An Integrated Approach to 3D Face Model Reconstruction from Video

Chia-Ming Cheng and Shang-Hong Lai

Department of Computer Science
National Tsing-Hua University
Hsinchu, Taiwan 30043
Email: lai@cs.nthu.edu.tw

Abstract

In this paper, we propose an integrated system to reconstruct 3D face models from monocular image sequences. Our approach is to adapt a generic 3D face model based on the sparse 3D geometric constraints recovered from a video sequence. The proposed face reconstruction system consists of face feature extraction/tracking, face pose estimation, structure from motion, structure from silhouette, model adaptation, and texture mapping. In the first step of our system, some face feature points in some representative frames in the image sequence are selected. Then these feature points are tracked for the entire sequence. An approximate face pose for each image frame is obtained by using an iterative pose estimation method with feature correspondences between each image and the 3D generic model. Then we use a robust bundle adjustment algorithm to recover 3D structure of a set of face feature points from the image sequence. The structure from silhouette algorithm is proposed to find the displacement vectors of the points on the 3D generic model corresponding to the sampled points on the silhouettes. Combining the 3D geometric constraints recovered from the bundle adjustment and structure from silhouette, we deform the 3D generic head model by using radial basis function interpolation to fit these geometric constraints, thus reconstructing a personalized 3D head model. After the 3D head model is reconstructed, we compute its texture map by integrating all the images in the sequence with appropriate weighting. Finally, experimental results of 3D face model recovery from the proposed system are shown in this paper.

1. Introduction

With the rapid growth of internet and multimedia communication, the demand for an easy and inexpensive solution to create three-dimensional models has been increasing. Human faces are the focus of attention to humans in most application domains. For example, the MPEG-4 standard [1] has designed the Synthetic Natural Hybrid Coding (SNHC) for efficiently transmitting video streams containing face images. A 3D sensor, such as a laser range scanner or a structured-light sensor has been extensively used to create an individualized three-dimensional head model. Unfortunately, using laser range sensors for 3D model reconstruction is expensive, time-consuming, and restricted to specific environmental settings. Due to these drawbacks, researchers have been pursuing the approach to reconstruct 3D models from multi-view images or video images. This approach has the advantages of inexpensive cost, short acquisition time, and general working requirement. In spite of these advantages, the imaging approach is in general provides less accurate measurements compared to the laser range sensor approach.

There exist many methods for reconstructing 3D head models from images. Some researchers developed methods to reconstruct a 3D model from a sequence of silhouettes [2-3]. This approach relies on accurate contour extraction or contour tracking in the images. Another popular approach [4-5] is to reconstruct an individualized head model from two orthogonal views. It is based on recovering the structure of selected feature points in the face and then adjusting a generic model by using these control points to obtain the individualized 3D head model. Kang [6] presented a structure from motion algorithm to reconstruct a 3D head model from a video sequence. His algorithm used a constrained deformable model to deform a generic 3D model to match the structure of the individual 3D head computed from the video. Fua and Miccio [7] integrated stereo matching, head silhouettes and 2D feature locations to recover 3D head models from a generic model. More recently, Fua [8] developed a regularized bundle-adjustment method to adapt a generic 3D head model from the 3D structure recovered from a video sequence. Blanz and Vetter [9] presented a morphable head model that was constructed from the principal component analysis of a large set of 3D head models to match the projection of the morphable model to a face image.

In this paper, we proposed a new individualized 3D head model reconstruction system that combines a robust bundle adjustment algorithm with a shape from silhouette approach to provide sparse 3D data locations and then deforms a generic 3D head model to interpolate the set of recovered 3D locations. The

0-7695-1074-4/01 $10.00 © 2001 IEEE

robust bundle adjustment algorithm involves robust estimation of 3D face structures from an image sequence. The shape from silhouette approach determines the 3D displacement vectors for the corresponding points on the generic model. Thus the generic 3D head model is deformed to the recovered 3D head structure by using the radial basis function interpolation. Finally, the texture map of the reconstructed 3D head model is computed by integrating all the images in the image sequence with appropriate weighting.

The remainder of this paper is organized as follows. An overview of the proposed system is given in the next section. Section 3 describes a robust bundle adjustment algorithm for recovering structure from an image sequence. The structure from silhouette approach is presented in section 4. We give the 3D generic head model adaptation method by using the radial basis function interpolation in section 4. The integrated texture mapping scheme is given in section 5. Some experimental results are given in section 6. Finally, we conclude this paper in section 7.

2. System Overview

The proposed 3D model reconstruction system consists of the following steps: feature extraction/tracking, face pose estimation, bundle adjustment, structure from silhouette, generic model adaptation and texture mapping. In this paper, the camera intrinsic parameters are first calibrated by using the easy calibration tool developed by Zhang [10]. However, it is possible to include the estimation of these parameters in the bundle adjustment minimization framework that we will describe in details in the next section. The flow diagram of the proposed 3D reconstruction system is shown in Figure 1.

In face feature extraction and tracking, the current system uses a semi-automatic way. Firstly, it requires manual selection of face feature locations from a sparse set of representative frames in the video sequence and the corresponding locations in the generic 3D head model. Then, a cross-correlation based feature tracking method is applied to track all these feature points in the rest of all images in the video sequence. In face pose approximation, we used a traditional pose estimation algorithm [11] to compute the pose for each frame in the video

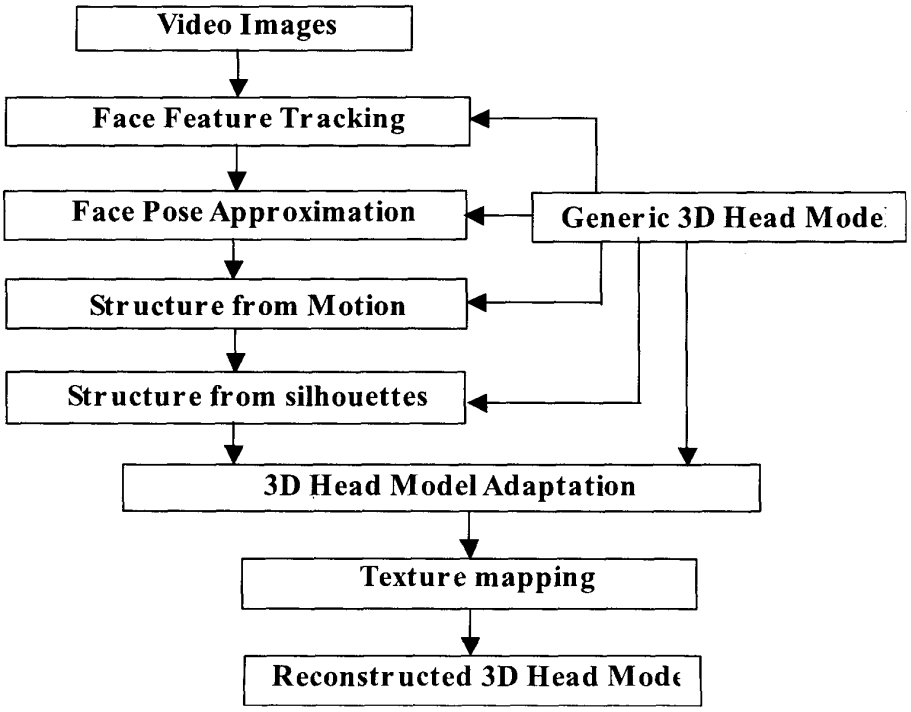


Figure 1. The flow diagram of the proposed 3D head model reconstruction system.

from the correspondences of the feature points in the image and the generic 3D model. Note that the 3D structure of the feature points in the generic model is simply an approximation to the true 3D structure of the head model, thus the computed pose estimate is a rough approximation. In the robust bundle adjustment algorithm, we apply a robust least-squares minimization algorithm to recover the individualized 3D face structure and pose for each frame. This robust least-squares formulation leads to solving a non-linear optimization problem, which requires a good initial solution to converge to the true solution. We used the previously estimated poses and the 3D locations of the feature points from the generic model as the initial guess. Then the structure from silhouette algorithm is applied to recover the 3D displacement vectors for the vertexes on the generic model corresponding to the silhouette on the face image. After the 3D structure of the face is recovered, we adapt the generic model to the recovered structure to obtain the individualized 3D face model. Finally, the texture map of the face is computed from all the images in the sequence.

3. Robust Bundle Adjustment

Let the 3D locations of the i-th feature point in a reference coordinate system be denoted by $\mathbf{X}_i$, where $i = 1, \mathrm{K}, N$ (N is the total number of feature points), and the corresponding 2D locations detected in the k-th frame be denoted by $\mathbf{x}_i^{(k)}$, where $k = 1, \mathrm{K}, M$ and M is the total number of frames in the sequence. The pose of the 3D face model in the k-th frame is represented by the rotation angles $\theta_x^{(k)}$, $\theta_y^{(k)}$, $\theta_z^{(k)}$ as well as the translation vector $\mathbf{T}^{(k)}$. The rotation matrix induced by the rotation angles $\theta_x^{(k)}$, $\theta_y^{(k)}$, $\theta_z^{(k)}$ is represented by $\mathbf{R}(\theta_x^{(k)}, \theta_y^{(k)}, \theta_z^{(k)})$. Let the 3D coordinate of the feature point $\mathbf{X}_i$ at the specified pose be denoted by $\hat{\mathbf{X}}_i^{(k)} = (\hat{X}_i^{(k)}, \hat{Y}_i^{(k)}, \hat{Z}_i^{(k)})$. Then they are related by the rigid transformation as follows:

$$\hat{\mathbf{X}}_i^{(k)} = \mathbf{R}(\theta_x^{(k)}, \theta_y^{(k)}, \theta_z^{(k)})\, \mathbf{X}_i + \mathbf{T}^{(k)} \qquad (1)$$

The perspective projection of the 3D face feature point $\hat{\mathbf{X}}_i^{(k)}$ under the associated pose is given by

$$\begin{pmatrix} \mathbf{u}_i^{(k)} & 1 \end{pmatrix}^T = \mathbf{K} \begin{pmatrix} \hat{X}_i^{(k)} / \hat{Z}_i^{(k)} & \hat{Y}_i^{(k)} / \hat{Z}_i^{(k)} & 1 \end{pmatrix}^T \qquad (2)$$

Note that the camera calibration matrix $\mathbf{K}$ is upper-triangular and assumed to be known from camera calibration. The robust estimation formulation leads to minimizing the following energy function

$$E'(\theta^{(1)}, \mathrm{K}, \theta^{(M)}, \mathbf{T}^{(1)}, \mathrm{K}, \mathbf{T}^{(M)}, \mathbf{X}_1, \mathrm{K}, \mathbf{X}_N) = \sum_{k=1}^{M} \sum_{i=1}^{N} \delta_i^{(k)} \rho(\mathbf{u}_i^{(k)} - \mathbf{x}_i^{(k)}) \qquad (3)$$

where $\delta_i^{(k)}$ is a Boolean variable indicating the presence of the i-th feature at the k-th frame and the robust ρ function is an extension of the Lorentzion (or Cauchy) function commonly used in robust statistics [12]. The minimization of the robust energy function in equation (3) can be achieved by the iteratively re-weighted least-square minimization [13]. The Levenberg-Mardquart algorithm [14] is employed to update the solution. The nonlinear optimization problem requires a good initial solution to converge to the true solution. In our algorithm, we used the corresponding locations of the face feature points in a generic 3D head model and the estimated head poses for each frame as the initial solution for the iterative minimization process.

4. Structure from Silhouette

After the bundle adjustment process, we get the estimated 3D locations for the feature points and the head poses for all the frames. To further improve the model, we use the silhouette information of the face image and the generic model with its corresponding pose to provide additional 3D constraints for the head model. These new data constraints will be integrated with the recovered 3D structure for the adaptation of the generic model.

In the beginning, we extract the occluding contour in the face image interactively by using the snake technique [15]. For each sampled points along the contour, the 3D line across the lens center and the contour point on the image plane in the world coordinates can be decided via perspective projection. Then, we have to determine an appropriate pair of corresponding vertices with one located on the line and the other on the generic model.

We used two criteria to initilaize the correspondence pairs from the silhouette information. The first criterion is the orthogonality between the surface normal of the point on the generic model and 3D line direction; and the other is the distance between these two correspondence points should be minimized. Therefore, we search the vertices among the mesh of the generic model that satisfy the first criterion as candidate vertices, then we choose the one from all possible candidates with the minimum distance from the 3D projection line as our initial guess.

To solve the ambiguity corresponding problem, we proposed a 3D active contour model to approach the solution. Let an occluding contour in an image be represented by a set of

ordered nodal points $\{(u_i, v_i) \mid i = 1,\text{K}\ ,m\}$. This occluding contour is extracted by a traditional snake technique. There exists an inversely perspective projected line in 3D for each 2D nodal point. Assume these corresponding 3D lines be denoted by $(U_{0,i} + U_i t, V_{0,i} + V_i t, W_{0,i} + W_i t)$, $i = 1,\text{K}\ ,m$, where each direction (U_i, V_i, W_i) is a unit vector. The 3D curve on the generic model is denoted by a set of ordered nodal points $\{(X_i, Y_i, Z_i) \mid i = 1,\text{K}\ ,n\}$. We used two criteria to determine the corresponding 3D curves. The first criterion is the orthogonality between the surface normal of the nodal point on the generic model and 3D inversely projected line direction; and the other is the distance between the corresponding points should be small. To achieve this, we used an energy-minimization spline approach to obtain the 3D curve. This energy function is given as follows:

$$\begin{aligned} & E(X_i, Y_i, Z_i \mid i = 1,\text{K}\ ,n) \\ & = \sum_{i=1}^{n} \{ \min_{1 \le j \le m} \left| (N_x(i), N_y(i), N_z(i)) \bullet (U_j, V_j, W_j) \right| \\ & + \min_{1 \le j \le m} \left\| (u_j, v_j) - (x_i, y_i) \right\|^2 \} \\ & + \alpha \sum_{i=2}^{n} \left\| (X_i, Y_i, Z_i) - (X_{i-1}, Y_{i-1}, Z_{i-1}) \right\|^2 \\ & + \beta \sum_{i=2}^{n} \left\| (X_{i-1}, Y_{i-1}, Z_{i-1}) - 2(X_i, Y_i, Z_i) + (X_{i+1}, Y_{i+1}, Z_{i+1}) \right\|^2 \end{aligned} \tag{4}$$

where $(N_x(i), N_y(i), N_z(i))$ is the surface normal direction for the point (X_i, Y_i, Z_i) on the generic model, (x_i, y_i) is the perspective projected coordinate of the point (X_i, Y_i, Z_i), α and β are regularization parameters controlling the degree of membrane and thin-plate smoothness of the curve, respectively. The energy function given in equation **(4)** is very nonlinear. Gradient-based minimization methods require a good initial guess to converge to an optimal solution. In our implementation, we applied a gradient decent algorithm to update the solution. The initial curve is computed by first finding the corresponding vertex in the generic model for each nodal point in the occluding contour such that its surface normal is approximately orthogonal to the inversely projected 3D line and the projected distance is minimal. To accomplish this, we find the triangles in the generic model that contain different-signed inner products of the surface normal directions at their vertices and the inverse projected 3D line direction and set all the vertices associated with these triangles as candidate vertices. From these candidate vertices, we find the vertex whose projection is closest to the nodal point in the occluding contour. After all the vertices corresponding to the nodal points in the occluding contour are determined, they can be re-parameterized and form an initial 3D curve on the generic 3D head model. Figure 2 depicts an example showing the 2D occluding contour and the 3D curve on the generic model computed by using the described 3D snake algorithm.

5. Generic Model Adaptation

In this section, we describe the algorithm for adapting the generic 3D model based on the 3D structure recovered from the bundle adjustment and structure from silhouette. Our algorithm is based on the radial basis function (RBF) interpolation in 3D space. Let the recovered 3D feature points be denoted by $\mathbf{V}^{(i)} = (V_x^{(i)}, V_y^{(i)}, V_z^{(i)})$ and the corresponding locations in the generic model be denoted by $\mathbf{P}^{(i)} = (P_x^{(i)}, P_y^{(i)}, P_z^{(i)})$ for $i = 1,\text{K}\ ,N$. Assume the vertex $\mathbf{P} = (P_x, P_y, P_z)$ in the generic model is related to the corresponding new vertex $\mathbf{P}' = (P_x', P_y', P_z')$ in the adapted model as follows:

$$\begin{aligned} \mathbf{P}' = [& \sum_{i=1}^{N} \alpha_{xi} F(\left\| \mathbf{P} - \mathbf{P}^{(i)} \right\|) + a_x + b_x P_x + c_x P_y + d_x P_z, \\ & \sum_{i=1}^{N} \alpha_{yi} F(\left\| \mathbf{P} - \mathbf{P}^{(i)} \right\|) + a_y + b_y P_x + c_y P_y + d_y P_z, \\ & \sum_{i=1}^{N} \alpha_{zi} F(\left\| \mathbf{P} - \mathbf{P}^{(i)} \right\|) + a_z + b_z P_x + c_z P_y + d_z P_z] \end{aligned} \tag{5}$$

where $F(\ )$ is a radial basis function that is given by $F(r) = e^{-r/R}$ with a constant R. The coefficients in the above transformation, i.e. $\acute{a}_{xi}, \acute{a}_{yi}, \acute{a}_{zi}$, for $1 \le i \le N$, $a_x, b_x, c_x, d_x, a_y, b_y, c_y, d_y, a_z, b_z, c_z, d_z$, are determined by solving a linear system that transforms all the points $\mathbf{P}^{(i)}$ to the new points $\mathbf{V}^{(i)}$ [12] in equation (5). After extracting these coefficients, we put the other non-feature vertices into **(5)** to calculate their new locations.

6. Face Texture Mapping

After the 3D geometric head model is reconstructed from the video, we can compute the texture map of the head model from the video, too. In this paper, we proposed a modified texture mapping method to integrate all the images in the sequence to produce the texture map. The integration is accomplished via an appropriate weighting scheme.

A cylindrical surface is used as an intermediate representation between the 3D model and images [16]. The cylindrical coordinate is represented as (θ, h, r). In the texture image coordinate, the horizontal axis is the rotation angle θ, and the vertical axis is the height value h. In order to create texture image, we need to combine several reference images that overlap each other. To solve this problem, we mainly adopt the image mosaic technique proposed by Burt and Adelson

[17]. In this paper, we extend the pyramid decomposition to merge several images into one.

Firstly, we create a temporary reference image and a reference mask as the same size with the texture image for each image. And then mapping the pixel information from image to the corresponding temporary reference image through the model coordinates transformation from the image pixel to the corresponding cylindrical coordinates. And the temporary reference mask keeps the weights for each corresponding pixel. Then, we utilize the downsampling and upsampling operators to obtain the Gaussian and Laplacian images for each transformed image, respectively. The Gaussian pyramid and Laplacian pyramid operators proposed in [17] are employed in this paper for image downsampling and upsampling. However, we used a rendering-based weighting function in a cylindrical representation to combine multiple face images at different views. The weights are applied to corresponding pixels in each Laplacian image to merge the weighted images as the resulting image at each level. Finally, the upsampling operator is applied to generate the final texture map.

The weighting scheme is a product of two terms. The first is based on the inner product of the unit surface normal on the 3D model and the perspective projection line direction for the corresponding pixel. The second term is a squared cosine function of the deviated cylindrical angle between the corresponding pixel and the image center at the corresponding view. To be more specific, this weighting function is given as follows.

$$w_i^j = (\vec{n}_i^j \cdot \vec{p}_i^j)^2 \times \cos^2(\theta_i^j - \phi^j) \qquad (6)$$

where $\vec{n}_i^j$ is the unit normal vector on the 3D model corresponding to the pixel i in frame j; $\vec{p}_i^j$ denotes the projection line through the optical center and the 3D location on the model for the same pixel; θ_i^j is the θ value in cylindrical coordinate for the above pixel; and ϕ^j is the θ value in cylindrical coordinates of the face model at the current frame j. The weight is set to zero when the inner product is negative. All the weights of the pixels in different images corresponding to the same pixel in the texture image are normalized so that the sum is equal to 1.

7. Experimental Results

In this section, we present some experimental results of recovering the 3D head model from a video sequence by using the proposed system. In our experiments, we fixed the camera and have the person move with rotation in front of the camera. Three representative images in an example video sequence with selected feature points are shown in Figure 2. We used 22 face feature points for structure from motion in the robust bundle adjustment algorithm. These face features are manually selected at representative frames and then tracked along the rest of the sequence by using a cross-correlation algorithm.

The generic 3D head model used in our experiment was provided by the Visualization Environment Lab of the Institute of Information Science, Academia Sinica, Taipei. It consists of 1414 vertices and 2576 triangles. The wireframe of this model and its original texture mapped image are displayed in Figure 3.

Similarly, the selected face feature points corresponding to the 3D generic model were manually chosen. After the structure from motion and structure from silhouette process, we obtained the displacement vectors as geometric constraints on the 3D generic model as depicted in Figure 4(a) and (b). After the RBF interpolation, the adapted model is shown in Figure 4(c).

Seven frames in the video sequence are used to compute the texture map of the 3D head model by integrating them with appropriate weighting. Figure 5 shows the computed texture map. Figure 6 gives three different views of the reconstructed head model rendered with the integrated texture map.

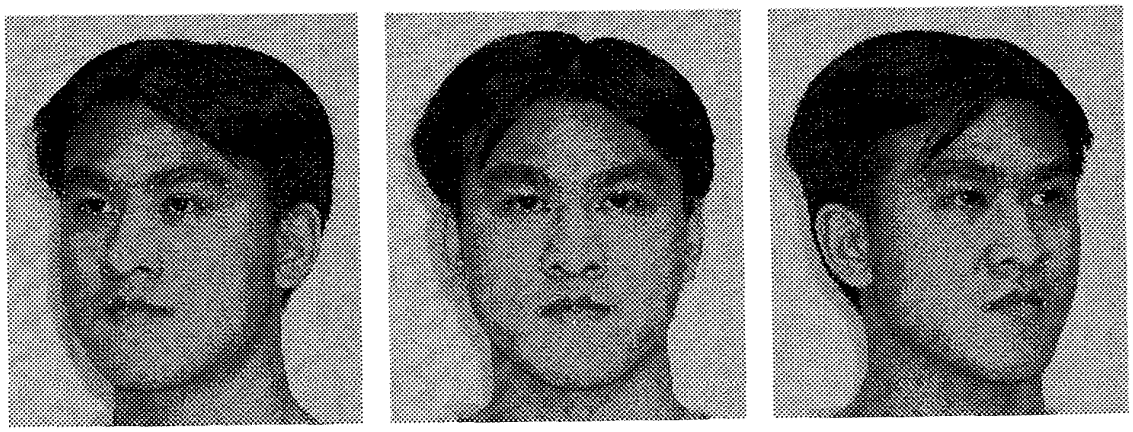


Figure 2. Three representative images in a video sequence with the selected face feature points indicated by white dots.

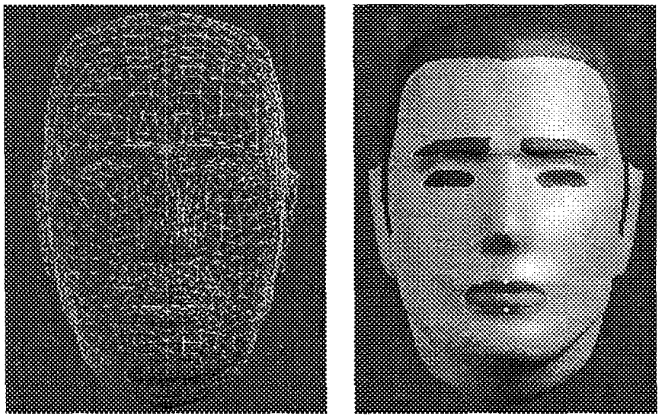
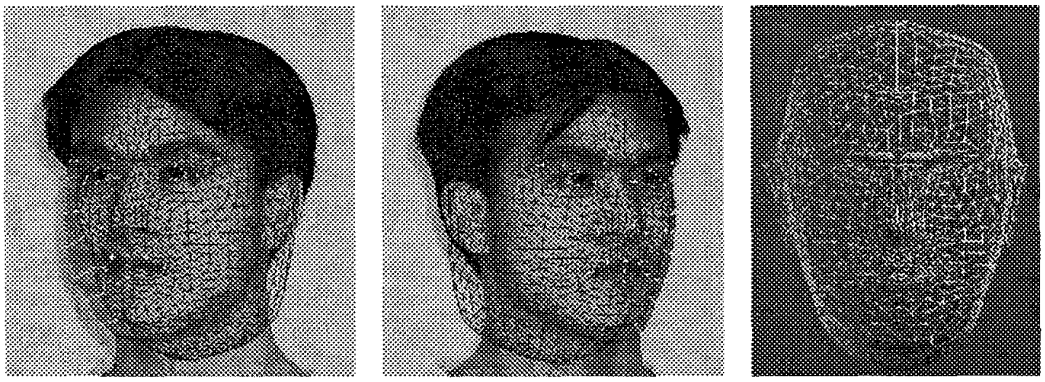


Figure 3. The generic 3D head model displayed with (a) a wireframe and (b) default texture mapped image with the feature points specified on the model.



(a) (b) (c)

Figure 4. The displacement vectors computed from the bundle adjustment and structure from silhouette are used to constrain the 3D generic head model as shown in (a) and (b). After model adaptation, the reconstructed 3D head model is shown in (c).

Figure 5. The integrated texture map.

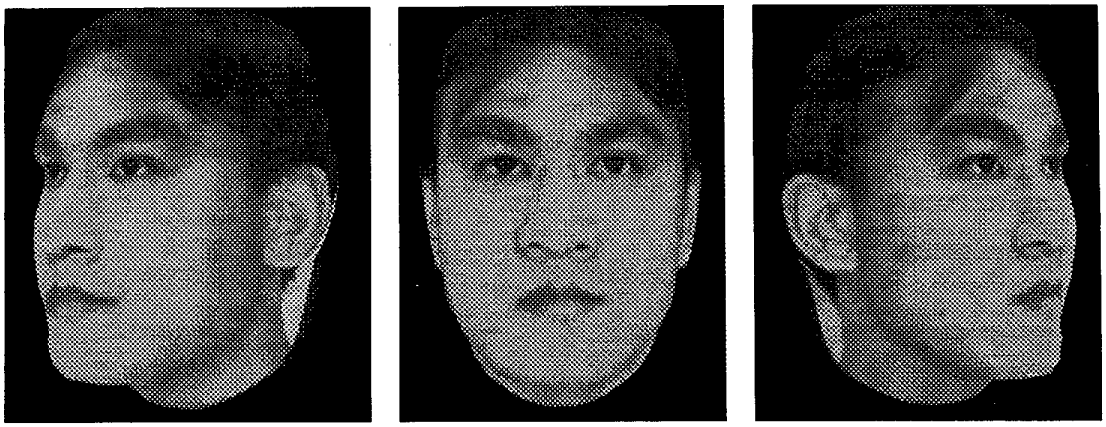


Figure 6. Three different rendered views of the computed 3D head model.

8. Conclusions

In this paper, we presented an integrated system for reconstructing an individualized 3D head model from a video sequence. Our reconstruction algorithm is based on the adaptation of a generic 3D head model. Three-dimensional geometric constraints on the head model are computed from the robust bundle adjustment algorithm and the structure from silhouette method. These 3D constraints are integrated to adapt the generic head model via radial basis function interpolation. Then the texture map of the reconstructed 3D head model is obtained by integrating all the images in the sequence through appropriate weighting. The proposed face model reconstruction method has the advantages of efficient computation as well as robustness against noises and outliers. The focus of future improvement of the current system lies on relieving user interaction in the feature extraction process. In addition, we would like to extract more feature points from the video sequence to better constrain the final reconstructed face model.

References

[1] F. Lavagetto, R. Pockaj, and M. Costa, "Smooth surface interpolation and texture adaptation for MPEG-4 compliant calibration of 3D head models," Image and Vision Computing, Vol. 18, No. 4, pp. 345-353, March 2000.

[2] J. Y. Zheng, "Acquiring 3D models from sequences of contours," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 2, pp. 163-178, 1994.

[3] T. Joshi, N. Ahuja and J. Ponce, "Structure and motion estimation from dynamic silhouettes under perspective projection," International Journal of Computer Vision, Vol. 31, No. 1, pp. 31-50, 1999.

[4] H.H.S. Ip and L. Yin, "Constructing a 3D individualized head model from two orthogonal views," Visual Computer, Vol. 12, pp. 254-266, 1996.

[5] W.-S. Lee and N. Magnenat-Thalmann, "Fast head modeling for animation," Image and Vision Computing, Vol. 18, No. 4, pp. 355-364, 2000.

[6] S. B. Kang, "A structure from motion approach using constrained deformable models and appearance prediction," Technical Report CRL 97/6, Camerage Research Laboratory, Oct. 1997.

[7] P. Fua and C. Miccio, "Animated heads from ordinary images: a least-squares approach," Computer Vision and Image Understanding, Vol. 75, No. 3, pp. 247-259, September 1999.

[8] P. Fua, "Regularized bundle-adjustment to model heads from image sequences without calibration data," *International Journal of Computer Vision*, Vol. 38, No. 2, pp. 153-171, July 2000.

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," Proceedings of SIGGRAPH, pp. 71-78, Los Angeles, CA., August 1999.

[10] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330-1334, 2000.

[11] R. M. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim, "Pose estimation from corresponding point data," IEEE Transactions on System, Man, and Cybernetics, Vol. 19, No. 6, pp. 1426-1446, 1989.

[12] S.-H. Lai and C.-M. Cheng, "Three-dimensional face model creation from video," Proceedings of SPIE Conf. On Three-dimensional Image Capture and Applications IV, Vol. 4298, Jan. 2001.

[13] G. Li, "Robust regression," in *Exploring Data Tables, Trends, and Shapes* (D. C. Hoaglin, F. Mosteller and J. W. Tukey Ed.), pp. 281-343, Wiley, New York, 1985.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C*, 2nd Edition, Cambridge University Press, 1992.

[15] M. Kass, A. Witkin, and D.Terzopoulos, "SNAKE: active contour models," *International Journal of Computer Vision*, Vol. 1, pp. 321-322, Jan.1988.

[16] F. I. Parke and K. Waters. *Computer Facial Animation*. AK Peters, Wellesley, Massachusetts. 1996.

[17] Peter J. Burt and Edward H. Adelson, "A multiresoluton spline with application to image mosaics," ACM Transactions on Graphics, Vol. 2, No. 4, pp. 217~236, Oct. 1983.