_{ik}, z_{ik}\}$ are the 3-D position of vertex k and L is the number of vertices.

PCA on the 3-D shape vector set $\{\boldsymbol{V}_i, i = 1, 2, \cdots, N\}$ gives the mean shape $\bar{\boldsymbol{V}}$ and the matrix $\boldsymbol{U}$ containing the first N_s significant eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{N_\text{s}}$.

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{N_\text{s}}] \quad (5)$$

Then, the shape $\boldsymbol{V}_i$ can be presented by a vector $\boldsymbol{S}_i$ in the PDM space:

$$\boldsymbol{S}_i = \boldsymbol{U}^\text{T}(\boldsymbol{V}_i - \bar{\boldsymbol{V}}) \quad (6)$$

The shape vector $\boldsymbol{V}_\text{r}$ can be obtained using a reconstruction formula with the PDM vector $\boldsymbol{S}_\text{r}$.

$$\boldsymbol{V}_\text{r} = \boldsymbol{S}_\text{r}\boldsymbol{U} + \bar{\boldsymbol{V}} \quad (7)$$

The 3-D PDM then describes the average shape by the mean position of the 3-D vertices and describes a number of variation modes by eigenvectors so as to compactly represent the deformation of the 3-D model with a small number of linearly independent parameters.

2.2.2 Construction of 3-D facial geometry PDM

The 3-D facial geometry models were reconstructed using the PDM training data from selected image pairs. The reconstruction was applied for all prototype image pairs. 28 landmarks were selected in the non-rigid facial part with 38 landmarks in the rigid facial part to reconstruct the 3-D face model. Epipolars were used to help the selection of correspondent landmark pairs. Then the 3-D positions of the landmarks were identified using binocular vision[14] as shown in Fig. 2.

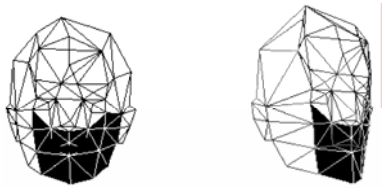

Fig. 2 3-D facial geometry model: the white and the black represent the non-rigid and rigid facial part that can be seen from both views

The 3-D positions of some landmarks that could only be seen with frontal view needed to be estimated. Given that the frontal image is actually frontal and eudipleural, the landmark depth can be replaced by the depth of its symmetrical landmark about the symmetry axis according to symmetry. The vertical perpendicular to the line between the two pupils was chosen to serve as the symmetry axis.

After reconstructing all the 3-D face models of the prototype image pairs, the non-rigid facial part vertices were used to construct the facial geometry PDM for

the vertices vector set $\{\boldsymbol{V}_i\}_{i=1}^{N}$ as described in Section 2.2.1. The eigenmatrix $\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2\cdots,\boldsymbol{u}_{N_s}]$ ($N_s=11$) was obtained with a 95% deviation. The facial shape is then represented by a relatively small number of eigenvectors.

Given the shape parameter $\boldsymbol{S}^{\text{synth}}=\{\boldsymbol{s}_i^{\text{synth}}\}_{i=1}^{N_s}$, the corresponding shape vector $\boldsymbol{V}^{\text{synth}}$ is

$$\boldsymbol{V}^{\text{synth}}=\boldsymbol{SU}+\bar{\boldsymbol{V}}=\sum_{i=1}^{N_s}\boldsymbol{s}_i^{\text{synth}}\boldsymbol{u}_i+\bar{\boldsymbol{V}} \tag{8}$$

2.3 Face model parameters set

The parameter set of the facial model can be represented by

$$\boldsymbol{C}=(\boldsymbol{S},\boldsymbol{T},\alpha,\beta,\gamma)^{\text{T}} \tag{9}$$

where $\boldsymbol{S}$ denotes the shape parameters $\{\boldsymbol{s}_i\}_{i=1}^{N_s}$, $\boldsymbol{T}$ denotes the texture parameters $\{\boldsymbol{t}_i\}_{i=1}^{2N}$, and α, β, and γ represent the three Euler angles of facial motion.

3 Model Fitting

The model fitting algorithm followed the principle of analysis-by-synthesis and searched for the optimal 3-D facial model parameters for a new facial animation image pair to be represented in the image space. The parameter set C of the new image was estimated by minimizing a cost function (as shown in Eqs. (10) and (11)) using Levenberg-Marquardt optimization[14].

$$C'=\arg\min\ (L(C)) \tag{10}$$

$$L(C)=\left[\boldsymbol{A}\frac{\boldsymbol{SU}+\bar{\boldsymbol{V}}}{(\boldsymbol{SU}+\bar{\boldsymbol{V}})(3)}-\boldsymbol{m}_{\text{view1}}\right]^2+\left[\boldsymbol{A}\frac{\boldsymbol{R}^{\text{T}}(\boldsymbol{SU}+\bar{\boldsymbol{V}}-\boldsymbol{T}_{\text{t}})}{\boldsymbol{R}^{\text{T}}(\boldsymbol{SU}+\bar{\boldsymbol{V}}-\boldsymbol{T}_{\text{t}})(3)}-\boldsymbol{m}_{\text{view2}}\right]^2+(\boldsymbol{I}'_{\text{view1}}-\boldsymbol{I}_{\text{view1}})^2+(\boldsymbol{I}'_{\text{view2}}-\boldsymbol{I}_{\text{view2}})^2 \tag{11}$$

where $\boldsymbol{A}$ denotes the matrix of camera intrinsic parameters, $\boldsymbol{R}$ the rotation matrix between the two cameras, and $\boldsymbol{T}_{\text{t}}$ the translation vector. $\boldsymbol{m}_{\text{view1}}$ and $\boldsymbol{m}_{\text{view2}}$ denote the vertex vectors of the frontal and half-profile pixel coordinates. $\boldsymbol{I}'_{\text{view1}}$ and $\boldsymbol{I}'_{\text{view2}}$ denote the synthesis images from Eq. (8).

The cost function contains the local fitting criterion around each landmark and the global fitting criterion for the synthesized appearance. Because the shape and texture parameters are independent and excessive computations will be required if a large number of image textures are involved, the shape analysis for 3-D geometry model in low-dimension was firstly achieved. Similar shapes corresponded to similar appearances, so the shape estimation result was used to guide initialization of the texture parameters by computing the shortest Mahalanobis distance in the PDM space.

4 Synthesis

The target facial animation image is synthesized based on $\boldsymbol{S}^{\text{synth}}$, $\boldsymbol{T}^{\text{synth}}$, α, β, and γ. In this work, the target facial animation image was synthesized using only the frontal prototype image textures, because in this work, the facial rotations are between the two camera angles and are not so large that the frontal prototype image textures can adequately provide the texture synthesis. When the target facial animation pose varies over a large range to a bigger extent and many obstructions occur, the other view prototype images must also be taken into account.

The dimension of the shape parameter S was 11 while the dimension of the texture parameter, t, was 32 by 2 (32 if only the frontal textures are involved). Together with the 3 dimensions of α, β, and γ, the facial image synthesis dimension is only 78 (46 if only the frontal views are involved). After 3-D vertex positions for the face model $\boldsymbol{V}^{\text{synth}}=[\boldsymbol{V}_x^{\text{synth}},\boldsymbol{V}_y^{\text{synth}},\boldsymbol{V}_z^{\text{synth}}]$ are computed from $\boldsymbol{S}^{\text{synth}}$ and Eq. (8), the model is rotated to a novel view using

$$(\boldsymbol{V}^{\text{synth}})^{\text{T}}=\boldsymbol{R}(\boldsymbol{SU}+\bar{\boldsymbol{V}})^{\text{T}}+\boldsymbol{T}_{\text{t}} \tag{12}$$

Then, the 2-D vertex positions $\boldsymbol{v}^{\text{synth}}=[\boldsymbol{v}_x^{\text{synth}},\boldsymbol{v}_y^{\text{synth}}]^{\text{T}}$ are obtained by projecting $\boldsymbol{V}^{\text{synth}}=[\boldsymbol{V}_x^{\text{synth}},\boldsymbol{V}_y^{\text{synth}},\boldsymbol{V}_z^{\text{synth}}]$ to the new view. The focus of the half-profile camera, f_2, was used as the projecting focus, so $\boldsymbol{v}^{\text{synth}}$ was calculated using

$$\boldsymbol{v}_x^{\text{synth}}=f_2\frac{\boldsymbol{V}_x^{\text{synth}}}{\boldsymbol{V}_z^{\text{synth}}},\quad \boldsymbol{v}_y^{\text{synth}}=f_2\frac{\boldsymbol{V}_y^{\text{synth}}}{\boldsymbol{V}_z^{\text{synth}}} \tag{13}$$

The 2-D facial mesh of the target image $\{\boldsymbol{S}_i^{\text{synth}}\}_{i=1}^{M_{\text{synth}}}$ (where M_{synth} is the number of the elements) is obtained based on the invariability of the 3-D PDM geometry topology in projection with estimations of the

obstructions.

The target image, $\boldsymbol{I}^{\text{synth}}$, can then be generated using

$$\boldsymbol{I}^{\text{synth}} = \rho_{\text{view1}}\boldsymbol{I}^{\text{synth}}_{\text{view1}} + \rho_{\text{view2}}\boldsymbol{I}^{\text{synth}}_{\text{view2}} \quad (14)$$

where ρ_{view1} and ρ_{view2} denote the weights of the frontal and half-profile synthesis images (if using only the frontal view texture, ρ_{view1} is 1 and ρ_{view2} is 0). $\boldsymbol{I}^{\text{synth}}_{\text{view1}}$ and $\boldsymbol{I}^{\text{synth}}_{\text{view2}}$ denote the synthesis images for the frontal and half-profile views.

The texture warping affine transformation can be based on the natural correspondence between the meshes for the target and frontal images and between the target and half-profile images.

Fig. 3 First line: original images in training data; second line: corresponding synthesis images; third line: synthesis images for the half-profile view; and the last line: synthesis images for a new view. The rough outlines of the face are due to a small number of vertices aligned along the edge of the geometry model which can be optimized by refining the model.

An operator Φ is defined as the texture warping operation. $\Phi(\boldsymbol{S}', \boldsymbol{S})$ represents a texture warping from mesh $\boldsymbol{S}$ to $\boldsymbol{S}'$. Then

$$\boldsymbol{I}^{\text{synth}}_{\text{view1}} = \sum_{i=1}^{N} \boldsymbol{t}_i \boldsymbol{I}_{i,\text{view1}} = \sum_{i=1}^{N} \boldsymbol{t}_i \sum_{j=1}^{M_{\text{synth}}} \Phi(\boldsymbol{S}^{\text{synth}}_j, \boldsymbol{S}^1_j) \quad (15)$$

$$\boldsymbol{I}^{\text{synth}}_{\text{view2}} = \sum_{i=N+1}^{2N} \boldsymbol{t}_i \boldsymbol{I}_{i-N,\text{view2}} = \sum_{i=N+1}^{2N} \boldsymbol{t}_i \sum_{j=1}^{M_{\text{synth}}} \Phi(\boldsymbol{S}^{\text{synth}}_j, \boldsymbol{S}^2_j) \quad (16)$$

Combining Eqs. (14)-(16), the synthesis can be represented by

$$\boldsymbol{I}^{\text{synth}} = \rho_{\text{view1}} \sum_{i=1}^{N} \boldsymbol{t}_i \sum_{j=1}^{M_{\text{synth}}} \Phi(\boldsymbol{S}^{\text{synth}}_j, \boldsymbol{S}^1_j) + \rho_{\text{view2}} \sum_{i=N+1}^{2N} \boldsymbol{t}_i \sum_{j=1}^{M_{\text{synth}}} \Phi(\boldsymbol{S}^{\text{synth}}_j, \boldsymbol{S}^2_j) \quad (17)$$

The Chinese phonemes can be classified into 29 species. After the text phoneme sequence was aligned to the corpus of the recorded images with the help of a look-up table, the trajectory distribution parameters were trained by using the approach described by Ezzat et al.[7]

5 Pre-Processing and Experimental Results

The rigid part of the 3-D facial geometry model was firstly projected to the desired view with the corresponding 2-D meshes for the rigid part then obtained. The background image was generated by texture warping followed by boundary smoothing and blurring.

The facial animation model was then tested for text driven facial animation. The results show that the approach can efficiently provide personal realistic multi-view facial animation as shown in Fig. 4. Notice now the right ear of the talking head gradually disappears as the talking head turns to the left. The tongue and teeth textures inside the mouth are also impressive when animated. The calculating time for one animation image synthesis was 150 ms in a Pentium 4 CPU 2.00 GHz processor.

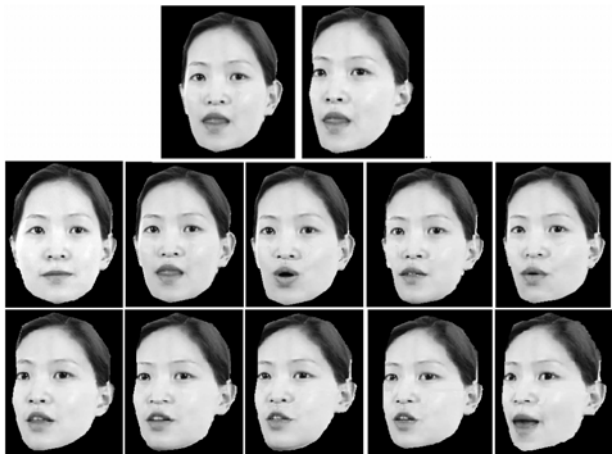

Fig. 4 An animated face images with realistic textures around the mouth. First line: same motion from two different views; second and third lines: animated faces for different syllables from a generated animation sequence for a head turning from front to left with the phonemes of null, /l/, /o/, /sh/, /u/, /j/, /i/, /ong/, /zh/, /ai/

Only 28 vertices were used in the non-rigid part of the facial geometry model, so then a few vertices aligned along the facial boundary resulting in a rough

edge on the animated face. The animation will be improved by refining the facial geometry model. The tests only analyzed the animation views between the two shooting views. However, a wider range of face animation can also be achieved by this approach.

Compared to existing face animation approaches such as Ezzat et al.[7], this approach is able to achieve vivid talking face sequence in various views. Thus, the talking heads are more attractive and lifelike. In addition, the computational burden is reduced by the vector representation. The vertices of the shape model play a very important role in enhancing the animation flexibility.

6 Conclusions

A multi-view face animation model was developed which combines a 2-D texture space with a 3-D facial geometry PDM using 2-D geometry mesh structures. The image-based method generates realistic images with the 3-D PDM efficiently describing various poses. The model can generate facial animation with realistic mouth movements while speaking from arbitrary viewing angles by processing two corpuses of training data. The mouth area texture including the teeth, lips, and tongue texture is very realistic. Moreover, the animation complexity is significantly reduced by the vector representation.

The animation approach learns from records of a specific subject to produce personal animations for that subject. An important future area of interest is to generalize the facial animations to persons other than this original subject without 3-D model construction.

References

[1] Choi C S, Aizawa K, Harashima H, Takebe T. Analysis and synthesis of facial image sequences in model-based image coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 1994, **4**(3): 257-274.

[2] Yau J F S, Duffy N D. A texture-mapping approach to 3D facial image synthesis. In: Proceedings of the 6th Annual Eurographics Conference. Sussex, UK, 1988: 129-134.

[3] Aizawa K, Harashima H, Saito T. Model-based analysis synthesis image coding (MBASIC) system for a person's face. *Signal Processing*: *Image Communication*, 1989, (1): 139-152.

[4] Beier T, Neely S. Feature-based image metamorphosis. *Computer Graphics*, 1992, **26**(2): 35-42.

[5] Oka M, Tsutsui K, Ohba A, Jurauchi Y, Tago T. Real-time manipulation of texture-mapped surfaces. *ACM Computer Graphics*, 1987, **21**(4): 181-188.

[6] Ezzat T, Poggio T. Mike talk: A talking facial display based on morphing visemes. In: Proceedings of the Computer Animation Conference. Philadelphia, USA, 1998: 96-102.

[7] Ezzat T, Poggio T, Geiger G. Trainable videorealistic speech animation. In: Proceedings of ACM SIGGRAPH. San Antonio, Texas, USA, 2002: 388-398.

[8] Black M, Fleet D, Yacoob Y. Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, 2000, **78**(1): 8-31.

[9] Cootes T, Edwards G, Taylor C. Active appearance models. In: European Conference on Computer Vision. Freiburg, Germany, 1998, (2): 484-498.

[10] Romdhani S, Gong S, Psarrou A. A multi-view nonlinear active shape model using kernel PCA. In: British Machine Vision Conference. Nottingham, UK, 1999: 483-492.

[11] Li Y, Gong S, Liddell H. Support vector regression and classification-based multi-view face detection and recognition. In: IEEE International Conference on Automatic Face & Gesture Recognition. Grenoble, France, 2000: 300-305.

[12] Cootes T F, Taylor C J, Cooper D H, Graham J. Training models of shape for sets of examples. In: Proceedings British Machine Vision Conference. Springer-Verlag, 1992: 9-18.

[13] Tony H, David H. Wormholes in shape space: Tracking through discontinuous changes in shape. In: Proc. of the 6th International Conference on Computer Vision. Bombay, India, 1998: 344-349.

[14] Wu Qing, Xu Guangyou, Wang Lei. A three-stage system for camera calibration. In: MIPPR'01. Wuhan, China, 2001: 22-24.