Automatic Face Modeling and Synthesis Based on Image Pairs

Peng Dai[1], Guangyou Xu[1], Thomas Riegel[2], Eckart Hundt[2]

[1] Key Lab. on Pervasive Computing, Ministry of Education, Tsinghua Univ.,
100084 Beijing, China
daip02@mails.tsinghua.edu.cn,
xgy-dcs@mail.tsinghua.edu.cn
[2] Siemens AG, Corporate Technology, CT IC 2,
81730 Munich, Germany
{Thomas.Riegel, Eckart.Hundt}@siemens.com

Abstract. Unlike traditional 3D model based or image based animation methods, in this paper a novel approach is presented to generate both facial actions and head rotations for photo-realistic facial animation based on one frontal and one half-profile facial image taken with an uncalibrated camera. We represent faces with 2D wire-frame models and use MPEG4 FAPs to encode basic facial actions. Hierarchical Direct Appearance Model is employed for facial feature localization. 3D deformable model is applied for pose estimation. By affine projection 3D deformable model and facial actions are mapped to 2D facial models and actions at various head poses. Coarse 2D models are refined with extracted facial features by RBF interpolation. Pose-variable facial animation is generated by synthesizing facial actions on 2D models and morphing facial textures between frontal and half-profile views. Experimental results demonstrate the effectiveness of our approach.

1 Introduction

Facial animation techniques have been applied in the field of Human-Computer Interface so as to provide intelligent and personalized services. Personalized avatars promise potential applications for new communication services. Video sequences can be encoded by avatars with high reduction of data transmission rate. Efforts have been made to introduce automatic facial modeling and animation techniques into interpersonal communication fields by offering web services or creating telecom applications on mobile devices. With individualized and simplified avatars, people can interact with each other more vividly at the cost of a lower network bandwidth.

Speech-driven facial animation and facial expression synthesis has been widely explored in previous literatures. Generally the approaches proposed in the field of facial modeling and animation fall into the following two categories: 3D model based approaches and image based methods.

Numerous research efforts have been made in the area of constructing 3D animated models of a specific person [1,4,6,7,8]. Some methods use Cyberware to extract 3D

shape data of human heads [3,4], and resort to texture mapping techniques to achieve the final 3D face model including both shape and texture information; some use 3D generic model and perform individualization based on a set of face images [1,6,8]. Based on 3D models, various techniques have been introduced for face tracking, analysis and synthesis along video sequences [2,5].

Image-based approaches can improve visual effects of facial animation, however, image based rendering techniques can not deal with pose variations quite effectively. An alternative to generate multi-view facial animation is to take a learning-based facial synthesis method based on a set of prototypical face images of different poses [9,10,11]. Within these approaches large facial image corpus is requested for the training of facial appearance models. Accurate detection of facial features is a constraint and prerequisite for the image-based facial modeling and animation systems. L. Yin [17] developed a complicated wire-frame model to represent human faces based on the frontal pose image. Detailed facial expressions are generated by method of modeling different human facial parts [13,14,16]. A. M. Tekalp et al. [15] extends the MPEG4 facial animation framework into 2D mesh animation, which saves the cost of 3D modeling process. A comparison between 3D model based method and image based method for facial image synthesis has been conducted in [18].

During interpersonal interactions, people tend to move their head slightly. Therefore how to synthesize talking heads with head movements becomes a significant challenge for generating vivid facial animation sequences.

User-oriented application scenarios request low computational costs. For instance, a user of the avatar system wants to construct a face model of his friend, however he neither has 3D data from Cyberware devices, nor does he have any idea of camera calibration technologies. How can we provide him a convenient and time-saving facial modeling and animation system?

In this paper we propose a novel method to generate pose variable facial animation based on facial image pairs. The image pair is taken from the same subject, with an uncalibrated camera. One image is at a frontal pose and the other is with a half-profile pose. Facial animation generated with our approach only presents slight pose changes, which is unlike 3D model based animation, however it still adds naturalness and vividness to the animation process.

The rest of the paper is organized as follows. In Section 2 Hierarchical Direct Appearance Model is introduced for facial feature localization. Section 3 presents our pose estimation method based on 3D deformable model. Section 4 describes the 2D facial model adaptation methods for our pose variable animation. Section 5 gives the facial action coding and facial animation method. Experimental results are presented in Section 6 and conclusions are drawn in Section 7.

2 Facial Feature Localization

We localize a set of facial landmarks on human faces so as to help construct face models. Hierarchical Direct Appearance Model (HDAM) [anonymous] is employed for the automatic facial landmark localization task. Direct Appearance Model (DAM)

[19] is proposed to predict shape information directly from texture information by building separate subspaces of texture and shape. HDAM extends DAM to a hierarchical model which imposes a two-layer analysis algorithm and performs better in algorithm convergence and initialization sensitivity. The feature points we choose in our implementation are mostly located in those facial areas with prominent visual information such as mouth, eyes, eyebrows, nose and lower-part face contour.
We train HDAM models separately for frontal faces and half-profile faces. Fig. 1 shows some HDAM based facial feature localization results, which are to be applied for head pose estimation and face modeling in the upcoming phases.

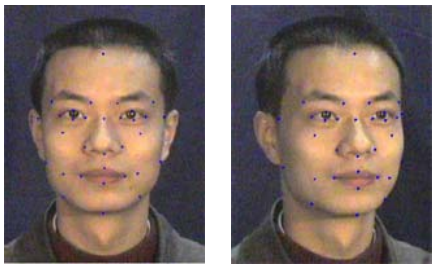


Fig. 1. HDAM based facial feature localization results

3 Pose Estimation

3.1 3D Deformable Face Model

In order to estimate head poses, we introduce a 3D generic face model. The 3D face model is given by a set of vertices $P_i = (X_i, Y_i, Z_i)\ (i = 1..n)$, which are a subset of a detailed MPEG4 compliant 3D head model. By triangulation we get a complete 3D wire-frame model. Fig. 2 illustrates the 3D model we use here.

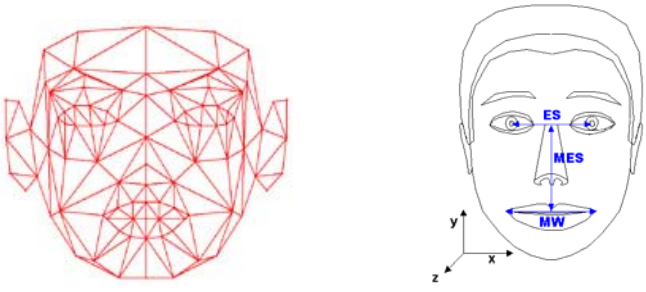


Fig. 2. Left: 3D Deformable Face Model. Right: Metrics applied for 3D model initialization

3.2 Model Initialization

Before estimating head poses with the 3D model, we need to individualize the 3D generic model for the specific person, that is, a coarse adjustment for the 3D model so as to approximately fit the face configuration indicated by the frontal image. The 3D model adjustment can be expressed as

$$\begin{pmatrix} X_s \\ Y_s \\ Z_s \end{pmatrix} = P_s = SP_g = \begin{pmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{pmatrix} \begin{pmatrix} X_g \\ Y_g \\ Z_g \end{pmatrix}. \quad (1)$$

where P_g and P_s denote vertices on the 3D generic model and specific model respectively, and the diagonal matrix S represents a scale transformation. The scale factors s_x, s_y and s_z are defined according to different benchmarks.

According to anthropometric knowledge, we consider a set of face metrics $FM = \{ES, MES, MW\}$, where ES, MES and MW denotes eye separation, eye-mouth separation and mouth width respectively. We calculate the metrics for the 3D model and the frontal face image respectively and adjust the 3D deformable model according to the ratios. Fig. 2 shows the face metrics described above. With the scale transformation described above, we change the geometric configuration of the 3D deformable face model.

3.3 3D Model Based Pose Estimation

In order to estimate head poses, we need to relate 3D deformable model to the image coordinate system. In this paper, weak-perspective projection model is adopted for 3D-2D mapping. A model based pose estimation method proposed in [22] requires camera intrinsic parameters and estimates head poses along video sequences, however, in our system camera parameters are not available. Therefore we propose an improved pose estimation approach based on two facial images. The mapping between 3D model vertices $P_i = (X_i, Y_i, Z_i)^T$ and 2D image feature points $p_i = (u_i, v_i)^T$ can be represented as

$$[u_i, v_i]^T = M(X_i, Y_i, Z_i, 1)^T. \quad (2)$$

where $M = \{m_{ij}\}\ (i = 1,2,3,4,\ j = 1,2)$ is a 4×2 matrix which includes both the camera intrinsic parameters and the rotation and translation of 3D deformable model. Let $R = \{r_{ij}\}\ \ (i, j = 1,2,3)$ and $T = (t_x, t_y, t_z)^T$ represent the rotation and translation between the camera and 3D model coordinate system.

Let f be the camera focal length, k_u and k_v be the pixel factors in horizontal and vertical directions, p_u and p_v represent the coordinates of the image center, s denotes the scaling factor, and Z_0 indicates the depth of the object. The mapping between 3D model vertices and image feature points can be expressed as

$$M = \begin{bmatrix} k_u fs\frac{r_{11}}{Z_0} & k_u fs\frac{r_{12}}{Z_0} & k_u fs\frac{r_{13}}{Z_0} & k_u fs\frac{t_x}{Z_0} + p_u \\ k_v fs\frac{r_{21}}{Z_0} & k_v fs\frac{r_{22}}{Z_0} & k_v fs\frac{r_{23}}{Z_0} & k_v fs\frac{t_y}{Z_0} + p_v \end{bmatrix}. \quad (3)$$

Then we consider the mapping of the 3D model onto the frontal image and half-profile image separately. Let $M = \{m_{ij}\}\ (i = 1,2,3,4,\ j = 1,2)$ be the projection matrix for the frontal face image, and $M' = \{m'_{ij}\}\ (i = 1,2,3,4,\ j = 1,2)$ be the projection matrix for the half-profile face image.

We choose eye contours and mouth contours as feature points for the 3D-2D mapping solution. Given these correspondence of 3D points and 2D points, we solve Equation (2) by SVD approach. Then we consider Equation (3) for the frontal and half-profile face image respectively. For the mapping between the 3D model and frontal face image, no rotation is necessary but only model translation, which means the projection matrix M can be expressed as

$$M = \begin{bmatrix} k_u fs\frac{1}{Z_0} & 0 & 0 & k_u fs\frac{t_x}{Z_0} + p_u \\ 0 & k_v fs\frac{1}{Z_0} & 0 & k_v fs\frac{t_y}{Z_0} + p_v \end{bmatrix}. \quad (4)$$

and the projection matrix M' can be expressed as

$$M' = \begin{bmatrix} k_u fs\frac{r_{11}}{Z_0} & k_u fs\frac{r_{12}}{Z_0} & k_u fs\frac{r_{13}}{Z_0} & k_u fs\frac{t_x}{Z_0} + p_u \\ k_v fs\frac{r_{21}}{Z_0} & k_v fs\frac{r_{22}}{Z_0} & k_v fs\frac{r_{23}}{Z_0} & k_v fs\frac{t_y}{Z_0} + p_v \end{bmatrix}. \quad (5)$$

Although f, k_u, k_v, s, p_u, p_v and Z_0 are not available in our system, we can figure out part of the rotation parameters, that is r_{11}, r_{12}, r_{13}, r_{21}, r_{22} and r_{23}.

The rotation matrix $R = \{r_{ij}\}\ (i, j = 1,2,3)$ can also be represented in the format of Euler angles α, β and γ which represent rotation angles around Z-axis, Y-axis and X-axis respectively. With part of the rotation parameters calculated previously, the three rotation angles between the frontal and half-profile faces are resolved. Hence the head pose corresponding to the half-profile face image is estimated approximately.
Fig. 3 shows the 3D model projection onto the frontal and half-profile face images respectively.

Fig. 3. 3D model projection at frontal view and half-profile view

4 Face Model Adaptation

During the pose estimation stage, we project 3D deformable model onto the frontal and half-profile facial images and get two distinct 2D wire-frame models accordingly. These 2D mesh-structure representations preserve the topology of the 3D model, however, these 2D models do not represent individual facial configuration and local facial features quite well. We need to perform geometrical deformation to the 2D models so as to differentiate specific faces from the generic face model.

4.1 2D Dynamic Models

With head pose changes, some areas of human faces are to be occluded and the corresponding 2D wire-frame models will change dynamically. In this paper we propose a dynamic 2D wire-frame model for human face modeling, within which we divide the entire model into interior area M_I and exterior area M_E. Most of the facial actions are expressed within the interior area, and for the exterior area there might be partial or complete occlusion circumstances during slight head pose variations. On account of this, we make a difference between the interior area and the exterior area in our implementation. For the interior area, we synthesize facial texture

by morphing between the frontal and half-profile facial images, while for the exterior area, we generate facial texture just by direct warping from the frontal image, without consideration about the half-profile image. Fig. 4 illustrates the interior area and exterior area of human face models.

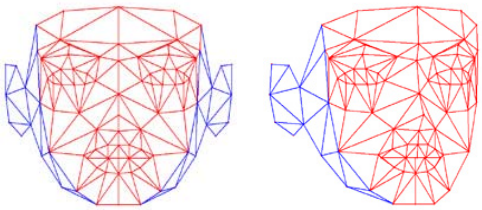


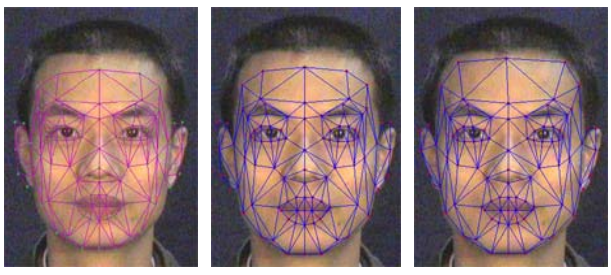
Fig. 4. Interior and exterior areas of face models. Left image shows the 2D face model at frontal pose, and right image expresses the 2D face model at half-profile pose. The red triangles represent interior area and the blue triangles denote exterior area

4.2 2D Model Refinement

The 2D model individualization process is based on the actual facial features extracted from the individual facial images. In this paper a two-step refinement method is proposed to adjust the 2D face models.

With the feature points extracted by HDAM method previously, we perform the first-step refinement to the coarse 2D models by Multi-Step Compactly Supported Radial Basis Function (MSCSRBF) based model adaptation [21]. MSCSRBF based approach ensures the independence of different facial areas during the interpolation process and performs with satisfying efficiency.

However these feature points do not reveal visual information concerning the forehead contour and the two ears, therefore the 2D models generated above are not sufficient for face contour representation. We adopt a modified snake algorithm [20] as the second-step refinement, which adapts the contours of forehead and two ears. Fig. 5 shows the refined models corresponding to the frontal and half-profile faces.



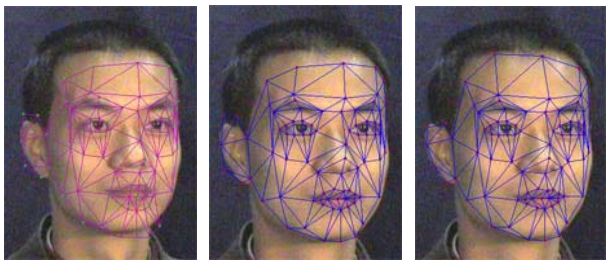


Fig. 5. Samples of two-step model refinement. Top row illustrates model refinement for frontal view, and bottom row shows model refinement for half-profile view

5 Facial Animation

Face and body animation specifications are included within MPEG4 standard. In MPEG4 facial animation framework, facial actions that represent visual speech and expressions are encoded with Facial Animation Parameters (FAP).

According to MPEG4 standard, there are totally 68 FAPs, among which the first two represent high-level animation parameters, visemes and expressions, whilst the left 66 FAPs encode facial actions on different parts of human faces. The FAP values are defined in Face Animation Parameter Units (FAPU), which are computed from spatial distances between major facial features on the model in its neutral state and account for faces of different sizes and proportions. Rules that define the deformation extent of the mesh-structure face models are represented with Face Definition Table (FDT) in MPEG4 framework. Facial animation can be generated with any parameterized face model for speech-driven animation if the visemes are known.

Based on the definition of MPEG4 facial animation, vertex displacements representing facial actions at different facial parts are driven by FAPs. According to the 3D-2D mapping acquired during pose estimation, displacements for the 3D model vertices at any poses can be projected and generate appropriate 2D displacements. Based on this basic idea, we project the MPEG4 3D facial animation rules onto image planes and derive different sets of 2D animation rules corresponding to various head poses. The mapping between 3D facial actions and 2D facial actions is given by

$$[\Delta u_i, \Delta v_i]^T = M(\Delta X_i, \Delta Y_i, \Delta Z_i, 1)^T \ . \quad \textbf{(6)}$$

Therefore a group of 2D animation rules are generated corresponding to different views. Since head pose changes in our implementation are slight, we only use three different sets of FDTs to represent the entire pose variable animation between the frontal view and the half-profile view.

After shape information acquired, facial texture information can be synthesized by image warping techniques. The image rendering process is controlled by the 2D wire-frame mesh at the corresponding head poses.

6 Results

The 3D deformable model we use here includes 94 vertices and 156 triangles. With a TTS system, we generate viseme streams as the input of our facial synthesis system.
We tested our pose variable facial animation system with a variety of persons. For each person we took one facial image at frontal view and one at half-profile pose. The image size is 640×480 and the camera is uncalibrated. During mouth movements, teeth patches are added into the mouth background. The animation is generated at normal video frame-rate on a PC with Pentium 4-M 1.29 GHz processor. Part of the animation sequences are shown in Fig. 6.

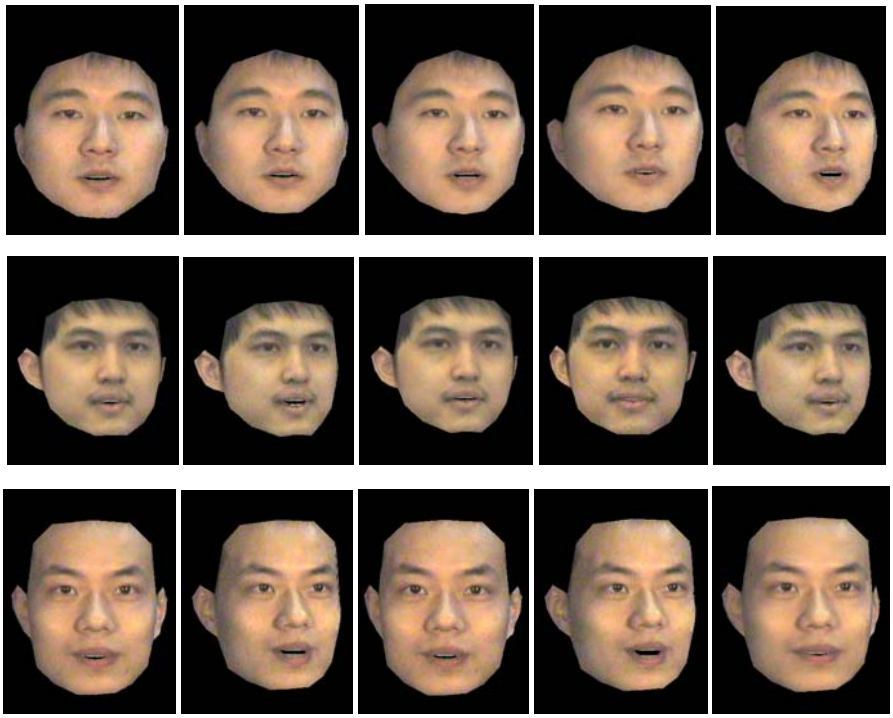


Fig. 6. Three samples of facial animation results. Images at each line are taken from animation sequences of various persons

7 Conclusion

In this paper, we have proposed a novel image-based facial animation method, which avoids the tedious work of 3D modeling process and does not need a large collection of facial image samples. Our method offers automatic facial modeling techniques and

generates photo-realistic facial animation with natural head rotations, which provides more convenient way for facial model construction and facial synthesis.
Our approach inspires new applications upon various platforms with lower computational power. The effectiveness and efficiency of our method promises feasibility of automatic facial modeling and animation implementations on mobile devices, which can provide better interactive applications concerning entertainment and intelligent communication services. Future work includes facial model enhancement by adding in hair, and porting our application onto smart mobile devices.

References

1. F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D.H. Salesin: Synthesizing Realistic Facial Expressions from Photographs. Proc. ACM SIGGRAPH'98, pp. 75-84, 1998.
2. Frederic Pighin, Richard Szeliski, David H. Salesin: Resynthesizing Facial Animation through 3D Model-Based Tracking. ICCV'99.
3. Volker Blanz, Thomas Vetter: A Morphable Model For The Synthesis Of 3D Faces. Proc. ACM SIGGRAPH'01.
4. Y. Lee, D. Terzopoulos, K. Waters: Realistic modeling for facial animation. Proc. ACM SIGGRAPH'95, pp. 55-62, 1995.
5. X. Wei, Z. Zhu, L. Yin, Q. Ji: A Real Time Face Tracking And Animation System. Proc. CVPR Workshop on Face Processing in Video (FPIV'04), 2004.
6. P. Fua, C. Miccio: Animated Heads from Ordinary Images - A Least Squares Approach. Computer Vision and Image Understanding, Vol.75, No. 3, pp. 247-259, 1999.
7. Yu Zhang, Edmond C. Prakash, Eric Sung: A New Physical Model with Multilayer Architecture for Facial Expression Animation Using Dynamic Adaptive Mesh. VCG'04.
8. W. Lee, N. Magnenat-Thalmann: Generating a population of animated faces from pictures. Proc. IEEE Workshop on Modeling People (ICCV'99 Workshop mPeople), Sep. 20, 1999.
9. T. Ezzat, T. Poggio: Facial analysis and synthesis using image-based models. Proc. Intl. Conf. on Automatic Face and Gesture Recognition, October 1996.
10. T. Ezzat, T. Poggio: Visual Speech Synthesis by Morphing Visemes. IJCV'00.
11. T. Ezzat, G. Geiger, T. Poggio: Trainable Videorealistic Speech Animation. SIGGRAPH'02.
12. T. Vetter: Synthesis of novel views from a single face image. IJCV'98.
13. Q. Zhang, Z. Liu, B. Guo, H. Shum: Geometry-driven photorealistic facial expression synthesis. Proc. ACM Symp. On Computer Animation, San Diego, CA, July, 2003.
14. E. Cosatto, H. P. Graf: Sample-based synthesis of photo-realistic talking heads. Computer Animation, 1998, pp. 103-110.
15. A. M. Tekalp, J. Ostermann: Face and 2D mesh animation in MPEG4. Image Communication Journal, Tutorial Issue on MPEG-4 Standard, Elsevier, 2000.
16. D. Fidaleo, U. Neumann: CoArt: Co-articulation Region Analysis for Control of 2D Characters. Proc. Computer Animation, vol. 00, pp. 17-22, 2002
17. L. Yin: Generating Realistic Facial Expressions with Wrinkles for Model-Based Coding. Computer Vision and Image Understanding, Vol. 84, No. 2, pp. 201-240, Nov. 2001.
18. D. Terzopoulos, Y. Lee, M. Vasilescu: Model-based and image-based methods for facial image synthesis, analysis and recognition. Proc. IEEE Conf. Automatic Face and Gesture Recognition (FGR'04).
19. X. Hou, S.Z. Li, H. Zhang, Q. Cheng: Direct Appearance Models. IEEE Proc. Conf. on Computer Vision and Pattern Recognition, Vol. 1, pp. 828-833, Dec. 2001.

20. C. Xu and J. L. Prince, Snakes, Shapes, and Gradient Vector Flow, IEEE Trans. Image Processing, pp. 359-369, March, 1998.
21. F. Lavagetto and R. Pockaj, The Facial animation Engine: Toward a High-level Interface for the Design of MPEG-4 Compliant Animated Faces, IEEE Trans. Circuits and Systems for Video Technology, vol. 9, no. 2, pp. 277-289, 1999.
22. F. Dornaika, J. Ahlberg: Face Model Adaptation using Robust Matching and Active Appearance Models. WACV'02: 3-7.