A Fully Automatic System To Model Faces From a Single Image

Zicheng Liu
Microsoft Research

August 2003

Technical Report
MSR-TR-2003-55

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

Abstract We present a system which automatically generates a 3D face model from a single frontal image of a face. Our system consists of two components. The first component is the feature feature detection, and the second component is the model fitting. We use an existing feature detection software for the first component. After we detect the face features, we fit a 3D face model by using a linear space of face geometries and assuming an orthogonal projection. Even though the depths of the resulting face models are usually not accurate, the models look recognizable as long as the view angle is not too far from the front. Our system has the advantage that it is fully automatic, robust, and fast. It can be used in a variety of applications for which the accuracy of depths are not critical such as games, avatars, face recognition on close-to-front-view images, etc.

1 Introduction

There has been a lot of work on face modeling from images. One technique which has been used in many commercial system is to use two orthogonal views [1, 6, 3]: one frontal view and one side view. Such systems require the users to manually specify the face features on the two images. Some of the commercial systems have tried to use some feature findings to reduce the amount of manual work.

Another type of face modeling system creates face models from a video sequence [4, 8]. Since it has images of multiple views, it can potentially compute the correct depth and can generate a texture image for the entire face. However, it requires the user to have a video camera. In addition, it usually requires some amount of user input to make it robust. In the situations where the user only has a single image, our system is more useful.

Blanz and Vetter [2] developed a system to create face models from a single image. Their system uses both a geometry database and an image database. It can only model the faces whose skin colors are covered by their database. Their system is computationally more expensive.

2 Linear class of face geometries

We the same representation for the face models as in [8]. A face is represented as a linear combination of a neutral face and some number of face *metrics*. A metric is vector that linearly deforms a face in certain way, such as to make the head wider, make the nose bigger, etc. Let us denote the face geometry by a vector $\mathcal{S} = (\mathbf{v}_1^T, \ldots, \mathbf{v}_n^T)^T$ where $\mathbf{v}_i = (X_i, Y_i, Z_i)^T$ $(i = 1, \ldots, n)$ are the vertices, and a metric by a vector $\mathcal{M} = (\delta\mathbf{v}_1, \ldots, \delta\mathbf{v}_n)^T$, where $\delta\mathbf{v}_i = (\delta X_i, \delta Y_i, \delta Z_i)^T$. Given a neutral face $\mathcal{S}^0 = (\mathbf{v}_1^{0\,T}, \ldots, \mathbf{v}_n^{0\,T})^T$, and a set of m metrics $\mathcal{M}^j = (\delta\mathbf{v}_1^{j\,T}, \ldots, \delta\mathbf{v}_n^{j\,T})^T$, the linear space of face geometries spanned by these metrics is

$$\mathcal{S} = \mathcal{S}^0 + \sum_{j=1}^{m} c_j \mathcal{M}^j \quad \text{subject to } c_j \in [l_j, u_j] \tag{1}$$

where c_j's are the metric coefficients and l_j and u_j are the valid range of c_j. The neutral face and all the metrics are designed by an artist, and it is done once. The neutral face (see Figure 1) contains 194 vertices and 360 triangles. There are 65 metrics.

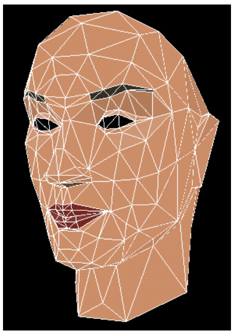

Figure 1: Neutral face.

3 Face modeling from a single view

3.1 Face feature alignment

Given an image of a face, to find the feature points on the face, we first use an existing face detector software [7] to detect the face. We then use the face alignment software by Yan et al [9] to find the face features. Figure 2 shows an input image and the alignment result.

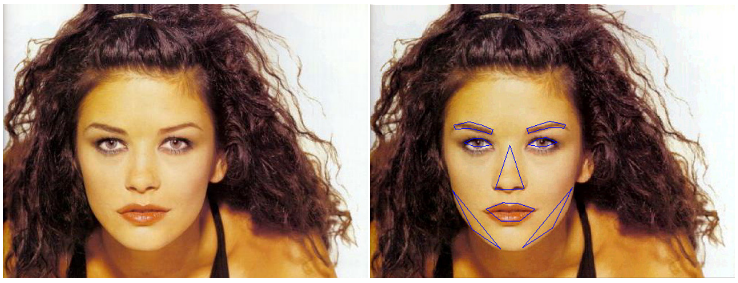

Figure 2: *Left: Input image. Right: The result from image alignment.*

3.2 Model fitting

We assume that the projection is orthogonal, and there is no out of the plane rotations for the face. Without loss of generality, let us denote $\mathbf{v}_i = (X_i, Y_i, Z_i)^T$,($i = 1, \ldots, f$) to be the feature points. Denote $\bar{v}_i = (X_i, Y_i)$ to be the projection of $\mathbf{v}_i$

on the XY plane. For each feature point $\mathbf{v}_i$, denote $\mathbf{m}_i$ to be its corresponding coordinate on the input image. Let $\mathbf{R}$ denote the 2x2 rotation matrix, $\mathbf{t}$ be the 2D translation vector, and s be the scale. We then have the following equation:

$$s\mathbf{R}\bar{\mathbf{v}}_i + \mathbf{t} = \mathbf{m}_i \tag{2}$$

From equation 1, we have

$$\bar{\mathbf{v}}_i = \bar{\mathbf{v}}_i^0 + \sum_{j=1}^{m} c_j \delta \bar{\mathbf{v}}_i^j \tag{3}$$

Therefore

$$s\mathbf{R}(\bar{\mathbf{v}}_i^0 + \sum_{j=1}^{m} c_j \delta \bar{\mathbf{v}}_i^j) + \mathbf{t} = \mathbf{m}_i \tag{4}$$

We solve this equation iteratively. We first estimate the s, $\mathbf{R}$, and $\mathbf{t}$ by assuming c_i to be zero vector. This is done by using the technique as described in [5]. Then we fix s, $\mathbf{R}$, and$\mathbf{t}$, and equation 4 becomes a linear system which can be solved by using a linear least square procedure. We can then re-estimate s, $\mathbf{R}$, and $\mathbf{t}$ by using the new estimates of c_i's, and so on. In our experiments, we find that one or two iterations are usually sufficient.

4 Results

Figure 3 shows the different views of the reconstructed 3D model based on the input image in Figure 2. We can see the frontal view (the image in the middle) looks very good as expected. There are quite large rotations for the images on the left and right. These two images still look quite recognizable. Once the models are constructed, we can immediately animate the face including generating different expressions and using text-to-speech to create lipsynced animation. We use the same mechanism as in [4, 8] to create facial animations. Figure 4 shows the three facial expressions generated by our system.

The images we use are 640x480. The total computation time for each image is about 7 seconds on a 1.7GHz PC. The main computation cost is the face alignment program.

5 Conclusion

We have presented a system to construct 3D face models for a single front image. The system is fully automatic. It is fast compared to the other face modeling systems. Furthermore, it is very robust. It can be used to construct personalized face models for games, online chat, etc. It can also be used as a tool to generate database of faces with various poses which are needed by the face recognition systems.

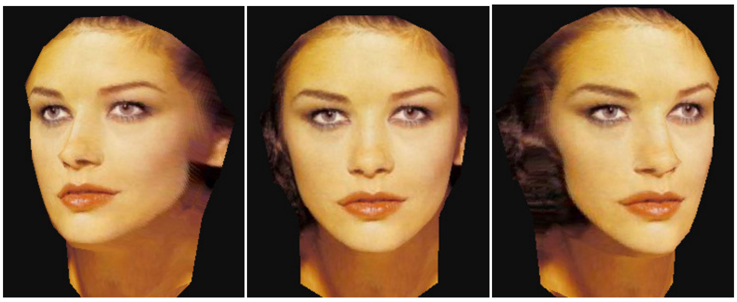

Figure 3: *The different views of the 3D model generated from the input image in Figure2.*

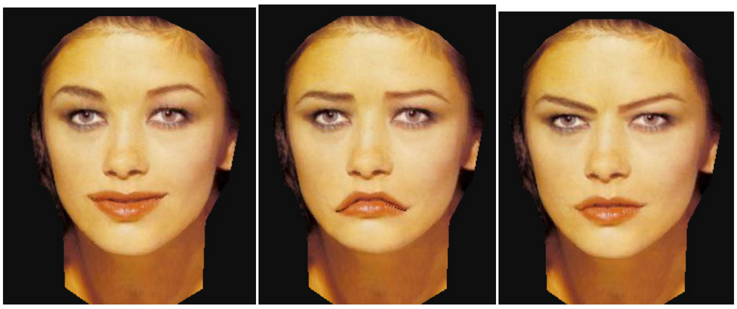

Figure 4: *Generating different expressions for the constructed face model.*

References

[1] T. Akimoto, Y. Suenaga, and R. S. Wallace. Automatic 3d facial models. *IEEE Computer Graphics and Applications*, 13(5):16–22, September 1993.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Computer Graphics, Annual Conference Series*, pages 187–194. Siggraph, August 1999.

[3] B. Dariush, S. B. Kang, and K. Waters. Spatiotemporal analysis of face profiles: Detection, segmentation, and registration. In *Proc. of the 3rd International Conference on Automatic Face and Gesture Recognition*, pages 248–253. IEEE, April 1998.

[4] P. Fua and C. Miccio. From regular images to animated heads: A least squares approach. In *Eurographics of Computer Vision*, pages 188–202, 1996.

[5] B. K. Horn. Closed-form Solution of Absolute Orientation using Unit Quaternions. *Journal of the Optical Society A*, 4(4):629–642, Apr. 1987.

[6] H. H.S.Ip and L. Yin. Constructing a 3d individualized head model from two orthogonal views. *The Visual Computer*, (12):254–266, 1996.

[7] S. Z. Li and L. Gu. Real-time multi-view face detection, tracking, pose estimation, alignment, and recognition. In *IEEE Conf. on Computer Visioin and Pattern Recognition Demo Summary*, 2001.

[8] Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen. Rapid modeling of animated faces from video. *Journal of Visualization and Computer Animation*, 12(4):227–240, Sep. 2001.

[9] S. Yan and et al. Ranking prior local confidence model for statistical shape localization. In *submitted*, 2003.