Recovering Non-Rigid 3D Shape from Image Streams

Christoph Bregler* Aaron Hertzmann Henning Biermann
Media Research Laboratory
Department of Computer Science
New York University
719 Broadway, 12th Floor
New York, NY 10003
{bregler,hertzman,biermann}@mrl.nyu.edu
NYU Computer Science Technical Report

June 16, 1999

Abstract

This paper addresses the problem of recovering 3D non-rigid shape models from image sequences. For example, given a video recording of a talking person, we would like to estimate a 3D model of the lips and the full head and its internal modes of variation. Many solutions that recover 3D shape from 2D image sequences have been proposed; these so-called structure-from-motion techniques usually assume that the 3D object is rigid. For example, Tomasi and Kanade's factorization technique is based on a rigid shape matrix, which produces a tracking matrix of rank 3 under orthographic projection. We propose a novel technique based on a non-rigid model, where the 3D shape in each frame is a linear combination of a set of basis shapes. Under this model, the tracking matrix is of higher rank, and can be factored in a three step process to yield to pose, configuration and shape. We demonstrate this simple but effective algorithm on video sequences of speaking people. We were able to recover 3D non-rigid facial models with high accuracy.

1 Introduction

This paper demonstrates a new technique for recovering 3D non-rigid shape models from 2D image sequences recorded with a single camera. For example, this technique can be applied to video recordings of a talking person. It extracts a 3D model of the human face, including all facial expressions and lip movements.

Previous work has treated the two problems of recovering 3D shapes from 2D image sequences and of discovering a parameterization of non-rigid shape deformations separately. Most techniques that address the *structure-from-motion* problem are limited to rigid objects. For example, Tomasi and Kanade's factorization technique [8] recovers a shape matrix from image sequences. Under orthographic projection, it can be shown that the 2D tracking data matrix has rank 3 and can be factored into 3D pose and 3D shape with

*Authors current address: Gates 138, 353 Serra Mall, Stanford University, Stanford, CA 94305-9010, bregler@stanford.edu

the use of the singular value decomposition (SVD). Unfortunately these techniques can not be applied to nonrigid deforming objects, since they are based on the rigidity assumption.

Most techniques that learn models of shape variations do so on the 2D appearance, and do not recover 3D structure. Popular methods are based on Principal Components Analysis. If the object deforms with K linear degrees of freedom, the covariance matrix of the shape measurements has rank K. The principal modes of variation can be recovered with the use of SVD.

We show how 3D non-rigid shape models can be recovered under scaled orthographic projection. The 3D shape in each frame is a linear combination of a set of K basis shapes. Under this model, the 2D tracking matrix is of rank $3K$ and can be factored into 3D pose, object configuration and 3D basis shapes with the use of SVD. We demonstrated the effectiveness of this technique on several data sets, including challenging recordings of human faces during speech and varying facial expressions.

Section 2 summarizes related approaches, Section 3 describes our algorithm, and Section 4 discusses our experiments.

2 Previous Work

Many methods have been proposed to solve the *Structure-from-motion* problem. One of the most influential of these was proposed by Tomasi and Kanade [8] who demonstrated the factorization method for rigid objects and orthographic projections. Many extensions have been proposed, such as the multi-body factorization method of Coseira and Kanade [4] that relaxes the rigidity constraint. In this method, K independently moving objects are allowed, which results in a tracking matrix of rank $3K$ and a permutation algorithm that identifies the submatrix corresponding to each object. More recently, Bascle and Blake [1] proposed a solution for factoring facial expressions and pose during tracking. Although it exploits the bilinearity of 3D pose and nonrigid object configuration, it requires a set of basis images selected before factorization is performed. The discovery of these basis images is not part of their algorithm.

Various authors have demonstrated estimation of non-rigid appearance in 2D using Principal Components Analysis [9, 6, 3].

Basu [2] demonstrates how the parameters can be iteratively fitted to a video sequence, starting from an initial lip model.

[7, 5] propose methods for recovering 3D facial models and their expressions from multiple images. These methods require key-frame images to be hand-selected, and the 3D reconstruction requires either user interaction or the placement of fiducials on the subject's face.

None of these techniques are able to estimate nonrigid 3D shape models from single-view 2D video streams without any initialization. In the next section, we demonstrate how this task can be solved by multiple factorization steps.

3 Factorization Algorithm

We describe the shape of the non-rigid object as a key-frame basis set $S_1, S_2, ...S_k$. Each key-frame S_i is a $3 \times P$ matrix describing P points. The shape of a specific configuration is a linear combination of this basis set:

$$S = \sum_{i=1}^{K} l_i \cdot S_i \qquad S, S_i \in I\!R^{3\times P}, l_i \in I\!R \tag{1}$$

Under a scaled orthographic projection, the P points of a configuration S are projected into $2D$ image points (u_i, v_i):

$$\begin{bmatrix} u_1 & u_2 & ... & u_P \\ v_1 & v_2 & ... & v_P \end{bmatrix} = R \cdot \left(\sum_{i=1}^{K} l_i \cdot S_i \right) + T \tag{2}$$

$$R = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \end{bmatrix} \tag{3}$$

R contains the first 2 rows of the full $3D$ camera rotation matrix, and T is the camera translation. The scale of the projection is coded in $l_1, ...l_K$. As in Tomasi-Kanade, we eliminate T by subtracting the mean of all $2D$ points, and henceforth can assume that S is centered at the origin.

We can rewrite the linear combination in (2) as a matrix-matrix multiplication:

$$\begin{bmatrix} u_1 & ... & u_P \\ v_1 & ... & v_P \end{bmatrix} = \begin{bmatrix} l_1 R & ... & l_K R \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ ... \\ S_K \end{bmatrix} \tag{4}$$

We add a temporal index to each $2D$ point, and denote the tracked points in frame t as $(u_i^{(t)}, v_i^{(t)})$. We assume we have $2D$ point tracking data over N frames and code them in the tracking matrix W:

$$W = \begin{bmatrix} u_1^{(1)} & ... & u_P^{(1)} \\ v_1^{(1)} & ... & v_P^{(1)} \\ u_1^{(2)} & ... & u_P^{(2)} \\ v_1^{(2)} & ... & v_P^{(2)} \\ & ... & \\ u_1^{(N)} & ... & u_P^{(N)} \\ v_1^{(N)} & ... & v_P^{(N)} \end{bmatrix}$$

Using (4) we can write:

$$W = \underbrace{\begin{bmatrix} l_1^{(1)} R^{(1)} & ... & l_K^{(1)} R^{(1)} \\ l_1^{(2)} R^{(2)} & ... & l_K^{(2)} R^{(2)} \\ & ... & \\ l_1^{(N)} R^{(N)} & ... & l_K^{(N)} R^{(N)} \end{bmatrix}}_{Q} \cdot \underbrace{\begin{bmatrix} S_1 \\ S_2 \\ ... \\ S_K \end{bmatrix}}_{B} \tag{5}$$

3.1 Basis Shape Factorization

Equation (5) shows that the tracking matrix has rank $3K$ and can be factored into 2 matrixes: Q contains for each time frame t the pose $R^{(t)}$ and configuration weights $l_1^{(t)}, ..., l_K^{(t)}$. B codes the K key-frame basis shapes S_i. The factorization can be done using singular value decomposition (SVD) by only considering the first $3K$ singular vectors and singular values (first $3K$ columns in U, D, V):

$$\textbf{SVD:}\quad W^{2N \times P} = \hat{U} \cdot \hat{D} \cdot \hat{V}^T = \hat{Q}^{2N \times 3K} \cdot \hat{B}^{3K \times P} \tag{6}$$

3.2 Factoring Pose from Configuration

In the second step, we extract the camera rotations $R^{(t)}$ and shape basis weights $l_i^{(t)}$ from the matrix $\hat{Q}$. Although $\hat{Q}$ is a $2N \times 3K$ matrix, it only contains $N(K+6)$ free variables. Consider the 2 rows of $\hat{Q}$ that correspond to one single time frame t, namely rows $2t-1$ and row $2t$ (for convenience we drop the time index (t)):

$$\begin{aligned} q^{(t)} &= \left[\begin{array}{ccc} l_1^{(t)} R^{(t)} & \ldots & l_K^{(t)} R^{(t)} \end{array}\right] \\ &= \left[\begin{array}{ccccccc} l_1 r_1 & l_1 r_2 & l_1 r_3 & \ldots & l_K r_1 & l_K r_2 & l_K r_3 \\ l_1 r_4 & l_1 r_5 & l_1 r_6 & \ldots & l_K r_4 & l_K r_5 & l_K r_6 \end{array}\right] \end{aligned}$$

We can reorder the elements of $q^{(t)}$ into a new matrix $\bar{q}^{(t)}$:

$$\begin{aligned} \bar{q}^{(t)} &= \left[\begin{array}{cccccc} l_1 r_1 & l_1 r_2 & l_1 r_3 & l_1 r_4 & l_1 r_5 & l_1 r_6 \\ l_2 r_1 & l_2 r_2 & l_2 r_3 & l_2 r_4 & l_2 r_5 & l_2 r_6 \\ & & & \ldots & & \\ l_K r_1 & l_K r_2 & l_K r_3 & l_K r_4 & l_K r_5 & l_K r_6 \end{array}\right] \\ &= \left[\begin{array}{c} l_1 \\ l_2 \\ \ldots \\ l_K \end{array}\right] \cdot \left[\begin{array}{cccccc} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 \end{array}\right] \end{aligned}$$

which shows that $\bar{q}^{(t)}$ is of rank 1 and can be factored into the pose $\hat{R}^{(t)}$ and configuration weights $l_i^{(t)}$ by SVD. We successively apply the reordering and factorization to all time blocks of $\hat{Q}$.

3.3 Adjusting Pose and Shape

In the final step, we need to enforce the orthonormality of the rotation matrices. As in [8], a linear transformation G is found by solving a least squares problem[1]. The transformation G maps all $\hat{R}^{(t)}$ into an orthonormal $R^{(t)} = \hat{R}^{(t)} \cdot G$. The inverse transformation must be applied to the key-frame basis $\hat{B}$ to keep the factorization consistent: $S_i = G^{-1} \cdot \hat{S}_i$.

We are now done. Given $2D$ tracking data W, we can estimate a non-rigid 3D shape matrix with K degrees of freedom, and the corresponding camera rotations and configuration weights for each time frame.

[1]The least squares problem enforces orthonomality of all $R^{(t)}$: $[r_1 r_2 r_3] G G^T [r_1 r_2 r_3]^T = 1$, $[r_4 r_5 r_6] G G^T [r_4 r_5 r_6]^T = 1$, $[r_1 r_2 r_3] G G^T [r_4 r_5 r_6]^T = 0$

4 Experiments

This work is motivated by our efforts in image-based facial animation. In order to test these methods, we collected video of people speaking sentences with various facial expressions. The video recordings contain rigid head motions, and non-rigid lip, eye, and other facial motions. We tracked important facial features with an appearance-based 2D tracking technique[2]. Figure 1 and 7 shows example tracking results for video-1 and video-2. For facial animation, we want explicit control over the rigid head pose and the implicit facial variations. In the following, we show how we were able to extract a 3D non-rigid face model parameterized by these degrees of freedom.

We applied our method to two different video sequences. The first is a public broadcast originally recorded on film in the early 1960's (video-1) and contains 1213 video frames. The second video was recorded in our lab (video-2) and contains 1000 video frames. Both recordings are challenging for 3D reconstructions, since they contain very few out-of-plane head motions. In a first experiment, we computed the reconstruction error based on the number of degrees of freedom (K) for video-1. We factorized the tracking data, and computed the back-projection of the estimated model, configuration, and pose into the image. Figure 2 shows the SSD error between the backprojected points and image measurements. For $K = 16$ the error vanishes. For the remainder of the paper, we set $K = 16$. Figure 3 and 4 shows for example frames of video-1 and the reconstructed 3D S matrix rotated by the corresponding $R^{(t)}$. To illustrate the 3D data better, we fit a shaded smooth surface to the 3D shape points.

We also investigated the discovered modes of variation. We computed the mean and standard deviations of $[l_1^t, ..., l_K^t]$ in video-1. Figure 5 and 6 shows ± 4 standard deviations of the second and third modes (S_1, S_2, S_3). Mode 1 covers scale change, mode 2 cover some aspect of mouth opening, and mode 3 covers eye opening. The remaining modes pick up more subtle and less intuitive variations.

Figure 8 shows the reconstruction results for video-2.

The results on these 2 video databases are very encouraging. Given the limited range of out-of-plane face orientations, the 3D details that we could recover from the lip shape is quite surprising. We are planning to record a "ground-truth" video of non-rigid objects. This will allow us to quantify the exact reconstruction error in 3D.

5 Discussion

We have presented a simple but effective new technique for recovering 3D non-rigid shape models from 2D image streams. It is a three step procedure using multiple factorizations. We were able to recover 3D models for video recordings of talking people. Although these are very concouraging results, we plan to evaluate this technique and its limitations on a larger data set of man-made articulated objects. This problem will be somewhat easier than the face database, but it will give us ground-truth values for performance evaluations. Reconstructing non-rigid models from single-view video recordings has many potential applications. In addition, we intend to apply this technique to our image-based facial animation system and to a model based tracking system.

[2] We used a learned PCA-based tracker similar to [6]

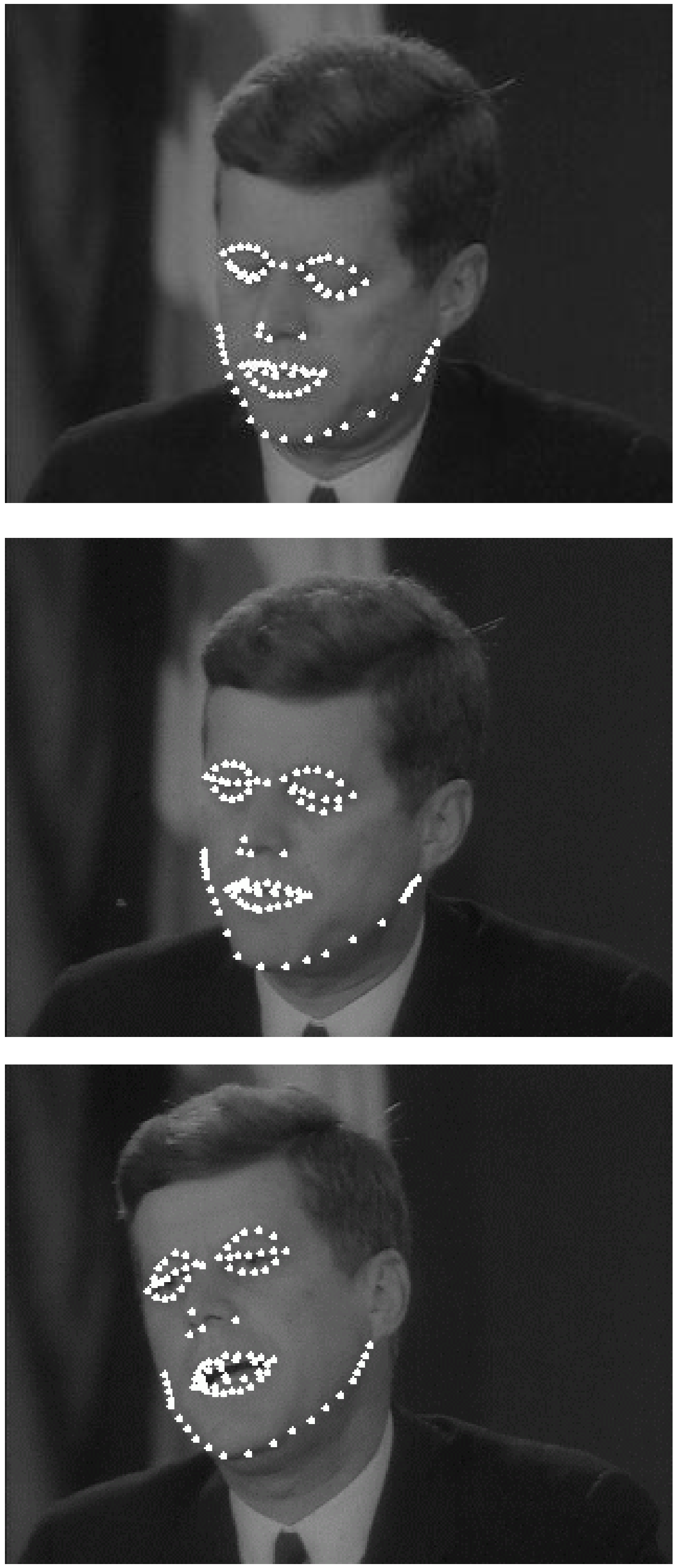

Figure 1: Example images from video-1 with overlayed tracking points. We track the eye brows, upper and lower eye lids, 5 nose points, outer and inner boundary of the lips, and the chin contour.

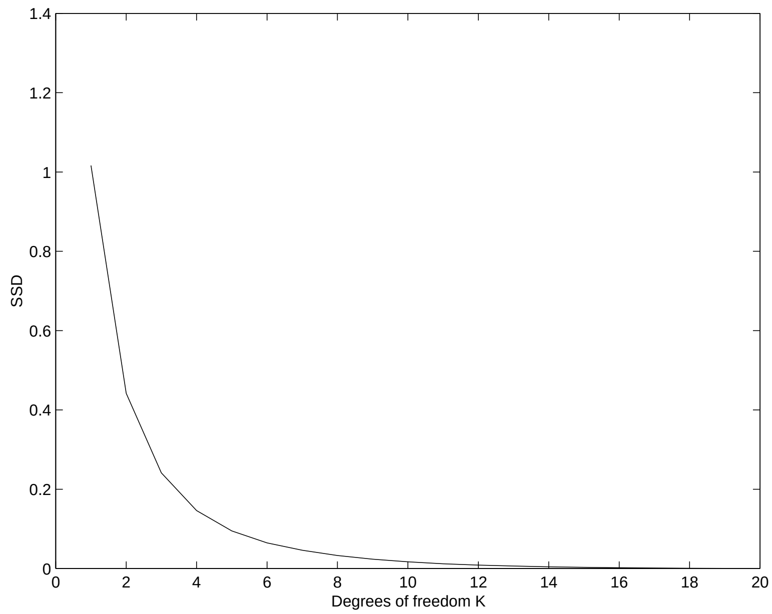

Figure 2: Average pixel SSD error of back-projected face model for different degrees of freedom: K

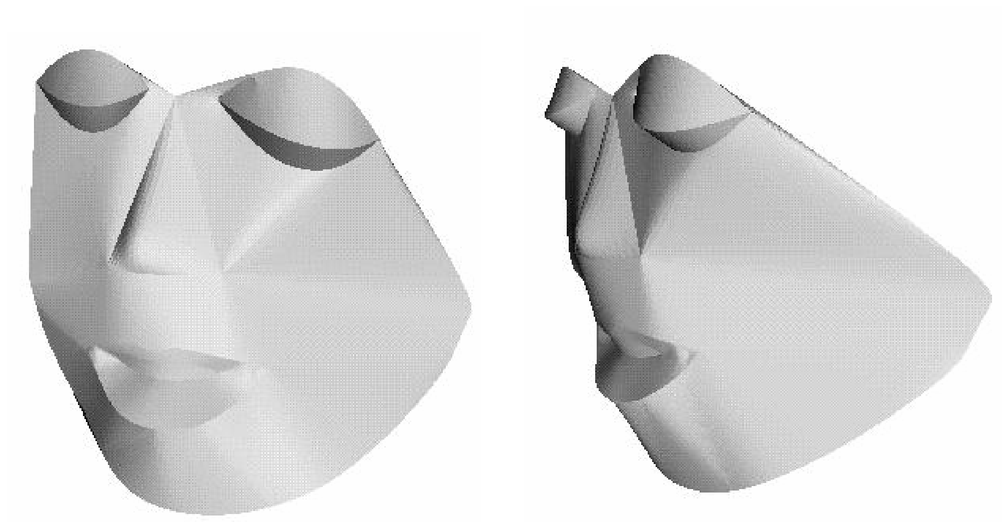

Figure 3: 3D reconstructed shape and pose for first frame of Figure 1

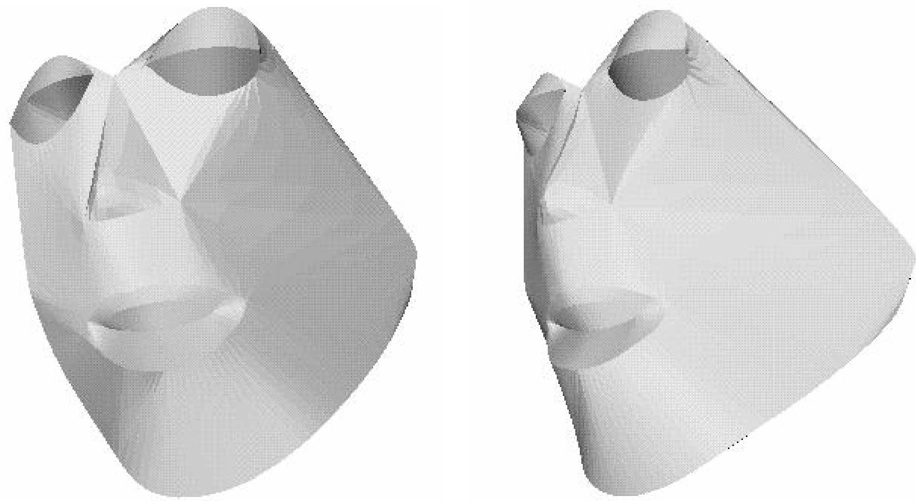

Figure 4: 3D reconstructed shape and pose for last frame of Figure 1

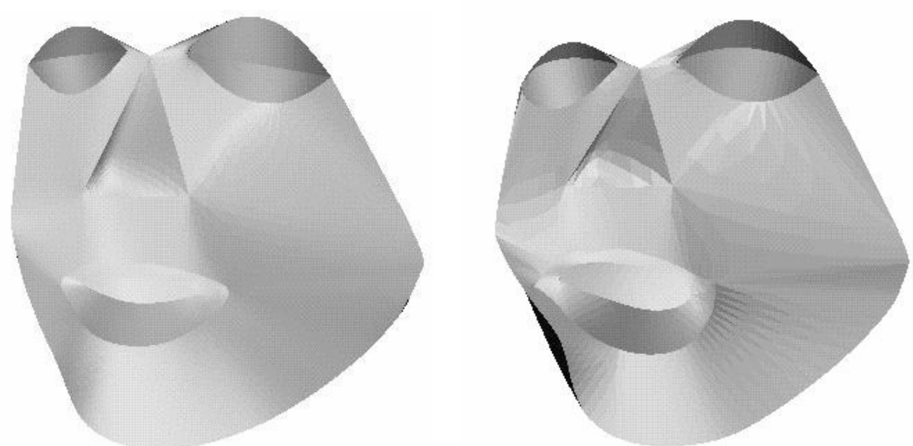

Figure 5: Variation along mode 2 of the nonrigid face model. The mouth deforms.

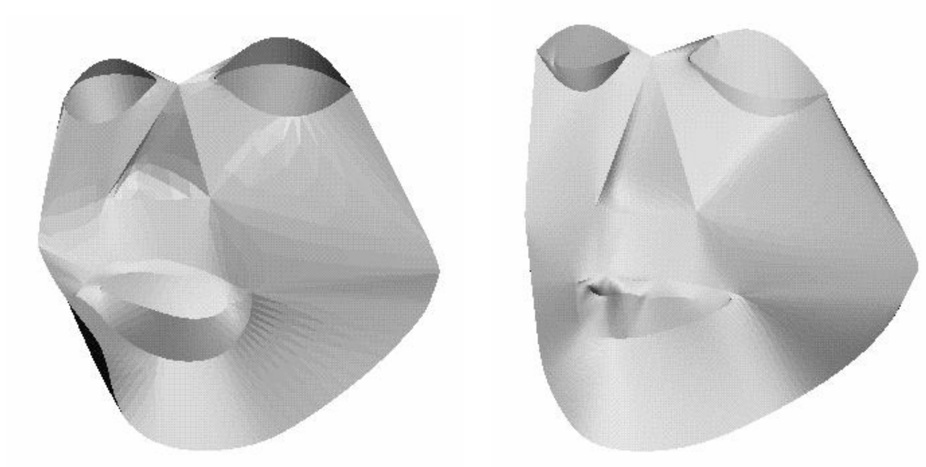

Figure 6: Variation along mode 3 of the nonrigid face model. The eyes close.

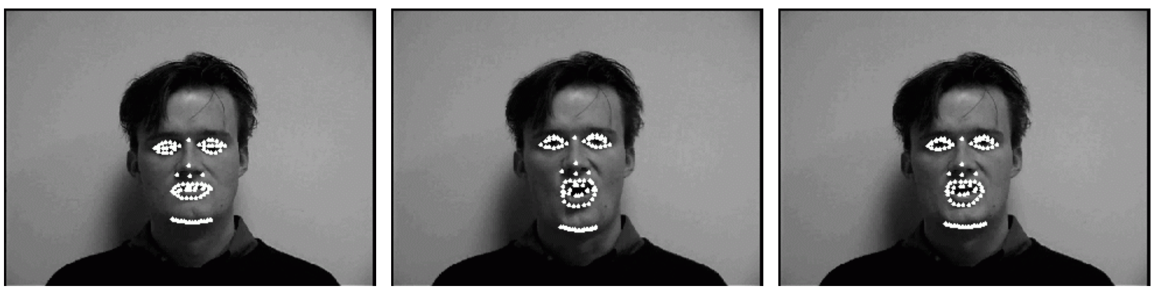

Figure 7: Example images from video-2 with overlayed tracking points.

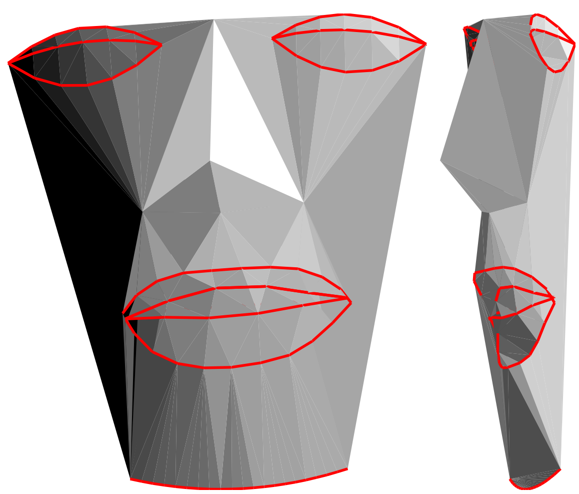

Figure 8: Front and side view for the reconstructions from video-2.

Acknowledgments

We like to thank Ken Perlin, Denis Zorin, and Davi Geiger for fruitful discussions, and for supporting this research, Clilly Castiglia and Steve Cooney for helping with the data collection, and New York University, California State MICRO program and Interval Research for partial funding.

References

[1] B. Bascle and A. Blake. Separability of pose and expression in facial tracking and animation. In *Proc. Int. Conf. Computer Vision*, 1998.

[2] S. Basu. A three-dimensional model of muman lip motion. In *EECS Master Thesis, MIT Media Lab Report 417*, 1997.

[3] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. In *J. Artificial Intelligence*, 1995.

[4] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. *Int. J. of Computer Vision*, pages 159–180, Sep 1998.

[5] Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Frédéric Pighin. Making faces. In Michael Cohen, editor, *SIGGRAPH 98 Conference Proceedings*, Annual Conference Series, pages 55–66. ACM SIGGRAPH, Addison Wesley, July 1998. ISBN 0-89791-999-8.

[6] A. Lanitis, Taylor C.J., Cootes T.F., and Ahmed T. Automatic interpretation of human faces and hand gestures using flexible models. In *International Workshop on Automatic Face- and Gesture-Recognition*, 1995.

[7] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. Synthesizing realistic facial expressions from photographs. In Michael Cohen, editor, *SIGGRAPH 98 Conference Proceedings*, Annual Conference Series, pages 75–84. ACM SIGGRAPH, Addison Wesley, July 1998. ISBN 0-89791-999-8.

[8] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. of Computer Vision*, 9(2):137–154, 1992.

[9] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.